PhD degree in Foundations of the Life Sciences and their Ethical Consequences

European School of Molecular Medicine (SEMM) and University of Milan

Faculty of Medicine

Settore disciplinare: FIL/02

BIOMEDICAL ONTOLOGIES:

EXAMINING ASPECTS OF INTEGRATION ACROSS BREAST CANCER KNOWLEDGE DOMAINS

Aleksandra Sojic

IFOM-IEO Campus, Milan

Matricola n. R07960

Supervisor: Dr. / Prof. Giovanni Boniolo

IFOM-IEO Campus, Milan

Anno accademico 2011-2012

Abstract

The key ideas developed in this thesis lie at the intersection of epistemology, philosophy of molecular biology, medicine, and computer science. I examine how the epistemic and pragmatic needs of agents distributed across particular scientific disciplines influence the domain-specific reasoning, classification, and representation of breast cancer.

The motivation to undertake an interdisciplinary approach, while addressing the problems of knowledge integration, originates in the peculiarity of the integrative endeavour of sciences that is fostered by information technologies and ontology engineering methods. I analyse what knowledge integration in this new field means and how it is possible to integrate diverse knowledge domains, such as clinical and molecular.

I examine the extent and character of the integration achieved through the application of biomedical ontologies. While particular disciplines target certain aspects of breast cancer-related phenomena, biomedical ontologies target biomedical knowledge about phenomena that is often captured within diverse classificatory systems and domain-specific representations. In order to integrate dispersed pieces of knowledge, which is distributed across assorted research domains and knowledgebases, ontology engineers need to deal with the heterogeneity of terminological, conceptual, and practical aims that are not always shared among the domains. Accordingly, I analyse the specificities, similarities, and diversities across the clinical and biomedical domain conceptualisations and classifications of breast cancer.

Instead of favouring a unifying approach to knowledge integration, my analysis shows that heterogeneous classifications and representations originate from different epistemic and pragmatic needs, each of which brings a fruitful insight into the problem. Thus, while embracing a pluralistic view on the ontologies that are capturing various aspects of knowledge, I argue that the resulting integration should be understood in terms of a coordinated social effort to bring knowledge together as needed and when needed, rather than in terms of a unity that represents domain-specific knowledge in a uniform manner. Furthermore, I characterise biomedical ontologies and knowledgebases as a novel socio-technological medium that allows representational interoperability across the domains.

As an example, which also marks my own contribution to the collaborative efforts, I present an ontology for HER2+ breast cancer phenotypes that integrates clinical and molecular knowledge in an explicit way. Through this and a number of other examples, I specify how biomedical ontologies support a mutual enrichment of knowledge across the domains, thereby enabling the application of molecular knowledge into the clinics.

Table of Contents

ABSTRACT **2**

TABLE OF FIGURES **6**

INTRODUCTION **8**

§1 A HISTORICAL BACKGROUND 8
§2 PROBLEMS IN CANCER RESEARCH 11
§2 ORGANISING KNOWLEDGE 14

CHAPTER I **21**

BIOMEDICAL ONTOLOGIES **21**

1.1. ELECTRONIC HEALTH RECORDS AND BIOMEDICAL CLASSIFICATION 21
1.2. BIOMEDICAL ONTOLOGIES AS A TOOL TO REPRESENT REASONING 27
1.2.1. PHILOSOPHICAL AND COMPUTATIONAL CONCEPT OF 'ONTOLOGY' 28
1.2.2. BIOMEDICAL ONTOLOGY AS A CLASSIFICATORY TOOL 38
1.2.3. REPRESENTING KNOWLEDGE AS LABELLING THE CLASSES 48
1.2.4. REPRESENTING REASONING OVER THE LABELS AND CLASSES 57
1.2.5. KNOWLEDGE REPRESENTATION AND KNOWLEDGEBASES 76
1.2.6. EPISTEMIC GROUPS INVOLVED IN ONTOLOGY BUILDING 82

CHAPTER II **89**

BIOMEDICAL REPRESENTATIONS: THE CASE OF BREAST CANCER PHENOTYPE **89**

2.1. DIVERSITY OF REPRESENTATIONS 91
2.1.1. PHENOTYPE ONTOLOGIES 95
2.2. HER2 PHENOTYPE ONTOLOGY 99
2.2.1. HER2 PROTEIN IN 'NORMAL' AND 'ABNORMAL' PHENOTYPES 100
2.2.2. CONCEPT MINIMISATION: NORMAL VS. ABNORMAL PHENOTYPES 103
2.2.3. IMPLEMENTATION OF HER2 MODEL 106
2.3. PLURALITY OF EPISTEMIC AND PRAGMATIC INTERESTS: THE 'HER2' CASE 107
2.4. CAPTURING KNOWLEDGE IN THE CONTEXT OF A DOMAIN 110

CHAPTER III **122**

CLASSIFICATORY KNOWLEDGE AND BREAST CANCER CLASSIFICATIONS **122**

3.1. PERSONALIZED MEDICINE EXPLICATING THE EPISTEMIC GAP 123
3.2. REPRESENTING CLINICAL VS. BIOMEDICAL KNOWLEDGE 126
3.3. THE LANGUAGE OF A DOMAIN: WITH AND WITHOUT 'PATIENT' 134
3.4. CLASSIFICATIONS: WHY, WHAT, AND HOW 140
3.5. CLINICAL CLASSIFICATION 152
3.5.1. TNM CLASSIFICATORY SYSTEM 153
3.5.2. CLINICAL TRIALS SHIFTING THE CLASSES 158
3.5.3. TNM AS A DUAL SYSTEM 161
3.5.4. AMBIGUITY IN CLASSIFICATION OF THE GRADE TWO TUMOURS 163
3.6. BIOLOGICAL CLASSIFICATION, HISTOLOGICAL AND MOLECULAR SUB-TYPES 167
3.7. RISK FACTORS AS THE CRITERIA FOR CLASSIFICATION 178

3.7.1. Identifying the risk factors 179
3.7.2. What do the risk factors explain? 181
3.7.3. The role of 'Age' across the classificatory domains........ 183
3.7.4. Representing 'Age' as an epistemic modifier 191

CHAPTER IV 199

THE BIOMEDICAL ONTOLOGIES FACING THE DISUNITY OF SCIENCE 199

4.1. The problems with merging domain knowledge........ 200
4.2. A revival of the unity vs. disunity debate........ 209
4.2.1. Examining the collaborative unity 219
4.2.2. The extended-meaning 'unity' 233

CONCLUSION 245

BIBLIOGRAPHY 253

Table of Figures

Figure 1 An example of an EHR interface 23

Figure 2 Standards for Health IT Interoperability 24

Figure 3 Kinds of Ontologies (Uschold, Gruninger 2004) 36

Figure 4 Diverse perspectives in classifying and representing protein-protein interactions (STRING) 41

Figure 5 Classification of evidence (STRING) 42

Figure 6 Distinguishing the ontology classification from other kinds of IT classifications (Rector 2006) 47

Figure 7 Representation of a semantic and a concept map 52

Figure 8 A visualisation of the ICD-11 information model in Protégé (Tu et al. 2010) 55

Figure 9 Knowledge claims as mediators in representation: from scientific objects to ontology models 65

Figure 10 Comparing the representational means of *LEX* and *ONT* 69

Figure 11 Abstraction from a microscopy to an iconic representation 70

Figure 12 Mapping of *PIC*, *LEX*, and *ONT* representations 71

Figure 13 Representing meaning of scientific terms and concepts 73

Figure 14 Knowledge organisation: epistemic groups and representation types 85

Figure 15 HER2 as a difference maker: ‘normal’ vs. ‘abnormal’ breast phenotype 101

Figure 16 HER2 protein detection by immunohistochemistry (IHC) 102

Figure 17 The inferred class hierarchy of the HER2 ontology as displayed in Protégé 105

Figure 18 Knowledge granularity 108

Figure 19 Various representations of a model, capturing different aspects of scientific reasoning 118

Figure 20 Clinical and biomedical reasoning 128

Figure 21 Semantic analysis - GOPubMed (*Mix* ‘patient’ +’cell line’; *Res* ‘cell line’ -’patient’; *Clin* ‘patient’ -’cell line’) 136

Figure 22 GOPubMed (‘genes ‘neoplasms’ -’patients’ vs. ‘genes ‘neoplasms’ -’patients’) 138

Figure 23 ‘Grey area’ in reasoning about grade two tumours 164

Figure 24 - Duck-rabbit 165

Figure 25 Additional information..166

Figure 26 Including preferences 166

Figure 27 Cancer vs. normal cells - Illustration by National Cancer Institute (Pat Kenny) 168

Figure 28 Histological special types of breast cancer, from Weigelt et. al, 2010, Molecular Oncology 4 (3):192-208...171

Figure 29 A plurality of breast cancer classifications..173

Figure 30 A classification of the parameters, which describe the molecular sub-types 176

Figure 32 The age related domains: risks, ‘normal’, and ‘abnormal’ biological processes 187

Figure 32 Age as a time dependent measure 187

Figure 33 The role of age across domains 189

Figure 34 ‘Age’ as an epistemic modifier..193

Figure 35 GO definition for GO:0003700 236

Figure 36 UniProt IDs for 'p53' across the species...240

Introduction

> One day, we imagine that cancer biology and treatment—at present, a patchwork quilt of cell biology, genetics, histopathology, biochemistry, immunology, and pharmacology—will become a science with a conceptual structure and logical coherence that rivals that of chemistry or physics (Hanahan and Weinberg 2000).

§1 A historical background

Breast cancer is recognized as one of the most common cancers in women in the Western World[1]. Since the mid-twentieth century, intensive breast cancer research has been conducted. However, the road to knowledge on breast cancer is not a short one. The first historical documents[2] describing breast cancer go back to Hellenistic times (Donegan and Spratt 2002). The term *καρκίνωμα* (karkinoma) was used to designate a malignant growth and the term *σκίρρος* (scirrhous) was used to designate solid tumours. Hippocrates (460-375BC) described in detail several breast cancer cases, the symptoms and disease progression, connecting them with a bad outcome. Around 30 AD, a Roman physician Aulus Cornelius Celsus, who noticed that women's breasts were a common site for cancer, described the four stages of breast cancer. He recognized as potentially curable only the first stage. Although breast cancer was very early recognized as a

[1] http://info.cancerresearchuk.org/cancerstats/types/breast/incidence/

[2] Disputably, the first mention of breast cancer dates 3000BC, in an Egyptian papyrus.

severe and often incurable disease, attempts at treatment have been diverse and numerous throughout history (Donegan and Spratt 2002). However, according to the available documents, the progress in understanding and the explanation offered for this disease was not dynamic until the nineteenth century. The first explanation that referred to natural causes was proposed by Hippocrates. He believed that the cause of breast cancer was an imbalance of the four bodily fluids: blood, phlegm, yellow bile and black bile. Hippocrates also noticed a correlation between a cessation in menstrual bleeding and the appearance of breast cancer. However, from this observation, he merely inferred that menstrual bleeding was beneficial for women. An explanation that connects menopause and breast cancer had to wait for the technological and scientific progress at the end of the nineteenth century. The role of hormones and a possibility that one organ can influence the function of another organ was unknown until the mid-nineteenth century.

Progress in the understanding of breast cancer pathology correlates with improvements in surgery and microscopy (Ignác Semmelweis, Louis Pasteur, Joseph Lister and William Morton). In 1838, Johannes Müller noted for the first time that cancer was composed of living cells. Müller also recorded a similarity between breast cancer cells and metastasis. He noticed the abnormal shape of cancer cells, describing it as a loss in the proportions compared to normal cells. However, knowledge about breast cancer hormone dependence played a crucial role in understanding the disease.

In addition to a direct therapeutic purpose, surgical interventions contributed to the understanding of breast cancer as a hormone-dependent disease. In 1872, a German surgeon, Alfred Hegar, and an American physician, Robert Battey independently of each other performed oophorectomy for the first time. Some years after, in 1889, Albert Schinzinger proposed surgical oophorectomy as a treatment for breast cancer. The proposal was related to the observation that the prognosis for breast cancer was better in older women than in younger women. Therefore, he reasoned that oophorectomy would make younger women prematurely old, thereby causing atrophy of the breast as well as cancer (Love and Philips 2002).

An important moment in the history of medicine occurred in the mid-nineteenth century when a new kind of interaction between clinical observations and experimental research fostered the progress of knowledge. In the 1850s, Claude Bernard introduced the term 'internal secretion' to explain his experimental observations that demonstrated secretion of sugar by the liver (Gruhn and Kazer 1989; Bernard 1999; Bernard, Greene, and Henderson 1957). Although Bernard did not use the term 'hormone', his idea of internal secretion and an interconnection of organs by the secreted substances marked the advent of endocrinology. In 1855, Thomas Addison described a syndrome of darkened skin and an experience of fatigue, nausea, and vomiting in patients, relating this to a malfunction of adrenals. Charles-Édouard Brown-Séquard supported Addison's discovery by replicating the described conditions. In his experiments, Brown-Séquard showed that the removal of adrenals in dogs was fatal. When Brown-Séquard was seventy-two, he tested the injected extracts of dogs' testicles on himself, claiming discovery of a rejuvenating effect that the testicles' extracts had on aged males. This led him to an 'organotherapy' theory, which was overstating the claim that every organ can be used for therapeutic purposes. However, even if overstated, the 'organotherapy' approach initiated advances in the therapeutic use of bodily extracts, one of which is hormonal therapy (Aminoff 2010).

Hormones played a therapeutic role in the treatment of cancer, even before the role of hormones was fully understood. Indeed, the therapeutic effects of 'the secreted substances' supported formation and testing of scientific hypotheses about cancer. Similar to other modern combinatorial therapeutic approaches, George Thomas Beatson, in 1895, combined surgical intervention with a gland extract treatment. He performed an oophorectomy on a woman with breast cancer and then continued the treatment with a thyroid extract that started a month earlier. Beatson reasoned that oophorectomy would cause fatty degeneration of the malignant cells (Beatson, Francis, and Eng 1896; Love and Philips 2002).

By the beginning of the twentieth century, the extracts secreted by organs were proved to be chemicals with significant functions in the body. Ernest Starling, in 1905, named these chemicals 'hormones', from the Greek *ὁρμή*, which designates rapid motion forwards or 'I exite'

(Henderson 2005; Tata 2005). According to Jamshed Tata, the history of science showed how ‘the introduction of a new word can act as a catalyst for research’, as the chain of events that followed the naming of hormones confirmed (Tata 2005). Sterling’s characterization of ‘the chemical messengers which, speeding from cell to cell along the blood stream, may coordinate the activities and growth of different parts of the body’ (Sterling quoted in Tata 2005) introduced also a reductionist picture in medicine. In addition to arousing the interest of medical practitioners, hormones were also quickly studied by chemists and molecular biologists (Tata 2005).

A rapid increase in knowledge in breast cancer biology has been enhanced by the development of twentieth century technologies, from imaging produced by X-rays to those represented as microarrays. This marriage between scientific research and technology resulted in a closer specification of breast cancer sub-types, improved therapy and decreased mortality. However, an increase in the amount of data that describe the disease on fine-grained levels has increased the complexity in the representation of the disease. The process of knowledge acquisition, which resulted in the acquisition of detailed knowledge in order to explain and predict breast cancer outcomes, also led to sub-disciplinary divisions.

§2 Problems in cancer research

The complexity of a phenomenon such as cancer causes many difficulties in the understanding, modelling, representing, explaining, and eventually, applying accumulated knowledge in clinics. Considerations such as which units and relations among the observed phenomena should be considered the most relevant for carcinogenic events led to the characterisation of cancer, in particular breast cancer, as a very heterogeneous disease (Fisher, Redmond, and Fisher 2008). Accordingly, heterogeneous models and representations of this disease have been developed. However, the reason for continuing with a plurality of representations, instead of giving a uniform or a single picture of the disease, does not seem to originate in the inability of scientists to identify the unique ‘biological objects’ responsible for the development of cancer in every single case. Quite the opposite, scientific research over recent decades resulted in an overwhelming body of knowledge about ‘the objects’ such as molecules, molecular complexes, and interaction networks shown to be significant in carcinogenesis.

In addition, numerous and heterogeneous data emanating from different levels of observation, such as the level of an organism, its environment, the cell, the cell’s environment, the vascular system, the endocrine and metabolic networks, seem to hinder a single representation of the disease. Since the complex task of understanding cancer and its treatment has been divided into many small problems and questions addressed by different sub-disciplines, the modes and scopes of related representations are expectedly dispersed within and across disciplines. Therefore, the problems related to cancer have been examined from various research perspectives, which have eventually resulted in a wide plurality of the disease’s representations. The representations of breast cancer may include images acquired by technologies such as ultrasound, X-ray, microscopy of histopathological samples. Moreover, the representations of the disease are not limited to visual representations of tumour, but may include mathematical equations, statistical graphs, molecular markers, microarrays data, and the phenotype-specific protein interactions, thus describing cancer according to the needs of and knowledge about a particular domain’s perspective.

Obviously, many of the problems in the representation of cancer come from the interpretation of data in the light of network complexity, static measurements that are often missing dynamic aspects of the disease, spatio-temporal heterogeneity, uncertainty, a combination of qualitative and quantitative data, and missing links between physiological and pathological data (Faratian et al. 2009). Thus, cancer is characterised in general terms as a complex, heterogeneous and dynamic disease, or a set of diseases (Vargo-Gogola and Rosen 2007). Concurrently, despite the vast amount of knowledge, cancer is difficult to manage in clinics and it continues to retain its label of a deadly disease.

The promising discovery of the oncogenes and tumour-suppressor genes in the 1980s and the identification of the hallmarks of cancer supported great hopes that cancer can be explained by simple principles.

> We foresee cancer research developing into a logical science, where the complexities of the disease, described in the laboratory and clinic, will become understandable in terms of a small number of underlying principles. Some of these principles are even now in the midst of being codified (Hanahan and Weinberg 2000).

In a seminal paper, *The Hallmarks of Cancer*, Douglas Hanahan and Robert A. Weinberg suggested that the cancer cell genotype was a manifestation of six essential alterations in a cell: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (Hanahan and Weinberg 2000). Even if these six hallmarks of cancer can be seen as the crucial units (terms) for the representation of cancer, they do not seem to be sufficient for a unifying scientific representation of cancer. The hallmarks of cancer, rather, have a character of general descriptive terms, loosely linked by a set of many possible mechanistic explanations, short of details and information as the actual directions of the disease development might be numerous. Therefore, the representation of cancer by means of the cancer hallmarks could be considered as a unifying but, nonetheless, oversimplified and too general description of pathology. The real problem for a detailed scientific representation seems to be a specification of 'the small number of underlying principles' that accurately represent carcinogenesis, which results in generally recognised hallmarks. Thus, the task is to specify the links that hold among the hallmarks of cancer, particular cell features, the context of an organism and its environment.

Contrary to the hopes of Hanahan and Weinberg, the highthroughput screenings of the cancer genomes revealed a significant number of functionally heterogeneous genes to be mutated in tumours. The poor overlapping between cancer genes has led to the conclusion that each tumour is a distinct disease with specific, and partly unrelated, genetic determinants.

So, the picture that connects the pathology of cancer with its molecular level is still fuzzy. Hopes that cancer can be explained by clear-cut principles in the case of breast cancer are particularly troublesome.

> Indeed, unlike colon cancer or pancreatic cancer, in which virtually all tumour mutation within a single pathway has a dominant role during tumour progression, in breast tumours no single dominant pathway or histological presentation has emerged. Characterization of the chromosomal aberrations, gene mutations and gene expression profiles of breast tumours has shown that breast tumorigenesis does not necessarily progress in a stepwise linear fashion from well-differentiated to poorly differentiated tumours (Stingl and Caldas 2007).

This heterogeneity in breast cancer specification results in obstacles in explaining, understanding, modelling and representing the disease. As expected, such a characterisation has resulted in a proliferation of molecular and clinical representations.

§2 Organising knowledge

An expansion of biomedical knowledge is especially evident at the fine-grained level, that is, examining various molecular aspects of the disease. However, the overwhelming volume of knowledge about molecular components and processes involved in carcinogenesis has resulted in an immense amount of information that has to be ordered and organised. Organisation of a huge amount of heterogeneous data, produced and analysed by various methodologies and diverse types of reasoning, necessitates collaboration across scientific domains. My aim is to address certain epistemological issues that emerge in the process of the organisation of the acquired knowledge into a comprehensive unit that integrates diverse knowledge domains. In particular, I examine how molecular biology knowledge can be integrated with clinical knowledge.

An emerging issue of knowledge integration concerns two apparently counteracting forces: disciplinary division that drives the specialisation of knowledge on one hand, and inter-disciplinary collaboration on the other. Disciplinary sub-specialisation allows a detailed analysis of the specific problems within a research group, focused on, for example, a particular protein that modifies gene expression, eventually resulting in the development of cancer. However, when knowledge acquired within a research group has to be integrated into the big picture of complex carcinogenic events, collaboration beyond the particular research domain is unavoidable. Moreover, collaboration among clinicians and molecular biologists will be shown to be crucial for the application of molecular knowledge into clinics.

A strong hope for the integrative endeavour in contemporary science is that by connecting dispersed pieces of knowledge produced within one domain which have not yet been connected with knowledge from another domain, when connected, can actually drive new inferences, producing new knowledge and a better understanding of disease. The well-established idea that knowledge can be expanded by connecting dispersed pieces of information has gained particular importance with the development of information systems such as databases, semantic web technologies, and the ontology tools that map and represent concepts and sets of data across different fields. Thereby, information technologies allow data storage, interconnection, and manipulation with a huge amount of heterogeneous types of data which goes beyond the cognitive abilities of any particular expert. As knowledge organised in databases represents the concepts in a computer readable form, the use of implemented algorithms, i.e. reasoners (e.g. FaCT++, Pellet, RacerPro)[3], supports an automated inference about the represented information. So, in the case of cancer management, the integration of the patient's clinical data with information about the patient's molecular profile, family history, presence of other diseases, life style habits etc. can help clinicians to assess the most appropriate and patient-specific treatment.

Information systems employed in the management of transferring health records into databases often use ontologies as a tool for information management. The term 'ontology' in the domain of information technologies is used as a technical term and has a different meaning from the philosophical term 'Ontology' (Section 1.2.1). In order to avoid confusion regarding the two

[3] A description of the most frequently used reasoners see http://www.cs.man.ac.uk/~sattler/reasoners.html

meanings, I accept the proposed distinction (Guarino, Oberle, and Staab 2009) to use the term 'Ontology', an uncountable noun with uppercase initial, as related to the philosophical discipline, which belongs to the scope of metaphysics[4]. On the other hand, the term 'an ontology', a countable noun with the lower case initial, I use to denote the computational artifacts (Gruber 1993, 1995; Guarino, Oberle, and Staab 2009). Since I address problems related to the methodological and epistemic aspects of the classification and representation of biomedical knowledge, I keep the discussion on the epistemological level, without committing to any particular metaphysical position. Thus, I primarily use the term 'ontology' in the latter sense, i.e. as it is used in computer science and ontology engineering.

Ontology engineering seems particularly promising in the domain of the life sciences, where a huge amount of data and heterogeneous types of knowledge have to be ordered, classified and represented. The ontologies that deal with the biomedical domain are named 'biomedical ontologies'. A systematic collection, annotation, and ordering of knowledge acquired in the local experimental settings facilitate the circulation of related knowledge claims as well as experimental data records, which can be further used and re-used across scientific communities (Leonelli 2008, 2009a). Indeed, information-based systems foster the circulation of knowledge that has been and still continues to be circulated through various channels of exchange, such as publications, printed images, and laboratory samples. However, the circulation of knowledge in this new medium provided by information technologies poses a huge new task for scientific communities.

Knowledge represented in a knowledgebase has to be explicit and interoperable. Therefore, it demands an explication of each term, experimental procedures and justifications that support the inferred claims. The explication that is required for the computational consistence of a knowledgebase has to face the ambiguities present in scientific language that can go unnoticed in the common use of language in daily communication among members of a

[4] http://plato.stanford.edu/entries/logic-ontology/

research group. For instance, the domain-specific meaning of a term might be used differently in another domain (Sections 2.3, 2.4, 3.7.3, 4.2.2). Furthermore, the research methods and justificatory criteria may differ across fields (Sections 2.3, 2.4, 3.2, 3.7.3). While working on alignments across numerous and heterogeneous fields, computer scientists collaborate with various domain scientists. In order to address interoperability problems, collaborative efforts make the inter-field disagreements explicit. In this way, efforts to align knowledge across the fields ideally lead to the integration of particular disciplines into a coherent science. Thus, Hanahan-Weinberg's visionary picture of cancer biology and treatment as a conceptually coherent science that integrates a patchwork quilt of various disciplines (Hanahan and Weinberg 2000) is widely shared within the emerging community of bioinformaticians, computer scientists, and knowledge engineers who are using and developing the tools to achieve the goals of an integrated science.

Pursuing the same idea, I try to take a rather critical stance and analyse the problems and obstacles on the way to knowledge integration. Hopefully, my analysis will also make a constructive contribution to the shared goal of an integrative science particularly focused around breast cancer-related problems. I characterise the relations, similarities and differences among the oncology sub-domains. In particular, I stress the implicit interdependence of classificatory categories and the particular roles that they play in the representation of knowledge of breast cancer. Such an analysis should clarify how a disciplinary context influences the meaning of a particular classificatory term and related concepts that should be positioned into a new context. Since the understanding of knowledge context preconditions efficient knowledge management, my task is to examine the context of the clinical and molecular classifications of breast cancer.

Concurrently, with its integrative tendency, contemporary medicine is going in a seemingly opposite direction, i.e. towards a particularisation of reasoning that is characteristic of personalised medicine. According to the personalised medicine approach, each patient is considered an individual with specific features that drive a particular and patient-specific response to therapy. Therefore, instead of applying a very general knowledge and reasoning that would classify the patients into large groups of individuals with similar features related to the disease, personalised medicine focuses on the patient-specific features that distinguish one patient from the others.

This switch of focus in medical reasoning makes explicit a need to specify the heterogeneous features that describe patient, disease, and environment through distinguishable modules or classes of features that can be used and re-used in particular cases of reasoning about patients' disease, preferences, and available therapies (Sections 1.1, 1.2.2, 3.1). A patient-tailored medicine is also moving clinical decision making from inferential reasoning, which is used to consider mostly large groups of patients while ignoring intra-group heterogeneities, to a clinical decision process that is patient-specific and context sensitive, as it considers subtle differences such as differences in gene and protein expression levels.

The organisation and inclusion of knowledge from molecular oncology into clinically useful modules of information could efficiently employ existing fine-grained knowledge into clinical practice. Bearing in mind the amount of data and molecular heterogeneity of breast cancer classes, the project of personalised medicine seems to depend, to a high degree, on the process of knowledge modularisation and the eventual integration of the represented modules. For this reason, I have decided to focus on the representation of breast cancer heterogeneity, diversity in its classification, and the domain-related descriptions, reasoning, and explanations.

The first chapter starts by presenting the relationship between biomedical classification and electronic health records (EHR). Furthermore, I discuss the basic distinctions between *classifications, thesauri,* and *ontologies*. In particular, the chapter illustrates how an ontology manages to capture knowledge and reasoning about a domain in an explicit way (Sections 1.2.2-1.2.5). As a heuristic tool to understand what ontologies and knowledge bases are, section 1.2.6 distinguishes certain specificities and interdependences of the epistemic groups involved in ontology building. Ontology building is described as an interdisciplinary endeavour that deals with the diversity of epistemic and pragmatic interests.

In the second chapter, I discuss the plurality of biomedical representations within and across the research domains that are capturing the breast cancer phenotypes. I point out that the diversity of phenotype representations, including phenotype ontologies, depends on the research questions and the particular aims for which a representation is designed (Sections 2.1-2.1.1). Throughout the chapter, I argue that a particular problem, such as the representation of a breast cancer phenotype, asks for a combination of various representational perspectives in order to cover the problem in its complexity. Nonetheless, I stress that a particular combination of perspectives takes place within a pragmatically driven framework that represents knowledge as needed and when needed. From the side of application, I propose a model that represents 'normal' and 'abnormal' phenotypes, based on the chosen clinical and molecular criteria, which describe HER2+ tumours (Sections 2.2-2.2.3). The HER2 model demonstrates how biomedical reasoning about HER2+ phenotypes is captured into an ontology model. The example shows how reasoning about HER2 as a cell component and as a tumour marker is targeted and represented within a formal model. In addition, the example illustrates how the demands of various communities that employ specific standards and methodologies are combined in a formal specification.

The third chapter focuses on classificatory knowledge. It starts by presenting certain distinct features of clinical and biomedical knowledge about breast cancer. The problems that arise from the different aims of the clinical and molecular domains, I characterise as an epistemic gap (3.1.). In order to understand the sources of the gap that may impact the formal knowledge representation, I explicate the epistemic needs that are driving diversities among various clinical and biomedical representations and classifications of breast cancer (3.2). Furthermore, I examine the demands that a well structured classification needs to satisfy (Section 3.4.). I point out the advantages and practical limitations of having a uniform classificatory system. Sections 3.5-3.7

present how particular clinical and biological classifications capture knowledge about breast cancer. The discussion of the classificatory systems outlines the specificities and interdependences of the classificatory categories that are often context sensitive. Having in mind an integration of clinical and molecular classificatory categories into an ontology model, I specify certain classificatory terms as the potential integrative links. In particular, I examine the role that 'age' plays across the clinical, molecular, and epidemiological domains. The example demonstrates the domain-specific conceptualisations of the 'age' category, which can be informative in understanding and representing breast cancer. Furthermore, I argue that the integration of the classificatory categories across the domains supports a multidirectional enrichment of classificatory knowledge.

The fourth chapter explores the ways in which the applied ontologies and knowledge bases can be considered as a new medium for knowledge integration. The chapter examines the aspects of novelty that information technologies bring to the field of molecular oncology and breast cancer research. In particular, the fourth chapter considers the impact of this new integrative endeavour on the philosophical debate on the unity and disunity of science. I argue that information technologies revive the debate in a particular way. Finally, I characterise the resulting 'unity' as *the extended-meaning unity* that is achieved through the process of co-ordinated decontextualisation of knowledge. Thus, I conclude that science might be considered as unified in terms of the established socio-technological standards that, on the various scales of society, support a co-ordination of the collaborative efforts in collecting, representing, and exchanging knowledge, while a weak form of the representational integration is eventually achieved by means of the interoperability across domain-specific representations.

Chapter I

Biomedical Ontologies

This chapter analyses the basic concepts and relations of biomedical ontologies and biomedical knowledge. It starts with a brief overview of the biomedical informatics domain, which deals with management of electronic health records. The remainder of the chapter should clarify what knowledge engineering, ontology tools, and knowledgebases mean from an epistemological perspective. In particular, I discuss how ontologies play the role of a classificatory tool that captures knowledge. The distinction between philosophical and computational ontologies is outlined in section 1.2.1. In order to clarify further how biomedical ontologies actually represent knowledge, I address the representational aspects of biomedical ontologies within sections 1.2.3-1.2.5. The concluding section characterises the process of a knowledgebase building, whereby various epistemic groups collaborate in order to capture knowledge formally.

1.1. Electronic Health Records and Biomedical Classification

An electronic health record (EHR) is an individual patient's medical record stored in digital form (Hristidis 2009). More precisely, an EHR is defined as a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting

(Farfán 2009). According to ISO (2004), “the EHR means a repository of patient data in digital form, stored and exchanged securely, and accessible by multiple authorized users. It contains retrospective, concurrent, and prospective information and its primary purpose is to support continuing, efficient and quality integrated health care” (ISO/DTR 20514).

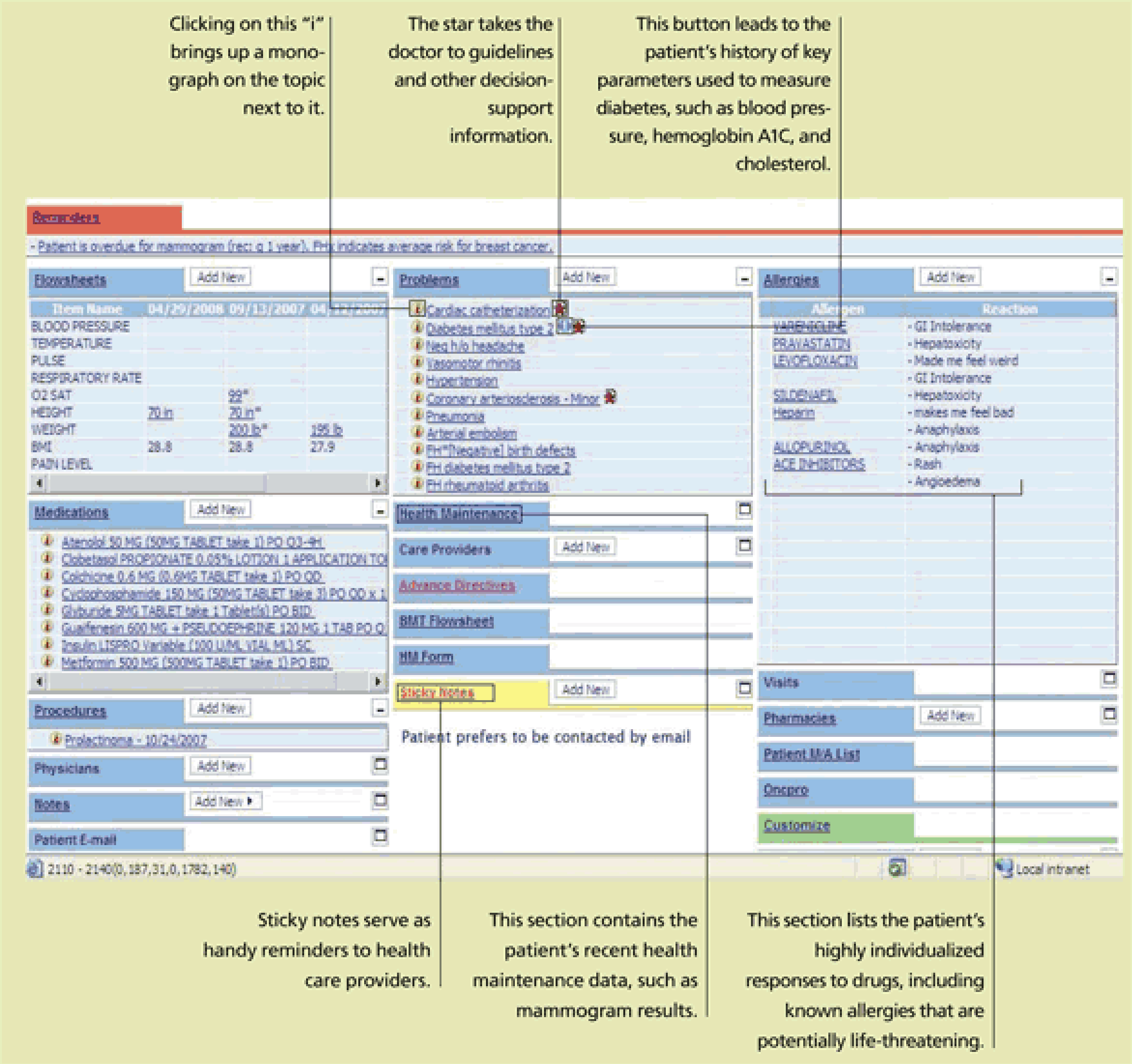
The EHR systems have been constructed in order to capture structured clinical information, organised in standardized formats and designed for a specific purpose (Rector, Nowlan, and Kay 1991; Rosenbloom et al. 2006). The main purpose of EHR is to support health care services. It has been recognised that use of EHR increases patient safety by reducing medical errors, and by improving general efficiency of medical care (Committee on Improving the Patient Record 1991; Rosenbloom et al. 2006).

In addition, EHR can be a useful source of information for biomedical research purposes. Systematically collected information about patients allows further analysis, which may produce new knowledge through various comparative methods, e.g. relating clinical findings, diagnosis and therapy related outcomes (Hristidis 2009). Use of EHR for research purposes is defined as secondary to the primary patient’s care delivery, and it is regulated by special legislatives that deal with the issues of privacy (El Emam, Michael Power, and Willison 2007; Van Der Linden et al. 2009; Guarda 2011).

The aim of this section is to give an overview of how medical information is structured within an EHR and how it is represented in order that it can be used and re-used in various settings of clinical practice and biomedical research. A more detailed analysis of epistemological and representational aspects of EHR is to come afterwards.

The EHR systems are designed to support multifunctional medical care settings. So, the EHR systems integrate heterogeneous information that can be relevant for a systematic and accurate treatment of a patient. Therefore, an EHR can contain several types of patient data, such as the patient’s demographic information, basic clinical data, signs and symptoms, medical history, immunizations, laboratory findings, radiology data, notes on problems and progress, billing records etc.

Concerning the structure of EHR, the first and starting point in EHR management is establishment of a unique identifier that will belong exclusively to one patient. The methods for



Clicking on this "i" brings up a monograph on the topic next to it.

The star takes the doctor to guidelines and other decision-support information.

This button leads to the patient's history of key parameters used to measure diabetes, such as blood pressure, hemoglobin A1C, and cholesterol.

Reeconders

- Patient is overdue for mammogram (rec: q 1 year). Pr[illegible] indicates average risk for breast cancer.

Flowsheets

Item Name	04/29/2008	09/13/2007	04/12/2007
BLOOD PRESSURE			
TEMPERATURE			
PULSE			
RESPIRATORY RATE			
O2 SAT		99*	
HEIGHT	70 in	70 in*	
WEIGHT		200 lb*	195 lb
BMI	28.8	28.8	27.9
PAIN LEVEL			

Medications

- Atenolol 50 MG (50MG TABLET take 1) PO Q3-4H
- Clobetasol PROPIONATE 0.05% LOTION 1 APPLICATION TO
- Colchicine 0.6 MG (0.6MG TABLET take 1) PO QD
- Cyclophosphamide 150 MG (50MG TABLET take 3) PO QD x 1
- Glyburide 5MG TABLET take 1 Tablet(s) PO BID
- Guaifenesin 600 MG + PSEUDOEPHRINE 120 MG 1 TAB PO Q
- Insulin LISPRO Variable (100 U/ML VIAL ML) SC
- Metformin 500 MG (500MG TABLET take 1) PO BID

Procedures

- Prolactinoma - 10/24/2007

Physicians

Notes

Patient E-mail

Problems

- Cardiac catheterization
- Diabetes mellitus type 2
- Neg h/o headache
- Vasomotor rhinitis
- Hypertension
- Coronary arteriosclerosis - Minor
- Pneumonia
- Arterial embolism
- FH*[Negative] birth defects
- FH diabetes mellitus type 2
- FH rheumatoid arthritis

Health Maintenance

Care Providers

Advance Directives

BMT Flowsheet

HM Form

Sticky Notes

Patient prefers to be contacted by email

Allergies

Allergen	Reaction
VARENICLINE	- GI Intolerance
PRAVASTATIN	- Hepatoxicity
LEVOFLOXACIN	- Made me feel weird
	- GI Intolerance
SILDENAFIL	- Hepatoxicity
Heparin	- makes me feel bad
	- Anaphylaxis
ALLOPURINOL	- Anaphylaxis
ACE INHIBITORS	- Rash
	- Angioedema

Visits

Pharmacies

Patient M/A List

Oncpro

Customize

2110 - 2140(0,187,31,0,1782,140)

Sticky notes serve as handy reminders to health care providers.

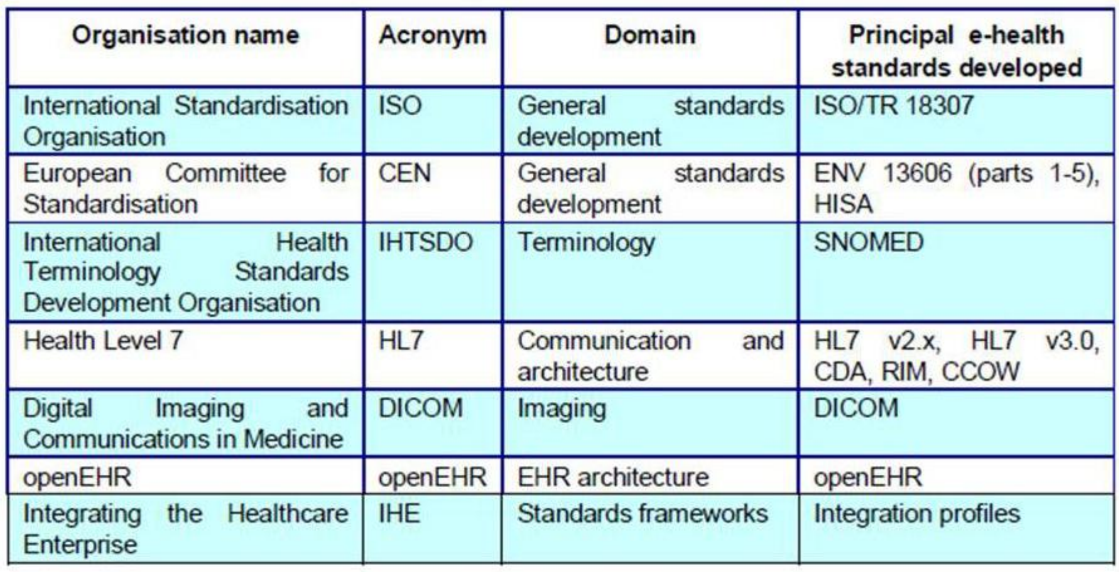
This section contains the patient's recent health maintenance data, such as mammogram results.

This section lists the patient's highly individualized responses to drugs, including known allergies that are potentially life-threatening.

Figure 1 An example of an EHR interface

specification of the patient identifier differ among the countries. The identifier is a digit calculated by a defined algorithm and generated from health insurance number, the social security number, or biometric characteristics of a patient (Kern, Fister, and Polasek 2009). This patent specific number will serve as a link to all other kinds of information related to the particular patient. Basic components[5] of an EHR are a patient's history, laboratory data, clinical findings such as biomedical imaging, diagnosis, medical procedures, and a diagnosis related group (DRG). The patient history is not a final product, but it should be available for a re-contextualisation in the light of new information during the patient's follow up (Kern, Fister, and Polasek 2009). Radiological data, medical images, and biomedical signals are usually recorded digitally, and described by a specialist. Diagnoses are coded according to the coding systems such as International Classification of Diseases – ICD-10 and SNOMED-CT, while drugs get usually coded by the ATC classification. Medical procedures are coded by a specific coding system, which often depends on the country and a particular purpose. A DRG is based on the diagnosis, age, and clinical data such as length of stay (hospitalisation), and treatment such as a surgical procedure.

As information contained within an EHR is heterogeneous and numerous, it is of crucial importance how it will be organised and structured. Structuring of information defines its usability. Certainly, the adequacy of represented information is a priority for both the health service users and the employees. A particular requirement of the frequent users of EHR, clinicians



Organisation name	Acronym	Domain	Principal e-health standards developed
International Standardisation Organisation	ISO	General standards development	ISO/TR 18307
European Committee for Standardisation	CEN	General standards development	ENV 13606 (parts 1-5), HISA
International Health Terminology Standards Development Organisation	IHTSDO	Terminology	SNOMED
Health Level 7	HL7	Communication and architecture	HL7 v2.x, HL7 v3.0, CDA, RIM, CCOW
Digital Imaging and Communications in Medicine	DICOM	Imaging	DICOM
openEHR	openEHR	EHR architecture	openEHR
Integrating the Healthcare Enterprise	IHE	Standards frameworks	Integration profiles

Source: empirica

Figure 2 Standards for Health IT Interoperability

[5] Figure 1, from http://www.hsph.harvard.edu/news/hphr/technological-and-computational-innovations/fall08healthrecord/index.html

and nurses, would be a well structured data entry and a user friendly interface (Kern, Fister, and Polasek 2009). However, as information about patient should be comprehensible beyond a non-local setting of a health care practice, i.e. available for re-use and interoperable with other healthcare systems, a work on standardisation related to EHR management have resulted in the internationally recognised recommendations. Figure 2 illustrates the most important standards and the organisations involved in development of the standards for interoperability among health care information systems.

As a significant part of information within EHR is expressed through medical language, a particular attention in the interoperability efforts has been focused on the standardisation of clinical terminology. Alan Rector in his paper *Clinical terminology: Why is it so hard?* presents the main reasons that have been obstacles for building of a 'common vocabulary', which is a precondition for the application of information technologies in the health care domain. According to Rector, the first reason why clinical terminology is hard to implement in EHR systems is its vast scale and the multiplicity of potential activities, tasks and users it is expected to serve (Rector 1999). As a result, an exponential explosion of diagnostic classes is expected. The second reason that makes terminology so hard originates in fundamental conflicts between the needs of users and the requirements for rigorously developed software. Namely, the terminology must be understandable to human healthcare professionals, expressed in their own language and so fit into the daily clinical routine. On the other hand, the terminology should behave in a 'rigorously predictable way for software engineers'(Rector 1999). For, only a rigorous classification can be expected to result in an accurate retrieval of information. The third reason that makes terminology hard is the complexity of clinical pragmatics that is an integral part of a practical use for data entry, browsing, and retrieval. A neglect of an appropriate understanding of clinical pragmatics has been criticised as a serious obstacle for implementation of IT in clinical practice (Wears and Berg 2005; Rector 1999). The fourth reason arises from the confusion of concepts and the words used to represent those concepts, i.e. confusion between the linguistic representations and the formal representations of concept[6]. Behind this lays a basic assumption of computer scientists that is expressed as a hypothesis of separability:

> For a clinical terminology, the representation of concepts and the relations between them can and should be separated from the linguistic knowledge about how these concepts are expressed in language and the pragmatic knowledge concerning how these concepts are used in dialogues with clinical users (Rector 1999).

The separability hypothesis seems to ask for a neglect of the pragmatics that has been previously criticised. However, the technical and computational needs for a formal representation of concepts justify the separability hypothesis (Rector 1999). Even so, it is questionable how the separability hypothesis and the inclusion of pragmatics can be reconciled. The fifth reason that makes medical terminology hard is that clinical conventions often do not conform to the usual logical or linguistic paradigms. Unusual linguistic construct that are established in clinical terminology do not fit a standard linguistic and logical interpretation. Therefore, there is a need for additional information on the terminology usage. Additional information should also define the cases in which terminological precision might be redundant. The sixth reason that makes terminology hard is that the formal concept representation, i.e. a logical form of the structured information, is a difficult task in itself. The seventh reason is a matter of terminological consensus. Health care professionals disagree among each other and achieving clinical consensus is difficult. One of the aims of the separability hypothesis is to minimise the difficulties which originate from a lack of consensus. However, as Rector points out, it is important to recognise in which areas a minimum level of consensus has to be established as a necessity. For example, a disagreement about which of the two terms, 'new growth' or 'malignant proliferation', has to be used in order to additionally specify the meaning of the term 'neoplastic' would be just a terminological disagreement. A mere terminological disagreement becomes redundant if there is an agreement that both terms can be mapped to the same concepts. On the other hand, in the cases where a

[6] A specification of what that means to represent a concept follows in section 1.2.

conceptual consensus cannot be reached, a formal representation can explicate the terminological similarities and differences. The eighth reason, which makes terminology hard, is a need for an alignment to the structure of existing classification systems such as ICD. The ninth reason lies in a coordination and coherence between terminology and models of the EHR. For example, information about 'local mastectomy' may be recorded as a surgical procedure, but a term 'local' may be assigned to the field for the extent (of mastectomy). As information should be recorded only once, a coordination of terminology with the EHR model is needed. The tenth reason, which makes clinical terminology hard, originates in the process of change in language. The changes may come from spontaneous changes in clinical practice as well as from the changes in the underlying system of concepts. A challenge for an accurate use of terminology within EHR systems lies in an incorporation of changes of language into the existing clinical records.

The ten reasons that make the terminology hard have additionally fostered development and standardisation of an underlying conceptual structure, which can serve as a stable ground for interoperability among EHR systems. Concepts are used as 'basic building blocks' rather than words, terms, or phrases (Cimino 1998). The 'basic building blocks' allow a multifunctional use of terminology, supporting representations in different languages, as well as easier quality evaluation.

1.2. Biomedical ontologies as a tool to represent reasoning

The following sections specify various aspects of biomedical ontology as a tool that is used to represent biomedical reasoning. I will first define the scope of my discussion by making a distinction between computational and philosophical understanding of ontology as a discipline. Then I will clarify the way in which biomedical ontologies classify and represent the selected domain by labelling the objects of interest. Concurrently, I characterise ontologies as a special kind of scientific representations, while distinguishing particular features that make the ontology representations so special. I conclude the chapter by considering ontologies in the context of knowledgebases, which are the final product of a collaborative effort of different interest groups involved in ontology building.

1.2.1. Philosophical and computational concept of 'ontology'

Within philosophical literature 'Ontology'[7] traditionally refers to the philosophical discipline which considers the modes of existence and the nature of reality. In simple terms, it tries to answer the question 'What is there?', as Quine would put it (Quine 1948). Hence, traditionally, Ontology belongs to the scope of metaphysics. Since Aristotle's time through the history of philosophy diverse methodologies have been employed as well as lines of thought in which Ontology was addressing its questions (Symons 2010; Quine 1948). Thomas Hofweber distinguishes four main directions within the Ontological debate (Hofweber 2011). According to Hofweber, Ontology might be understood as

> (O1) the study of ontological commitment, i.e. what we or others are committed to;
>
> (O2) the study of what there is;
>
> (O3) the study of the most general features of what there is, and how the things there are relate to each other in the metaphysically most general ways;
>
> (O4) the study of meta-ontology, i.e. saying what task it is that the discipline of ontology should aim to accomplish, if any, how the questions it aims to answer should be understood, and with what methodology they can be answered. (Hofweber 2011)

[7] I keep in line with the Guarino, Oberle, and Staab lexical distinction (introduced in §2) where an uncountable noun with capital letter 'Ontology' designates the philosophical discipline, while a countable noun 'an ontology' designates the computer scientist's use of the term the Guarino, Oberle, and Staab (2009)

Although meta-ontology has a scope of its own, taking a certain meta-ontological position determines philosophical position on the three other fields of Ontological studies. While some meta-ontological positions were framed within a meta-metaphysical debate, developing foundations of metaphysics (Chalmers, Manley, and Wasserman 2009), some other meta-ontological positions have led towards the rejection of metaphysics and Ontology in the traditional sense. For example, Rudolf Carnap's meta-ontological position, consisting in assembling philosophy as a discipline which should give a framework useful to scientists, led him to reject (O2) and metaphysics in general. According to Carnap, philosophers should ask useful and non trivial questions such as questions concerning the correspondence between linguistic framework and empirical reality. He argued that questions about the existence of abstract entities such as 'Are there numbers?' make sense only within a linguistic framework which has already set up the discussion about numbers, asking a question internal to the framework. Without its linguistic framework that question would be meaningless. However, within the framework, such a question becomes trivial (Carnap 1950). Therefore, in Carnap's view, questions of *the 'what there is' kind* in the metaphysical sense should be left out of the philosophical debate. Although Carnap's rejection of metaphysics and the (O2) kind of questions has been widely criticised (Quine 1951; Hofweber 2011) it demonstrates how a meta-ontological position can formulate an Ontological position within a philosophical debate.

In general, philosophical debate concerning any of the (O1)-(O4) questions is focused on philosophical problems related to, for example, the existence of entities denoted by linguistic terms, types and tokens, natural kinds, abstract entities such as numbers, fictive characters such as Pegasus, relations between particulars and universals, and the way in which the entities in question should be understood. The debates often tackle the questions about realism and the existence of the external world as well as of the mentioned entities.

Computer scientists are also using the term 'ontology' while considering 'entities', describing them and building relations among them. In addition, a useful ontology for computer scientists aims to be an accurate representation of the domain of interest. So, a similarity between the philosophical and computational use of the term ‘ontology’ is not a coincidence. In particular, a significant part of professionals working on the development of ontologies are logicians, and likewise some are philosophers who are dealing with the Ontological questions. However, as the use of the same term within philosophy and computer science domains has produced numerous confusions and misunderstandings, explicit distinctions have been made between the two uses.

> The word “ontology” is used with different senses in different communities. The most radical difference is perhaps between the philosophical sense, which has of course a well-established tradition, and the computational sense, which emerged in the recent years in the knowledge engineering community, starting from an early informal definition of (computational) ontologies as “explicit specifications of conceptualizations” (Guarino, Oberle, and Staab 2009).

Guarino et al. are particularly concerned with the distinction between computational ontologies and Ontology in the metaphysical sense, that is, addressing the question of ‘being qua being’, thereby trying to identify attributes of things that belong to them because of their very nature. Contrary to such an essentialist approach to the modelling of reality, Guarino specifies ontology engineering as a perspective driven approach. So, he says ‘[...] we refer to an ontology as a special kind of information object or computational artifact’ (Guarino, Oberle, and Staab 2009). Referring to the similar position previously presented by Thomas Gruber (1993), Guarino specifies that the account of existence in this case is a pragmatic one, i.e. for an artificial intelligence (AI) system, ‘what ‘exists’ is that which can be represented’ (Gruber 1995; Guarino, Oberle, and Staab 2009).

Had we, however, tried to develop a unifying framework which would allow an appropriate use of the term ‘ontology’ in both domains, reconciling philosophical and computational approaches while respecting all relevant differences, it could have been done by means of the specification of a meta-ontological position. Namely, ontology within the computer science domain can be understood as related to a particular meta-ontological position that is focused on the functional and representational aspects of entities that are computational artifacts, asking particular kinds of ontological questions. For, the computational meta-ontological position specifies what kind of ontological questions are to be addressed, avoiding confusion with the traditional metaphysical questions of existence.

While building an ontology, computer scientists and logicians are not asking if there are numbers, or whether there are natural kinds independent of the observer. What there is, for a computational ontologist, is that which has to be or can be represented (Gruber 1993; Guarino, Oberle, and Staab 2009; Boniolo 2012). The task of an ontologist is to develop a representation, written in a logical or programming language, which functions as a pragmatic tool designed for a particular purpose. Thus, from a technical point of view, designing an ontology of Conan Doyle's characters will not be of much different kind than working on a cell cycle ontology.[8] The represented entity of Sherlock Holmes, with the detailed representation of his traits, will not have an ontologically different status than the represented molecule included in the cell cycle ontology. Both represented entities will be a kind of (formal) representation. Moreover, the evaluation of the two ontologies will not be crucially different, despite the fact that the target systems against which the evaluation has to be done might be considered as very much different from an Ontological point of view. However, the questions of Sherlock Holms's existence are left to those philosophers who are interested in Ontology, while computer scientists and logicians who are developing ontologies have already enough real questions to address. These real questions concern the adequacy of ontology as related to the domain of interest, i.e. its usability

[8] In this respect I disagree with Guarino's note (2009, p. 2) that a discourse on ontology of fictional entities belongs just to the domain of a metaphysical speculation. While a discussion on the Ontological status of Sherlock Holmes would belong to metaphysics, a technical design of Sherlock Holmes' traits in a formal ontology has a wide source of references, such as Conan Doyle's novels and the related cinematography. For the purposes of a Holmes ontology design, the metaphysical questions of the existence of Sherlock Holmes need not be addressed at all.

as a representational artefact, in particular its purpose of supporting various formal reasoning tasks serving particular domain-specific interests, whilst the questions concerning the Ontological nature of the represented domain entities stay out of the computer scientists' scope.

I single out five questions as particularly relevant for an ontology design. The first question is what the aim of representation is. The second question asks for distinguishing who are potential users. Thus, it should be considered what kinds of demands are posed on the representation regarding the established aim and the users' needs. The third question is what kinds of features have to be selected and what kind of granularity has to be employed in the representations. The fourth one is how the features should be related in order to function best for the established purpose. The fifth question concerns evaluation of the ontology, i.e. how well the ontology represents its target and how efficiently it performs the defined purposes.

I consider these questions, which are relevant for the ontology modelling and evaluation, as primarily epistemic questions. Accordingly, the epistemological discourse concerning ontology engineering involves an analysis of the issues such as classification, representation, modelling, and evaluation. Even if the same questions can be addressed from an Ontological perspective, I leave such a debate out of the scope of my interest.

The very process of ontology design starts with observation. The target system that has to be represented should be carefully observed and analysed before the relevant features and the relations among them, which will eventually become the components of the representation, are selected (Guarino, Oberle, and Staab 2009). The relation of observation and representation in sciences has been profoundly discussed among philosophers (Hanson 1958; van Fraassen 1980, 2008; Hacking 1983; Boniolo 2007). I will bring some aspect of the debate in the section which considers representational aspects of ontologies. However, what is obvious is that the processes of observation and the selection[9] of relevant features involve collaboration among the ontology

[9] The selection of relevant features I separately address when I discuss how the aims of a research domain influence the labelling and classification (Sections 2.2, 3.4).

engineers and the domain experts such as molecular biologists and clinicians, who will also be the users of represented knowledge (Ekins et al. 2011; Leonelli 2008). Accordingly, an ontology is also defined as 'formal descriptions of shared knowledge in a domain', i.e. 'formal specification of a shared conceptualization'(Borst 1997). A shared conceptualization is of a particular importance for exchange of knowledge defined through the use and re-use of ontologies.

Like for the term 'ontology', there is a need for an additional distinction of how the terms 'concept' and 'conceptualisation' are used within the computational literature. Since concepts are conceived as constituents of thoughts, philosophical discussion about concepts mainly belongs to the scope of philosophy of mind, which is addressing the questions related to mental representations. However, within the computer scientists' literature, including the aforementioned definitions of ontology, 'concept' is mostly conceived as something that can be explicitly and inter-subjectively represented. The concepts that are objects of an individual mind are not of an interest for an ontologist as 'there must be agreement on the conceptualization that is specified' (Borst 1997). As soon as an agreement has been reached, a concept might be represented. Such a representation of a conceptualisation, however, is not a mental representation, but an inter-subjective representation that can get a linguistic, logical, or graphical form[10]. Therefore, we can speak about 'concept representation' as a technical term used in computational terminology to denote a symbolic representation of a concept that was agreed upon within a community.

In his paper 'What is an ontology?' Guarino specifies in which sense an ontology is a conceptualisation and how concepts can be represented by means of a formal language. He agrees that complete conceptualisation, which is in an individual mind, cannot be shared, but that what can be shared is an approximation of conceptualisation (Guarino, Oberle, and Staab 2009). While the individual conceptualisations, which are private to the mind of the people, are *implicit*

[10] The importance of an inter-subjective linguistic representation shared by a community was emphasised by Putnam (Putnam 1973) and critically refined by Dupré (Dupré 1981); for a relation of concepts and scientific linguistic representations see also (Boniolo 2001, 2007).

conceptualisations, shared conceptualisations can be made *explicit* by means of language. An agreement on a conceptualisation can be reached by an acceptance of a common domain as a universe of discourse (D), a shared vocabulary (V), and the relations (R) that hold among the objects[11] belonging to the segment of the world (W) which is in the focus of the discourse. The segment of the world that is specified through the discourse can be also labelled as a target domain, a target system (S).

> Two different agents (outside the observed system) will share the same meaning of "cooperating" if, in presence of the same world states, will pick up the same couples as instances of the cooperates-with relation. If not, they will have different conceptualizations, i.e., different ways of interpreting their sensory data. For instance, an agent may assume that sharing a goal is enough for cooperating, while the other may require in addition some actual work aimed at achieving the goal. (Guarino, Oberle, and Staab 2009)

However, an explicit specification of conceptualisation which is fixed by its extensions has been criticized as too case-dependent, and therefore unable to persist the changes within its extensions. Such a case-specific conceptualisation would not be open for future re-use. In order to avoid a specification of conceptualisation that is case-dependent Guarino introduces a formal specification that defines conceptual relation as an intensional relation (Guarino, Oberle, and Staab 2009). Accordingly, a conceptualisation is understood as an *intensional relational structure*.

For example, in the case of the electronic health records (EHR) the universe of discourse may contain breast cancer patients. As each of the patients has been assigned a personal identifier that is a digit, the domain of discourse would contain the set of related numbers, i.e. D={1111, 1112, 1113,...2111}. However, D for the EHR would need to include other elements such as a patient's history, laboratory data, clinical findings, biomedical imaging, diagnosis, medical procedures, and a diagnosis related group. Moreover, as the patient is followed by an oncologist, the doctors' name is also an element of D. The relevant relations would compose a set of

[11] The object is understood in the most general way as anything that can be thought of and thereby represented in a symbolic form.

relations, e.g. R={diagnosis, name, patient, doctor, followed_by..}. While some of the relations are unary conceptual relations (e.g. patient, doctor) others are binary relations (e.g. <patient, diagnosis>, <patient, name>, <patient, doctor>).

The represented concepts are ordered in hierarchies according to the degree of generality. So, both the concept 'patient' and 'doctor' belong to a more general category that is 'person'. Likewise, the health records assigned to a patient are ordered in a hierarchy of classes and subclasses. The class of 'clinical findings' would contain the sub-classes such as 'laboratory' findings, 'imaging' records etc. The 'diagnosis' would have a subclass such as 'neoplasm', which is further subdivided into various types of cancer. The modularisation of the represented concepts is important for 1) ordering of knowledge that has to be represented, 2) re-use of the represented information across different domains and for various purposes, 3) automated reasoning on the represented information.

However, not all of the approaches to the modularisation fulfil these three tasks in the same respect. The approaches that are more focused on a shared vocabulary have less formalism involved in the representation of concepts. Having the form of glossary and thesauri, such kind of ontology can fulfil (1) its task to a certain degree, but it will not be able to perform much in the way of (2) and nothing in the way of (3). On the other hand, ontologies with explicit specification of the conceptual relations that are related to a term employ formal representations that allow automated reasoning about the represented terms.

According to that what kind of language and degree of formalisation is used for specification of the represented information, different types of ontology can be distinguished. Michel Uschold and Michael Gruninger have characterised various approaches to ontology concerning the way in which the meaning of terms gets specified (Uschold and Gruninger 2004). The degree of how explicit the meaning-specification is corresponds to the degree of logical formalism used in the representation, making a continuum (figure 3).

The Uschold-Gruninger classification of the ontologies also corresponds to the degree to which the meaning of a term is made explicit by means of an explicit specification of the relational conceptual structure. So, terminology and thesauri specify the meaning of concepts only by means of the lexical definition. Even if a terminological definition is a kind of explication of a concept, the representation of a concept is expressed in natural language. Therefore, when represented as a terminology, the conceptual relations have only an *implicit* representation, i.e. the meaning of the term given through its definition gets representation only in the mind of a person who is assigning the meaning to the terms contained in the definition. On the other hand, within the ontologies in which the concepts are also formally represented the meaning of the terms is explicitly specified through the representation of the concepts and their relations with other concepts.

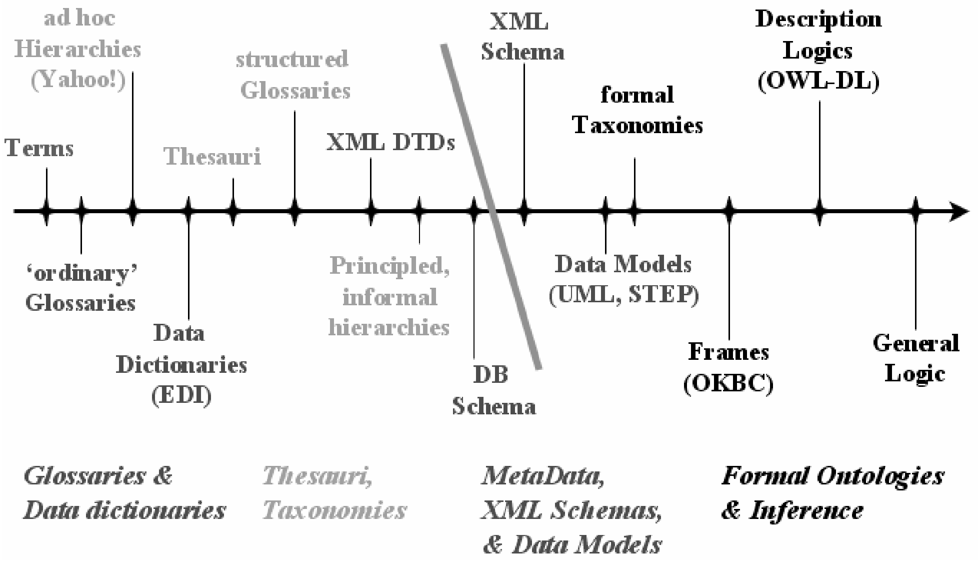


Figure 3 Kinds of Ontologies (Uschold, Gruninger 2004)

For example, the medical term 'hepatitis' can be defined as 'an inflammatory disease of liver'. But the lexical definition does not represent the meaning of the terms that compose the definition. Precisely, the semantics of the terms is implicit in the sense that it is left to the interpreter to assign the meaning to them by a cognitive process which assigns referents and relations to the terms. However, if the conceptualisation of the meaning of 'hepatitis' gets formally represented, the composite terms and the relations among them will be represented through the representation of an explicit conceptual structure. In description logics (DL), a fragment of first order logic which is often used as a formal language to represent ontological structure, the symbol 'Π' is used as the operator 'and', '∃' stands for the existential quantifier, and '≡' for the relation of equivalence. The conceptual representation of 'hepatitis' represented in DL have the form

Hepatitis ≡ Inflammatory disease Π ∃ has_location.Liver

The reasoning about the relations that hold between the definitions' constituents would be as following

> For example, the statement "Hepatitis ≡ Inflammatory disease Π ∃ has_location.Liver" (which uses the equivalence operator ≡) denotes that (i) every particular instance of Hepatitis is also an instance of Inflammatory disease that is located at some instance of Liver, and (ii) that every instance of Inflammatory disease that is located at some Liver is an instance of Hepatitis. Hence, in any situation, the term on the left can be replaced by the expression on the right and vice versa. Concepts that are defined using the equivalence operator (≡) are called fully defined concepts. (Schulz, Cornet, and Spackman 2011)

The logical form that represents reasoning about the concept can be also represented in the ontology languages which explicitly define the classes, subclasses and instances. Thus, the semantics of the represented concepts is provided and represented. One of the most frequently used languages to represent ontology reasoning is the Web Ontology Language (OWL). The instances of liver disease in OWL will be represented as the members of the class 'liver disease'. Another ontology class in OWL will be 'inflammatory disease'. Obviously, clinical knowledge confirms that not all inflammatory diseases are liver diseases. For example, influenza is an inflammatory disease that is not a liver disease. Likewise, fatty liver is a liver disease, but it is not an inflammatory disease. That means, when represented in OWL, the instances of 'liver disease' and 'inflammatory disease' classes will not fully overlap. But only those instances of the two classes which are at the same time members of both sets will be defined as the members of the ‘hepatitis’ class. In a similar way ‘liver’ as the constitutive term of the ‘hepatitis’ definition is represented by an explicit specification of its meaning. The concept related to the term ‘liver’ is represented within a hierarchical structure of body parts which have been assigned specific functional and anatomical features.

Within a formally represented conceptual structure, automatic inferences can be performed about the instances assigned to the classes. As a simple example, we can consider an instance ‘A’ as having been assigned an instance of the class ‘hepatitis’ by the relation ‘_hasHepatitis’. Since ‘disease’ class is a super-class of the class ‘hepatitis’, ‘A’ can automatically get assigned a membership in the class ‘disease’. So, it can be inferred that ‘A has a disease’. If ‘A’ belongs to the class ‘person’, it can be inferred that ‘A is a person who has disease’. Having in mind that ‘liver’ is defined as a sub-class of the class ‘organ’, it can be inferred that ‘A is a person who has a diseased organ’. This simplified inferential structure is just an example of reasoning with ontologies which can be applied in more complex contexts.

The illustrated characterisation of computational ontologies as specification of conceptualisation will be discussed from various aspects in the following sections. The first task is to explain the classificatory aspects of ontologies.

1.2.2. Biomedical ontology as a classificatory tool

This section examines in which way biomedical ontologies can be considered as classificatory tools. Classifications in sciences play various important roles. The first and perhaps the most important role of classification relates to organisation of knowledge. By means of classification, knowledge within a domain gets fixed, organised and specified[12]. Classificatory categories in that way may be communicated, compared and analysed. Accordingly, the choice of

[12] This issue, with the examples from molecular oncology, is addressed in more detail within the second chapter, particularly in the section 2.3.

classificatory units and classificatory principles reflects the objectives of a particular domain (Dupré 1981, 2002; Hacking 2002). A parallel between the ontology aims as introduced in the previous section and the classificatory aims of the sciences is more than apparent. Ontologies are used to represent the domain knowledge by means of labels and classes, represented in various degrees of expressivity and formalism (Uschold and Gruninger 2004). However, I will try to develop an analysis of the relations between biomedical ontologies and biomedical classifications in some more detail.

The biomedical domain consists of heterogeneous fields some of which are directly related to clinical management of patients, while some others focus primarily on the scientific research. As the aims of each of the domains have its specific demands, e.g. to represent, predict or explain particular phenomena, diverse classificatory systems have been developed in order to fit best the domain needs[13]. A need for the development of domain specific classifications originates in the very nature of classification as a methodology that is fixing the objects of interest into classes (Dupré 1981, 2002; Hacking 2002; Boniolo 2012). While some classifications need less detailed and more general classes (e.g. ICD which has been primarily used for billing purposes), some other classifications will show interest in the fine grained details for representing knowledge of interest (e.g. classification of genes and proteins).

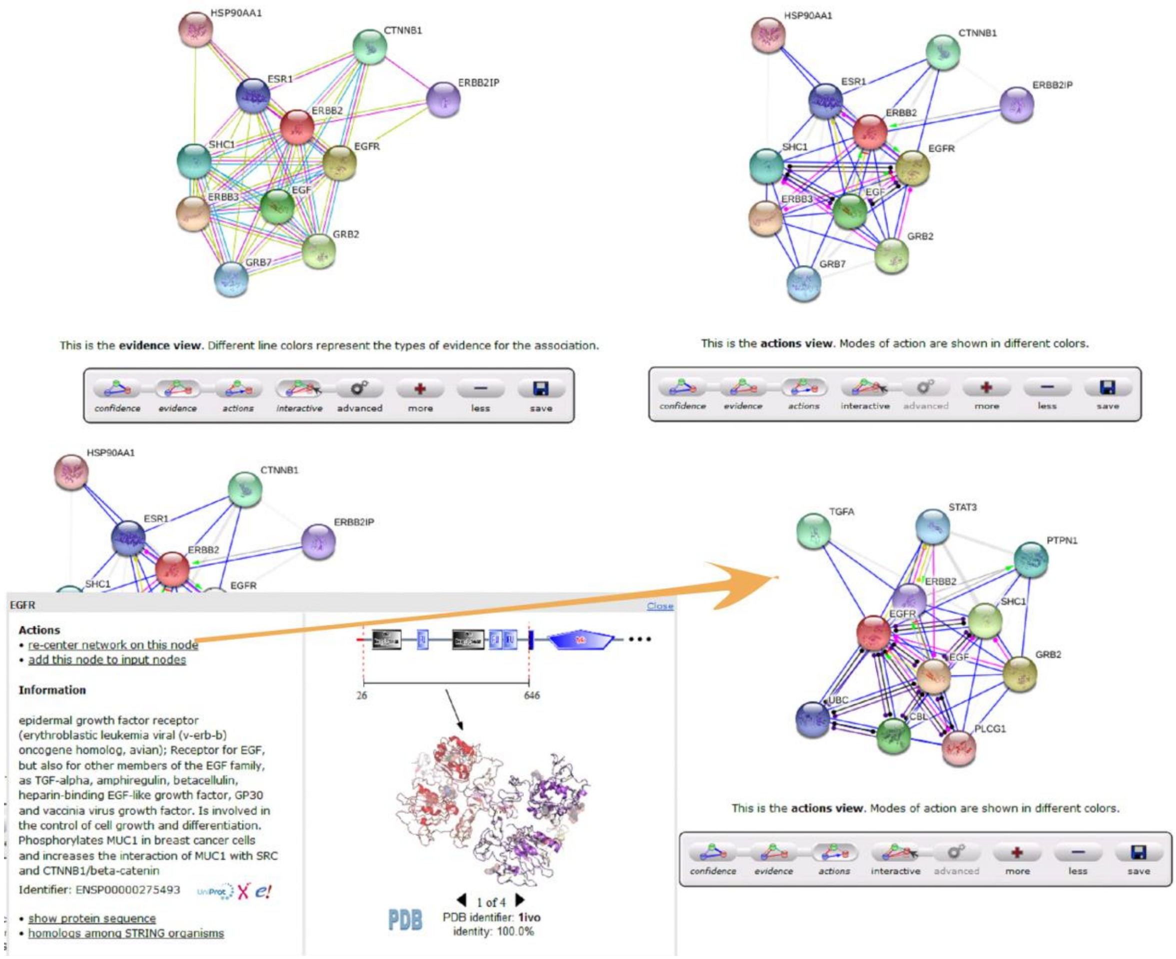
Corresponding to the demands of biomedical knowledge, biomedical ontologies are as heterogeneous as biomedical domains (Kutz 2011). The ontologies for the management of patients and clinical decision making will be designed according to the classificatory knowledge that is already applied in clinical practice. On the other hand, numerous domain specific ontologies have been developed as suitable to represent laboratory findings, genomic data, microarray data, RNA data, biological processes, procedures of biomedical investigation etc. For the aim of better descriptions of genomic investigations, the Genomic Standards Consortium (GSC) has introduced the minimum information about a genome sequence (MIGS) specification.

[13] A detailed analysis of the breast cancer classificatory systems is developed in the second chapter.

Such a specification serves a coordinated and standardised classification of genomic data. Requirements regarding quality of genomic data are not a novelty to scientific communities. Scientific publishing policies are the best example of how scientific results have to satisfy certain standards in order to be accepted as reliable data. However, computational technologies and in particular introduction of ontologies has played a significant role in the coordination of scientific practices (Ekins et al. 2011). This has been recognised as a novel and complex interaction of scientific knowledge and technology (Ekins et al. 2011; Leonelli and Ankeny 2012).

Before I move further to my analysis of ontologies as classificatory tools, I would like to stress that not every classification will consider the same kinds of objects as the objects relevant for classification. While clinicians classify patients, diseases and available therapy, biochemists and molecular biologist classify proteins, genes, and processes such as the relations among the interacting gene products. In the case of classification of protein-protein interactions, the recognised interaction will be the object of classification. The *interaction* among the proteins is, from an epistemic perspective, a different kind of object than *proteins* which are involved in the process. An epistemic goal focused on the representation of a protein may be accomplished by a representation of the protein structure. However, the interaction centred approach shows an interest in the processes and not only in the structure of proteins involved in the processes. The epistemic focus defines a process, e.g. protein-protein interaction, as the object of classification. Even so, protein domain structure can serve as a classificatory criterion, reflecting the classificatory aim but only in an indirect way. The protein structure has often been a classificatory choice because of its tractability and its association with protein function (Orengo et al. 1997). In addition, information on the structure can be a useful tool for predictions of possible protein-protein interactions, modelled and represented as the interaction networks (ibid.). Modelling networks of protein functions and interactions is a context sensitive process. Not only is it sensitive to the experimental procedure in the context of a particular organism, but it is also sensitive on the research questions. For example, The Search Tool for the Retrieval of Interacting

Genes (STRING)[14] database organises knowledge[15] on molecular interactions including over five millions of proteins (Szklarczyk et al. 2011). Knowledge about the interactions is organised by diverse classificatory criteria, so that a research question may get the relevant output through the query. The existing classifications of the model organisms, experimental systems, and the types of evidence are included in the database as a classification of the research questions presented into the queries.



This is the **evidence view**. Different line colors represent the types of evidence for the association.

This is the **actions view**. Modes of action are shown in different colors.

EGFR

Actions

- re-center network on this node
- add this node to input nodes

Information

epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian); Receptor for EGF, but also for other members of the EGF family, as TGF-alpha, amphiregulin, betacellulin, heparin-binding EGF-like growth factor, GP30 and vaccinia virus growth factor. Is involved in the control of cell growth and differentiation. Phosphorylates MUC1 in breast cancer cells and increases the interaction of MUC1 with SRC and CTNNB1/beta-catenin

Identifier: ENSP00000275493

- show protein sequence
- homologs among STRING organisms

PDB identifier: 1ivo
identity: 100.0%

This is the **actions view**. Modes of action are shown in different colors.

Figure 4 Diverse perspectives in classifying and representing protein-protein interactions (STRING)

Figure 4 shows how a question changes the representation of the interactions. The two upper figures represent the HER2 (erbb2) associated interactions in *Homo sapiens:* the figure on the upper left represents the network of evidence about the interactions; the figure on the right

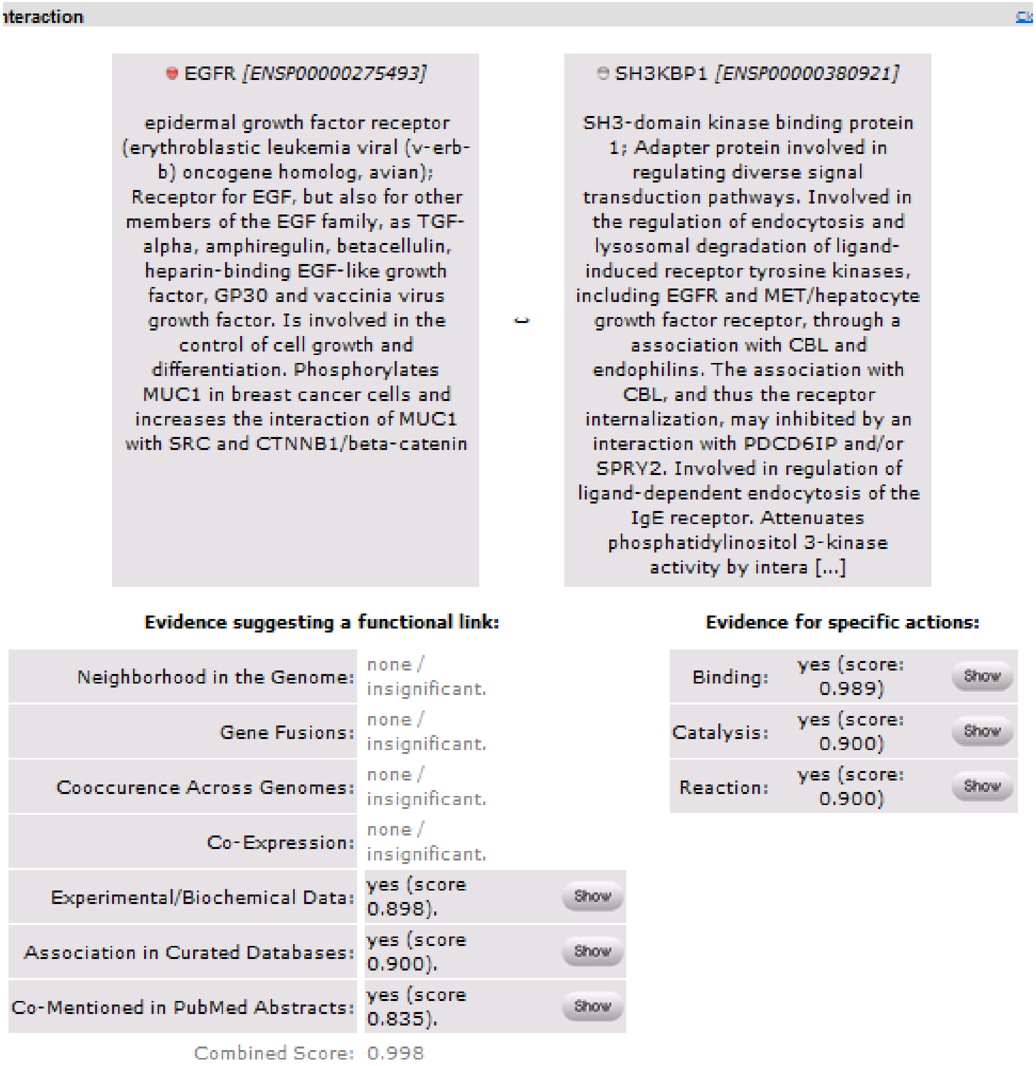
[14] http://string-db.org

[15] I do not go into the details of the STRING database structure, organisation, and maintenance. I rather use the STRING database as an example to illustrate a plurality of classifications, representations, and the research questions.

represents kinds of interaction such as ‘activates’, ‘suppress’ etc. Moreover, the lower two figures show how a specification of the question changes the representation of interactions.

If a researcher interested in ERBB2 needs more information about interactions among ERBB2 and EGFR, the network will be re-centred and represented in the way that differs from the two representations above.

The objects of classification in biomedicine are not only biomedical objects recognised as organisms, molecules, biological and molecular processes. The experimental procedures, methodologies, publications, and the various kinds of data used as evidence in assertion of scientific or clinical claims are getting classified as well. The STRING database presents one



iteraction

● EGFR *[ENSP00000275493]*

epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian); Receptor for EGF, but also for other members of the EGF family, as TGF-alpha, amphiregulin, betacellulin, heparin-binding EGF-like growth factor, GP30 and vaccinia virus growth factor. Is involved in the control of cell growth and differentiation. Phosphorylates MUC1 in breast cancer cells and increases the interaction of MUC1 with SRC and CTNNB1/beta-catenin

● SH3KBP1 *[ENSP00000380921]*

SH3-domain kinase binding protein 1; Adapter protein involved in regulating diverse signal transduction pathways. Involved in the regulation of endocytosis and lysosomal degradation of ligand-induced receptor tyrosine kinases, including EGFR and MET/hepatocyte growth factor receptor, through a association with CBL and endophilins. The association with CBL, and thus the receptor internalization, may inhibited by an interaction with PDCD6IP and/or SPRY2. Involved in regulation of ligand-dependent endocytosis of the IgE receptor. Attenuates phosphatidylinositol 3-kinase activity by intera [...]

Evidence suggesting a functional link:

Neighborhood in the Genome:	none / insignificant.	
Gene Fusions:	none / insignificant.	
Cooccurence Across Genomes:	none / insignificant.	
Co-Expression:	none / insignificant.	
Experimental/Biochemical Data:	yes (score 0.898).	Show
Association in Curated Databases:	yes (score 0.900).	Show
Co-Mentioned in PubMed Abstracts:	yes (score 0.835).	Show
Combined Score:	0.998	

Evidence for specific actions:

Binding:	yes (score: 0.989)	Show
Catalysis:	yes (score: 0.900)	Show
Reaction:	yes (score: 0.900)	Show

Figure 5 Classification of evidence (STRING)

example of how evidence can be classified. The example shows assignment of a confidence score to a protein-protein interaction, which is calculated by looking at the classes of evidence such as publications, curated databases, experimental data about specific interactions (which are again classified according to the type of interaction, e.g. binding, catalysis etc.). I will not go here into the analysis of social and epistemic factors that influence acceptance of evidence. The issue has already been addressed in a social context (Leonelli 2012c). I rather want to stress that classification in science also includes objects such as 'evidence'.

Namely, everything that can be communicated is classified in a more or less explicit way. An explication of classification, however, is a crucial demand of computational ontology. Represented terms, concepts and classes which aim to communicate certain information could not be processed by machines or comprehended by humans across research communities if classification had not been explicitly represented. The STRING database integrates knowledge acquired through the interpretation of heterogeneous experiments, data and literature mining techniques (Szklarczyk et al. 2011). Information stored and organised in model organism databases, or manually curated[16] functional classification scheme of the Kyoto Encyclopedia of Genes and Genomes (KEGG)[17] database can be integrated into one single database, which is unlike its sources in that it is focused primarily on the protein interactions. Such an integration of heterogeneous information is possible, in the first place, thanks to a shared vocabulary.

A shared vocabulary is a ground for 'light' kind of ontologies (Uschold and Gruninger 2004; Bodenreider 2008) which serve as the links to the related terms and information stored across distributed database systems. Terminologies are 'light' by means of a specification through a very general term definition, established across the research communities as a shared meaning. The defined meaning supplies a minimal condition for communication and linkage of the terms represented in one database with the equivalent terms represented in other databases. Regarding the context of model organism databases, the establishment of shared terminology

[16] For a description of the curation process see (Leonelli 2008, 2009b, 2012c)

[17] www.genome.jp/kegg

was recognised as playing a pioneering role in the development of ontologies (Leonelli and Ankeny 2012). However, terminologies are 'light' connections as they do not represent conceptualisation in a formal language, even if they can be mapped to the representation of concepts and datasets (Uschold and Gruninger 2004). Thanks to their generality, terminologies have some advantages over formalised ontologies. The main advantage of terminology is a high interoperability and re-usability across contexts. The relation of equivalence is a sufficient condition for connecting the related data sources. However, a disadvantage of 'light' ontologies is a lack of computational and inferential possibilities (ibid.).

An example of ontology that has a form of terminology which is enriched with 'is' and 'part of' relations is The Gene Ontology (GO)[18]. The GO has played an important role in a co-ordinated annotation of genes and gene products. The project of GO started in 1998 as a collaborative effort of three model organism databases, FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Informatics (MGI) project (Harris et al. 2004; Leonelli 2008, 2009b).

> Collaborating databases provide data sets comprising links between database objects and GO terms, with supporting documentation. Every annotation must be attributed to a source, which may be a literature reference, another database or a computational analysis; furthermore, the annotation must indicate the type of evidence the cited source provides to support the association between the gene product and the GO term. A standard set of evidence codes qualifies annotations with respect to different types of experimental determinations. For example, a direct assay to determine the function of the exact gene product being annotated is more reliable than a sequence architecture comparison (Harris et al. 2004).

By providing a controlled and shared vocabulary, GO contributes to communication and exchange of knowledge across communities (Leonelli 2009a). In that way, shared vocabularies also facilitate exchange of data associated with the terms. However, when clarifying what kind of classificatory tools ontologies are, we should keep in mind that GO, even if one of most frequently used

[18] http://www.geneontology.org

ontology sources, has a limited and very basic formal structure. The equivalence and parthood relations build a hierarchical structure that is associated with the datasets stored in a relational database. Besides the retrieval of the linked documents and data sources, very little or no inference can be automatically performed. A more specific information retrieval would ask for a cautious application of complex algorithms for data analysis (Yon Rhee et al. 2008). So, GO, as it is now, performs only one part of a high array of possible applications that an ontology can have (Bodenreider 2008; Kutz 2011).

When we speak about ontologies as classificatory tools, another distinction which has to be considered is a distinction between ontologies and databases. It is a distinction between a classification of data into the data structures and a specification of ontology classes. Within a database, data can be organised and structured even without the use of ontologies. Indeed, a long lasting tradition of database architecture concerns relational databases (Chen and Sidhu 2007; Martinez-Cruz, Blanco, and Vila 2011). Within a relational database, data are structured into the modules according to the classificatory criteria. However, a classification of data stored into a relational database significantly differs from an ontological classification. Data models are distinguished form ontology models. The main difference between the two consists in the locality of data models and their potential re-usability. While ontology design aims to a shared representation of conceptualisation that can be re-used in various contexts, classification of data into the relational database fulfils its goal within a local application.

> A data model, on the contrary, represents the structure and integrity of the data elements of the, in principle "single", specific enterprise application(s) by which it will be used. Therefore, the conceptualisation and the vocabulary of a data model are not intended a priori to be shared by other applications (Spyns, Meersman, and Jarrar 2002).

Section 1.2 characterises ontologies as an explicit specification of conceptualisation, which is a formal representation of the shared meaning. Accordingly, ontology as a classificatory tool can be understood as a tool that supports an explicit specification of what kind of objects are getting classified. Unlike the data structures which lack an explicit interpretation, the meaning of ontology classes may be comprehensible thanks to an explicit representation of the conceptual relations that are modelled.

> But, unlike task-specific and implementation-oriented data models, ontologies, in principle and by definition - see above - should be as much generic and task-independent as possible. The more an ontology approximates the ideal of being a formal, agreed and shared resource, the more shareable and reusable it becomes (Spyns, Meersman, and Jarrar 2002).

For its general and formal mode of classifying, ontologies are also said to represent knowledge that formally specifies agreed logical theories for an application domain.

> Ontological theories, i.e. a set of formulas intended to be always true according to a certain conceptualisation (ibid.).

Sabina Leonelli also defends a view on ontologies as theories, but drawing on the example of Open Biomedical Ontologies (OBO), she undertakes a less formal account of ontology as a *classificatory theory* (Leonelli 2012a, 2012b). However, what kind classificatory theory it is, remains open. In particular, it remains an open issue how Leonelli's view of ontology as a classificatory theory, which aims towards generality by providing common labels without involving (law-like) axiomatic statements (Leonelli 2012b), relates to the traditional view of ontology that concerns logical theory with a full axiomatic specification.

Michael Gruninger, for example, criticises ontologies which are specified as taxonomies or class hierarchies, without justification of the classification (Gruninger 2004). I hold that such a view of ontology as a statement of logical theory that demands the explication of the classificatory reasoning might be aligned with the view presented in (Leonelli 2012b). More precisely, had we understood an ontology as the agreed *logical theory* (Gruninger 2004; Spyns, Meersman, and Jarrar 2002), where the classes and the axiomatic relations that hold among them represent (formal) models of the theory, Gruninger's demand for justification of classification could be provided only from the empirical sources and the established biomedical knowledge. So, a ‘good ontology’ will be an ontology that satisfies only intended models of its conceptualisation, while a ‘bad ontology’ maps onto unintended models as well (Guarino, Oberle, and Staab 2009). Thus, such an ontological theory that is enriched with justification seems to have the status of a peculiar kind of a *hybrid theory* that combines logical theories and formal models with empirical knowledge, various domain-specific conceptualisations, scientific models, and theories.

Independently of the peculiar theoretical status of ontology, the importance of ontology as a classificatory tool is doubtless. Therefore, I clarify here the relation between ontology and classification, in order to stress the point that ontologies classify intensional structures. Figure 6 (Rector 2006) gives an example of mapping among an ontology model, a data structure model (applicable on relational databases), and an information model (e.g. represented in OWL). The Figure shows how ontologies can be applied and communicated within and across databases, while classifying different things. The differentiation between ontologies as models of meaning and other kinds of computational models, which has been made therein, clarifies the specific character of ontologies as classificatory tools. Data structures classify data within a database. The codes represent classified data from the data structure as well as the instances of the ontology

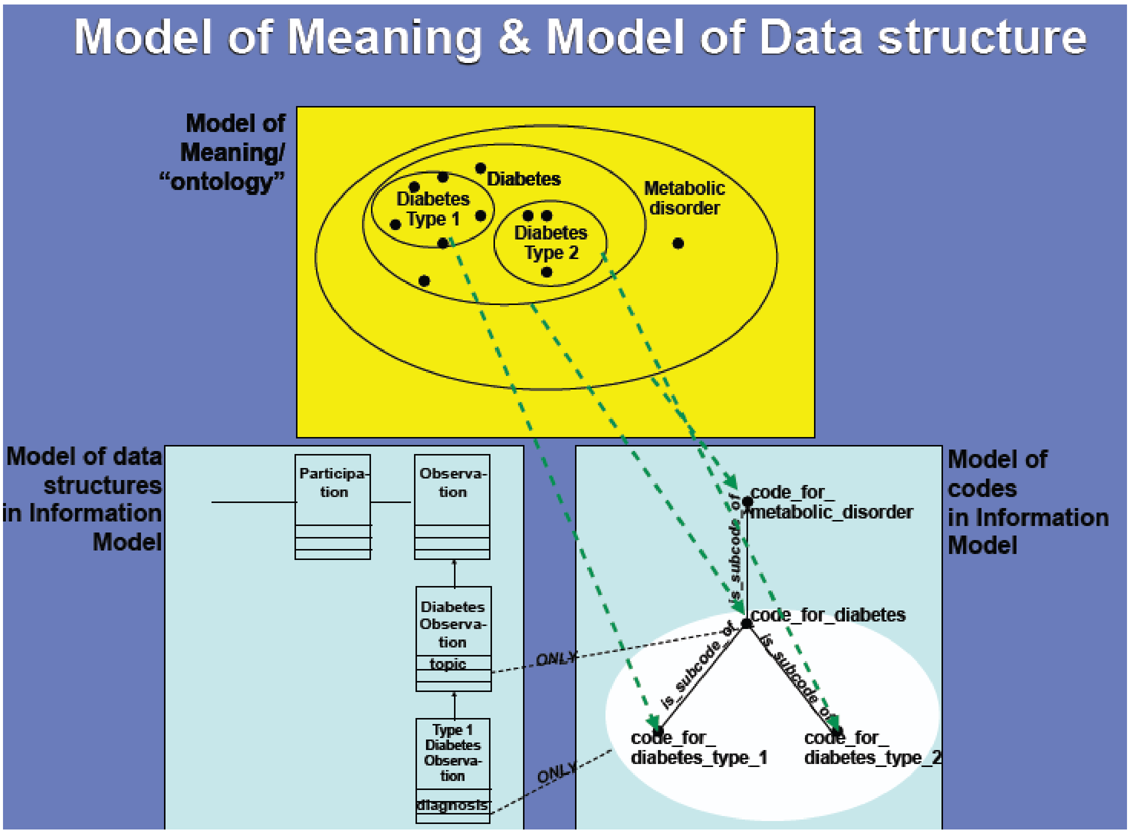

Model of Meaning & Model of Data structure
Model of Meaning/ “ontology”
Diabetes
Diabetes Type 1
Diabetes Type 2
Metabolic disorder
Model of data structures in Information Model
Participa-tion
Observa-tion
Diabetes Observa-tion
topic
Type 1 Diabetes Observa-tion
diagnosis
ONLY
ONLY
Model of codes in Information Model
code_for_ metabolic_disorder
is_subcode_of
code_for_diabetes
is_subcode_of
is_subcode_of
code_for_ diabetes_type_1
code_for_ diabetes_type_2

Figure 6 Distinguishing the ontology classification from other kinds of IT classifications (Rector 2006)

model. The instances represented as codes are computational artifacts of an information model, which can be implemented on the reasoning over data structures stored in a database. The figure shows how an ontology model classifies types and sub-types of a disease, representing its instances as the members of particular classes. So, what is an ontology classifying are the classes and instances of the concepts, represented as the terms with explicitly specified meaning within a conceptualisation. Such an ontology-classification can be captured by various representational means, e.g. a graphical (figure 6) or formal specification, each of which organises the agreed intersubjective and shared meaning.

1.2.3. Representing knowledge as labelling the classes

The previous sections have stressed the importance of shared terminology in knowledge representation and communication. The principles of ontology modelling have additionally emphasised the need for explicit specification of meaning which has just an implicit form in the terminology representation. This section provides an analysis of how is meaning made explicit in the ontology modelling thanks to the process of labelling. Moreover, it should clarify in which way an ontology model actually represents knowledge.

The meaning of a term is characterised as implicit if it is not explicitly represented (Guarino, Oberle, and Staab 2009). Implicit meaning belongs to the scope of individual cognition whereby a person is assigning the meaning to a term. More precisely, the meaning of a term is implicit as long as its semantics lacks an explicit representation which can be intersubjectively comprehended. Semantics of a term gets specified through an assignment of the term’s value, i.e. through a specification of what the term denotes. While a terminology represented as the lexical specification of a term through its definition directs the interpreter towards comprehension of the defined term within a cognitive and pragmatic[19] process of semantic assignment, the ontological mapping of the term to its reference explicitly represents the meaning by representing the mapping relation between the term and its reference. Therefore, an ontology succeeds in representing a term explicitly because the term's semantics is incorporated into the formal representation, thereby becoming an explicit representation of meaning.

The representation of semantic mapping-relation explicates the term's meaning. After reaching an agreement about the domain of discourse, the aim of representation, and a shared vocabulary, a team of ontology experts in collaboration with domain scientists may establish a shared conceptualisation of what has to be represented. The conceptualisation represented in an explicit form specifies the meaning of terms that belong to the domain of discourse by labelling the objects of reference. The process of labelling assigns the terms not only particular instances but also the reference classes. Thus, an assignment of a reference class to a term specifies a set of instances which the term denotes.

For example, the term 'breast cancer' within an ontology of breast cancer is a label which maps onto the classes of breast cancer such as 'luminal A', 'luminal B', 'HER2 positive' by a mapping relations '_is' and 'subclass_of'. What 'breast cancer' labels is a superclass which consists of subclasses labelled as 'luminal A', 'luminal B', 'HER2 positive' etc. Each of the classes contains instances which represent individuals diagnosed with the corresponding type of breast cancer. Therefore, every person diagnosed with breast cancer will be represented as an instance belonging to a particular class. There will not be instances of breast cancer that belong directly to

[19] Philosophy of language debates have stressed the significance of cases in which reference is fixed by what or who the speaker has in mind which need not coincide with the semantic reference of the expression that she is uttering to that effect. This distinction was first drawn by Keith Donnellan in his seminal paper *Reference and Definite Descriptions* (Donnellan 1966) with respect to definite descriptions, but others have also applied it to other terms. See, also Kripke, *Speaker's Reference and Semantic* (Kripke 1977). While Donnellan seems to think that both such uses of appropriate terms are a matter of semantics, Kripke seems inclined to think that the referential use, unlike the attributive one, is a matter of pragmatics.

the ‘breast cancer’ superclass, because every instance of ‘breast cancer’ will be a member of a ‘breast cancer’ subclass, which represents a particular type of breast cancer.

This kind of reasoning about ‘breast cancer’ corresponds to biomedical reasoning that has to be represented. For, any breast cancer in order to get characterised as breast cancer must be first classified by its type according to the established biomedical classificatory criteria. Without specification of the cancer type it would not be possible to make a biomedically justified claim that it is a case of breast cancer. While pre-diagnostic reasoning based on observed signs of breast cancer, e.g. ultra sound imaging, may produce a hypothesis for classification, the diagnosis of breast cancer is always given through a specification of the cancer type. Therefore, only because a cancer is characterised by additional histopathological and molecular analysis as a kind of breast cancer it is also characterised as a breast cancer. The issue of biomedical reasoning and breast cancer classification is addressed in the second chapter. However, this example illustrates how biomedical reasoning influences the representation and mapping of the labels onto the classes and instances within an ontology model. In that way, the labelling of classes represents established biomedical knowledge.

The semantics within computational discourse stays on the level of representation. What a label denotes is a computational artifact, a class or an instance. An instance within a computational model, as illustrated in the foregoing example, may represent an individual that is a real patient. However, the represented instance does not denote the individual within the representation. While the labels denote, the instances represent. On the other hand, the same instance may have been assigned a proper name that is a label which at the same time can denote the instance within the ontology as well as an individual represented as the instance of a ‘breast cancer’ class.

The process of labelling in ontology engineering plays an important role. Labels represented as the terms of a shared vocabulary may be comprehended by humans. Thanks to the computational techniques the labels and the labelled data can also be processed by machines.

The terms used in daily communication of knowledge serve as links for exchange of information. Also, by means of language it is possible to reach an agreement on a shared conceptualisation. The agents involved in the building of a shared conceptualisation communicate their own view on the problem, trying to find an agreement among each other. As a result, a shared conceptualisation uses terms to represent the agreed upon understanding of a domain.

However, not every representation of conceptualisation is of the same kind. A representation of the shared conceptualisation may consist of a map that represents the terms within a network of related terms (figure 7). Such a map is sometimes labelled as *semantic map*.

> Semantic mapping, also known as idea mapping, is used to explore an idea without the constraints of a superimposed structure. A semantic map visually organizes related concepts around a main concept with tree-like branches. [...] This technique facilitates communication between end-users and system analysts in support of information requirements analysis (Montazemi 2009).

A semantic map is a model which in an informal way represents how the terms within a domain of discourse relate to each other. It is a starting step in establishing a shared conceptualisation. However, such a map does not incisively specify the way the terms relate to each other or what are they the labels for. A map that is enriched with a semi-formal specification of the relations that hold among the terms is labelled as a *concept map* (Montazemi 2009).

> Concept mapping is a useful tool for organizing and representing concepts (events or objects) and their interrelationships in a particular domain. Each concept is designated with a label. The relationship between two concepts in a concept map is referred to as a proposition; propositions connect concepts to form a meaningful statement. Relationships between concepts are associative (ibid, p. 169).

The map which represents conceptual relations[20] consists of the terms which are the labels that denote the concepts. Within a concept map, as described above, the labels stand for the concepts, while the relations among the labels represent the shared understanding of what those concepts mean. So, 'cancer' will mean a type of 'disease' which can be further distinguished from other types of diseases. The statements represented through the relations that hold among concepts may be as detailed or specific as a domain and the aim of representation requires. For example, a conceptualisation within a clinical domain will ask for a representation relevant for clinical outcome and the treatment of patients. Therefore, information about aggressiveness of tumour will often be included into the representation.

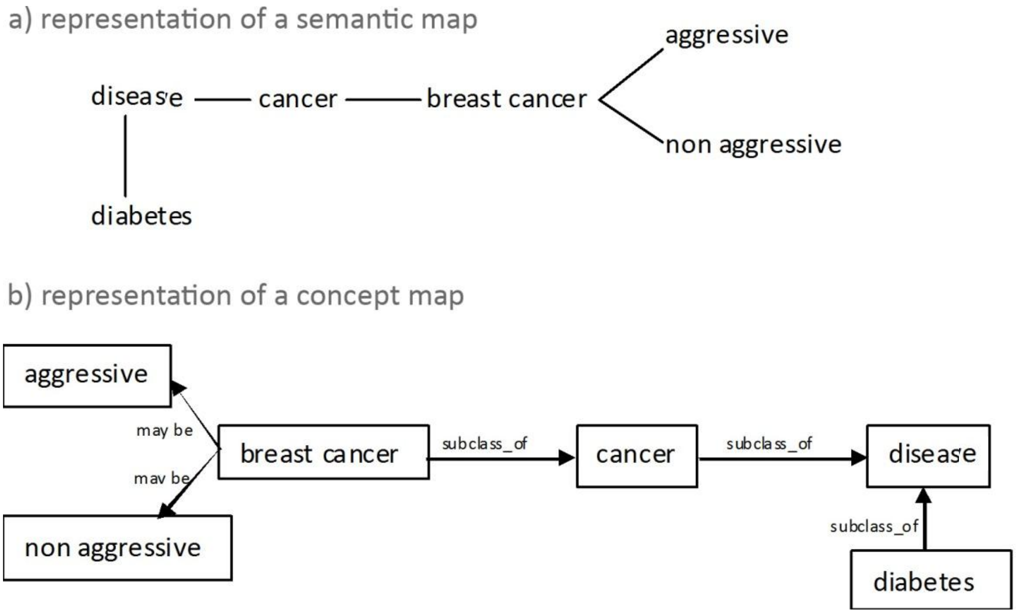


Figure 7 Representation of a semantic and a concept map

However, within a semi-formal representation of related terms which are labels for the concepts, represented semantic relations have an intensional rather than a denotative character. In other words, had we accepted the received view holding that labels denote the concepts (Guarino, Oberle, and Staab 2009; Montazemi 2009) we would also have accepted the view holding that denotation of the labelled concepts stays in the mind of epistemic agents. So, the

[20] Note that conceptual relations presented as a concept map provide a semi-formal specification of conceptualisation which differs from the formal definition of conceptualisation introduced in 1.2.1 and developed in Guarino et. al (2009).

concept as object of reference cannot be intersubjectively grasped. But, the agents can communicate their view of the meaning of a concept, establishing a shared view on how the concept should be represented. Thereby, the meaning of a concept can be represented through a description of its relation with other concepts that are labelled with agreed upon terms.

For instance, figure 7-b may be considered as a directed acyclic graph where the arrows represent an asymmetric 'is' relation among the descriptive and categorical terms that stand for the concepts with various degrees of generality. The relation is asymmetric as 'cancer' is a 'disease', but not every 'disease' is 'cancer'. Such a relation is descriptive as it describes 'cancer' as a disease, and 'breast cancer' as a type of disease which can be aggressive. The terms are categorical as they are general terms used in biomedicine to classify types of disease. The semi-formal concept map represents a conceptualisation where the related terms clarify the meaning of other terms within the network of related terms. However, within the map, the term 'disease' does not denote 'cancer', it just specifies that 'cancer' can be comprehended as a 'disease'. Such a map represents a shared understanding of how the terms should be understood by describing how they relate to each other. Even so, those terms are labels which semantics has not yet represented explicitly. They do not denote instances or classes, although they might be perceived as the models for representation of conceptual relations among the instances and classes.

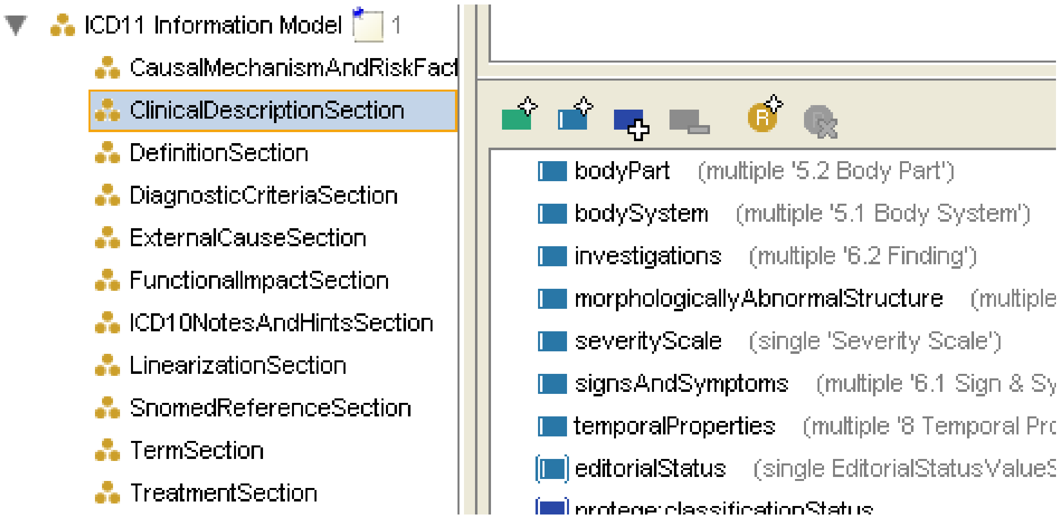
In other words, a representation of the labels gets gradually enriched with a representation of semantic relations, scaling the models according to their level of explicity from a semantic map, through a concept map to an ontology representation.

An ontology model represents the labels mapped onto the classes which the labels denote. In addition, all terms from a vocabulary (V) get assigned the semantic values that also belong to the domain of discourse (D) (see 1.2.1). Consequently, knowledge represented as an ontology is explicitly represented. Knowledge is represented explicitly as it represents the claims such as 'breast cancer may be an aggressive disease', accompanied with a justification for the mapping relations between the represented labels and classes. Therefore, the represented claims accompanied with the justification are the knowledge claims. Justification for the claim may be provided as a reference to the publications or experimental data that are evidence for the claim (Leonelli 2009a). The representation of the mapping relations also explicates the reasoning about the labels and the classes. An example of how biomedical reasoning gets represented through the labelling of breast cancer classes has been briefly introduced in this section. I will now give some more examples of how biomedical knowledge gets represented within an ontology.

International Classification of Disease (ICD) is one of oldest and most commonly used health care classificatory systems. Its first version, ICD-1, was realised in 1900 to record causes of death. Since then ICD has been going through revisions every ten years, while certain recommendations are realised yearly. The later versions have extended ICD applications to morbidity classification, oncology, primary care, and billing purposes. ICD belongs to the World Health Organisation's Family of International Classifications (WHO-FIC). The main purpose of ICD is to provide global standards for organisation and exchange of information about disease and health related problems (Fritz 2000; World Health Organization 2005). Following the goals, ICD provides standards that can be used at various health care delivery points, allowing interoperability and communication of information across the health care systems on the global level. While the last revision, i.e. ICD-10 realised in 1990, was not able to cover significant developments on biomedical and technological level, production of ICD-11, which should be released in 2014, includes development of ICD ontology and a dynamic revision platform (WHOFIC Network 2007). The introduction of the ICD ontology together with an organised web based collaborative effort which includes various stake holders makes a radical change in the approach to the updates and revisions of ICD classification. The aspect on which I will focus on regards a transparent and explicit representation of biomedical knowledge that is getting built into ICD-11, as it

> [...] will allow linking classifications and clinical terminologies through their proper knowledge representation i.e. the diagnostic formulations as formal operationalization of any diagnosis including signs, symptoms, laboratory findings etc. in standard vocabularies (WHOFIC Network 2007).

The information model of ICD-11 is a three-layer model. The first two layers compose the Foundation layer, which is divided into (a) the Ontology layer and (b) the Category layer. The ontology layer has been planned to get aligned with SNOMED, while the category layer contains the descriptions of ICD categories. The third layer is the Linearizations layer, which is a generalization of the traditional ICD. It specifies the inclusions, exclusions, and residual categories, and it also supports new uses for ICD. The Linearizations layer requires an automated generation of a linear representation of the information represented as a poly-hierarchy in the Foundational layer (Tu et al. 2010). According to the demands of a particular domain a specific linearization will be produced. For example, as primary care classification requires a less granular classification it may result in one output class code, while a morbidity report may have as a result two codes, e.g. influenza with pleural effusion (J10.1), influenza virus identified (J.10) (ibid).



ICD11 Information Model 1
CausalMechanismAndRiskFact
ClinicalDescriptionSection
DefinitionSection
DiagnosticCriteriaSection
ExternalCauseSection
FunctionalImpactSection
ICD10NotesAndHintsSection
LinearizationSection
SnomedReferenceSection
TermSection
TreatmentSection

bodyPart (multiple '5.2 Body Part')
bodySystem (multiple '5.1 Body System')
investigations (multiple '6.2 Finding')
morphologicallyAbnormalStructure (multiple
severityScale (single 'Severity Scale')
signsAndSymptoms (multiple '6.1 Sign & Sy
temporalProperties (multiple '8 Temporal Pro
editorialStatus (single EditorialStatusValueS
protege:classificationStatus

Figure 8 A visualisation of the ICD-11 information model in Protégé (Tu et al. 2010)

Protégé was selected as a platform to support the Content model[21] for ICD-11, and the Web Ontology Language (OWL) has been selected for its implementation. The OWL Content model formalises the three layer structure as a conceptualisation. Such a web based platform has enabled a collaborative workflow to support an interdisciplinary and a cross-cultural production of ICD-11.

[21] The WHO ICD-11 Revision Steering Group (RSG) appointed a Health Informatics and Modeling Topic Advisory Group (HIM-TAG) to develop the ICD-11 Content Model (Tu et al. 2010).

> In the Protégé implementation, an ICD category is represented as a class whose details are determined by a set of metaclasses. Each metaclass (e.g., a ClinicalDescriptionSection metaclass), groups a set of related properties (e.g., body part, body system, signs and symptoms, and severity scale) that an ICD category may have. By associating different metaclasses with an ICD category, we can flexibly specify different sets of properties with it (ibid.).

Figure 8 shows the Protégé interface for ICD-11, where a metaclass 'ClinicalDescriptionSection' is represented as a section that groups the set of properties, i.e. descriptive classes, used in clinical practice to characterise a disease. Thus, each category of ICD is represented as a class, which is specified by its metaclass, e.g. 'ClinicalDescriptionSection'. Each of the categories has also been assigned two types of terms. *Linguistic terms* describe the content of category in the model, and as the ordinary language terms they are placed at the text fields. On the other hand, the *references terms* belong to the ontology layer. The reference terms are specified by codes which are imported from the external terminology sources or other established ontologies (e.g. SNOMED-CT).

> Reference terms essentially represent coded information that expresses the meaning of a category in a computer-interpretable way. By contrast, linguistic terms are language-specific terms meant to help human users interpret the meanings of ICD categories (Tu et al. 2010).

Thus, the reference terms are the links that make a representation to be interoperable with other terminology and ontology representations. By linking the codes, the information and the sources related to the codes can be linked as well. This feature brings significant and novel advantages to the classification. While most health care classifications used to be designed for a narrow use, e.g. billing records or a very broad classification of a disease type, the new technology allows the enrichment of classification by various categories that have not been linked before.

Although ICD-11 does not employ complex formal ontology tools for the representation and reasoning, the Content model is still grounded on the Foundational layer, i.e. ontology and categories. For, ICD-11 is an example of explicit representation of biomedical knowledge that can be used in various settings. The classes and their labels are linked in the information model, while the justification for the mapping of the classes relies on the available published scientific evidence which is reviewed by the curators. Moreover, the process of ICD revision based on the web workflow platforms includes numerous experts coordinated around special task groups. The classification of oncology knowledge has been presented as one of the ICD-11 priorities, including genomic classifications.

1.2.4. Representing reasoning over the labels and classes

In the previous section I explained how biomedical ontologies represent knowledge through the process of labelling. The labels mapped onto the classes and their instances explicitly represent the shared meaning of represented concepts. The aim of this section is to clarify how biomedical ontologies *represent* reasoning. The representation of reasoning put forward in this section is addressed from a perspective of the debate about scientific representation. So, I characterise here the *representational features* of ontology as a particular kind of scientific representation, while I address the inferential processes associated with the representations in section 1.2.5 .

Throughout this section I extend my discussion from the previous section in order to show that the represented mapping relations within an ontology not only represent a term's meaning, but they also represent the reasoning about the domain. I specify reasoning about a domain as a primary target of ontology models. In that way I present the general criteria for understanding ontology models as representations of reasoning that map the domain of interest. By comparing the representational means and forces of lexical, pictorial, and ontological representations, I demonstrate particular features of ontology models, which explicitly represent the structure of reasoning about a domain.

On the one hand, I structure my argument by looking at the ontology modelling practice, which represents a term's meaning as a particular aspect of domain knowledge that maps and labels the classes (Guarino, Oberle, and Staab 2009; Gruber 1995). Namely, I analyse the reasoning of ontologists who map and label the classes within a particular context, for a particular purpose. The example which I give at the end of the section should clarify how corresponding representation, depending on the knowledge context and a pragmatic interest that the ontology is designed for, instantiates the reasoning employed to represent knowledge. On the other hand, I will be arguing that such an approach to represent knowledge actually results in *the explicit representation of meaning, which is also an explicit representation of reasoning.* In order to support my claim I invoke a relevant philosophical discussion on representation in science. I will also try to make a distinction between ontology representations and other kinds of scientific representation, claiming that explicitness makes an ontology a particular kind of representation. This is to say that unlike some other kinds of representation, an ontology *explicitly represents reasoning*.

The role of representation in science has been characterised as crucial for acquisition and organisation of scientific knowledge (Frigg 2006; Suárez 2010). Even so, philosophers widely disagree on how scientific representation should be understood. Namely, philosophical debates have been focused on various problems regarding representations in science, usually in the context of models in science (Knuuttila 2011; Frigg and Hartmann 2006). Accordingly, various accounts of scientific representation have been developed. For instance, some debates primarily focus on the representational, i.e. 'stand for', role of representation, questioning the character of the relationship between the representations and the world that the representation 'stands for' (see section 1.2.5).

The analyses of a dyad[22] between a model and its target (Knuuttila 2010; Suárez 2010) have involved an array of questions, ranging from the Ontological questions on realism (Chakravartty 2007; Frigg 2010) to the questions on idealisation, abstraction (Cartwright 1983; Boniolo 2007; Woods and Rosales 2010), or the relation of similarity between a representation and its target system, additionally questioning if 'similarity' and 'isomorphism' can be applied at all to such a relation (Knuuttila 2011; Suárez 2010; Suárez 2003). Regarding the issue of representational validity, the appropriateness of using the terms 'truth'(Mäki 2010), 'adequacy' (van Fraassen 1980, 2008) and 'successfulness'(Knuuttila 2011) for a representation have also been addressed.

I will return to some of the issues concerning representation in sections 1.2.5, 2.1 and 2.4, while in this section I rather focus on one of the approaches to scientific representation which I find the most relevant for positioning my argument to characterise ontology as the representation of reasoning. In what follows, I consider a pragmatic approach presented by Mauricio Suárez (2004, 2010).

In his paper *An inferential conception of scientific representation* (Suárez 2004), Suárez proposes a deflationary or minimalist strategy for understanding representation, which, according to him, gives a general account of representations in science. While the substantive approaches to representation are those which aim to characterise robust properties and relations between representations (i.e. 'sources' in Suárez's terminology) and targets, the deflationary approaches emphasise just *functional dependencies* of a representation and its target, within a particular context of inquiry[23]. The deflationary strategy nicely handles the problems linked with substantive

[22] Note that a representational dyad has been often conceived of as just one part of a more complex 'representational vehicle' (Suárez 2010), including an interpreter and/or other contextual and pragmatic elements (Suárez 2010; van Fraassen 2008; Giere 2010).

[23] The label 'deflationary' for the related account of representation in science comes from the analogue deflationary and contextual approaches to truth and knowledge in epistemology, as for instance in (Williams 1996; Wright 1992; Horwich 1998). A pragmatic approach to scientific representation which may be labelled as 'deflationary' is not completely new in philosophy of science. Some views that can be considered as historical predecessors of the deflationary approach are presented, for instance, in Boniolo (2007).

theories of representation (French 2003; Aronson, Harré, and Way 1995) that defend similarity (for short, *SIM*)[24] or isomorphism (for short, *ISO*) as the most appropriate account of the scientific representation. For instance, the deflationary strategy avoids a substantivists' demand for necessary and sufficient conditions[25] that a representation needs to satisfy. The argument that *SIM* and *ISO* cannot be a necessary condition for representation is grounded on the counterexamples which demonstrate that *SIM* and *ISO* fail in some cases of successful representation (Suárez 2003). For instance, like Picasso's painting Guernica, a mathematical equation need not contain similar or isomorphic relations to its target in order to successfully represent a targeted phenomenon. The arguments against the position holding that *ISO* and *SIM* are sufficient for representation is grounded on the criticism that *ISO* and *SIM* leave out the directionality of representation, which is rather a necessary condition for any representation to satisfy its role (Suárez 2003, 2004). Namely, with certain exceptions, a representation and its target can rarely be considered as symmetric. The *representation* represents its *target* only if the two stay in a representational relation whereby representation leads to its target. However, *ISO* does not assign any role to this kind of directionality. Suárez has presented a number of cases in which, without including representational directionality, the very representational function loses its meaning. The most convincing case might be exactly from the field that *ISO* is most confident with: The case of a differential equation which should represent a random walk motion in a phase-space structure, deflates the *ISO* account of representation as the equation actually represents another representation, i.e. the motion of the vector in Hilbert space that corresponds to the state of the particle (Suárez 2003). Moreover, the substantive accounts usually employ a reductive strategy in claiming that any representation can be reduced to the relation of isomorphism *ISO* and/or similarity *SIM* that holds among the representation and its target.

[24] The labelling of isomorphism and similarity, respectively, as [iso] and [sim] was introduced in (Suárez 2003). Instead, I use annotation *ISO* (for isomorphism) and *SIM* (for similarity), just as a matter of convenience in reading.

[25] The demand for necessary and sufficient conditions for scientific representation advocated by the similarity (a) and isomorphism (b) proponents may generally be presented as a) *SIM*: A represents B if and only if A is similar to B; b) *ISO:* A represents B if and only if the structure exemplified by A is isomorphic to the structure exemplified by B. For more details see Suárez (2003, 2004).

However, as Suárez (2004, 2010) points out in his argument from variety, the substantive accounts cannot deal with the problem of many possible relations, which may be suitably drawn between a representation and its target. For instance, a representation of tumour in an animal model, which is a particular kind of material representation (Morgan and Morrison 1999; Blasimme and Maugeri 2011; Ankeny and Leonelli 2011), may represent the anatomic structure of the tumour, the structure of the molecular components, the protein interaction pathways, the distribution of colours within its microscopy image, the physical forces that drive an asymmetric cell division. Which particular one of many possible representations the model will instantiate depends on the purpose the model will be used for. Obviously, *ISO* cannot cope with such a demand. Moreover, the deflationists avoid the problems of so-called 'logical argument', originally developed by Nelson Goodman (Goodman 1968). While considering representations in art, Goodman points out that a representational relation cannot be properly considered in terms of resemblance (i.e. equivalence) as it lacks the properties of being transitive, reflexive, and symmetric. The recent philosophical debates have revived Goodman's argument, applying it to scientific representations. So, Suárez, for instance, gives an example of an architectonic representation of a bridge, showing the exceptions in which the transitivity, reflexivity, and symmetry between the representation and its target cannot hold (Suárez 2004). Eventually, Suárez proposes an *inferential conception of representation*, specifying that

> [Inf]. A represents B only if (i) the representational force of A points towards B, and (ii) A allows competent and informed agents to draw specific inferences regarding B. (Suárez 2004, p. 773)

Such a characterisation of scientific representation is minimalistic in specifying only two very general, but necessary conditions that any scientific representation needs to satisfy. Namely, the representation A (source) needs to have *some* representational force, and a competent user (e.g. a scientist) can *use* representational forces to draw inferences about B (target).

Indeed, the inferential account of representation seems general enough to cover all kinds of scientific representations, yet, it is specific enough to distinguish them from non-scientific representations thanks to its demand for providing specific inference for a competent user. Thanks to this minimalistic and quite a pragmatic request, which does not ask for any special universally applicable relation (e.g. isomorphism) between a representation and its target, the inferential account seems to avoid the outlined problems that *ISO*, and *SIM* are struggling with.

However, some opponents of the inferential account may still argue that such a specification is not specific enough. Even if the inferential account grasps a general character of scientific representation, 'representational force', 'specific inferences', and 'competent user' remain unspecified. According to Suárez, every further specification depends on a particular context. However, without giving specific criteria of how context influences a representation, the very inferential account of representation is open to the objections of indetermination and relativism. Moreover, if we accept Chakravartty's criticism that the intended use of representation and its related triadic[26] form (Chakravartty 2010) actually conflates the representational means and ends[27], we can then criticize Suárez' account for precluding any specification of representation understood in terms of a final product. An implication of Suárez' view would be that representation is never finite because any representation may have unpredictably many uses. Therefore, the representational forces of a representation cannot be characterised or spelled out independently of a particular use. The opponents of Chakravartty's view rightly point out that a distinction between what scientific representations are and what we do with them cannot be easily, if at all, separated (Knuuttila 2011). Still, it seems that there is something more to say about representation. After all, a scientific representation is a product of modelling, designed for a particular purpose. With this in mind, independently of our acceptance or rejection of Chakravartty's point, we might try to provide some more specific characterisation of the representational forces and the related inferential potentials. My aim in what follows is to provide such a specification for ontology representations. In particular, I will show that ontologies

[26] The representational dyad of the source and its target is enriched with the intended use (an agent).

[27] Chakravartty critics distinguish between what scientific representations are and what we do with them (Chakravartty 2010).

are a special kind of representation because they capture ‘specific inference’ that brings into the representation (source) an explicit and intersubjective representational force. First, I will characterise certain features of ontology models that allow them to represent explicitly the targets.

I want now to specify ‘representational vehicle’ in ontology building. I consider representational forces not only as a final product that will be used by scientists, but also as an intentionally designed construction built into a representation. I explain how the context of ontology building influences design of the sources in order to represent its targets. I will show that the designing strategy constitutes ontologies as a special kind of representation, because it explicitly represents the related knowledge context, while representing reasoning about a domain. Thereby, unlike some other scientific representations, ontologies represent explicitly the contextual reasoning.

Still, I acknowledge a huge variety of needs in biomedical domain, which indicate a demand for deflationary treatment of representation in the ontology domain as well. Thus, I argue for a deflationary account of ontology representation that avoids objections for being undetermined and relativistic, because, as the following analysis shows, the pragmatic and contextual specification is an integral component of ontology representation.

The condition that representational forces should allow an inferential potential in competent users, proposed by Suárez as a necessary condition for any scientific representation, is also recognised in terms of ‘surrogative’ reasoning (Swoyer 1991; Contessa 2007; Suárez 2010). As Suárez points out, surrogative reasoning is a complex phenomenon of scientific practices employed in modelling.

> First of all, for the source to have this capacity it needs to be endowed with some internal structure: it must be the case that the source can be divided into parts and the relations between the different parts can be outlined. Secondly, the source’s parts and relations are in some way

> interpreted in terms of the target's own parts and relations. This is an implicit condition without which surrogative inference would be impossible. (Suárez 2010, p.19)

This means that the modelling practice distinguishes some 'parts' that are chosen to be represented, and the representation of the 'parts' should in some way correspond to the relations recognised in the target system. This condition, however, need not entail *ISO* or *SIM* as a pragmatic aim and the context of a research question may distinguish various kinds of parts and the relations to represent. The correspondence between the representation (of parts and their relations) and the parts and relations of the target can be perceived in more general terms as an *interpretation*, which has a pragmatic function in drawing inferences. Moreover, a set of norms that provide criteria of what should be considered as a valid inference of an interpretation is also context dependent (Suárez 2010, p.19, Boniolo 2007, p. 78, 188).

How does such a view of representation relate to the ontology models? Ontology models have been characterised as domain specific (page 23- 50). The context of a domain and the aim of an ontology model condition establishment of a particular conceptualisation (Guarino, Oberle, and Staab 2009). The experts collaborating on the establishment of a shared conceptualisation define the domain to model, a shared vocabulary, and the aim of the intended conceptual model (Guarino, Oberle, and Staab 2009; Ekins et al. 2011). While it can be argued that the aim and context influence any kind of representation (Giere 2010), the context in an ontology model seems to play a specific role, because through its designing process the context of modelling becomes a representational force in itself. I will now try to explain in what way knowledge context is a representational force in an ontology model.

The target of an ontology model may be characterised as twofold[28]. What is actually captured in an ontology is *knowledge* about how a domain is perceived by the domain scientists. Whereby capturing knowledge of domain-scientists about a domain-problem, an ontology model

[28] The distinction of a twofold character of the ontology model's target is just an analytical distinction as in the modelling practice the two appear as concurrent.

specifies and explicitly represents: 1) *how* the scientists conceptualise the domain; 2) what is knowledge about, i.e. *what* are the objects of particular interest to the scientists (see also 1.2.5). By means of capturing a shared conceptualisation of a knowledge domain, the model depicts the shared understanding (conceptualisation) of a domain, which is a representational target. Since ontologies primarily aim at capturing conceptualisation (1.2.1-1.2.3), I label the shared conceptualisation a *direct* representational target, while the objects of scientific knowledge will be an *indirect* target of the ontology representation.

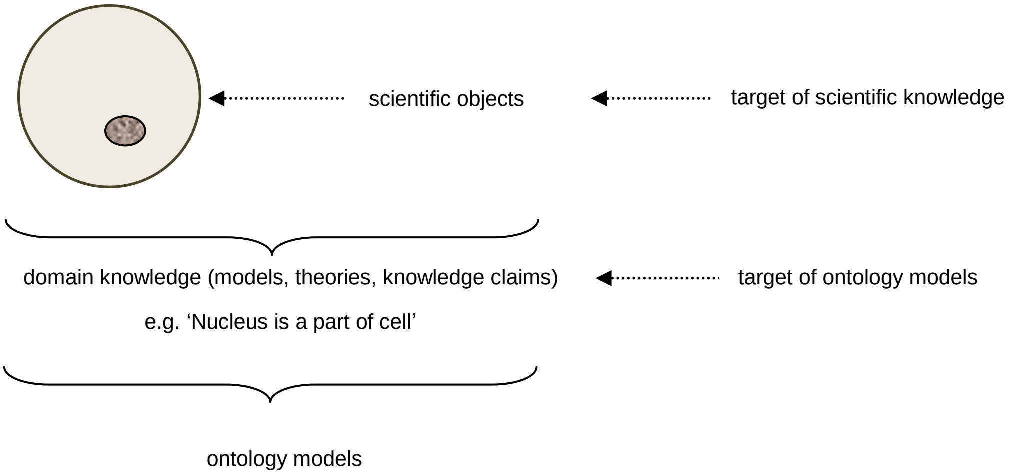

scientific objects
target of scientific knowledge
domain knowledge (models, theories, knowledge claims)
e.g. 'Nucleus is a part of cell'
target of ontology models
ontology models

Figure 9 Knowledge claims as mediators in representation: from scientific objects to ontology models

For example, a conceptualisation of a domain which considers that 'nucleus is a part of the cell' will be a direct target of an ontology model, while the objects such as nucleus and cell are just an indirect target of the model. The terms 'direct' and 'indirect' are in a way an arbitrary choice, as it can be argued conversely that a direct target of an ontology are the scientific (empirical) objects. Indeed, in a certain respect the modelling starts with the consideration of the scientific objects that are going to be represented. However, I decide to label conceptualisation of knowledge about a domain as a direct target as I want to stress the top-down approach of the ontology modelling, which goes from the representation of knowledge about a domain to the representation of the domain. Namely, the main reason for giving the priority to the top-down approach originates in the analysis of the modelling practice, wherby ontology engineers who are building models do not need to poses a scientific expertise that consists of detailed knowledge about some particular biological system. While building the models, the ontologists consult the domain-scientists and available literature in order to capture how the scientists think about the problem of interst.

Thus, an ontology is a result of collaborative effort and it captures a shared conceptualisation of a domain, which is usually communicated through the knowledge claims. While an ontologist is considering a conceptualisation to represent domain knowledge, the knowledge of how a domain may be conceptualised is a direct target of the intended model, while the content of what the conceptualisation is about constitutes an indirect target of the model. In other words, ontology modelling considers that what is believed to be a justified claim about a domain. A claim 'Nucleus is a part of cell' or a claim 'HER2 overexpression characterises an aggressive breast cancer phenotype' are the claims that play a role of mediators in the ontology modelling (figure 9). Those claims communicate a shared view, i.e. the understanding of a particular domain that should be represented. Of course, these claims, as presented here, can be also considered as a lexical (for short, *LEX*) representation (of knowledge)[29].

The represented claims also contain 'parts' or the terms which represent 'objects' that knowledge is about. Consider the terms 'cell' and 'HER2' that represent[30] the objects of scientific knowledge, characterised as the empirical objects, and described through a set of features. The ontology modelling does not consider directly these kinds of objects, but rather the features of the features used to describe the objects of a scientific domain. While biologists do not care so much if the cell should be represented as a concept, a type or a relation, an ontology modelling group is attempting to specify the meaning of the 'parts' of scientific claims, which are used to describe a domain. So, for an ontologist who wants to conceptualise and represent a domain, the

[29] It is questionable if a claim can be considered as a knowledge claim out of the context where an agent assigns to it a belief accompanied with its justification. That would be another important difference between a lexical representation of a claim and its ontological representation. An ontology represents the conceptual structure of the claim that is believed to be justified (established as a shared conceptualisation), whereby the sources of evidence can also be included in the representation. The evidence for the claim may be represented as another ontology representation or it can be given through the linked references to the evidential sources such as the research papers or the curators' statements.

[30] The section 1.4 focuses on a denotative force of representation, including lexical representations.

understanding of how the scientists think about a domain will be important contextual information, which will influence the resulting domain representation. Of course, besides a variety of knowledge communication channels, one of the most frequently used means to communicate knowledge are the lexically represented claims.

Therefore, we shall consider how a lexical *LEX* representation of knowledge, i.e. a knowledge claim, and a representation of an ontology model (for short, *ONT*) relate to each other. While the represented 'parts' of a *LEX* source are interconnected by means of an implicit representational (linguistic) structure, *ONT* explicitly represents the conceptual relations that hold among the represented parts. For the ontology model explicitly represents a shared view on the 'parts' meaning, representing the concepts and their relations (1.2.3). While the terms are a constitutive part of *LEX*, the terms are just labels that denote represented concepts in *ONT*. It might be argued that the same holds for *LEX*, i.e. the terms are just the labels that denote concepts. But the difference among the two is that the *LEX* related concepts stay in the mind[31], while the *ONT* related concepts are represented within *ONT*. The *LEX* terms get an assigned meaning within a cognitive process whereby an individual mind 'represents' the concepts. Likewise, *LEX* term's semantics gets specified within a cognitive process of the term's interpretation, while the *ONT* semantics is explicitly defined when an ontology gets applied to a domain. This difference in the representational means and targets entails a difference in representational forces of the two kinds of representation.

I continue my characterisation of the differences between the *LEX* and *ONT* representational forces by quoting Suárez' specification of the forces as a capacity of the sources:

> [For] the source to have this capacity it needs to be endowed with some internal structure [...]
>
> (Suárez 2010, p.19)

[31] The cognitive representation of concepts is a huge topic in the philosophy of mind. *Inter alia*, a connection between the linguistic and cognitive representations has been discussed, for instance, by Woodfield (2007).

We may examine ‘the capacity’ of *LEX* and *ONT* sources, comparing how the representation of the internal structure empowers its representational forces. The aforementioned specification has shown that the *LEX* ‘parts’ and their relations manifest a deficiency in explicit representation of the ‘internal structure’. However, a high successfulness of the claim ‘Nucleus is a part of cell’ to communicate information and support inferences cannot be neglected because a lexical form is one of the most common means used for knowledge exchange. Given that the ‘internal structure’ of *LEX* does not represent explicitly the relations among the ‘parts’, the question is what brings an apparently strong inferential force to this particular *LEX* source. The answer is more than obvious. Our knowledge of the ‘language game rules’ enriches this source with a strong inferential force. We can quickly grasp what might be the ‘parts’ of the claim, what the parts might denote, and what kind of (mereological) information they communicate. However, this inferential capacity is a capacity that an interpretation brings into the source. The inferential force of the source is not obviously an inbuilt feature of the source. Rather, the force would be forceless without the extra-linguistic rules that bring to the *LEX* source its ‘internal structure’. That is to say, the internal structure of *LEX* is external to it. Take a reader who does not understand the English vocabulary, the grammar, or the meaning of the represented terms. He might be simply considered as an incompetent user. Surely, he might rightly appear incompetent as he is not familiar with the particular ‘language rules’. On the other hand, he might be a very competent scientist, even an expert in the cell biology. However, despite his expertise the representational force of *LEX* will not be accessible to him. That does not mean *LEX* loses its representational force in principle, but it shows that in this particular case, without the acquaintance with the external rules the internal structure and the representational force of *LEX* source stays obscure. Contrary to *LEX*, the conceptualisation which *ONT* source represents does not depend on knowledge of extra-linguistic rules, at least not in the same sense as *LEX* does. The *ONT* source represents a shared understanding of a domain, and which particular language will be used to label the represented concepts becomes irrelevant for the very representation of the source. The represented concepts, associated with the terms as the labels may equally well be associated with the terms of any possible language. For, the synonymous terms coming from various languages might be mapped onto the same concept represented within an ontology. Moreover, 'the internal structure' of *LEX* is indeed an explicitly represented structure that constitutes *ONT*. The syntactical rules that vary among many natural languages will be irrelevant for *ONT*, as the structure represented in *ONT* primarily depends on the understanding of the domain. The *ONT* structure represents the relations among the represented concepts of a shared conceptualisation. Thus, in the case of an informal or a semi-formal model of an ontology, graphs can function as a convenient tool for representing a conceptual model (*ONT* in figure 10). Such an *ONT* source will preserve its internal structure independently of the syntactical variety through which a related *LEX* can be expressed.

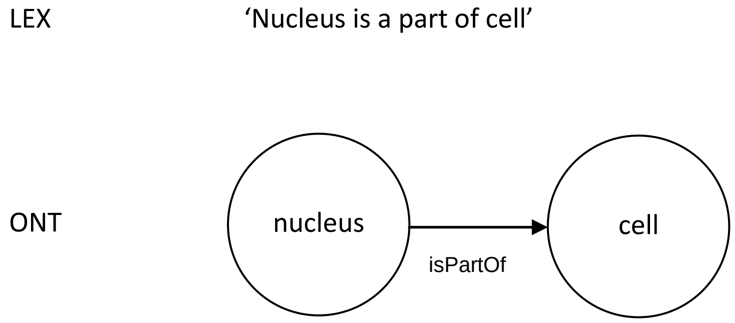


Figure 10 Comparing the representational means of *LEX* and *ONT*

Hence, a lexical and an ontology representation of the same claim differ in, at least, three aspects. They have different representational targets, means of representation, and a diverse degree of explicitness. Accordingly, the representational forces vary among the two representations. Before I get further into distinction of *LEX* and *ONT*, I will briefly introduce a third kind of representation.

Iconic or pictorial (for short, *PIC*), representations are frequently used in science. So, knowledge that 'Nucleus is a part of cell' may be represented by means of a picture, depicting nucleus and cell. The picture of nucleus drawn within the cell would represent a mereological relation of parthood. Such a representation would satisfy the inferential criteria of having a representational force by representing the parts (nucleus and cell) in a certain relation. Thereby it allows us to infer that in a biological, i.e. empirically observed target (mediated by a microscopy picture of the cell), the nucleus is positioned within the cell. The iconic representations belong to a wider group of visual representations, and a microscopy image is just one of them. The microscopy images usually represent a heterogeneous structure (figure 11) which may be dissected into the parts. Some parts, i.e. the segments of the image, will be labelled as ‘nucleus’, while some other segments will be labelled as ‘cell’. In *PIC*, the represented parts employ a certain degree of abstraction[32] from a microscopy image, as many observed segments of the image are ignored in a simplified representation of the nucleus drown within the cell (figure 11).

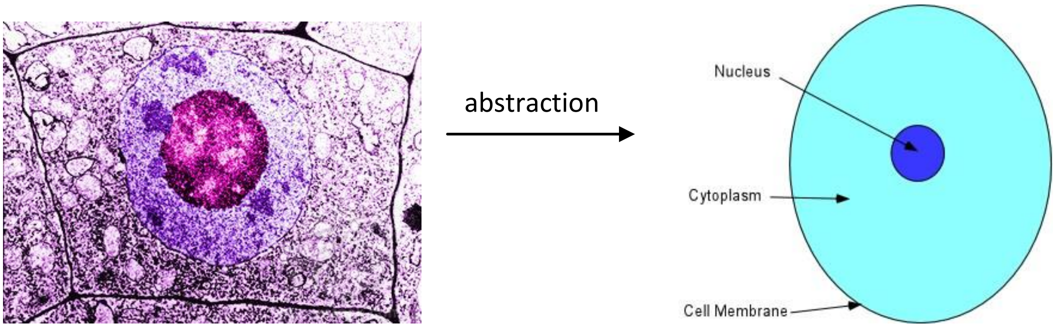


Figure 11 Abstraction from a microscopy to an iconic representation

In a certain respect, the microscopy image corresponds to *PIC*, but it also corresponds to the *LEX* and *ONT* representation. Namely, all three representations distinguish ‘nucleus’ and ‘cell’ as the parts which stay in a certain relation. Moreover, *LEX*, *ONT*, and *PIC* have a common effect of their representational forces providing the same inference: Nucleus is a part of cell. The represented relational structure of the sources’ parts allows inference to the target, an empirical phenomenon, usually labelled as a ‘biological entity’ (the cell)[33]. Some other, more complex, representations may represent hypotheses or theories (e.g. cancer stem cell), providing scientists

[32] The role of abstraction in science has been profoundly discussed, see for instance (Woods and Rosales 2010; Cartwright 1983). I do not go here into the debate, but I stay in line with the received view on the heuristic and pragmatic roles that a simplified representation of a system can have.

[33] Inference from a representation such as a microscopy image has been criticised in the context of the debate about the observable and unobservable in science (van Fraassen 1980). Without going into the related metaphysical discussion, I accept the view that I find epistemologically justified, i.e. a microscopy image is an observation, which allows inference about the observed phenomenon (Boniolo 2007).

with an inference about the target, which can be further tested through various experimental strategies.

Having introduced lexical, pictorial, and ontological representations, which employ specific representational means to represent a target, I will further compare how these various representational means influence representational forces[34]. In the previous sections (page 45-50) I have already made a distinction between lexical, semantic map, conceptual and ontology models, where they differ according to the degree to which the semantics of the represented terms was made explicit. I now extend my distinction to the pictorial model in order to clarify how these various models represent reasoning.

In order to compare the inferential representational force of *LEX*, *ONT*, and *PIC*, I will start by drawing a map (Figure 12), which represents how these representations interrelate. For

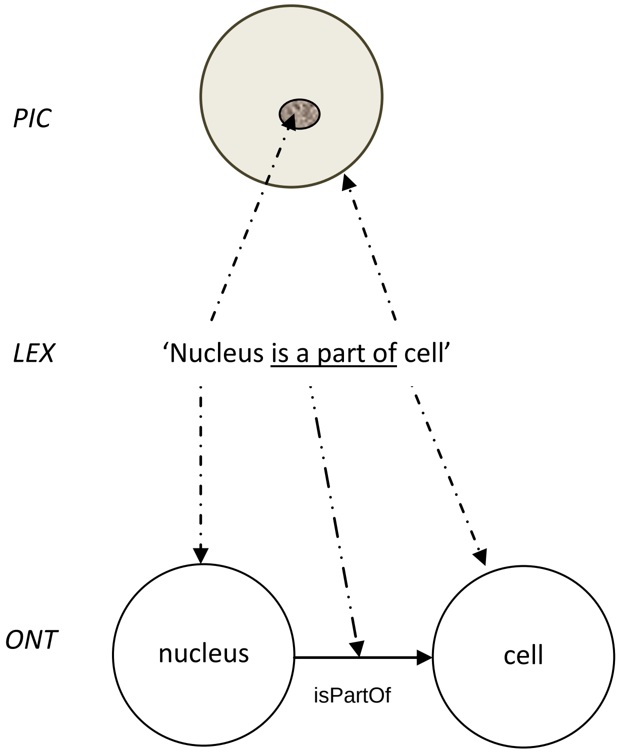


Figure 12 Mapping of *PIC*, *LEX*, and *ONT* representations

[34] While section 1.4 focus on the distinction among various kinds of representational forces, I consider them here mainly in the context of representation of reasoning.

reasons of simplicity, I use here just a simplified version of *ONT* that is a semi-formal representation, which may be further formalised.

Figure 12 shows how *LEX*, *ONT*, and *PIC* may be cross-related thanks their common target, i.e. the objects of scientific knowledge that are represented as the 'parts' within each of the representations. However, as the representational means differ, the internal structure, which represents how the parts are related, differs as well. The structure within *PIC* uses visual means to represent the mereological relation between 'nucleus' and 'cell'. Likewise *ONT*, the represented *PIC* relation does not depend on the 'language rules', but it stays universal across the languages by means of which the labelling of the parts will be performed. However, thanks to its generality, the represented *PIC* structure may equally well serve the purpose of representing any kind of phenomena that have the same relation. For example, the smaller circle may represent a domain of cancer research which is a subset of the big circle, e.g. the life sciences research. Only by the process of labelling of the *PIC* components, the structure will represent the aimed domain, i.e. nucleus as a part of cell. Thus, the representational force of *PIC* highly depends on its interpretation in a particular language employed to communicate knowledge of a domain. However, its structure does not depend on that which particular language is used for the labelling. English, Italian or German can be used interchangeably, while without making any change to the source its representational force stays preserved.

On the other hand, *LEX* is embedded in a chosen natural language. Depending on that which language is used in *LEX*, the source will change. The represented terms will vary as well as the syntactical structure. However, the representational force of *LEX* does not ask for the same kind of interpretation as *PIC* does. The interpretation of *LEX* does not need additional labelling as the represented parts are already labels. However, the interpretation of the labels depends on knowledge of the particular linguistic rules.

It is also worth briefly noting another difference among the sources that concernes the 'stands for' role of representation. The *PIC* may be considered as a simplified image of its target, while neither *LEX* nor *ONT* can. For, *PIC* functions to represent how we imagine a target, while *LEX* and *ONT* represent how we think about a target. Of course, what we imagine and how we think of something is not mutually exclusive, but it is rather a complementary cognitive ability (Woodfield 2007), which is not my focus here.

Finally, we may get back to the ontology models that represent knowledge about a domain, while summarising what makes the ontology representation to be a particular kind of representation. The ontology representation explicitly represents reasoning by means of representing the relations among the represented concepts.

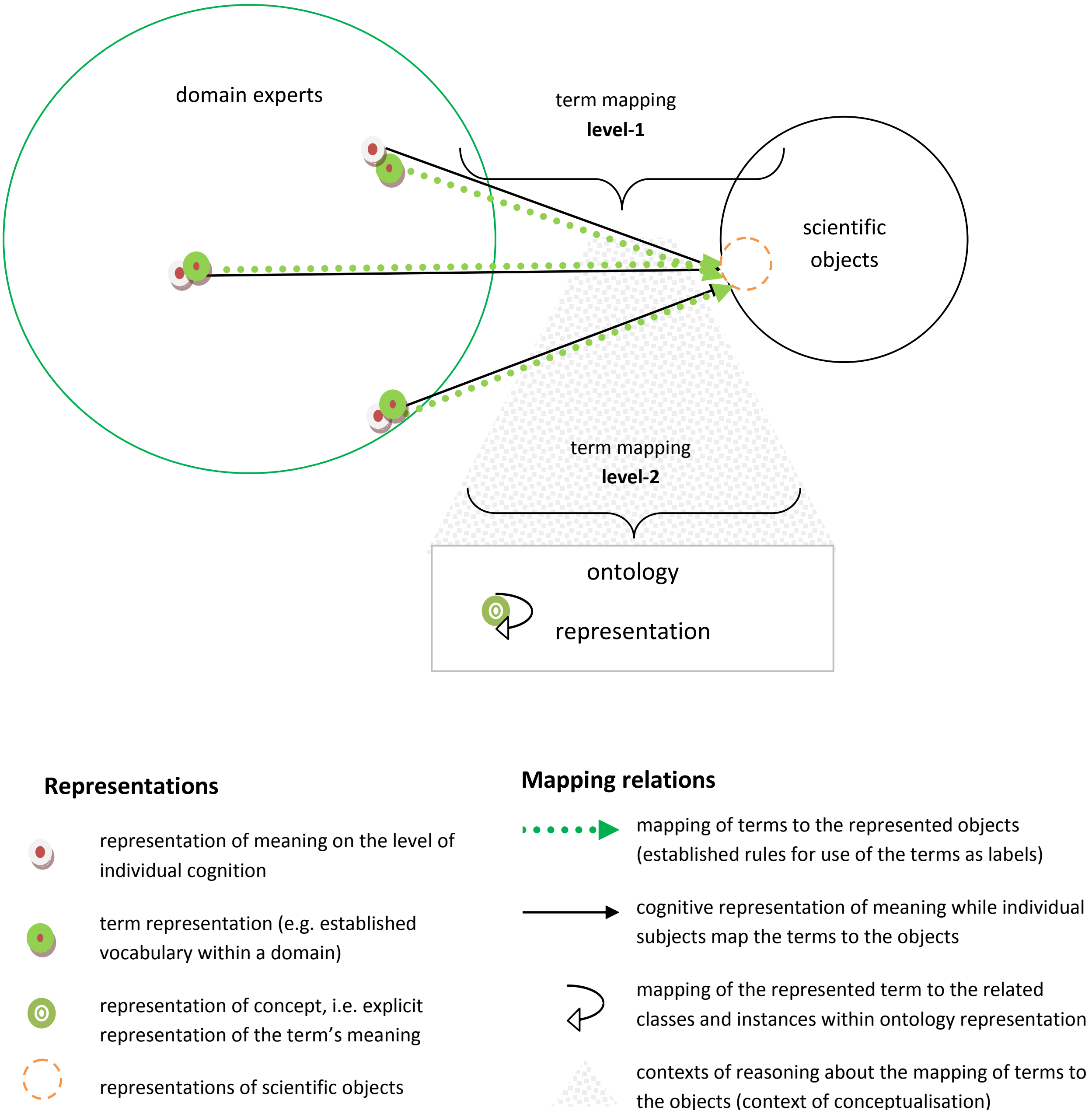

domain experts
term mapping
level-1
scientific
objects
term mapping
level-2
ontology
representation
Representations
representation of meaning on the level of individual cognition
term representation (e.g. established vocabulary within a domain)
representation of concept, i.e. explicit representation of the term's meaning
representations of scientific objects
Mapping relations
mapping of terms to the represented objects (established rules for use of the terms as labels)
cognitive representation of meaning while individual subjects map the terms to the objects
mapping of the represented term to the related classes and instances within ontology representation
contexts of reasoning about the mapping of terms to the objects (context of conceptualisation)

Figure 13 Representing meaning of scientific terms and concepts

So, a particular domain ontology considers the reasoning about a domain. The reasoning is built into an ontology through the representation of concepts and the explicitly represented relations that hold among the concepts. The represented concepts are the parts of an ontology representation, while the relations may be represented as the arrows with a specific relational function (a conceptual model), or they can be defined through a set of axioms in a formal ontology representation. In that way by representing a conceptualisation an ontology explicitly represents reasoning about a domain. The labels can be mapped onto the concepts, but they do not change the meaning of the represented concepts, as multiple labels may be mapped to the same concept. In order to conclude my argument on how the knowledge context gets built into an ontology model, I distinguish the representational field of the ontology modelling from the representational field of other kinds of scientific representations.

Figure 13 summarises the discussion from this and previous sections. In addition, the figure distinguishes two kinds of mapping relations. On the level-1 mapping is used to label and represent the scientific objects (an indirect target of ontology model). However, on the level-2 mappings represent how the scientists reason about their targets (a direct target of an ontology model). So, a conceptualisation of a domain, labelled and represented by various means within the level-1, is the main target of the level-2 (ontology) representations. In that way, an ontology model targets the context of scientific reasoning that is going to be represented in an ontology model. Eventually, an ontology model can be applied as a top level of a knowledge base, organising information stored in a database.

In order to clarify how the ontology represents the reasoning in practice, we can consider, for example, a (clinical) breast cancer ontology, which aims to represent an aspect of clinical knowledge (see also 2.2-2.4). By representing *that* a tumour's classification, diagnosis, and corresponding therapy relate to each other in a certain way, the representation will actually represent *how* the clinicians reason about the classification and other represented concepts. For, the labels and classes will be selected to represent clinical knowledge on how certain categories relate to each other. On the other hand, a molecular representation may aim to represent molecular mechanisms involved in carcinogenesis. The relevant ontology will therefore represent the reasoning about the molecular processes that result in the cancer phenotype. Obviously, these two representations will have diverse criteria for what has to be represented. The labels used in the clinical representation are most often excluded from the molecular representation (see Section 3.2, 3.3). At least, the clinical labels are not explicitly present in the molecular representations, which as such stay on the molecular level of observation.

The granularity of the representations will also be of diverse kinds. While the clinical representation may consist of the classes labelled as ‘patient’, ‘organ’, ‘therapy’, ‘diagnosis’, and ‘prognosis’, the molecular representation will consist of the fine-grained representations of the molecular structures, functions, and interactions, labelling the classes of ‘genes’ and ‘proteins’ (Chapter III). Moreover, as clinical reasoning mainly relies on the statistical evidence satisfactory for its clinical applicability, the statistical inference used in molecular domain is used to justify explanatory relations that hold among the interacting genes and gene products. Therefore, the criteria of what makes evidence satisfactory may differ in clinical and molecular reasoning[35].

The process of ontology design, while representing a domain, considers a variety of reasoning that is present in scientific practice. Of course, in a final model, the variety of reasoning about a domain gets represented according to the aim for which the ontology is designed for. Like any kind of representation, an ontology representation also abstracts from many possible ways to represent a domain. In Chapter II I provide an example that illustrates how scientific reasoning and abstraction is captured by an ontology model and its various representational forms. The following section introduces some basic ideas of what knowledgebases are, while the concluding section characterises the epistemic groups involved in an ontology and knowledgebase building. A collaboration among the groups of experts coming from the fields that an ontology is dealing

[35] Chapter three develops further distinctions between clinical and molecular reasoning. The focus of this section is rather on how representation of a knowledge domain actually represents the reasoning of the domain.

with try to share their knowledge and understanding of a domain in order to achieve a shared conceptualisation and model a representation for a particular purpose (Section 1.2.6).

1.2.5. Knowledge representation and knowledgebases

Knowledge representation (KR) is studied as a subfield of computer science aiming to capture human knowledge formally (Lifschitz, van Harmelen, and Porter 2008). The study of knowledge bases (KBs), knowledge based systems, and their applications is closely related to the field of artificial intelligence (AI) that deals with KR and automated reasoning systems (Brachman and Levesque 2004; Sowa 2000). In this section I present some basic distinctions that clarify how KR and KBs are used as a tool and a medium to represent knowledge formally. The discussion should specify the roles of KR and KBs in organising and representing biomedical knowledge.

In answering the question 'What is a Knowledge Representation?' Randall Davis et al. distinguish five crucial but distinct roles of KR (Davis, Shrobe, and Szolovits 1993). According to the authors, KR is

1) a *surrogate* for things that enables reasoning about the world[36]
2) a set of *ontological commitments*, which determine in what terms we think about the world
3) a *fragmentary theory of intelligent reasoning*, composed of (i) the representation's conception of intelligent reasoning, (ii) the set of inferences which representation *sanctions*, (iii) the set of inferences which representation *recommends*
4) a *medium* for pragmatically efficient *computation*
5) a *medium of human expression*

[36] The Davis et al. approach does not question metaphysical issues of status of the 'world'. Neither will I go into the metaphysical discussions here. By the term 'world' I designate phenomena that are the representational target. Therefore, I interpret represented 'objects' within KR as standing for the targeted phenomena.

Each of the five roles imposes specific demands on a representation which formalises and represents knowledge. I discuss the roles of KR, as presented in (Davis, Shrobe, and Szolovits 1993), while considering the biomedical context. I also propose certain modifications in the above outlined view on KR.

(1) *KR as a surrogate for things* captures the *things* that are the object of interest for the agents who are representing knowledge. Thus, one of the roles of KR is to represent 'things' (Ibid.). Moreover, such a surrogate for 'things' should also enable (an automated) reasoning about the world. We should next specify the character of those 'things' that are captured within a KR in order to enable reasoning. Within the clinical domain the 'things' of interests are, for example, the patients, diagnostic and prognostic terms, and other disease related phenomena, which are often described as signs and symptoms. While characterising KR as a surrogate that *stands-for* the things in the world, Davis et al. consider a variety of 'things' that KR might represent. However, for heuristic purposes I will slightly reframe their view of KR as a surrogate that 'stands-for things'[37].

The heterogeneous kinds of 'things' that we meet in a biomedical domain does not include just objects described as material things, but also a huge array of normatively and conventionally established 'objects'. So, I use the term 'object', while denoting anything that can be represented[38]. In other words, KR also plays a role of surrogate for 'objects' such as 'poor prognosis', which is obviously a normative concept that is used in clinical practice to describe expected clinical outcome of patients. The assignment of 'poor prognosis' depends on various socio-cultural and scientific aspects present at the time of prognosis assessment. For example, had a new therapy been discovered, the prognosis assessment would have switched from 'poor' to 'good' prognosis.

[37] See also the discussion in 1.2.4 and the critics of *stands-for* role of representation.

[38] For a similar treatment of 'object' see, for instance, (Boniolo)

According to the view of KR as a surrogate for 'objects', KR stands-for the targeted objects, whereby the KR denotes either material objects in the *real world* or it represents the *abstractions* such as mathematical 'objects'. For short, I label every *object captured in KR* as *KRobj*, while the targeted 'objects' I label *KRtarg*. Since *KRtarg* can be 'real world objects' such as 'poor prognosis' and 'diagnosis', I interpret all *KRtarg* as phenomena that a KR is designed to capture. Such an approach to *KRtarg*, being metaphysically neutral, avoids the problems that a characterisation of the Ontologically heterogeneous targets can have (Section 1.2.1)[39]. When understood as a phenomenon, *diagnosis* is treated as an equally appropriate target of KR as is any other kind of phenomenon (described as a material object, for example). While material objects are phenomena that can be described through their physical features, phenomena such as diagnosis can be described through various standards, classificatory, and therapeutic criteria, which play a certain role in diagnosis assessment.

In addition, KR is a formal representation of reasoning. Therefore, a surrogate role of KR, pointed out by Davis, should be extended to the phenomenon of reasoning. In the case in which a KR captures the reasoning of the clinicians about a diagnosis, *KRobj* stands for the *KRtarg* that is *the reasoning of the clinicians*. That is to say, not only KR is a surrogate for 'objects', but also *a surrogate for the reasoning* of the domain experts about the domain 'objects' (see 1.2.4 and the examples in Sections 2.2.1-2.4). The features of KR described through the roles (2)-(5) disentangle how KR serves a surrogate for reasoning.

Another point I need to make concerns the semantics of KR. Davis et al.'s account, according to which *KRobj* denote the real world 'objects' (which I interpret as phenomena), provides an explanation of the KR's semantics only within the framework in which KR is a surrogate that stands for the *KRtarg* that are the 'objects' in the world. However, the formal semantics of KR that is applied to the reasoning over data stored in a database actually denotes not only *KRtarg*, but also the computational artefacts that are the instances of information model

[39] The treatment of the KR targets as the phenomena that KR aims to capture, no matter how 'real' they are, avoids a number of problems such as the Ontological status of the abstract and fictional entities, measurement records etc. (Section 1.2.1.).

and data structures within a database (Section 1.2.4). So, an instance such as X, which is recorded in a database, stands for a patient, while denoting a real person that can get assigned various types of predicates, which are recorded in a database, such as a name (having a value 'Mary'), a diagnosis (having a value 'HER2 positive breast cancer'), etc. Thus, the semantics of KR is twofold because the interpretation of *KRobj* can be twofold. The interpretation of *KRobj* can denote both the phenomena in the world and the instances within a database that stand for the objects that a KR aims to capture and reason about.

Biomedical ontologies are designed to support structuring and organising of biomedical knowledge, while representing terms and related classes according to that how they are understood and what they 'mean' in a biomedical domain. The basic structure of a knowledgebase consists of *A boxes* and *T boxes* (Brachman and Levesque 2004; Sowa 2000). The collected data are stored in a database as *instances* within A boxes, while the classes, which are designed according to an ontology model, are *types* stored within T boxes. By assigning a mapping relation from the instances to the classes, the data acquire an interpretation.

Consider the case of the GO knowledgebase[40] where the stored information about genes and gene products is organised into the classes such as *cellular component*, *biological process* and *molecular function*. A particular class, labelled as 'cell proliferation' has assigned and represented the instances of genes and proteins which are associated with the biological process of cell proliferation. For example, a human HER2 gene, labelled as 'HER2', is described as a member of the GO class 'cell proliferation'. By its membership in the class, the term 'HER2' has actually acquired the meaning of *'the gene involved in cell proliferation'*. Accordingly, the stored data (e.g. the information about the HER2 coding region) also acquire the meaning that associates the stored data and the cell proliferation. Such a structuring of data associated with the same label, actually assigns the same meaning to the instances of the class, thus facilitating the information retrieval, data comparison, and re-use of the recorded data across the research context (see e.g. (Leonelli 2008)). As the terms with the same or similar meaning are getting aligned by the

[40] http://www.geneontology.org/

ontology matching tools, ‘HER2’ from GO can also be mapped to the SNOMED term ‘HER2’ as well as to the ‘HER2’ from various model organism databases (Section 4.2.2). Consequently, the researchers working with various model organisms can use the labels as the links in order to retrieve represented knowledge about the gene of interest.

GO database is just one example of *knowledgebase* (KB). Knowledgebases are a special kind of databases because they are not mere data storages. Since KBs represent knowledge about a domain as structured in the inter-related categories, they *store and represent knowledge*. Knowledge within a KB is structured according to the conceptual structures (Section 1.2.4.) that explicitly connect the pieces of the stored information. The conceptual structures of a knowledgebase are ontologies, which represent the meaning of the concepts as the types, classes, and relations among them.

For example, since ‘intramembrane protein’ is understood as an integral part of the cell membrane, one of the meanings for ‘intramembrane protein’ in an ontology model will be ‘cellular component’. However, since the ‘intramembrane protein’ can be also described through its molecular functions and certain biological processes, a member of the class ‘intramembrane protein’ will also get assigned the meaning of a protein that is associated with ‘cell proliferation’.

The decision on which relations and mappings among the types and instances are the most appropriate is getting achieved through the specification of the ontological commitments. As discussed in Section 1.2.1, the ontological commitments of KR and KBs, although related to the metaphysical questions about the world, are primarily the result of the pragmatic endeavour of the ontology engineers that are considering how to represent a domain for a particular purpose. Thus, the agreement about that which kinds of ‘objects’ and the relations will be represented is to be understood as an engineering task rather than as a metaphysical endeavour (Lord and Stevens 2010).

Nonetheless, while specifying *KRobj*, KR provides a set of formal specifications that explicates the kinds of objects and the relations that a KR commits to represent. Metaphorically speaking, the ontological commitments are ‘a strong pair of glasses’ that bring into the focus certain aspects.

> If, as we argue, all representations are imperfect approximations to reality, each approximation attending to some things and ignoring others, then in selecting any representation, we are in the very same act unavoidably making a set of decisions about how and what to see in the world. That is, selecting a representation means making a set of ontological commitments. The commitments are, in effect, a strong pair of glasses that determine what we can see, bringing some part of the world into sharp focus at the expense of blurring other parts. (Davis, Shrobe, and Szolovits 1993)

The choice of what will be in the focus of a KR determinates the commitments that will be formally specified. Furthermore, the specification of ontological commitments accumulates in the layers

1) choice of a technology that is used to represent the commitments (e.g. frames, rules, etc.);

2) having chosen the representational technology, the choice needs to be made about the kinds of entity and the relations that are going to be represented (e.g. after taking decision of representing disease by means of the frames, the selection is made on the structuring of the disease entities and manifestations, providing a taxonomy structured around organ systems[41]);

3) the third layer of the commitments includes the instantiation, i.e. the choice is made on that which disease will be represented and in which branches of the hierarchy they will appear.

Obviously, making an ontological commitment as if, for example, alcoholism, homosexuality, and chronic fatigue syndrome should be included as the disease entities is a complex socio-cultural task which can have severe consequences for society (Kitcher 2001;

[41] Obviously, this is just an example of the commitments about the disease representation and an alternative choice might focus on some other aspects such as the cell type that characterises cancer.

Hacking 2002). Thus, the debate on the level of society and the established standards obviously influence the choices of the entities that are going to be represented formally (see group 1, Section 1.2.6).

1.2.6. Epistemic groups involved in ontology building

This section presents a comparative analysis of the distributed knowledge of various epistemic groups involved in the integration of heterogeneous types of representation into a knowledgebase (KB). In particular, I distinguish *where*, *how*, and *by whom* knowledge is represented by characterising six epistemic groups, and by discussing how membership to a group impacts the representation as well as knowledge (base)[42] types. Note that these groups exhibit rich interdependencies and partially overlap.

1) The characterisation of the epistemic groups starts with the societal demands for problem solving, such as, for example, the need for personalised breast cancer therapy (Gurwitz, Lunshof, and Altman 2006; Hamburg and Collins 2010). The demands may be represented in the form of standards, platforms and funding policies (Jasanoff 2005; Kitcher 2001; David 2002; Hamburg and Collins 2010). In a democratic society, knowledge on this level can be represented as common or shared knowledge available to the members of society; knowledge can be distributed through various channels or common-sense KBs.

2) The second epistemic group to be discussed is at the level of an individual scientist whose 'knowledge base' is a collection of relevant background knowledge, here to be understood as cognitive representations placed in the mind, arguably, in the form of conceptual maps (see (Medin and Rips 2005)). As far as the conceptualisation of a problem has not been

[42] The term 'knowledge base' is used in deferent ways by different research communities. While in cognitive sciences background knowledge of individuals constitute a cognitive knowledge base, in AI 'knowledge base' designates databases within which knowledge about a domain is structured and explicitly represented.

communicated in an inter-subjectively accessible way, e.g. by means of language, the related representations stay private to the mind of a particular epistemic agent. Thereby, the representation of a cognitive conceptualisation is *implicit* (Guarino, Oberle, and Staab 2009). Likewise, whilst an individual may assign a referent (an object, a term) to such an implicit representation, the semantics of the relation between the referent and the related cognitive content stays implicit, i.e. accessible only to the individual mind.

3) As the third epistemic group, I specify the scientific communities, each of which is composed of the specific disciplinary domain scientists (clinicians, molecular biologists, bioinformaticians etc.). This epistemic group establishes knowledge within a scientific community as a received view, having the form of explicit and inter-subjective representations expressed in the respective scientific languages, circulated through publications. Like in group (1), knowledge can be distributed in various ways. Contrary to the implicit conceptualisations (group 2), a community establishes shared conceptualisations that are made explicit by means of language. A shared conceptualisation within a community, thanks to the shared terminology and the context of its use, enables distinctions between domain specific terms and their relations as an agreed upon meaning to represent domain knowledge. However, these shared representations are still local and specific to particular communities. Like in (2), the semantics of the terms and the mapping relations among them stay implicit. Namely, it is left to the interpreters of language to assign meaning to terms by a cognitive process which assigns referents and relations to the represented terms.

4) The fourth group comprises scientific communities formed around a particular problem (e.g. breast cancer). As the group contains multidisciplinary teams focused on a particular problem, knowledge will need to be coordinated in such a way that the used scientific terms and reference classes will conform with knowledge within diverse domains. For instance, the biomedical terms might be structured into networks of terms that represent how these terms are interrelated in the domain knowledge. Thus, collaboration here results in merging

knowledge from different domains. The representation of the merged knowledge coming from different perspectives on the same problem might be a 'unified semantic map' (see group (2)) that serves as a semi-formal conceptual model and an intermediate step towards the KB and the formal ontology to be employed in KR (see e.g. (Montazemi 2009)). Note that group (4) is heterogeneous in itself because it is composed of experts from various scientific fields (group 3), thereby producing a variety of breast cancer related representations.

5) The fifth is the communities of logicians and ontologists who are formalising ontologies according to the needs and specificities of a particular field. Domain knowledge and the merged domain knowledge will be expressed as ontologies written in various formal languages (e.g. refining foundational ontologies such as DOLCE (Masolo et al. 2003), BFO[43], or GFO[44] etc. formalised in OBO[45], OWL[46], or first-order logic, etc.) As a shared conceptualisation gets enriched with formal specifications that have well-defined formal semantics (Uschold and Gruninger 2004), the semantics of the represented domain is *explicit*. Accordingly, related KBs will contain explicitly represented knowledge.

6) The sixth group involves computer scientists, programmers and engineers, who are designing databases and applying formal ontologies as well as various reasoning tools to large data sets. Technically, a representation built on top of a database involves types and mapping relations structuring the data, and can be considered as meta-data. Here, the representation integrates the types and mappings with instances (data). Epistemic accuracy of the mappings depends on how well the mappings correspond to the scientific knowledge and the empirical findings of the represented domain (e.g. breast cancer). In contrast to groups (2) and (3), knowledge in a KB is not scattered over various representational spaces or layers, but integrated into one. As the ontology mapping terms to instances provides the representational reference within a KB, both semantics and representation are explicit.

[43] http://www.ifomis.org/bfo/

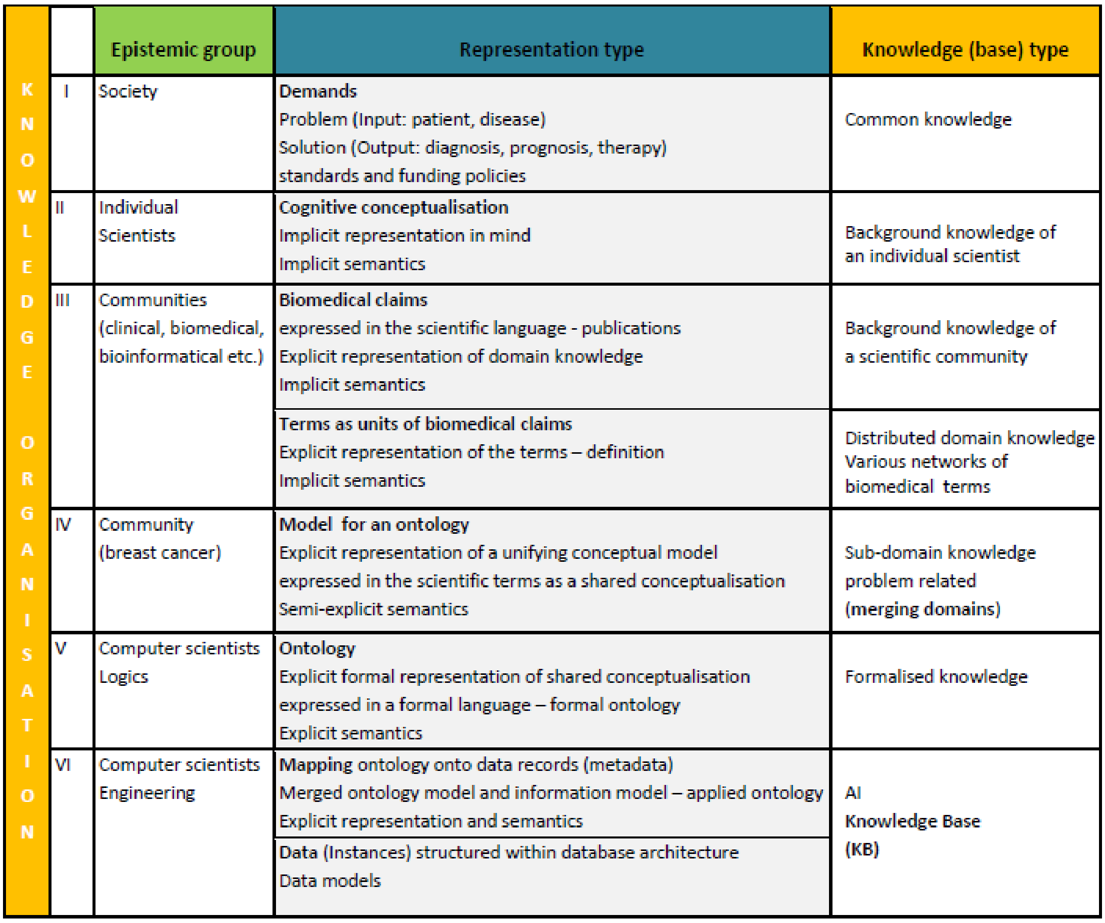
[44] http://www.onto-med.de/ontologies/gfo/

[45] http://www.geneontology.org/GO.format.obo-1_4.shtml

[46] http://www.w3.org/TR/owl2-overview/

Table 14, shows the epistemic agents as they are organised into the six groups, illustrating to what extent they are involved in ontology and KB building. The hierarchical grouping of the agents is just an approximation, according to the representational means and the level of explicitness employed to represent knowledge. Of course, a real time process of knowledge organisation and distribution is much more complex than what this table can summarise.

However, the table sketches some basic distinctions distributed across the specific representational domains. The epistemic groups are ordered according to the degree to which knowledge representations are made explicit, ranging from implicit and less formal to the most explicit formal representation of knowledge. Collaboration among the domain experts is presented as crucial for knowledge integration. A group of experts with a common interest is collaborating in establishing a shared conceptualisation (Borst 1997) (group 4), which gets



		Epistemic group	Representation type	Knowledge (base) type
KNOWLEDGE ORGANISATION	I	Society	**Demands** Problem (Input: patient, disease) Solution (Output: diagnosis, prognosis, therapy) standards and funding policies	Common knowledge
	II	Individual Scientists	**Cognitive conceptualisation** Implicit representation in mind Implicit semantics	Background knowledge of an individual scientist
	III	Communities (clinical, biomedical, bioinformatical etc.)	**Biomedical claims** expressed in the scientific language - publications Explicit representation of domain knowledge Implicit semantics	Background knowledge of a scientific community
			Terms as units of biomedical claims Explicit representation of the terms – definition Implicit semantics	Distributed domain knowledge Various networks of biomedical terms
	IV	Community (breast cancer)	**Model for an ontology** Explicit representation of a unifying conceptual model expressed in the scientific terms as a shared conceptualisation Semi-explicit semantics	Sub-domain knowledge problem related (merging domains)
	V	Computer scientists Logics	**Ontology** Explicit formal representation of shared conceptualisation expressed in a formal language – formal ontology Explicit semantics	Formalised knowledge
	VI	Computer scientists Engineering	**Mapping** ontology onto data records (metadata) Merged ontology model and information model – applied ontology Explicit representation and semantics	AI Knowledge Base (KB)
			Data (Instances) structured within database architecture Data models	

Figure 14 Knowledge organisation: epistemic groups and representation types

formalised (groups 5 and 6) according to the established standards that help them label and describe the domain of interest in an interoperable way (Ekins et al. 2011).

Knowledge levels, groups, or layers have been discussed previously in the AI literature. For instance, Newell introduced an agent-based distinction between the 'knowledge level' and the 'symbol level' in (Newell 1982), and (Brachman 1979; Guarino 1994, 2009), analysed layers in formal ontology design. In more detail, Brachman, in 1979, introduced a classification of the primitives used in KR systems at the time (Brachman 1979), distinguishing the following four levels: (i) 'Implementational', (ii) 'Logical', (iii) 'Conceptual', and (iv) 'Linguistic'. Guarino added to these four layers yet another layer, namely the 'Epistemological Layer' for the primitives, situated between the 'Logical' and the 'Conceptual' layers (Guarino 1994, 2009). Unlike these approaches, I mainly aim at distinguishing *human agents as individuals and groups focused around particular epistemic interests*, whilst analysing the corresponding impact on representation types.

Regarding the perspectivism in ontology representation, the distinction of the epistemic groups captures some basic similarities and differences among the agents who aim at conceptualising and representing a problem. Namely, consider the following snippets from Gruber's definition (Gruber 1995) of ontology as:

- 'simplified view of the world that we wish to represent for some purpose': an ontology as a technical artefact is not intended to cover the world in its entirety, but only chosen aspects of the world, on specific levels of abstraction, and for given purposes - largely independent of particular metaphysical positions such as realism and antirealism; here, group (4) will typically informally specify the relevant domain knowledge (e.g. a conceptual map of breast cancer phenotypes that integrates structural and functional features of a protein involved in carcinogenesis), whilst group (5) is in charge of establishing an agreement on how to formally codify this knowledge.

- ‘committed to some conceptualisation’: ontologies presuppose various decisions concerning ontological commitments. These originate partly in common sense knowledge (group (1)), precisifications given by members of group (2), and agreements as they are established in groups (3) and (4). Finally, the formal implementation of the ontological commitments is again left for groups (5) and (6), merging collaborative interests of (1)-(6).

- ‘“what exists” is that which can be represented’: ontological commitments are dependent on the expressive capabilities of selected representational formalisms. The choice of an adequate formal language can only be established as an interplay between logician (group (5)), computer scientist (group (6)), and the domain experts of (3) and (4).

- ‘representational vocabulary’ and ‘human-readable text’: there is a tension between the logical vocabulary used, and the natural language concepts and terms it is meant to capture, and, in the case of e.g. breast cancer, various forms of scientific representations such as graphs, mathematical equations, images, 3D models etc. Reconciling this tension requires deep interaction between the various groups of domain experts and formal logicians and computer scientists.

- ‘an ontology is the statement of a logical theory’: on a technical level, an ontology is seen as equivalent to a logical theory, written in a certain formalism. Clearly, this task is for group (5), respecting the requirements of group (6).

Heterogeneity of formal languages is particularly important in the life sciences, where size of ontologies and needed expressivity vary dramatically. For example, whereas weak (i.e. sub-Boolean) DLs suffice for the NCI thesaurus (containing about 45.000 represented concepts) which is intended to become the reference terminology for cancer research (Sioutos et al. 2007), other medical ontologies such as GALEN[47] require the full expressivity of the OWL language (a decidable fragment of first-order logic), while foundational ontologies typically require at least full first-

[47] http://www.opengalen.org/

order logic (Kutz 2011). Alternatively, new and more expressive languages to model complex biological processes have also been proposed (Boniolo, D'Agostino, and Di Fiore 2010; Boniolo et al. 2012), while the application of these languages to the reasoning with biomedical ontologies is an open possibility.

An example of a heterogeneous combination of formalisms is discussed in (Hastings, Kutz, and Mossakowski 2011), where it is shown that in order to adequately represent the spatial structure of molecules as they are described in chemical ontologies such as ChEBI (De Matos et al. 2010), ontology languages need to be combined with formalisms such as Marydic second-order logic.

Chapter II

Biomedical Representations: the Case of Breast Cancer Phenotype

Scientists involved in biomedical research, ontology and the knowledgebase building need to face a plurality of epistemic motivations, on a conceptual, formal, and technical level. The analysed distinctions among the six categories of *human agents as individuals and groups* (1.2.6) focused around particular epistemic interests have illustrated the impact of these groups and individuals on representation types, mapping and reasoning scenarios. This chapter addresses further a plurality of representations, related formalisms, expressivities and aims, as they are found across diverse scientific communities.

While discussing the phenotype representations, I keep in line with the position of perspectivism, as developed in the philosophy of science (for instance in (van Fraassen 2008; Callebaut 2012)). Perspectivism accepts the position that human agents can never have a completely neutral and perspective-independent picture of 'the world', i.e. 'a view from nowhere' (van Fraassen 2008; Goodman 1978; Giere 2006; Giere 2010; Wimsatt 2007; Callebaut 2012). Rather, every scientific representation is defined by its research context and particular questions that scientists address (Boniolo 2005; Sintonen 2005; Dupré 2002). The position of perspectivism reconcile with a broader pluralistic stance in philosophy of science and epistemology (Kellert, Longino, and Waters 2006; Dupré 1993; Mitchell 2003; Kitcher 2001). I label my pluralistic approach as perspectivism because I believe it enables me to give a fruitful account of scientific practices, without committing to any particular metaphysical position. Therefore, following van Fraassen's view that

> What scientific representation is and how it works is everyone's concern, and there we may find a large area where more general philosophical differences need make no difference (van Fraassen 2008, p. 3).

I focus my analysis on the level of perspective-driven scientific representations.

As a case study of this chapter I analyse the breast cancer phenotype representations. Through a number of examples from biomedicine, in the first part of the chapter I examine how molecular, clinical, and ontology researchers capture phenotypes from a specific perspective. I demonstrate a heterogeneity of representation types for breast cancer phenotypes and stress that the characterisation of a tumour phenotype often includes parameters that go beyond the representation of a corresponding empirically observed tumour, thus reflecting significant functional features of the phenotypes as well as epistemic interests that drive the modes of representation. Accordingly, the represented features of cancer phenotypes function as epistemic vehicles aiding various classifications, explanations, and predictions.

In order to clarify how the plurality of epistemic motivations gets captured in phenotype representations while integrating various representational forces, in the reminder of the chapter I distinguish semantic, cognitive, and pragmatic functions of representation. I explain a successful use of representations, which serves multiple purposes in various contexts, by distinguishing specific representational aspects. Eventually, I conclude that ontology models gain a representational advantage over other kinds of representation as they successfully integrate explanatory, representational, predictive, and computational aspects of representation.

2.1. Diversity of representations

In this section I discuss a representational diversity of breast cancer phenotypes. In general terms, a phenotype is defined as a set of features of an organism that emerges as a result of interactions of its genetic material (specified as genotype) and the environment (Lewontin 2004; Dupré 1993). While the genotype of an organism is the class to which that organism belongs as determined by the description of the physical material made up of DNA passed to the organism by its parents, the phenotype of an organism is the class to which the organism belongs as determined by the description of the organism's physical and behavioural characteristics (Lewontin 2004). So, a human's genotype belongs to the class 'human' thanks to its inherited genetic material, description of which, despite of individual variations, fits to the general description of genome that is typical for the species 'homo sapiens'[48]. A human's phenotype includes features such as its size, its shape, its metabolic activities and its patterns of movement and behaviour. Some of these features can be selected to describe a *typical* human phenotype. In the case of anatomical descriptions, a canonical human anatomy represents a human phenotype as the organism composed of parts such as legs, arms, organs, cells etc. However, phenotypic features highly vary among individuals, and which particular features will be selected to classify individuals depends on classificatory aims. For example, a phenotype can be described through its structural, functional, or dispositional features, or through some quality such as eye colour (Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010).

The representation of phenotypes plays an important role in clinical and biomedical knowledge, aiming at describing disease, assigning diagnoses and recommending therapies. Therefore, phenotypes are of particular interest to the biomedical expert if they include features that describe aberrances from what is considered a 'normal' phenotype. Moreover, exactly the

[48] The long lasting debate on species classification illustrates best that an agreement on what *the* typical features are, even on a gene level, is not a trivial task. We therefore consider features in terms of a conventional agreement at the current stage of scientific understanding.

aberrant features play a crucial role in the classification of a disease. A disease usually gets characterised through a distinction between 'normal' and 'abnormal' phenotypes, where 'abnormal' phenotypes serve as the marks of disease. The 'abnormal' phenotypes associated with a disease are labelled as *phenotypes of disease* (PD). However, the questions of what is 'abnormal' and *what* should be considered as a phenotype of a disease and *how* such a phenotype should be represented are rather contentious (Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010). Clearly, the choice of how a PD should be represented is *normative* and *context dependent*. Consider the case of breast cancer and BRCA1 and BRCA2 gene mutations. In the age of genomic medicine, the very definition of disease has changed introducing a new kind of asymptomatic diagnosis. So, the carriers of BRCA mutations, without having developed any signs of breast cancer, still have a likelihood of 40-80% for developing an aggressive cancer phenotype during their life span (Fackenthal and Olopade 2007). The establishment of preventive treatments such as prophylactic surgery, chemoprevention and screening designed for the BRCA mutation carriers demonstrates that carriers of the mutated genes are indeed treated as patients (Metcalfe et al. 2008). Genomic medicine, thus, shifts the focus of PD from a traditional organ level approach to the gene level, treating apparently healthy people as 'patients'. For, the 'normal' breast phenotype in a BRCA mutations carrier will be irrelevant in the light of knowledge about 'abnormal', fine-grained phenotypes related to the gene expression patterns of the mutated genes. Since such a phenotype is classified as PD based on a disposition to develop the disease, the BRCA mutations carriers can get assigned a *dispositional* PD, according to the phenotype classification proposed in (Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010).

Although these new directions in biomedicine aim at an integration of clinical and biological knowledge, the requirements across biomedical sub-domains significantly vary. So, a clinician will have different criteria for the representation of a phenotype than a molecular biologist. Regarding the goals of a discipline and the research context, information that is relevant for a clinician does not need to satisfy the needs of a molecular biologist who is mostly interested in phenotypic information about the molecular mechanisms associated with a disease. Likewise, features of a phenotype such as ‘obesity’, although clinically significant for breast cancer risk assessment, are excluded from the molecular description of a phenotype. Moreover, as I will illustrate by an example (HER 2 protein representation), the representations that a molecular biologist is typically interested in prioritise the explanatory role of the selected features, while the clinical representations usually aim towards clinical usability such as diagnosis assessment and therapy choice. As a result, heterogeneous representations and classifications of breast cancer phenotypes are employed in clinical and biomedical practice (Gospodarowicz, O'Sullivan, and Sobin 2006; Perou et al. 2000; Faratian et al. 2009).

As previousely discussed (see section 1.2.4), the minimal conditions that any scientific representation needs to satisfy are that

> A represents B only if (i) the representational force of A points towards B, and (ii) A allows competent and informed agents to draw specific inferences regarding B. (Suárez 2004, p.773)

Accordingelly, a mathematical, a pictorial, or a logical representation are the potential sources for an inference about a target. If A leads a scientist to a specific inference about B, which is in our case a phenotype, then A will be a representation of B (the phenotype).

In light of this view of representation, concerning ways phenotypes can be captured by various technological and scientific tools, *representations of PDs* will include images acquired by technologies such as ultrasound, X-ray, and microscopy of histopathological samples. Moreover, representations of PDs are not limited to visual representations of a tumour, but include mathematical equations (Enderling et al. 2007), statistical graphs, molecular markers, microarrays data (van't Veer et al. 2002), and the phenotype specific protein interactions (Chuang et al. 2007), thus describing PDs according to the needs of and knowledge about a particular domain aspect.

That is to say, a representation reflects which aspects of knowledge have been targeted by the representation and for what purpose. Differential equations of a mathematical model aim at representing cancer phenotypes whilst modelling, for example, carcinogenesis’ dynamics. A PD is sometimes represented by the equations that model the response to a particular treatment, thereby playing a *predictive* role (Enderling et al. 2007). Some other representations have primarily a *heuristic* role, using mechanistic and causal models to represent various aspects of a PD, which can support the *understanding* of carcinogenesis or a specific response to a therapy (Nahta et al. 2006). Accordingly, a representation reflects a scientist's choice to model a certain subset of the domain knowledge for a particular purpose. Not every PD representation can equally satisfy all possible explanatory, predictive, and pragmatic needs, even within a single domain. Rather, various types of representations have been employed to model diverse aspects of a domain problem. Therefore, 'choosing a representation' might be considered a highly intentional act (Giere 2010).

However, a representation such as a histopathological image will not, itself, represent any knowledge unless it gets interpreted. Knowledge within a domain is *explicitly* represented only if the representations get systematically connected with related interpretations, knowledge claims, and reasoning over the representations. As a representation may have various interpretations, expressing diverse aspects of knowledge about what is being represented, it can also get assigned diverse knowledge claims and mapping relations.

Therefore, besides the heterogeneity of PD representations, biomedical ontologies have to deal with a heterogeneity of reasoning about PDs, comprising different kinds of formal (or logical) representations as well as various types of reasoning (see section 2.2.). Conversely, the intended reasoning methods or types over PDs also influence the choice of representation of PDs because such representations are mediated by domain specific methods and interventions, employed in the imaging, measuring of the gene expression and other diagnostic and experimental techniques (Hacking). Consider, for example, clinical representations of breast cancer that go beyond the tumour imaging representation. According to the standards of the TNM classificatory system (Gospodarowicz, O'Sullivan, and Sobin 2006), the clinical classification of tumours considers tumour size (T), lymph nodes involvement (N), and presence of metastasis

(M). Of course, tumour size is just one feature and is not sufficient for the characterisation of the tumour type. Cancer is a dynamic and complex disease of an organism and the PD representations, therefore, go beyond the characterisation of a tumour captured in a static picture. For example, knowledge about lymph nodes' status or proliferation marker KI-67 provides additional information about a tumour's phenotype. Likewise, tumour markers provide a view on the PDs through the specific interventions on the representation. A detection system can target a gene or a protein of interest, staining the samples in order to produce a clinically useful PD representation. Had the estrogen receptor (ER) been detected, the PD would have been described as an ER positive tumour, which significantly differs from an ER-negative tumour, which does not respond to the endocrine therapy (Goldhirsch et al. 2011). Thus, the therapeutic criteria often play a crucial role in the specification of the tumour phenotypes.

2.1.1. Phenotype ontologies

Phenotype ontologies aim at capturing phenotypes formally. Developing phenotype ontologies has been recognised as a particularly challenging task for ontologists because of the complexity and heterogeneity of features used to describe a phenotype. Several approaches have been proposed to describe phenotypes formally, some of which use 'qualities' as attributes assigned to 'entities' (Mungall et al. 2007; Mungall et al. 2010). In particular, phenotype ontologies encounter the problem of having to represent features that are not 'normal' in terms of canonical anatomy. For instance, Hoehndorf et al. (Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010) have introduced two disjoint predicates, C (canonical) and NC (non-canonical) in order to support 'abnormal' phenotype representations. This approach is part of a general framework that categorises the concept of phenotype as 'Phene' into the classes of 'object phene' and 'process phene', while 'object phene' is subdivided into the classes 'structural phene', 'qualitative phene', 'dispositional phene' and 'participatory phene', including further subdivisions

(Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010). The phenotypic relations of the Phene ontology are formalised using the OWLDEF method (Hoehndorf et al. 2010) in order to endow phenotype representations with explicit semantics and to enable interoperability with existing domain ontologies. An advantage of such an approach compared to the entity-quality (EQ) characterisation of phenotypes (Mungall et al. 2007) is that it can support inference among the pheno classes that are imported from domain ontologies such as FMA, while EQ has limited interoperability and inferential potentials as it considers all phenotypic features as qualities (see (Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010)). The introduced Phene relations such as *CC-pheneOf-lacks-part* nicely describe phenotypes that are 'abnormal' due to absence of anatomical parts. In contrast to this, in this chapter I will address the case where the quantity of a part (e.g. the HER2 protein which is present in the cell membrane in both 'normal' and 'abnormal' phenotypes) makes a qualitative difference, resulting in an 'abnormal' phenotype. In collaboration with Oliver Kutz, I have been dealing with this problem by using a non-monotonic mereology specification (see (Sojic and Kutz 2012)), which I present in section 2.2.

Regarding biomedical ontologies in general, there have been particular efforts in creating alignments among the internationally recognised biomedical vocabularies (Rector 2003; Hartung et al. 2012; Milian et al. 2010), e.g. as represented in the form of a thesaurus (e.g. NCI, SNOMED), metathesaurus (UMLS), and biomedical classifications such as the International Classification of Disease (ICD)[49], with OBO Foundry ontologies and other formal ontologies such as General Formal Ontology GFO (Herre 2010), GALEN (Rector and Nowlan 1994), and the Dolce ontology (Masolo et al. 2003). The OBO foundry includes various ontologies each of which covers some aspects that are relevant for the representation of phenotypes. For instance, the Human Disease Ontology (DO)[50] aims to represent disease related concepts. For the representation of breast cancer phenotypes some of the key terms from DO include 'cancer', 'breast cancer', 'Her2-receptor

[49] Respectively, http://ncit.nci.nih.gov/, http://www.ihtsdo.org/snomed-ct/, http://www.nlm.nih.gov/research/umls/, http://www.who.int/classifications/icd/en/

[50] http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology

positive breast cancer’, ‘female breast cancer’, ‘male breast cancer’, ‘estrogen-receptor positive breast cancer’, ‘estrogen-receptor negative breast cancer’, ‘progesterone-receptor positive breast cancer’, ‘progesterone-receptor negative breast cancer’, and ‘triple-receptor negative breast cancer’[51]. The represented concepts in DO are organised in a hierarchical structure using the *is_a* relation. However, the hierarchy of DO concepts is not sufficient to represent breast cancer phenotypes. Although ‘cancer’ and ‘breast cancer’ within DO are specified by their lexical definitions, DO does not represent explicitly the defined concepts. That is to say, definitions of cancer as ‘disease of cellular proliferation that is malignant and primary, characterized by uncontrolled cellular proliferation, local cell invasion and metastasis’ and of breast cancer as ‘a thoracic cancer that originates in the mammary gland’ do not represent how, for example, ‘cancer’, ‘breast cancer’, and ‘Her2-receptor positive breast cancer’ relate to each other in a more specific way than what the relation of subsumption can communicate. A more specific representation of a breast cancer phenotype asks for concepts and relations that DO does not contain. As it has already been recognised in the case of prostate cancer (Overton, Romagnoli, and Chhem 2011) and phenotype ontologies in general (Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010), a breast cancer phenotype ontology needs to combine various other ontologies each of which is designed for a specific purpose. The Foundational Model of Anatomy (FMA) aims to represent anatomical concepts and relations in a canonical way. Thereby, FMA does not include ‘abnormal’ anatomical features. However, FMA can provide a phenotypic characterisation of a canonical breast anatomy and its lymphatic system that is lacking in DO. Likewise, the Phenotypic Quality Ontology (PATO)[52] provides representations of the qualities that are lacking in FMA and DO. Even if DO includes some of the quality terms such as ‘aggressive’, the particular qualities are not represented as separate concepts, rather they are just included in a definition or they make an inseparable part of a more complex concept such as ‘aggressive periodontitis’. Particular qualities from PATO (e.g. ‘abnormal’, ‘increased concentration’, ‘poorly

[51] The DO terms’ identifiers listed are, respectively, DOID:162, DOID:1612, DOID:0060079, DOID:0050671, DOID:1614, DOID:0060075, DOID:0060076, DOID:0060077, DOID:0060079, DOID:0060081

[52] http://www.bioontology.org/wiki/index.php/PATO:Main_Page

differentiated', 'aggressive', 'progressive')[53] can be easier reused in various contexts, facilitating automated inference. The Units of Measurement Ontology (UO), on the other hand, provides concepts such as 'count', 'molecule count', 'percent', 'length unit'[54], which are important for molecular and clinical descriptions of breast cancer phenotypes. In a similar manner, an ontology that aims to represent breast cancer phenotypes needs to consider, *inter alia*, the concepts 'assay' and 'specimen'[55] from the Ontology for Biomedical Investigation (OBI)[56], 'HER2 signaling pathway' and 'regulation of Neu/ErbB-2 receptor activity'[57] from the Gene Ontology (GO), 'Tyrosine kinase-type cell surface receptor HER2' (CCO:B0005540) from the Cell Cycle Ontology (CCO) or from the Protein Knowledgebase (UniProtKB:P04626), and 'tyrosine kinase inhibitor' (CHEBI:38637) from Chemical Entities of Biological Interest (ChEBI) (De Matos et al.). Concerning the clinical management of patients, the information from a drug bank should also be considered in the description of a phenotype, allowing a specification of how, for instance, HER2 positive breast cancer behaves in response to the 'Herceptin' (DB00072) treatment.[58]

The presented variety of ontologies properly illustrates the very definition of ontology as a specification of a conceptualisation that represents a selected aspect of the world for a particular purpose, as given by Gruber (Gruber 1995). Accordingly,

> [...]When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. [...] In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other

[53] Respectively, PATO:0000460, PATO:0001162, PATO:0002106, PATO:0000871, PATO:0001818

[54] Respectively, UO:0000070, UO:0000192, UO:0000187, UO:0000001

[55] Respectively, OBI:0000070, OBI:0100051

[56] http://obi-ontology.org/

[57] Respectively, GO:0038128, GO:0060726

[58] http://www.drugbank.ca/drugs/DB00072}; Herceptin in the Drug Bank database has not yet been organised in an ontology, but the references to the related ontologies are provided.

> objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. (Gruber1995, 908-909)

Each of the above mentioned ontologies (e.g. FMA, PATO, GO, ChEBI etc.) is designed for a particular purpose. Therefore, the represented concepts are selected and related among each other in a way that best fits the aims of the particular ontology. While some of the concepts from diverse ontologies can be mutually aligned, in other cases it is likely that the concepts will stay represented only in one domain (universe of discourse) and not in the others. An introduction of new concepts (e.g. introduction of GO and OBI concepts to FMA) would unnecessarily increase complexity and most likely it would result in numerous redundancies and inconsistencies. Therefore, a combination of various modules from domain ontologies, as proposed in e.g. (Kutz, Mossakowski, and Lücke 2010; Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010; Kutz 2011), seems as the most viable solution when a particular task, such as the representation of breast cancer phenotypes has to be achieved. In our paper, we tackle just a segment of the problem as to how certain modules of FMA can be used in the representation of a breast cancer phenotype. Before we propose our model, we analyse how the diversity of domain interests influences the ontology design. The analysis of the epistemic groups involved in the ontology building emphasises a plurality of epistemic needs among the groups that our proposal attempts to deal with.

2.2. HER2 phenotype ontology

In this section, I discuss some aspects of the ‘mereology of HER2’, and the issue of ‘normal’ and ‘abnormal’ breast phenotypes. The Human Epidermal growth factor Receptor 2 (HER2) is an intramembrane protein that belongs to the family of epidermal growth factor receptors. HER2 positive breast cancer (HER2+ BC) has been recognised as a very aggressive form of cancer, characterised by amplification and overexpression of the ErbB2 (HER2) gene as well as the genes of the ErbB2 amplicon. Accordingly, the HER2 protein, coded by the ErbB2 gene, is present in a high concentration within the cell. The related cell phenotype is characterised through various detection methods, e.g. immunohistochemistry (IHC), as showing a high concentration of the intramembrane receptor HER2, while the cancer tissue phenotype has been described as 'poorly differentiated' and associated with a poor prognosis. As the predictive power of HER2 has been demonstrated in clinical practice, HER2 has acquired additional descriptions such as tumour marker (Kumar 2008) and a drug target[59]. 'Poor differentiation' of the examined tissue sample is still one of the main histopathological criteria for cancer assessment. However, as I focus here on the HER+ breast cancer, the feature that marks the disease is 'HER2 overexpression'.

2.2.1. HER2 protein in 'normal' and 'abnormal' phenotypes

I here describe a parthood relation in the context of a breast cancer subtype, where 'HER2' represents the intramembrane protein and a difference maker for the typical HER2+ breast cancer phenotypes. Each of the phenotype segments (i.e. the cell membrane, the cell, the tissue and the organ) acquires 'cancer phenotype' features due to overexpression of HER2.

Through a fine-grained representation of protein-protein interaction networks it can be captured how HER2 overexpression relates to carcinogenic processes, such as cancer cell growth, proliferation, migration, and adhesion that eventually result in a poorly differentiated tissue and an aggressive cancer phenotype. In our model, we abstract from the details of molecular processes, while using 'HER2' as a phenotype marker across the mereological segments.

[59] The impact that scientific change has on the ontology structure has been addressed in (Leonelli et al. 2011).

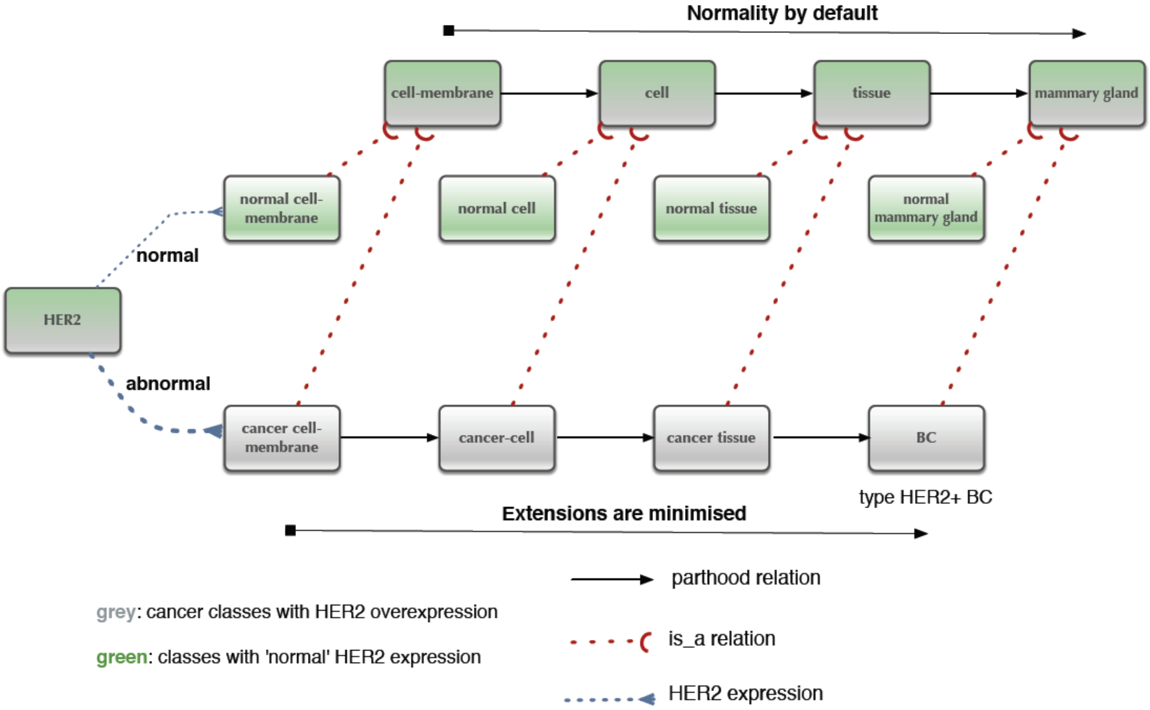


Figure 15 HER2 as a difference maker: ‘normal’ vs. ‘abnormal’ breast phenotype

In terms of the Phene ontology (Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010), the model of HER2+ breast cancer phenotype includes *object-phene* structural classes of the cancer affected mammary gland (HER2+ breast cancer). Apparently, the *relevant quality* of the phenotype is *overexpression* of the HER2 protein. A detection system by a method such as immunohistochemistry (IHC) stains a tissue sample with a colour dye (see DAB in Figure 16). The produced image represents HER2 overexpression as ‘brown colour’ of the membrane.

Of course, the colour does not inhere in the HER2 protein. Rather, it marks the presence of the protein within the membrane. Thus, the *quality* of ‘brown’ membrane within the HER2+ breast tissue sample actually represents the *quantity* of the HER2 proteins marked by the dye. As the quantity in this case describes the amount of proteins, we treat ‘overexpression’ in terms of detected HER2 proteins.

The IHC method itself does not provide the number of HER2 proteins, but it just detects HER+ cells in a tissue sample. However, it does not impact our model as quantitative proteomics and certain IHC image processing, in principle, can detect the amount of a protein (see e.g. (Shi et al. 2006; Di Cataldo et al. 2010)). For the same reason, our formal model, which considers IHC scores, assigns the 'cancer' feature to a phenotype on a tissue level.

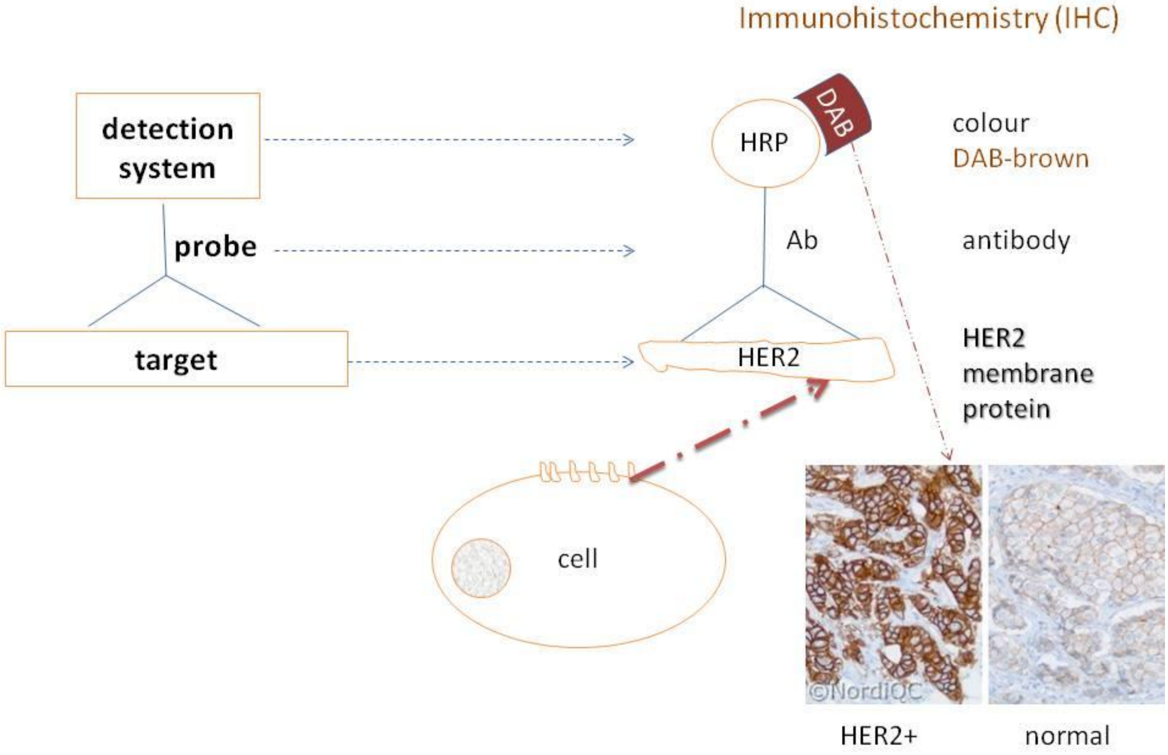


Figure 16 HER2 protein detection by immunohistochemistry (IHC)

The detected number of HER2+ cell membranes corresponds to the number of HER2+ cells. A tissue sample having more than 30% of HER2+ cells acquires 'HER2+ cancer' phenotype (Goldhirsch, Ingle, Gelber, Coates, Thürlimann et al. 2009), which is then also assigned to the mammary gland. Such an approach allows us to describe 'overexpression' as mereology of the HER2+ phenotype.

2.2.2. Concept minimisation: normal vs. abnormal phenotypes

The Figure 15 sketches just a fragment of HER2 related breast cancer phenotypes. And indeed, note that the classification of organ parts etc. into 'normal' vs. 'abnormal', as shown in Figure 15 is not a classical dichotomy as suggested by these terms. Indeed, attaching the property of abnormality to e.g. a particular tissue depends non-monotonically on the presence of information concerning overexpression of HER2. That is to say, unless such information is explicitly known, per default tissue will be considered normal. Therefore, adding information to the formal modelling (more instances of HER2 are part of some tissue) results in retracting the property of 'normality'.

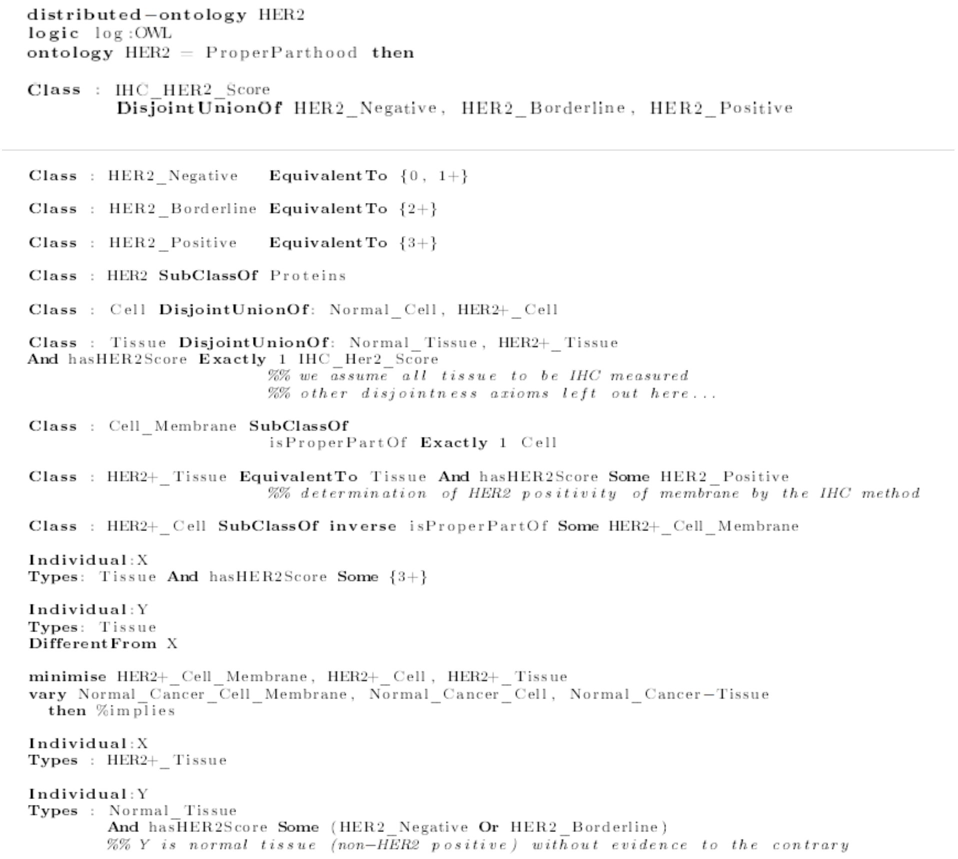
Modelling this formally is not straightforward in standard ontology languages such as OWL or FOL. Or rather, it can only be simulated by explicitly listing exceptions. However, it can be done elegantly by using non-monotonic formalisms such as default rules (Reiter 1980), autoepistemic logic (Moore 1985), or circumscription (McCarthy 1980). These formalisms were all devised in the early 80-s to overcome exactly such shortcomings of classical first-order logic in modelling exceptions and common sense rules.

Indeed, the necessity of supporting non-monotonic rules and abductive reasoning for ontology design has been noted many times (Bonatti, Lutz, and Wolter 2006; Elsenbroich, Kutz, and Sattler 2006; Hoehndorf et al. 2007). In connection with biomedical ontologies, and phenotype ontologies in particular, Hoehndorf et al. (Hoehndorf et al. 2007) discuss the problem of combining an ontology for anatomy such as the Foundational Model of Anatomy (FMA) (Rosse and Mejino 2003), which encodes a *canonical* view on its subject matter, with a phenotype ontology such as the Mammalian Phenotype Ontology (Smith, Goldsmith, and Eppig 2004), in which phenotype description often directly contradict the canonical definitions. A simple example is given by defining the anatomy of a mouse, which, canonically, always has as part a tail, whereas a 'mouse without a tail' is surely a reasonable description of a phenotype.

Unlike the approach of (Hoehndorf et al. 2007), using the Distributed Ontology Language (DOL) (Lange et al. 2012; Sojic and Kutz 2012) allows to declaratively specify the non-monotonic extension of an ontology (or a combination of several ontologies) in one ontology specification. In the context of DOL, a multi-logic modelling environment, the non-monotonic framework of choice to be initially supported is *circumscription*, invented by John McCarthy in 1980. The reason for this choice, on a technical level, is simple enough to explain: the basic idea on which circumscription is based is the *minimisation* of (the extension of) concepts or relations (terms for short), assuming the extensions of some terms are fixed, whilst the extension of certain other terms may freely vary. To give formal semantics to this idea, all that is needed is the possibility to define a pre-order on models, which can be done in a logic-independent way.

I will not go here into details of the semantics for non-monotonic extensions, but I rather sketch how the HER2 modelling illustrated above in Figure 16 can be captured in a formal specification.

Formal specification of the HER2 ontology:



```
distributed-ontology HER2
logic log:OWL
ontology HER2 = ProperParthood then

Class : IHC_HER2_Score
        DisjointUnionOf HER2_Negative, HER2_Borderline, HER2_Positive
```

```
Class : HER2_Negative    EquivalentTo {0, 1+}

Class : HER2_Borderline EquivalentTo {2+}

Class : HER2_Positive    EquivalentTo {3+}

Class : HER2 SubClassOf Proteins

Class : Cell DisjointUnionOf: Normal_Cell, HER2+_Cell

Class : Tissue DisjointUnionOf: Normal_Tissue, HER2+_Tissue
And hasHER2Score Exactly 1 IHC_Her2_Score
                        %% we assume all tissue to be IHC measured
                        %% other disjointness axioms left out here...

Class : Cell_Membrane SubClassOf
                        isProperPartOf Exactly 1 Cell

Class : HER2+_Tissue EquivalentTo Tissue And hasHER2Score Some HER2_Positive
                        %% determination of HER2 positivity of membrane by the IHC method

Class : HER2+_Cell SubClassOf inverse isProperPartOf Some HER2+_Cell_Membrane

Individual:X
Types: Tissue And hasHER2Score Some {3+}

Individual:Y
Types: Tissue
DifferentFrom X

minimise HER2+_Cell_Membrane, HER2+_Cell, HER2+_Tissue
vary Normal_Cancer_Cell_Membrane, Normal_Cancer_Cell, Normal_Cancer-Tissue
  then %implies

Individual:X
Types : HER2+_Tissue

Individual:Y
Types : Normal_Tissue
        And hasHER2Score Some (HER2_Negative Or HER2_Borderline)
        %% Y is normal tissue (non-HER2 positive) without evidence to the contrary
```

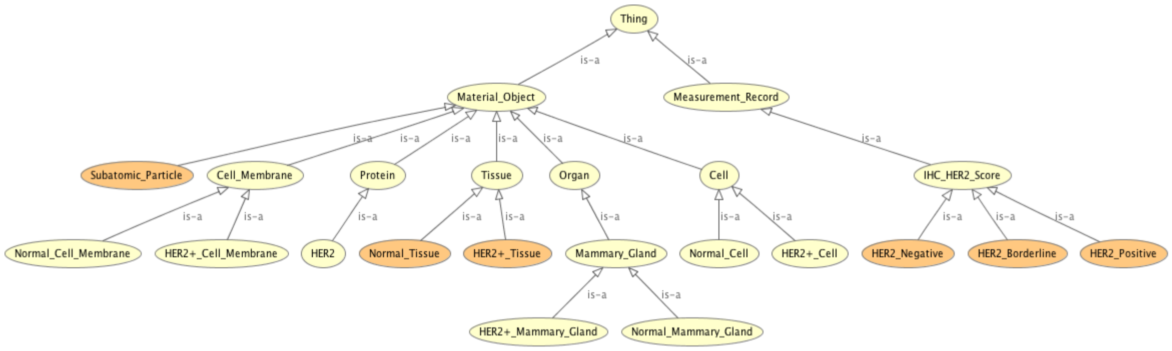


Figure 17 The inferred class hierarchy of the HER2 ontology as displayed in Protégé

Moreover, the HER2 ontology demonstrates how demands of the epistemic groups (see Section 1.2.6) influence each other on a formal level. According to the St. Gallen International Expert Consensus recommendation (Goldhirsch, Ingle, Gelber, Coates, Thürlimann et al. 2009), the threshold for the determination of HER2 positivity by the IHC method is an intense membrane staining of >30% of the tumour cells. However, percentages cannot directly be modelled in OWL. Since percentage of staining is scored by pathologists and clinical oncologists, according to the panel of experts recommendations (Carlson et al. 2006), into one of the classes 0, 1+, 2+,3+, we have used a qualitative abstraction. I.e. we explicitly introduce the IHC HER2 Score as an enumerated concept.

The main idea now is to minimise the extension of the 'abnormal' concepts. I.e. unless there is evidence for overexpression of HER2 in, e.g., a tissue, it will be classified as 'normal'. The following specification sketches this idea.

This ontology in particular assumes that any tissue is assigned an IHC HER2 Score, i.e. any particular tissue, if it were measured, would be assigned a specific value from IHC_HER2_Score. Here, the non-monotonicity comes into play. The minimisation of, e.g., the concept HER2+_Tissue means that any individual, here Y, which does not explicitly meet the defining criteria of HER2+_Tissue will be classified as belonging to its negation, i.e. being a Normal_Tissue. Without circumscription, no such new information could be derived (note that the statements following

%implies are marked as following logically from the above ontology specification). Indeed, consider Figure 17 showing the class hierarchy of the HER2 ontology as implemented in OWL without circumscription. Whilst it does already infer that individual X must belong to the class HER2+_Tissue, it does not infer any information about individual Y.

Unrestricted use of circumscription easily yields undecidable reasoning when applied to even rather weak DLs. Whilst (Bonatti, Lutz, and Wolter 2006) discusses several such undecidability results and fragments of OWL for which it is decidable, there are also interesting restrictions that keep reasoning decidable even over full OWL (Sengupta, Krisnadhi, and Hitzler 2011).

2.2.3. Implementation of HER2 model

I have discussed the importance of non-monotonic modelling for bio-medical ontologies, and have sketched a non-monotonic HER2 modelling, combining the theories of parthood and circumscription of concepts to minimise the 'abnormal' cancer-related concepts. The Hets system (Mossakowski, Maeder, and Lüttich 2007), which provides a prototypical implementation of DOL, is currently being extended to also support non-monotonic inference based on circumscription. The HER2 ontology is a work in progress, and, in collaboration with Kutz, I intend to include further criteria to describe 'abnormal' phenotypes, while reusing various OBO Foundry ontologies such as e.g. the Units of measurement ontology[60], and the Phene ontology (Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010) classes and relations. In particular, we aim at performing an experimental verification of the usefulness of the model. Existing ontology repositories such as BioPortal lack the ability to host heterogeneous ontologies such as the HER2 ontology just

[60] http://obofoundry.org/cgi-bin/detail.cgi?id=unit

sketched. We therefore host the HER2 ontology at OntoHub[61], a new ontology repository currently under active development that will support DOL ontologies and their features in their full generality. Users of OntoHub can upload, browse, search and annotate basic ontologies in various languages via a web frontend. OntoHub accesses the Heterogeneous Tool Set Hets via a RESTful web service interface for having the structure of ontologies analyzed. Hets already supports a large number of basic ontology languages and logics, and is capable of describing the structural outline of an ontology from the perspective of DOL, which is not committed to one particular logic (see (Lange et al. 2012) for more information on OntoHub). Beyond basic ontologies, OntoHub supports linking ontologies across ontology languages, and creating distributed ontologies as sets of basic ontologies and links among them. An important difference to the mapping facilities of e.g. BioPortal is that links in OntoHub have formal semantics, and therefore enable new reasoning and interoperability scenarios between ontologies.

2.3. Plurality of epistemic and pragmatic interests: the 'HER2' case

The epistemic groups involved in ontology building and knowledge representation have been characterised as dealing with a problem according the group specific representational means and specific demands (Section 1.2.6). Nonetheless, the characterised groups (1-6) collaborate in establishment of a common goal. In this section I present a plurality of epistemic and pragmatic interests, while examining the case of merging clinical and molecular reasoning[62] about HER2 protein.

In biomedical ontologies, metadata in the form of tags, annotation, or more generally documentation, is of particular importance. Indeed, many biomedical ontologies have an extremely shallow logical structure, namely consist only of taxonomies, or even just of sets of

[61] http://ontohub.org/

[62] The distinction between clinical and biomedical knowledge, which is relevant for the diversity of clinical and molecular reasoning, I particularly address in Chapter II (Section 3.2).

represented concepts, however accompanied with a rich set of metadata. It is clear that the separation of the epistemic groups from Section 1.2.6 has a direct impact on the kinds of annotations and metadata that can be expected to be generated. For instance, the particular scientific communities (groups (2) and (3)) need not associate identical sets of concepts as related to a term in use. When the Human Epidermal growth factor Receptor 2 (HER2, also known as ErbB2) is used as a tumour marker in the community of clinical oncologists, 'HER2' is related to the diagnostic terms. E.g. over-expression of *HER2* supports diagnosis of HER2 breast cancer, which is described as an *aggressive tumour* with a *poor clinical outcome and a low likelihood of a long term survival*. On the other hand, among the group of molecular biologists 'HER2' is mostly used to characterise molecular processes such as the HER2 related *protein-protein interactions* that trigger the *carcinogenic events*. Of course, 'HER2' can serve as a link between the two domains.

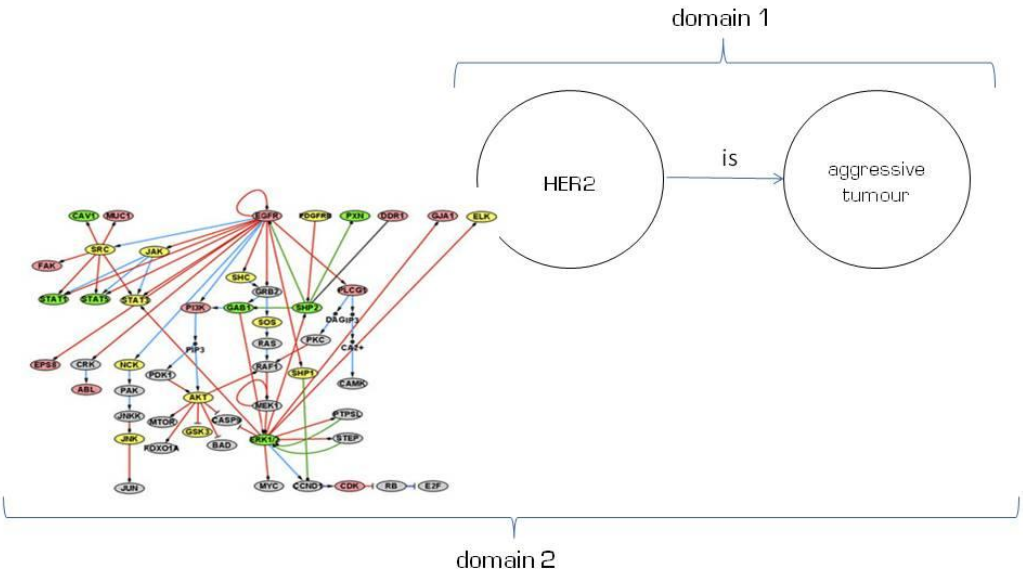


Figure 18 Knowledge granularity

However, as interests diverge among and within disciplines concerning ways of describing a phenotype, distinguishing similarities and difference makers will vary among knowledge domains. So, HER2 will not be the same difference maker for a clinician and for a biologist. The main difference that will be relevant for a clinician will be a difference in the patients survival associated with the expression of HER2 (Slamon et al. 1987). The biologist who focuses on the cellular signalling pathways looks for, for example, a differential expression of the ErbB2 gene while comparing the phenotypes of two types of cell lines (Lacroix and Leclercq 2004). Consequentially, a justification of asserted similarities and generalisations asks for a different kind of evidence in diverse domains. Clinical evidence is acquired through survival analysis and clinical trials while biologists provide evidence through diverse experimental and explanatory methodologies (La Caze 2010). Accordingly, the reasoning of the groups (2)-(4) (characterised in Section 1.2.6) influences the related mappings and justifications implemented by groups (5) and (6).

Furthermore, research interests within a domain evolve over time. So, what had been previously established as terms reference class might shift according to the changes in research focus (Kitcher 2001; Hacking 1993; Griffiths 2009).

A relation between a term and its reference class gets its justification within domain knowledge of a community as an adequate mapping relation. The justification is expressed through claims that support the mapping relations. Regarding the previous example, 'HER2' will be mapped onto a 'bad prognosis' within clinical knowledge (see Figure 18, Domain 1), and the mapping will be justified by the statistical data retrieved from survival analyses. Likewise, biological knowledge provides an alternative mapping relation and a related justification for a mapping between 'HER2' and 'tumour aggressiveness', e.g. protein interaction pathways that result in cell proliferation and tumour aggressiveness (see Figure 18, Domain 2), captured, for instance, in GO, UniProt, and KEGG[63].

Capturing particular instances of clinical and molecular reasoning about a problem, e.g. HER2+ breast cancer, provides recognisable patterns that can be considered as typical to a certain knowledge domain (see Sections 3.2 and 4.2.2). These diverse patterns of clinical and biomedical reasoning can be perceived as domain specific (Evans and Patel 1989).

A detailed analysis of the mappings within and between knowledge domains asks for a multidisciplinary approach involving a community-based process of knowledge production (Gaudillière and Rheinberger 2004; Leonelli and Ankeny 2012). A group of experts with a common interest is collaborating in establishing a shared conceptualisation (Borst 1997) (groups 4 and 5), which gets formalised (group 5) and implemented (group 6) according to the established standards that help them label and describe the domain of interest in an interoperable way (Leonelli 2008; Ekins et al. 2011).

The following section examines how a term such as 'HER2' can be used in order to capture knowledge in different contexts.

2.4. Capturing knowledge in the context of a domain

In this section I address the representational diversity and the flexibility of language that enables capturing knowledge in context. While the first chapter presented how the term's meanings get represented explicitly through the semantic maps and ontology models, this section considers the explication within the contexts that need to deal with the issues such as vagueness, inconsistency, and ambiguity of the reference classes. I return to this issue in Chapter IV, where I

[63] http://www.genome.jp/kegg/

show how knowledge captured by terms supports knowledge integration by connecting various aspects of knowledge associated with the same term.

Obviously, the information captured in a scientific term is *uninformative* when presented outside of a relevant context. A specification of the background knowledge has been offered, *inter alia*, in terms of The Principle of the Presumption of Knowledge and The Principle of Relevance (Strawson 1964).

The Principle of the Presumption of Knowledge states that

> when an empirically assertive utterance is made with an informative intention, there is usually or at least often a presumption (on the part of the speaker) of knowledge (in the possession of the audience) of empirical facts relevant to the particular point to be imparted in the utterance. (p. 97).

As for the Principle of Relevance, it concerns the fact that

> [w]e do not, except in social desperation, direct isolated and unconnected pieces of information at each other, but on the contrary intend in general to give or add information about what is a matter of standing or current interest or concern. (p. 115).

In the case of vagueness it seems even more important to stress how the content of information depends on its use. While disambiguation can be achieved by supplying a definition that specifies the contexts in which a term can be used, the vagueness cannot be so constrained. That is because a term, although defined, has a range of applications that can stretch or shrink depending on the criterion of the user. For example, if a hormone receptor status is defined as positive, it is a scientific context, hypothesis, theory, interrogative interest that supplies a criterion of acceptability etc. In order to state that a tumour is HER2+, it is not sufficient that the receptor is expressed. Various research domains can be interested in different aspects of the protein expression, thus looking for different kinds of thresholds.

The movements of the thresholds (as standardised values) have been particularly significant in the case of the estrogen receptor positive tumours (ER+). Not so long ago the standard for the ER+ assessment has been settled as a condition that at least 20% of the whole amount of the cells in a tumour specimen express ER. The standard has gone down to 10% only to rich 1% nowadays' (Clark 1994; Harvey et al. 1999).

The second, even more troubling example of vagueness demonstrates an absence of standards. A labelling in science, as well as in the ordinary language, often does not have a fixed reference class. For instance, when a tumour specimen observed by a pathologist is getting classified, it is also getting assigned a prognostic label belonging either to the class of 'well differentiated cells' or 'poorly differentiated cells' (Section 3.5.4). Since the reference class for the 'well differentiated' is a subjective qualitative description, which is uneasy to fix by general standards, the labelling and classification turn out to be vague. The reference class that determines the label will depend on the individual pathologist's experience, his background knowledge and theories that he supports, which may eventually depend on his intuition. So, around 30% of the specimens, mainly belonging to the G2 patients, will suffer of the uncertainty caused by the vagueness in the clinical histopathological classification (Ivshina et al. 2006; Sotiriou and Pusztai 2009). It has been reported that there is 50% discordance among pathologist's classification of the grade two (G2) breast cancers.[64]

The third kind of problem with fixing the reference classes of the classificatory terms follows the attempt to restrain the scope of the meaning of a term by standardisation. For instance, the conventional threshold for the 'positivity' assessment will not suit equally well all research contexts, especially those dealing with complex molecular interactions. For example, in the case of a conditional amplification of a protein that inhibits HER2 pathway, a single snapshot

[64] In the third chapter I discuss to which extent the vagueness in the scientific classification is a mix of the empirical questions accompanied with a lack of standardisation and to which extent it depends on the theoretical and experimental context.

of HER2 expression cannot be a general guide for the prognosis assessment. The relevant information might be a ratio between HER2 expression and the expression of another protein that regulates HER2 expression through a feedback loop. So, within the established range of HER2 expression might be another threshold that induces the switching of the expression of the protein that regulates HER2 expression itself (Le, Pruefer, and Bast 2005; Nahta et al. 2006). Thus, dealing with the vagueness of the classificatory labels by limiting the scope of the term's meaning through a simple rounding of the result in a yes-no form, according the established criteria, can have a trade-off in the exploratory, predictive and explanatory power.

Nonetheless, the scientific language, in order to satisfy the requests for precision and disambiguation, employs numerous standards, which aim at establishing the precise reference values as well as the lexical meanings of the terms as the domain dependent (Section 4.1). In that way, the terms are used to describe a domain of interest so that the scientist can communicate and exchange knowledge (Section 4.2.2).

Alongside describing a domain, the domain specific terms acquire particular semantic, cognitive and pragmatic significance. I have presented how the terms are used to label and represent scientific knowledge by means of the semantic maps and ontology models (Section 1.2.2.-1.2.4). I here discuss how a distinction of semantic (S), cognitive (C) and pragmatic (P) values that the terms acquire within a particular representation elucidates which aspects of domain knowledge have been selected to represent and for what purpose.

The semantic value[65] of a term is traditionally understood as its reference, i.e. an object (of reference). In the case of ontology models, the terms are used as labels, while the representational target is scientific reasoning (discussed in 1.2.4. and exemplified by the case of reasoning about HER2 captured within a scientific model such as the IHC detection system). Before discussing the case of HER2 ontology representation, I introduce some basic and distinct

[65] On the other hand, the semantic value of a claim composed of the ordered terms is a truth value (Higginbotham). Instead of speaking in terms of true and false claims, outside of a formal logic context I rather use modes such as acceptable and inacceptable claims, sorting them according to the criteria of adequacy for the relevant empirical context.

features of cognitive, semantic, and pragmatic values that are relevant in representing a biomedical domain.

The cognitive value of a term consists in its heuristic and explanatory roles, which are closely related to the particular semantic and pragmatic specifications (e.g. how a particular term is used within a context)[66]. It has been shown that 'HER2' does not have the identical cognitive value in clinical and biomolecular domains (Section 2.3), because it is not used in the same way within the two domains. Instead, it seems that 'HER2' is used to denote and explain different things (Section 2.3, see also Chapter III).

Concurrently, while the cognitive value of a term plays an important role in the reasoning of individual scientists (group, Section 1.2.6), the collaborative work (groups 1-6, Section 1.2.6) on the ontology building aims at capturing that what is perceived as a shared cognitive value, i.e. conceptualisation of a problem (Section 1.2.6). In other words, the semantic maps and ontology models are capturing explicitly how domain scientists think about a problem. Therefore, through the semantic maps and ontology specification, the cognitive value of a term gets its explicit form.

The particular semantic and cognitive roles that a term plays within a domain becomes explicit through the specification of its semantics, and vice versa, the pragmatic role of the term (what it is used for) will decide its semantics. The term 'HER2' as 1) a tumour marker *and* as 2) a protein with specific biochemical features might be considered as complementary (Section 4.1 also considers 'HER2' ambiguity). However, these two meanings of 'HER2' are not identical, because the same term 'HER2' does not have the same semantic and cognitive value in the context when it is used to denote either a tumour marker or a biochemical component. In practice, the context of use directs the language users to select the most appropriate meaning of

[66] For a contextual interpretation of semantics that regards knowledge claims and their use in practice see for instance (Stanley 2007, 2005). Without going into the theoretical debate that involves different accounts of the semantics and pragmatics in the philosophy of language and psychology, I just aim to illustrate the interdependences of the semantic and pragmatic roles of a term that are apparent in the case of 'HER2' example within the biomedical context.

the term (Section 2.4.). Yet, the distinction of various semantic and cognitive values will be important in an explicit representation of meaning.

For instance, 'HER2' as a tumour marker plays the role of a diagnostic label after getting assigned a measurement value. The measurement value might be provided through various methods (e.g. FISH, IHC). In the case of FISH method, the signal of HER2 *gene expression in nucleus* provides a measurement value for the HER2 marker, while in IHC method the measurement captures expression of *HER2 intramembrane protein* (Sections 2.2.1, 2.2.2). Even so, it seems that within the clinical use of 'HER2' as a tumour marker, the application of language is quite ambivalent about the semantics that might capture either the gene expressed in nucleus or the proteins within the cell membrane. Both methods of measurements are a reliable source of information for a clinician (Carlson et al. 2006). The choice of method and the 'entities' that it captures is less relevant for a clinician than for a biologist[67], because the semantics of 'HER2' in clinical domain primarily aims to capture (and denote) the patients to whom the HER2 positive marker is assigned. Since the reference class of 'HER2' can belong either to a (referential) domain of tumour markers (denoting also the patients) or to a (referential) domain of genes and gene products (denoting also gene expression in a tissue culture), the semantic roles of the term will diverge across the domain contexts. Likewise, the cognitive value of 'HER2' will either explain tumour aggressiveness while denoting the patients involved in the survival analysis, or by denoting the molecular pathways that induce carcinogenic processes (Section 2.3).

In other words, the cognitive value specified by the role that the term plays in the two domains depends on its pragmatic context. As I have outlined in Section 2.3 (and I elaborate further in Chapter III), a primary goal of a biologist is to explain, for instance, how 'HER2' induces cell proliferation, the goal of a clinician is to assign diagnosis and prescribe the most appropriate treatment (Chapter III). Accordingly, the semantic value of the term in a context depends on its

[67] A biologist, on the other hand, might care very much about the method dependent difference, e.g. in the case when the research interest targets the interactions of the intramembrane proteins, 'HER2' semantic value specifically denotes the protein expressed in the cell membrane.

cognitive and pragmatic values. The pragmatic interests of a research domain direct the choice on what explanatory scope for the term is most desirable and which reference class can suit it best. However, when the pragmatic interests are shared among the domains, these various cognitive roles can be merged[68] (Section 2.2.3).

Accordingly, the (pragmatic) fixing of what a scientific term represents (as its explicit meaning when the term is used), integrates the *goal oriented contextualisation* of meaning and the *lexical* aspects[69] of language, which shows a flexibility to the potential applications, i.e. the lexical meaning of terms defined in a domain terminology (Section 4.2.2).

This twofold aspect of the terms' specification is exemplified by a divergence between the initial specification of a term and the re-use of the term within contexts that differ from the original one ('HER2' in biological and clinical context). For instance, the pragmatic goal of 'HER2' specification within an initial labelling (naming a protein) fits the terminological specification that has been employed in solving certain problems in the context of molecular research (Kumar 2008). However, as the problems and possible applications evolve, the meaning of the term evolves as well. Historically speaking, the term 'HER2' had been firstly used to describe and denote a protein, i.e. Human Epidermal growth factor Receptor 2 (an intramembrane receptor that belongs to the family of growth factor proteins). Sequentially, HER2 acquired an additional description of a tumour marker, as soon as its predictive power in clinics has been demonstrated (Kumar 2008; Griffiths 2009).

The HER2 ontology is an example of how the two meanings, HER2 conceptualised as a tumour marker and HER2 conceptualised as a protein, get merged into a representation. Moreover, Figure 19 demonstrates that various representational forms of an ontology model capture different aspects of reasoning about the targeted phenomenon.

[68] We have seen in the case of HER2, an ontology model supports the integration of these various semantic and cognitive roles.

[69] By lexical aspect I consider that how a term is conventionally defined. For the discussion about flexibility of the lexical definitions see Section 4.2.2.

The representations 1-3 in Figure 19 are three different ways to target the scientific reasoning, which uses IHC method to asses HER2 positivity. These models that are targeting scientific reasoning about HER2 (Figure 19, right hand side) reflect various representational (S, C, P) values captured within its various forms. Accordingly, the meaning of 'HER2' captured within a particular representational form can be valued as relevant (semantically, cognitively and pragmatically) according the particular representational context.

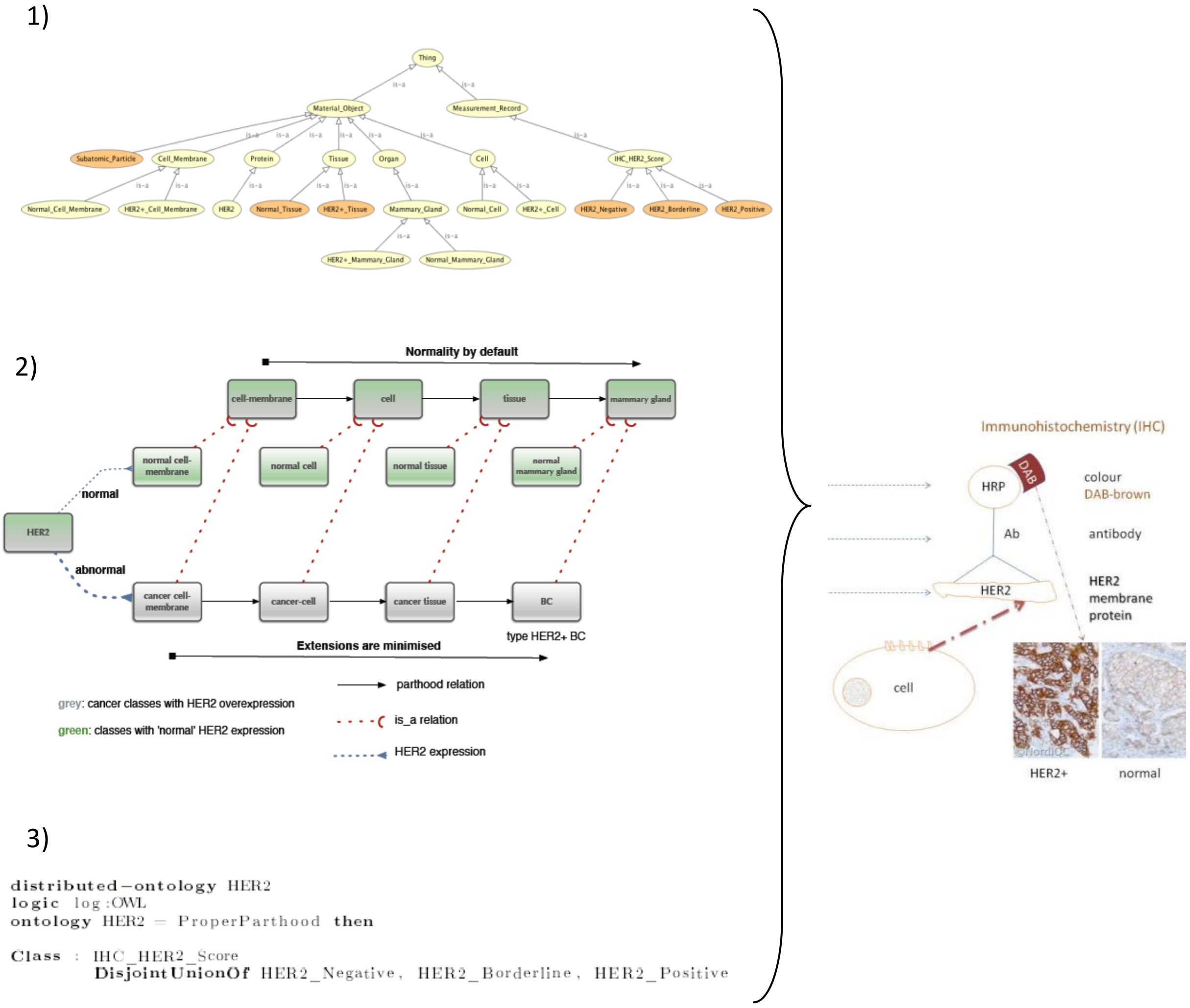
The visualisation of the model in Protégé captures HER2 as a protein that is described as a kind of material object. More precisely, it captures HER2 as a type that is sub-type of particular material objects that are proteins. The semantic value of 'HER2', represented as being of a HER2-type, regards its denotative capability to capture the set of proteins described as HER2.

While the hierarchical structure of the model (represented in 1, Figure 19) nicely serves its heuristic purpose to depict the represented entities according to their generality, such a hierarchy also shows certain heuristic limitations.

First, the Protégé representation of the HER2 model (1, Figure 19) cannot depict simultaneously both the hierarchical subsumtion relations (*is_a*) and the parthood relations (*part_of*) that describe the mereological features of protein, cell membrane etc.. Second, while representing 'normal' and 'abnormal' types of cell (tissue etc.) as disjoint classes in the hiararchy, the visualisation of the model explicates the problem in capturing simultaneously these two mutually exclusive sub-types. Namely, a cell is exclusively either 'normal' or 'abnormal' (i.e. HER2+). Yet, both 'normal' and 'abnormal' cells are conceptualised as a kind of cell.

While 'normal' and 'abnormal' cell are both conceptualised as belonging to a common type (e.g. both normal and abnormal cells are a kind of cell), the assignment of the instances (semantic values in particular cases) will produce an inconsistency in the reasoning. In other words, in the case in which the multiple instances are assigned to the class cell, the reasoning with 'normal' and 'abnormal' (HER2 over-expressed) instances may imply that a cell is both the 'normal' and 'abnormal' cell. Thus, the cognitive value of the Protégé (1, Figure 19) representation also serves to explicate the problem with reasoning about ‘normal’ and ‘abnormal’ features that both can be used to describe a type, but not every instance of it at a time.

Accordingly, the model 2 in Figure 19 complements the reasoning represented in (1) by explicating that the assignment of ‘normal’ and ‘abnormal’ features (as dependent on the HER2 expression and assigned on various level, e.g. cell, tissue) are disjoint at a time. In that way, the ‘normal’ and ‘abnormal’ types are all represented as disjoint, while still describing properly the relevant parthood and *is_a* relations.



```
distributed−ontology HER2
logic log:OWL
ontology HER2 = ProperParthood then

Class : IHC_HER2_Score
        DisjointUnionOf HER2_Negative, HER2_Borderline, HER2_Positive
```

Figure 19 Various representations of a model, capturing different aspects of scientific reasoning

The third representation in Figure 19 is a snapshot of the formal specification of the HER2 ontology that integrates the reasoning represented in (1) and (2) by using DOL, OWL, the default logic and the rules for non-monotonic reasoning that is outlined in (Sojic and Kutz 2012).

Moreover, the reasoning about the instances (and HER2 semantics) that was implicit in (1) and (2) becomes explicit in its formal specification.

Since,

```
Class : Tissue DisjointUnionOf : Normal_Tissue , HER2+_Tissue

And hasHER2Score Exactly 1 IHC_Her2_Score
```

The reasoning of the pathologists about the particular instances of HER2 (together with the metadata, e.g. %% determination of HER2 positivity of membrane by the IHC method), will support the automated reasoning that assigns HER2 positivity to the instances in case HER2 score is provided.

```
Individual : X
Types : Tissue And hasHER2Score Some v
Individual : Y
Types : Tissue
DifferentFrom X
minimise HER2+_Cell_Membrane , HER2+_Cell , HER2+_Tissue
vary Normal_Cancer_Cell_Membrane , Normal_Cancer_Cell , Normal_Cancer-Tissue
   then %implies
Individual : X
Types : HER2+_Tissue
Individual : Y
Types : Normal_Tissue
And hasHER2Score Some ( HER2_Negative Or HER2_Borderline )
      %% Y is normal tissue (non-HER2 positive) without evidence to the contrary
```

In this way the formalised HER2 ontology provides explicit semantics for the terms such as ‘HER2’, while integrating clinical reasoning about the assessment of HER2 positivity with the molecular knowledge about HER2 mereology and hierarchical classification. Surely, the model is open to embrace and integrate the additional descriptions of the HER2molecular features.

I here briefly *summarise* some of the main points addressed in this chapter. I argued that the plurality of biomedical representations of breast cancer phenotypes originate in a variety of research interests distributed across the research domains. I showed that representation of breast cancer phenotypes requires a combination of various representational perspectives and a combination of several ontologies. In addition, I stressed an important problem in modelling a disease phenotype that most of the established ontologies at the moment cannot capture. Therefore, I have addressed certain directions in which an ontology develops while dealing with the particular tasks that are at stake[70], combining segments (modules) of various reference ontology as well as heterogeneous logical formalisms. I also proposed a model, developed in collaboration with Oliver Kutz, which represents 'normal' and 'abnormal' phenotypes in order to describe the HER2+ tumours (Section 2.2.1-2.2.3). The model captures overexpression of the HER2 protein, while representing *mereology* of the HER2+ phenotype. Since HER2 protein is present in both normal and abnormal phenotypes, the model employed non-monotonic reasoning that specifies 'abnormality' according to the available information about the IHC scores. Unless such information is explicitly known, the default tissue is labelled as 'normal'. Accordingly, adding information to the formal modelling results in retracting the property of 'normality'. Since such information (the ranges of values) is a matter of convention and the best available knowledge at a current time, the HER2 example demonstrates that the classification and modelling of a disease phenotype is not fixed once forever by some intrinsic features, but is rather the result of a pragmatic approach to modelling that is open for the revisions in the light of new information, technological and scientific advancements. I also argued that the representation of HER2 involves various kinds of reasoning (clinical, molecular), each of which brings fruitful insight into the problem. Therefore, I argued for a pluralistic view of the ontology modelling that needs to be open for a variety of perspectives and formalisms, which a platform such as OntoHub can support in practice. Finally, I specified how various representational forms (hierarchy, mereology, formal specification) of the HER2 ontology contribute to the understanding of certain aspects of the model, whereby the formal specification integrates the reasoning, captured by the model, in the most comprehensive manner.

In the following chapter, while examining the classificatory knowledge about breast cancer, I discuss how domain knowledge needs to define the roles of the classificatory terms.

[70] For a detailed study that concerns the ontology evolution, in particular GO, see (Leonelli et al. 2011).

Having diverse pragmatic and epistemic aims, various knowledge domains will demonstrate different predictive, explanatory, and evidential criteria that define the relevance of the terms in use.

Chapter III

Classificatory Knowledge and Breast Cancer Classifications

This chapter addresses a number of problems that accompany the attempts of merging knowledge from different domains. In Chapter II, I have explained an existing plurality of disease representations as a result of the perspective driven approaches to the practical problems in biomedicine. A variety of perspectives on a problem seems to play a particularly important role in the case of breast cancer, which is recognised in clinical and molecular research not as a single disease but as a set of diseases with distinct features (Vargo-Gogola and Rosen 2007). In this chapter, I shall further examine the epistemic and pragmatic reasons, which drive domain specific characterisation of this heterogeneity, in order to see if and how the variety of the approaches across the clinical, molecular, and epidemiological domains can be mutually aligned.

I focus on the domain specific characterisations, which have resulted in distinct breast cancer classificatory systems. Therefore, my aim in this chapter is to discuss and sort out similarities and differences of these assorted classificatory systems, each of which represents a particular aspect of knowledge about breast cancer. In the following sections I attempt to elucidate in which respect clinical and molecular knowledge are diverse, and how the domain specificities influence classification within each of the particular sub-domains. In the second part of the chapter, starting from section 3.5, I particularly stress the interdependencies among the clinical, molecular and epidemiological classifications.

As the understanding of a domain and the related classifications is a precondition for an ontology building, my analysis of the classificatory concepts and their use in biomedical practice should provide a conceptual framework for a breast cancer ontology, which can support an integration of the breast cancer related domains.

3.1. Personalized medicine explicating the epistemic gap

Scientific knowledge importantly relies on generalisations (Reichenbach 1951). I here examine scientific generalisation in the context of breast cancer research, while arguing that a sort of epistemic gap emerges when the biomedical generalisations, and in particular those expressed as classificatory knowledge, need to face reasoning about particular cases. The main motivation to undertake an analysis of biomedical reasoning about general and particular cases, which has already been addressed within numerous studies in philosophy of science and medicine, lies in the particular need to explicate reasoning of a domain for the purposes of ontology modelling. Having this in mind, I will show how personalised medicine explicates this epistemic gap between classificatory knowledge, which provides perspective driven generalisations, and application of that knowledge into practice.

A practical reason for making generalisations in science is due to an attempt to uncover common features and relations that describe phenomena of interest. Knowing that something holds in most cases of a research context, scientists can make inferences about new phenomena that are of similar kind. Thus, generalisations are useful when they can support predictions, generation of new hypothesis, and testable implications (Hempel 1966).

Likewise, knowledge claims about cancer could not be easily reused in various contexts if they would not include distinctions of some general cancer related features. In the case of breast cancer, knowledge about cancer features, which is also utilised in clinical assessment of the patients, support structuring of the classificatory systems into categories such as tumour size, nodal status, presence of metastasis, hormonal status, gene expression, tubule formation, poorly differentiated tissue etc. (see 3.5, and 3.6). Consider, for example, categories that include hormonal status of tumour, such as estrogen receptor positive (ER+), estrogen receptor negative (ER-), and HER2+ classificatory categories. ER positivity is a useful category, which describes the tumours that are responsive to hormonal therapy (Weigel and Dowsett 2010; Ludwig and Weinstein 2005). A knowledge claim that *ER+ tumours are responsive to hormonal therapy* is an example of generalisation. This generalisation has also its predictive value and it can generate testable implications, expressed, for example, in a claim that *since the patients with ER+ status are likely to respond to hormonal therapy, they also have a good prognosis*. Obviously, such a prediction is probabilistic and it expresses likelihood that something will be the case (Sadegh-Zadeh 2011). Similarly, the very generalisation about the features of ER+ tumours has been inferred by means of probabilistic reasoning and statistical methods, which compared ER status and therapy response in patients, in order to analyse how likely is that ER+ tumours respond to hormonal treatment (Weigel and Dowsett 2010; Ludwig and Weinstein 2005). As the correlation between ER positivity and the treatment response has shown a high statistical significance, the generalisation was attained as justified. Moreover, this generalisation became very useful tool that helps clinicians to classify patients and recommend therapy.

However, although most ER+ patients will indeed respond to the hormonal therapy, and, consequently, they will have a high likelihood of survival, in some cases it will not be the case. That is to say, generalisations are abstractions, which in the case of probabilistic reasoning abstract from the statistical outliers. However, the statistical outliers represent the real patients, whose tumours behave differently than most of the tumours with seemingly similar features. For example, clinical studies show that not all of ER+ patients respond well to the hormonal therapy.

> Patients with ER+ and serum HER2 positive metastatic breast cancer are less likely to respond to hormone treatment and have a shorter duration of response than ER+ and serum HER2 negative patients. Their survival duration is also shorter (Lipton et al. 2002).

Thus, information about presence of HER2 positive status changes significance of the generalisation, which characterises the ER+ category as the one that indicates good prognosis and prolonged survival. Since a good response to the hormonal therapy and a prolonged survival does not hold for those patients who also have HER2 positive tumours, the above stated generalisation about ER+ tumours is just conditionally valid. The same conditional validity can be shown for almost every other classificatory category. Thus, general knowledge about the tumour features and their grouping into the particular classes that share some common features is always context dependent. Before a treatment for a patient gets tailored, the patient gets classified into one of the existing classificatory groups, each of which is a kind of generalisation. Of course, clinical oncologists are aware of personal differences in therapy response, and many of the parameters are indeed taken into account in clinical reasoning and decision making (Goldhirsch et al. 2011). Nevertheless, a gap between classificatory knowledge as a form of abstraction and its application to the reasoning about particular cases does not lose its significance.

The idea of personalised medicine which should fit particular patients has made explicit this epistemic gap between general knowledge and its applicability to the particular cases. It has been recognised that generalisations of classificatory categories cannot fit every patient (Hamburg and Collins 2010; van 't Veer and Bernards 2008; Davis et al. 2009). According to this view, each patient should be rather treated as an individual with particular combination of features that characterise his particular tumour and its behavioural patterns. Still, scientific knowledge continues to have a form of generalisations, which now need to be carefully stratified and combined in order to fit reasoning about particular cases. Apparently, dealing with patient specific features involves combination of parameters that involve family history (Tyrer, Duffy, and

Cuzick 2004), particular risks considered in the age context (Anderson, Jatoi, and Sherman 2009), genomic profiling (Ivshina et al. 2006; Correa Geyer and Reis-Filho 2009; Pusztai et al. 2006; Sotiriou and Pusztai 2009), and a variety of data aquiered through the methods used in systems biology (Gonzalez-Angulo, Hennessy, and Mills 2010) etc.

On this path of integration of numerous and heterogeneous parameters derived from various sources, knowledge organisation that uses ontologies plays an important role (Dumontier et al. 2010; Gurwitz, Lunshof, and Altman 2006; Hoehndorf, Dumontier, Gennari et al. 2011). Biomedical ontologies aim at structuring categories and instances by explicitly represented relations. In this way, the reasoning with a huge amount of data and heterogeneous information that is relevant for the patients' assessment can become explicit, comprehensible to humans, and processable by machines.

In this chapter I undertake an analysis of biomedical reasoning about breast cancer classificatory systems and related categories, each of which includes parameters that need to be combined in personalised medical reasoning. My analysis of specificities and interdependences of these parameters should provide an initial insight into how classificatory parameters can be combined and explicitly represented in a breast cancer ontology.

3.2. Representing clinical vs. biomedical knowledge

For the purposes of my analysis, which should identify the clinical and biomedical representational domains, this section focuses on a distinction between clinical knowledge and biomedical knowledge. Although the relationship between clinical and biomedical knowledge is particularly intertwined, I specify certain distinctions in order to clarify (further in Sections 3.5 – 3.6) how these two domains address classification and specification of breast cancer. I use a descriptive method, while referring to the published studies on clinical and biomedical reasoning, in order to pursue a comparative analysis, which should explicate the differences and similarities among the two domains. Starting from a description of the disciplinary focus, which produces diversities between clinical and biomedical reasoning, I distinguish some important features that are specific for each of the knowledge domains.

In cognitive psychology *clinical knowledge* is defined as a type of knowledge associated with clinicians who are able to ascribe the attributes to sick people (Evans and Patel 1989; Boshuizen and Schmidt 1992). This type of knowledge includes the description of the ways in which a disease can manifest itself in the patients, the kind of complaints that are expected rgarding the given disease, the nature and variability of the signs and symptoms, and the ways in which the disease can be managed (Evans and Patel 1989). Thus, clinical knowledge is descriptive when a diagnosis needs to be assessed[71] and it is prescriptive when a therapy has to be recommended.

Clinical descriptions capture a disease manifestation through the description of signs and symptoms. The recognition that the description of a patient's signs and symptoms *fits* the 'sickness' attributes, which are characteristic for a disease, helps the clinician to diagnose the disease (Sadegh-Zadeh 2011; Schön 1991). So, a diagnostic act takes place when the clinician's knowledge about the disease, described through the set of attributes, *fits* the set of attributes, acquired through the examination, describing a) the patients' complaints (*symptoms*) and b) the related clinical findings (*signs*).

The cognitive processes involved in the disease assessment, as well as the problems that clinicians need to face in practice, are far more complex than what this brief specification of clinical knowledge can grasp (Evans and Patel 1989; Sadegh-Zadeh 2011; Schön 1991). Still, this distinction gives a general picture of an important aspect that constitutes clinical reasoning as an application driven approach, where the background knowledge (of an individual clinician ${}_{cb}K_{a}$ and clinical community ${}_{Cb}K_{a\ i}$ in Figure 20) about disease manifestation, signs, and symptoms that

[71] A clinical description of the disease attributes also contains normative elements. In the following sections I actually characterise every process of classification and attribute ascription as a normative act. However, I here abstract from such an analysis in order to present some other general features of clinical knowledge.

characterise a disease, is getting matched with an adequate description of a particular patient. So, the diagnostic process involves a *comparison* of 1) the data associated with a disease, in the context of *general* knowledge on that what is known about a disease and how the disease is described in medical literature, and 3) the data about a patient, i.e. *particular* knowledge based on description of the patients' signs and symptoms.

> Clinical interaction requires the understanding of particulars to be integrated with the understanding of universals. When medical knowledge generated from groups is applied to individuals, careful negotiations with the specific patient and situation are essential for adequate understanding and management [(Miller 1992; McWhinney 1989; Hollnagel 1999)]. (Malterud 2001)

In the diagnostic process, the application of general knowledge about the disease thus comprises a particular knowledge by acquaintance with the patients' signs and symptoms, whereby the direction of the diagnostic act goes from general to particular. That is to say,

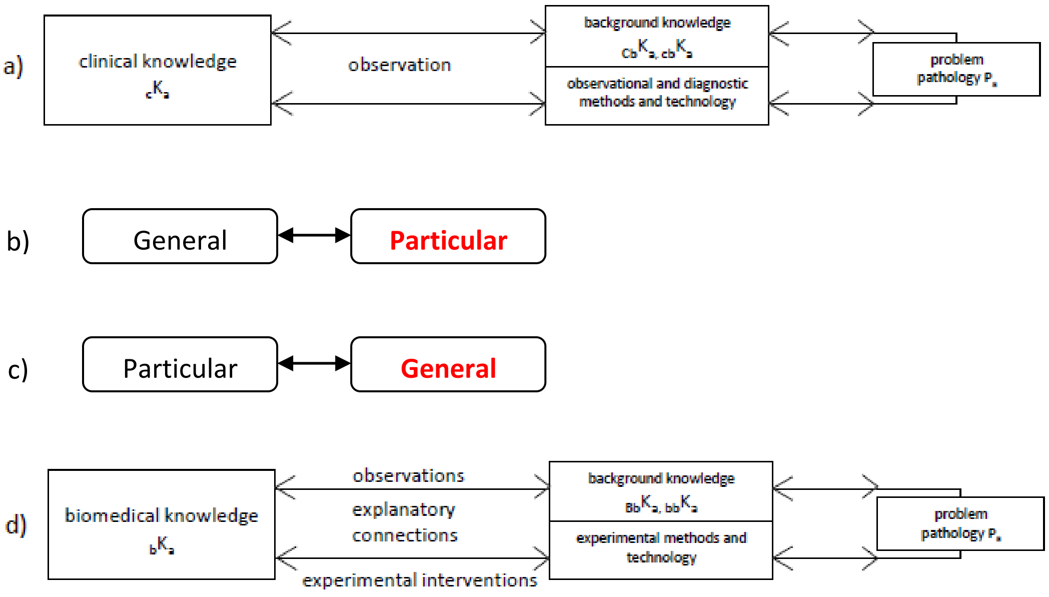


Figure - Clinical and biomedical reasoning
a,b) clinical reasoning: $_cK_a$- clinical knowledge, $_{cb}K_a$ - background knowledge of an individual clinician, $_{Cb}K_a$ - the medical community knowledge; c,d) biomedical reasoning: $_bK_a$- the biomedical knowledge, $_{bb}K_a$ - background knowledge of an individual biomedical scientist, $_{Bb}K_a$ - the biomedical research community knowledge.* Red colour marks a dominate aspect

Figure 20 Clinical and biomedical reasoning

diagnosis is directed from knowledge about the disease to the application of this knowledge onto a particular instance of the disease. The *directedness* of clinical reasoning *towards the instances* (in Figure 20 marked in red colour) has the opposite direction than the one I will indentify as crucial for achieving some important targets of biomedical reasoning.

Biomedical knowledge is knowledge about the pathological principles, mechanisms, or processes underlying the manifestations of disease (Patel 1989; Boshuizen and Schmidt 1992). Thus, similarly to clinical knowledge, biomedical knowledge also employs descriptive terms in capturing its targets, i.e. pathological processes and mechanisms. However, the focus of a biomedical description is on the *relations* that hold among the scientific targets, which are labelled by descriptive terms[72]. That is to say, a description of the mechanisms and processes emphasises the causal connections that can *explain* why and how a disease develops (Schaffner 1993; Thagard 1998). Such a description is, in a certain respect, more complex and eventually more general in its character than the description of signs and symptoms that are attributed to a particular patient. The causal connections that scientists are looking for, in order to *explain* why and how a disease develops, usually do not aim at capturing one particular instance of a causal event. Instead of explaining an *instance* of a disease, biomedical scientists are looking for an explanation of a *type* of disease. Therefore, the mechanisms and processes, which are often targets of a biomedical explanation, are those mechanisms that are common for most instances of a disease type. Even if knowledge acquisition starts from the observations of particular pathological cases in an experimental setting (molecular, histological etc.), the acquired observations would not make a significant contribution to biomedical knowledge, if they would not also support a generalisation about the observed phenomena, either confirming or refuting the proposed claims about the phenomena. Indeed, various experimental methodologies and experimental tools are used in achieving such a goal that drives scientific inference from particular

[72] Note that in section 1.2.4 I specified exactly this *relation* among the terms (and related concepts) as the main target that an ontology model attempts to capture (Figure 9).

claims about experimental results, to the general claims, which can be applied on to many other cases that represent similar kind of phenomena.

While a set of descriptions composed of (clinical) signs and symptoms does not always need to have an explanatory role in order to provide a clinically useful disease characterisation, such a description within biomedical knowledge is utilised just as a hint in the search for the related causal relations and explanations. A simple example is fever and a flue diagnosis. In the absence of information about any other severe condition that manifests itself with fever, the fever as a sign of flue will support a diagnosis of flue. Indeed, a clinician does not need to go into details, which would explain the mechanisms of immunological reactions resulting in fever, or into those mechanisms that explain how a particular pathogen causes fever. Likewise, although the symptoms are even more difficult to understand and explain than signs (Malterud 2000), they are often a useful mark for a clinician in recognising a disease.

Within biomedical knowledge there is a tendency for an integration of knowledge about signs and symptoms with knowledge about the causal relations that explain their presence, connecting disease manifestation with the biological, chemical, and physical processes involved in a disease (Evans and Patel 1989; Schaffner 1993).

Obviously, biomedical knowledge is not far apart from clinical knowledge, but it is rather closely linked with it. Indeed, the scheme of clinical reasoning in Figure 20 represents biomedical knowledge as a constitutive component of clinical knowledge and reasoning. A clinician, having acquired medical education, uses biomedical knowledge, which helps him to understand the disease mechanisms in order to asses an adequate diagnosis (Sassower and Grodin 1987). Even so, a difference in disciplinary focus explains a difference in clinical and biomedical reasoning, which influences the related domain-specific representations.

Contrary to biomedical students and researchers, the understanding of the disease mechanisms is not the main target of a practitioner, but rather a *useful tool* in achieving his main goal, which is the treatment of a patient. Accordingly, signs and symptoms are often used by clinicians as ‘shortcuts’ in reasoning about a disease. The empirical studies, which measured velocity and correctness in solving complex cases (Evans and Patel 1989; Boshuizen and Schmidt 1992; Patel 1989), demonstrate that knowledge acquired through the clinical practice (background knowledge of an individual clinician (Boshuizen and Schmidt), ${}_{cb}K_a$ - in Figure 20) helps clinicians to perform better than biomedical students who have not yet acquired practical experience with patients. The difference in performance has been explained, *inter alia*, by causal reasoning utilised by biomedical students in problem solving. According to the cognitive psychologists who designed the studies in order to test and compare the two types of reasoning, the use of causal reasoning in problem solving was, most likely, slowing down the inferential process from the description of signs and symptoms to diagnosis (Patel 1989; Patel, Kaufman, and Magder 1991). The studies also show that the clinicians were widely using their background knowledge acquired in practice, while making 'shortcuts' in reasoning.

Thus, regarding the directionality of clinical and biomedical reasoning, we can distinguish two paths: 1) from particular to general and 2) from general to particular (Figure 20). While in the process of learning these two paths highly interact, in the process of knowledge application and knowledge production one of the paths usually has the priority. The recognition of this difference in prioritising one of the directions in reasoning about disease is particularly relevant for the characterisation of conceptualisation that an ontology should capture. For this purpose, I outline a general distinction of the domain-specific directionality that influences the disease representation and conceptualisation within the clinical and biomedical domains.

Learning from the case studies, i.e. problem based learning (PBL), is one of the methods often utilised in medical education (Donner and Bickley 1993; Griffiths 2009; Kljakovic 2001). PBL is evaluated as an efficient method in training students to apply their knowledge to the real cases (Donner and Bickley 1993). Namely, students learn how to apply efficiently general knowledge on the cases, while by studying the cases they also learn some general features of the disease as well as the exceptions and variations that they may encounter in clinical practice. However, when the knowledge needs to be applied to a particular diagnosis, the process of reasoning eventually goes from general to particular. The interplay of the reasoning about general and the reasoning about particular seems to be common also to the biomedical domain. Even if handbooks typically present general knowledge about phenomena, learning by doing experiments rather resembles PBL in medicine (Scharfenberg, Bogner, and Klautke 2007). Likewise, application of biomedical knowledge to particular instances, when a single phenomenon needs to be understood and explained, requires application of previously acquired general knowledge to the instances. However, as the aims of biomedical knowledge go beyond understanding of singularities, the *understanding of the singular* processes and relations involved in a disease, *rather serves as a tool* in production of *(general) biomedical knowledge* that can explain the *types of phenomena* of similar kind.

Moreover, while an encounter with *similar* kinds of disease phenomena in the clinical domain (which produces clinical knowledge (${}_{cb}K_{a}$) acquired through experience), has a direct applicability in classifying patients's signs and symptoms (Brooks, Norman, and Allen 1991), an encounter with *similar* kinds of phenomena in the biomedical domain cannot directly acquire the status of biomedical knowledge (${}_{Bb}K_{a}$). It will rather contribute background knowledge of a scientists (${}_{bb}K_{a}$) who further needs to investigate the similarities. In other words, unless an explanation for the similar manifestations has been provided and supported with scientific evidence, the recognition of similarity among phenomena will maintain the status of intuition, which can nevertheless support generation of hypothesis.

According to the distinctions stressed above, an ontology representation, which aims at capturing the connections and relations among observational (descriptive) terms of *biomedical* knowledge, actually *explicitly represents the explanatory connections* among those terms[73]. Instead of making an *implicit* connection of signs, symptoms and the disease manifestation, biomedical knowledge uses the scientific methods and the experimentation in order to *explicate*

[73] In that way, when connected through the explicit explanatory relations, the observational scientific terms can be also considered as the theoretical terms. The explanatory connections among the scientific terms are justified within a theory, and accepted by scientific community as a received view (${}_{Bb}K_{a}$). I need not go here into the philosophical debate on the observable-theoretical distinction, the theory-ladeness and the ontological commitments, since my argument is focused on another level. What I want to stress here is that *the explanatory connections that hold among the terms play a crucial role in the biomedical knowledge representation*.

these connections. An ontology, on the other hand, provides a tool to explicitly represent both, knowledge claims and the justification of the claims, by representing the experimental methods, interventions, the difference makers and the use of the controls (Brinkman et al. 2011; Karp et al. 2004).

In the context of ontology engineering, the distinctions I have made between clinical and biomedical knowledge can be informative in explaining[74] why

1) the domain of discourse and represented concepts do not overlap more significantly across the clinical and biomedical ontologies (Milian et al. 2010);
2) clinical ontologies either have a form of thesauri (SNOMED CT), being less formal, or, if formalised, then, they are rather designed to support clinical decisions (Bodenreider 2008) than to provide a formal representation of the explanatory connections among the represented concepts, e.g. disease mechanisms;
3) biomedical ontologies include formal representations that explicate causal relations, which can explain various aspects of diseases and support knowledge discovery (Hoehndorf, Schofield, and Gkoutos 2011; Röhl ; Gangem, Catenacc, and Battaglia 2004);
4) clinicians resist to accept use of ontologies and the information based systems for the decision support and rather rely on their own experience and intuitions (Wears and Berg 2005);
5) indeed, biomedical ontologies aim at integrating biomedical and clinical knowledge in order to support evidence based medicine (Hadzic and Chang 2005; Gangem, Catenacc, and Battaglia 2004).

[74] The conclusions I draw have support in the literature, though additional empirical testing of the domain ontologies needs to be done.

In the following sections I will first analyse some general aspects of language that illustrate distinctions among the domains. Then I will make the connections between language and classification. The introduced distinctions between clinical and biomedical knowledge will be particularly relevant for the discussion of clinical, molecular, and epidemiological classifications of breast cancer in the sections 3.5-3.6, where I show how the disciplinary aims direct particular classificatory choices.

3.3. The language of a domain: with and without 'patient'

An ontology captures a conceptualisation of a domain, communicated and represented by means of language that labels the classes, instances and relations among the objects of interest. Thus, defining a vocabulary and the domain of discourse is important for ontology design (Chapter I). For this reason, the distinctions I will make in this section are relevant for the later discussion about biomedical classification (Sections 3.5-3.6), which a breast cancer ontology aims to capture.

I present a few examples which demonstrate a terminological diversity, which I explain as focus dependent. Unlike in the previous section, I will not proceed by specifying particular domains as the distinct ones. I rather use an empirical test in order to show how terminology reflects the domain diversity. So, I perform my analysis by looking at the terminological distinctions among the research domains. First I explain my criteria for selecting 'patient' as a difference making term.

The pragmatics of classificatory language relates to the practical side of language to be used as a *tool for marking the objects of interest*. As the objects are marked by means of language it is possible to look at the language the other way round, because the language also reflects *which generalisations are established* within a community and therefore *what is the (representational) target of domain knowledge*. For example, the language of the clinical domain includes, *inter alia*, the terms 'patient', 'therapy', 'clinical outcome', and 'diagnosis'. In section 3.2 I have shown how clinical knowledge depends on the pragmatic interests of clinicians to help people improve their health conditions. According to this aim, the choice of classificatory categories in clinical practice is very much guided by a search for a framework that can endorse the best available *diagnosis* and *treatment* assessment (Charlin, Tardif, and Boshuizen 2000). Taking into consideration the goals of a clinical domain, the organisation of clinical knowledge in terms of diagnosis, clinical outcome, and therapy becomes evident. Therefore, some of the central terms within the clinical discourse (e.g. ‘patient’, ‘disease’, ‘diagnosis’, ‘prognosis’, ‘clinical outcome’, and ‘therapy’) are also used in the clinical classifications, where they play a crucial role in structuring clinical knowledge.

In order to provide a justification for my claim that language reflects the domain interests, which highly vary, I will show that the same terms do not have the same status within the research fields whose focus goes beyond the patients’ treatment, e.g. biochemistry and cell biology.

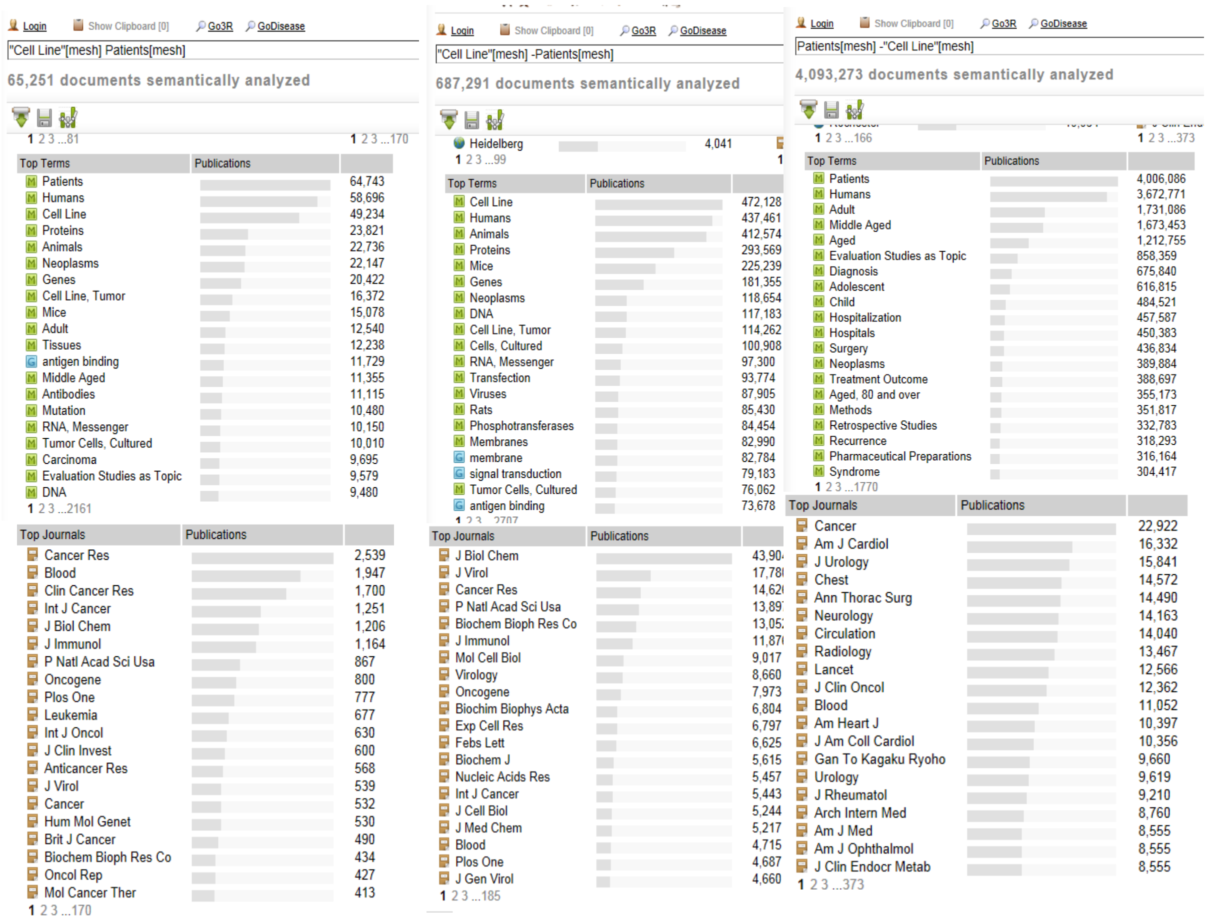
I present here the data retrieved from GOPubMed[75], which is a knowledge-based search engine for biomedical texts, The Gene Ontology (GO), and Medical Subject Headings (MeSH) (Doms and Schroeder 2005; Dietze and Schroeder 2009). It structures the millions of articles from the MEDLINE database, allowing the filtering and analysis of results according to the needs of a user.

I selected the term ‘patient’ as one of the key terms frequently used in the clinical domain. The goal is to test how often the term appears in biomedical literature as well as in which kind of literature. In order to test the variety of the contexts in which the term appears I combined a set of terms that included the terms from laboratory research, such as ‘cell line’, as well as the terms ‘neoplasm’ and ‘genes’. I performed a query that resulted in the groups of papers, those which abstracts contained the term ‘patient’ and those which did not contain it. In addition, the documents were semantically analysed, showing the list of terms that are present

[75] http://www.gopubmed.org

within each of the groups. The terms' list is ranked by the number of publications which contain them.

By filtering the documents for those ones which contained both terms, 'Cell line' AND 'Patients' (Cell Line"[mesh] + Patients[mesh]), the semantic analysis was performed on 65,251 documents. On the other hand, the 'Cell line' documents which did not contain the term 'Patient' (Cell Line"[mesh] -Patients[mesh]) were semantically analysed in 687,291 documents, while 'Patients' documents which exclude 'cell line' (Patients[mesh] -"Cell Line"[mesh]) were analysed in 4,093,273 documents. Figure 21 shows ordering of the results for each search respectively.



"Cell Line"[mesh] Patients[mesh]

65,251 documents semantically analyzed

1 2 3 ...81   1 2 3 ...170

Top Terms	Publications	
Patients		64,743
Humans		58,696
Cell Line		49,234
Proteins		23,821
Animals		22,736
Neoplasms		22,147
Genes		20,422
Cell Line, Tumor		16,372
Mice		15,078
Adult		12,540
Tissues		12,238
antigen binding		11,729
Middle Aged		11,355
Antibodies		11,115
Mutation		10,480
RNA, Messenger		10,150
Tumor Cells, Cultured		10,010
Carcinoma		9,695
Evaluation Studies as Topic		9,579
DNA		9,480

1 2 3 ...2161

Top Journals	Publications	
Cancer Res		2,539
Blood		1,947
Clin Cancer Res		1,700
Int J Cancer		1,251
J Biol Chem		1,206
J Immunol		1,164
P Natl Acad Sci Usa		867
Oncogene		800
Plos One		777
Leukemia		677
Int J Oncol		630
J Clin Invest		600
Anticancer Res		568
J Virol		539
Cancer		532
Hum Mol Genet		530
Brit J Cancer		490
Biochem Bioph Res Co		434
Oncol Rep		427
Mol Cancer Ther		413

1 2 3 ...170

"Cell Line"[mesh] -Patients[mesh]

687,291 documents semantically analyzed

Heidelberg 4,041
1 2 3 ...99

Top Terms	Publications	
Cell Line		472,128
Humans		437,461
Animals		412,574
Proteins		293,569
Mice		225,239
Genes		181,355
Neoplasms		118,654
DNA		117,183
Cell Line, Tumor		114,262
Cells, Cultured		100,908
RNA, Messenger		97,300
Transfection		93,774
Viruses		87,905
Rats		85,430
Phosphotransferases		84,454
Membranes		82,990
membrane		82,784
signal transduction		79,183
Tumor Cells, Cultured		76,062
antigen binding		73,678

1 2 3 ...2707

Top Journals	Publications	
J Biol Chem		43,90
J Virol		17,78
Cancer Res		14,62
P Natl Acad Sci Usa		13,89
Biochem Bioph Res Co		13,05
J Immunol		11,87
Mol Cell Biol		9,017
Virology		8,660
Oncogene		7,973
Biochim Biophys Acta		6,804
Exp Cell Res		6,797
Febs Lett		6,625
Biochem J		5,615
Nucleic Acids Res		5,457
Int J Cancer		5,443
J Cell Biol		5,244
J Med Chem		5,217
Blood		4,715
Plos One		4,687
J Gen Virol		4,660

1 2 3 ...185

Patients[mesh] -"Cell Line"[mesh]

4,093,273 documents semantically analyzed

1 2 3 ...166   1 2 3 ...373

Top Terms	Publications	
Patients		4,006,086
Humans		3,672,771
Adult		1,731,086
Middle Aged		1,673,453
Aged		1,212,755
Evaluation Studies as Topic		858,359
Diagnosis		675,840
Adolescent		616,815
Child		484,521
Hospitalization		457,587
Hospitals		450,383
Surgery		436,834
Neoplasms		389,884
Treatment Outcome		388,697
Aged, 80 and over		355,173
Methods		351,817
Retrospective Studies		332,783
Recurrence		318,293
Pharmaceutical Preparations		316,164
Syndrome		304,417

1 2 3 ...1770

Top Journals	Publications	
Cancer		22,922
Am J Cardiol		16,332
J Urology		15,841
Chest		14,572
Ann Thorac Surg		14,490
Neurology		14,163
Circulation		14,040
Radiology		13,467
Lancet		12,566
J Clin Oncol		12,362
Blood		11,052
Am Heart J		10,397
J Am Coll Cardiol		10,356
Gan To Kagaku Ryoho		9,660
Urology		9,619
J Rheumatol		9,210
Arch Intern Med		8,760
Am J Med		8,555
Am J Ophthalmol		8,555
J Clin Endocr Metab		8,555

1 2 3 ...373

Figure 21 Semantic analysis - GOPubMed (*Mix* 'patient' +'cell line'; *Res* 'cell line' -'patient'; *Clin* 'patient' -'cell line')

An analysis of the results of this search can take various directions[76]. The result that is most relevant for my argument concerns the research focus captured through language, whereby

[76] For example, a publication bias, the research funding policies, and the search engine design are for sure interesting points that should be considered in explaining some of the search results. Having in mind many

presence of the selected set of terms modifies 1) number of retrieved documents; 2) ordering of the terms that are associated with the input terms, according to the number of documents in which they appear, and 3) the ordering of the journals that the related documents are published in.

The first and obvious insight comes from a number of documents retrieved for each search. The papers that consider 'patients' out of the cell line research context are far more numerous (4,093,273) than those which do include 'cell line'(65,251). The result is obvious because not every research related to the patients considers also cell lines. Likewise, as it is well known that there are many research fields that are related to the use of cell lines without considering patients (e.g. tissue culture of experimental cell lines), the retrieved result (687,291) is not surprising. However, the result is still informative and it provides evidence that some research context in biomedical domain do not focuses (explicitly[77]) on patients. Moreover, it shows that the fields that exclude the term 'patient' also exclude other patient related terms.

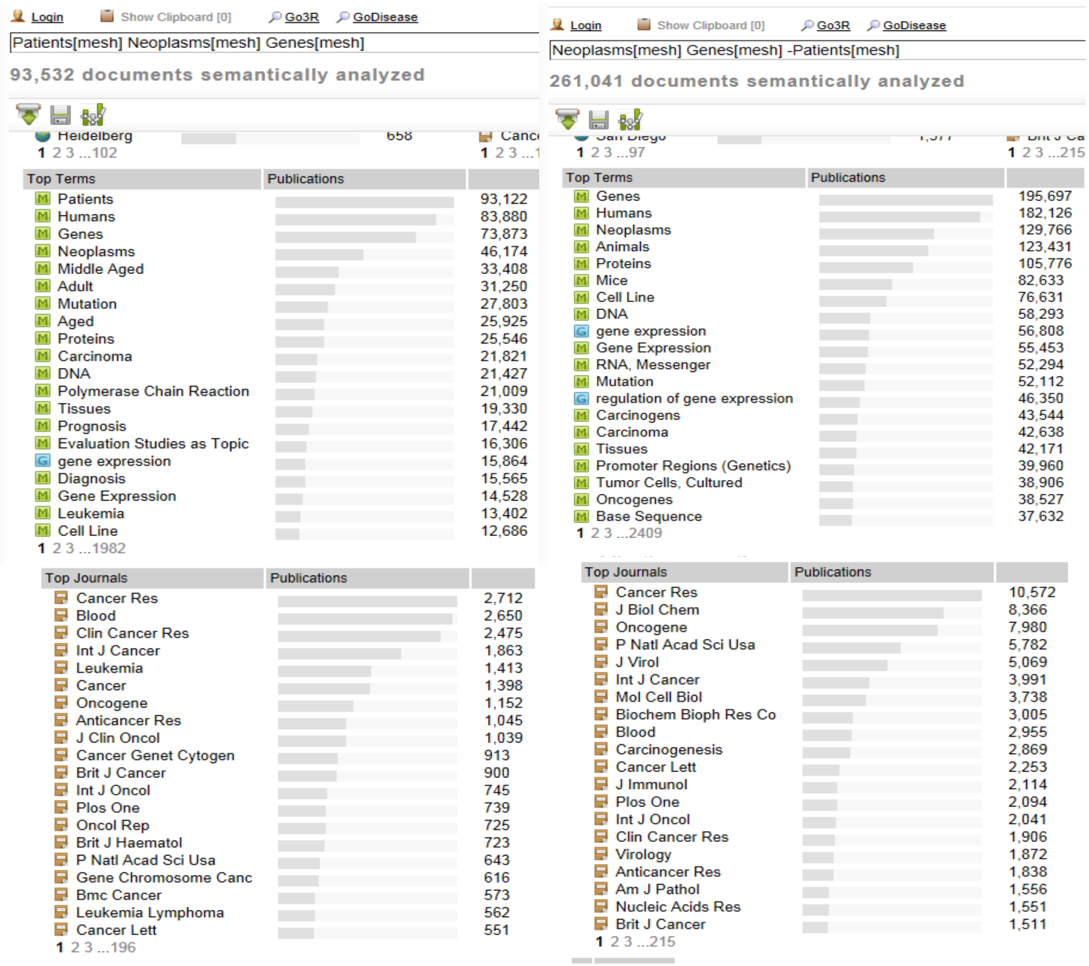
When we compare the three columns, the first one ('patients '+'cell line') is the most heterogeneous with the respect to the terms that appear in it (e.g. 'adult', 'antigen binding', 'genes' etc.), as well as the types of the sours journals (*Journal of Clinical Investigations, Oncogene*, etc.). The group is a mix of publications that contains clinical and experimental research reports. Therefore, I label this group as *Mix*, because it constitutes a *mixed* class of terms and publications. That does not mean that the other two columns are not mixed groups, but it means that they show less heterogeneity among the key terms retrieved by the semantic analysis than the first column. The middle column ('cell line'- 'patients '), as expected, includes more (basic) research oriented publications than the other two. The focus terms describe a rather basic

possible reasons that can induce a bias result of the semantic analysis, e.g. the search limited to the abstracts and not to the whole text, I forego from making a strong claim about an unquestionable validity of the test performed on GOPubMed, which I use rather as an operative hint to illustrate my point.

[77] That is to say, terminology of a published research captures the domain of discourse explicitly, whereby certain terms are used more frequently than some others, depending on particular needs to communicate the message. Thus, the terms emphasised within a domain-discourse reflect the domain-specific representational focus (captured terms) and, arguably, the research focus in a broader sense.

research domain, while those terms that are patient related such as ‘age’ and ‘diagnosis’ do not appear in the result, at least not in the top positions. Also, the journal ranking here includes, as the top ranked, the basic research journals (e.g. Molecular Cell Biology, Nucleic Acids Research etc.), which are either absent or not so highly ranked in the other two columns. So, I label this column as *Res*, because it is *research* oriented. The third column, which is on the right hand side (‘patients ‘-’cell line’), shows as significantly associated with the clinical terms and clinically relevant journals. Thus, I label it as *Clin*, which stands for a *clinically* cantered column. The terms highly ranked in *Clin* contain a variety of age related terms, as well as the terms ‘diagnosis’, ‘hospitalisation’, ‘treatment outcome’, ‘reoccurrence’, ‘syndrome’ etc. Accordingly, the top ranked journals in *Clin* are, broadly speaking, the journals that cover clinical domain.

I have performed a similar test for the set of terms ‘genes’, ‘neoplasms’ and ‘patients’, where the first column (left) presents results of semantic analysis that includes all three terms,



Patients[mesh] Neoplasms[mesh] Genes[mesh]

93,532 documents semantically analyzed

Heidelberg 658
1 2 3 ...102

Top Terms	Publications
Patients	93,122
Humans	83,880
Genes	73,873
Neoplasms	46,174
Middle Aged	33,408
Adult	31,250
Mutation	27,803
Aged	25,925
Proteins	25,546
Carcinoma	21,821
DNA	21,427
Polymerase Chain Reaction	21,009
Tissues	19,330
Prognosis	17,442
Evaluation Studies as Topic	16,306
gene expression	15,864
Diagnosis	15,565
Gene Expression	14,528
Leukemia	13,402
Cell Line	12,686

1 2 3 ...1982

Top Journals	Publications
Cancer Res	2,712
Blood	2,650
Clin Cancer Res	2,475
Int J Cancer	1,863
Leukemia	1,413
Cancer	1,398
Oncogene	1,152
Anticancer Res	1,045
J Clin Oncol	1,039
Cancer Genet Cytogen	913
Brit J Cancer	900
Int J Oncol	745
Plos One	739
Oncol Rep	725
Brit J Haematol	723
P Natl Acad Sci Usa	643
Gene Chromosome Canc	616
Bmc Cancer	573
Leukemia Lymphoma	562
Cancer Lett	551

1 2 3 ...196

Neoplasms[mesh] Genes[mesh] -Patients[mesh]

261,041 documents semantically analyzed

1 2 3 ...97
1 2 3 ...215

Top Terms	Publications
Genes	195,697
Humans	182,126
Neoplasms	129,766
Animals	123,431
Proteins	105,776
Mice	82,633
Cell Line	76,631
DNA	58,293
gene expression	56,808
Gene Expression	55,453
RNA, Messenger	52,294
Mutation	52,112
regulation of gene expression	46,350
Carcinogens	43,544
Carcinoma	42,638
Tissues	42,171
Promoter Regions (Genetics)	39,960
Tumor Cells, Cultured	38,906
Oncogenes	38,527
Base Sequence	37,632

1 2 3 ...2409

Top Journals	Publications
Cancer Res	10,572
J Biol Chem	8,366
Oncogene	7,980
P Natl Acad Sci Usa	5,782
J Virol	5,069
Int J Cancer	3,991
Mol Cell Biol	3,738
Biochem Bioph Res Co	3,005
Blood	2,955
Carcinogenesis	2,869
Cancer Lett	2,253
J Immunol	2,114
Plos One	2,094
Int J Oncol	2,041
Clin Cancer Res	1,906
Virology	1,872
Anticancer Res	1,838
Am J Pathol	1,556
Nucleic Acids Res	1,551
Brit J Cancer	1,511

1 2 3 ...215

Figure 22 GOPubMed (‘genes ‘neoplasms’ -’patients’ vs. ‘genes ‘neoplasms’ -’patients’)

while the second column (right) excludes the documents that do not contain the term ‘patient’ (Figure 22). The results of the test confirms that the focus of a field, reflected through the language, changes a whole set of related terms that appear in one field and get marginalised in another one.

The ‘patient’ group contains also ‘middle aged’, ‘adult’, ‘aged’, ‘prognosis’, ‘diagnosis’ (corresponds to *Clin*), while those terms are marginalised in the other group (right column, corresponds to *Res*). A variation in the journals ranking, also confirms a disciplinary variety that is reflected through language and the terms used in addressing a problem. For example, while *Molecular Cell Biology* is highly ranked among the results which do not include ‘patient’ (Figures 21 and 22), it does not appear as a central journal for the research papers which include the term ‘patient’. Accordingly, a whole set of the related terms demonstrates a diversity in the disciplinary discourse, which prioritise a domain specific terminology. Thus, in both tests, the term ‘patient’ was a difference maker, which supported a distinction between clinical and other biomedical research domains.

Apparently, the domains such as cell biology often consider just those terms that describe normal functioning of biological organisms. So, only if the cell biology *also* addresses the disease related processes, clinical terminology can be of some importance. The missing clinical terms in our test, therefore, might be explained by the domain dependent differences in research focus. Accordingly, whilst the basic clinical terms might be present in the vocabulary of cell biology, they appear less frequently than many other terms, which are used within the cell biology discourse.

Another point worth mentioning regards the role of clinical terms in the applications for research funding, whereby they refer to the potential medical applications. In such a context, justification of the research goes beyond an epistemic need for understanding biological phenomena (Kitcher 2001; Kincaid, Dupré, and Wylie 2007), while emphasising usability of the project for society.

In the following sections I analyse in more detail how the diversity in the key terms choices influences structure of classification.

3.4. Classifications: why, what, and how

Knowledge and classification are unavoidably interdependent. As soon as something gets settled as an object of knowledge, a classification is made (Goodman 1978). Classification is a grouping of features or objects that are perceived as similar. But, identification of similarity already assumes classification as establishment of criteria for that *what* will be, consciously or unconsciously, 'classified' as similar. As Nelson Goodman phrases it, '[...] the response to the question 'Same or not the same?' must always be 'Same what?'' (Goodman 1978) (see also (Wiggins 1980)). So, a choice of 'what' is going to be compared is the classificatory criterion which actually precedes a classification aware comparison of objects that will afterwards be grouped as being the same, similar or diverse. Classification in this broad sense concerns firstly the processes of perception as cognitive act that is structuring content in a categorical form (Hanson 1958; Goodman 1978; Harnad 1990). The aspect of knowledge that I address as classificatory knowledge has a somewhat different scope from the perceptual cognitive categorisation. Even so, in my approach to classification, I emphasise a cognitive role in the process of organising and structuring knowledge into the classificatory categories.

I specify classificatory knowledge as knowledge expressed by means of language that is in use within a community (Putnam 1975; Dupré 1981; Kitcher 2001). Classificatory knowledge through its linguistic expression acquires an external representation[78] of that what is known about a domain and captured in classificatory terms within the structure of a classificatory system. So, the classificatory terms can be intersubjectively related to the reference classes (Putnam 1973),

[78] For a variety of representational means by which knowledge that is captured in language, can acquire an external form see Section 1.2.6.

while grouping and labelling the objects of interest. A diversity of interests, however, influences particular classificatory choices about *what* will be classified and *how* it will be done (Dupré 1981).

In this section I discuss how domain interests influence classification. I also address the consequences of the domain shaped classification to the biomedical ontologies. While discussing Ludger Jansen's paper 'Classifications' (Jansen 2009), I point out important aspects of scientific classifications presented in the paper as well as certain problems that are neglected therein.

Jansen attempts to characterise what a good classification is by a critical analysis of a 'classificatory parody', which exemplifies a bad classification. The classification in question is presented by Jorge Luis Borges (1981), as coming from a certain Chinese Encyclopaedia. I follow Jansen in his naming of this classification as CAT, which stands for *Chinese Animals Taxonomy*.

CAT:

(1) those that belong to the Emperor

(2) embalmed animals

(3) trained animals

(4) suckling pigs

(5) mermaids

(6) fabulous animals

(7) stray dogs

(8) those animals included in the present classification

(9) animals that tremble as if they were mad

(10) innumerable animals

(11) animals drawn with a very fine camelhair brush

(12) others

(13) animals that have just broken a flower vase

(14) animals that from a long way off look like flies

Obviously, this list is composed of various kinds of categories which are not systematically ordered and organised. Thus, as Jansen points out, the list can serve as a good heuristic example of mistakes and problems that a classification needs to face. Indeed, the author shows that some of 'comical' features of CAT also appear in the National Cancer Institute Thesaurus (NCIT), which also contains the type 'others' and a mixture of heterogeneous classificatory kinds (Jansen 2009).

Instead, the features that a good classification should have, according to Jansen, are a) Ontological Grounding, b) Structure, c) Disjointness, d) Exhaustiveness, e) No ambiguity, f) Uniformity, g) Explicitness and precision, h) No meta-types (Jansen 2009).

I shall now discuss each of the demands for a good classification, while examining to what extent a fulfilment of these demands is feasible and advisable.

a) ***The Ontological Grounding***, according to Jansen, supports *classification of things based on traits that belong to those things*. According to this criterion the category 'others' (12) is an unsatisfactory classificatory category, because, according to the author, the trait *being other* does not belong to the thing but it is a meta-type. The same criticism was directed against (14) because it includes a characterisation of things as they appear to us. I agree with Jansen that, unlike other CAT categories, (12) and (14) are meta-types, which include classificatory criteria in the characterisation of the category. A good classification should make a distinction between types and meta-types, keeping consistency in the selected classificatory criteria (see Uniformity principle), while distinguishing between *why* something is classified and *what* is classified. In particular, I will argue that because a good classification of *what* is classified depends on *why* it is getting classified, the classificatory *what-why* dependence should be made explicit. I shall now clarify a few points related to *The Ontological Grounding* principle, whereby I stress consequences of neglecting the *what-why* dependence.

From the formulation of *The Ontological Grounding* principle, it is apparent that the demand for grounding classification on 'traits that belong to those things' assumes a certain ontological (and metaphysical) position. Thus, we can try to specify a meta-ontological position that explicates an ontological position from which this principle is derived. In Section 1.2.1 I presented a meta-ontological approach for distinguishing ontologies according to a particular view on what the discipline of ontology should accomplish. So, meta-ontology is

> O4) [...] saying what task it is that the discipline of ontology should aim to accomplish, if any, how the questions it aims to answer should be understood, and with what methodology they can be answered. (Hofweber 2011)

Even if *The Ontological Grounding* principle does not specify any meta-ontological position about what the aims and methods of an ontological analysis should be, by looking at the principle, we can distinguish a meta-ontological position that provides a framework for formulation of this particular (ontological) principle. Since the principle is presented in terms of a demand for classification of 'things' according to the 'traits that belong to those things' we can recognise an ontological statement, which accepts that the task of Ontology is a (metaphysical) analysis of the things in the world. A search for the traits that *belong* to things entails a position which accepts that the ontological analysis is

> (O2) the study of what there is in the world, what kinds of objects
>
> (O3) the study of the most general features of what there is, and how the things there are relate to each other in the metaphysically most general ways (e.g. what kind of features *belong* to the things)

When the aims of ontology are defined as (O2) and (O3), the classifications and distinctions of objects and their ontological features often assumes that there are objects with properties, which belong to them independently of the role of the observer. Such a position usually entails a version of ontological realism (Merrill 2010; Smith and Ceusters 2010), while the features of objects are often characterised as *intrinsic* or *essential* properties. Therefore, a good classification just needs to capture properly those features in classificatory categories. Consequently, (14), i.e. animals that from a long way off look like flies, not only has a meta-type status, but it also breaks the Ontological principle, while classifying the objects as they look to us and not as they are. In other words, (14) does not classify things according to the properties that *belong* to those things. Therefore, by looking at that how the term 'belong' is used in the formulation of *The Ontological Grounding* principle, it is likely that the principle is actually grounded on a version of the metaphysical position outlined above.

However, the existence of essential and intrinsic properties, which aim at capturing traits that *belong* to things, is highly disputable. It is difficult to define *what* the intrinsic properties of a species or of a person would be (Sterelny and Griffiths 1999; Dupré 2002). Moreover, even if we would accept the existence of some essential features, the description of those features would necessarily involve our epistemic apparatus (natural, technological or social) through which we describe the world. Consider a classification of a cell as belonging to the taxa *Eukaryote*. The cell will be characterised as a eukaryotic cell through the description of a sample observed through a microscope. Therefore, the cell is classified as the eukaryotic cell, *because* of the way it looks to us when we are capturing it by some experimental apparatus such as the microscope. Of course, this observation alone is not sufficient for a scientific classification, but it is accompanied and supported by the background knowledge and language accepted within a scientific community. As a result of the acceptance of scientific conventions, the process of classificatory reasoning is often just implicit in the produced classification, and if explicit, then it constitutes meta-data on *why* and *how* the things are classified as they are.

Nevertheless, scientific knowledge significantly relies on systematic observations that describe phenomena (Ankeny 2000; Dupré 2002). Having in mind that every description involves cognition of an agent who is capturing phenomena, Jansen's demand for a strict separation of ontology and epistemology (Jansen 2011) seems faint. Of course, mixing the levels of ontological and epistemological arguments in philosophy is a severe mistake, but the classificatory practice demonstrates an important interaction between the two levels, i.e. a *what*-level and a *why*-level. In classificatory practice and ontology engineering, the descriptions of the objects of interests necessarily include epistemic elements (Newell 1982). The classificatory reasoning is a *process* that results in a particular *classification*. A description of the classificatory reasoning provides information that will support an understanding (cognitive act) of the described objects as being of certain (*what*) kind. Thanks to the description, certain features will be selected and grouped according to the criteria that are recognised (cognitive act) as similar or diverse. Thus, a distinction of classificatory categories, as well as ‘objects’, ‘predicates’ and ‘relations’ in an ontology, always gets captured through an epistemic filter. It is hard to imagine a characterisation of an object from a neutral perspective, without involving a cognitive act of the observer, i.e. the agent who is describing something in a certain way. In other words, every description includes cognition, and therefore, it is a result of an epistemic activity. As far as description is an epistemic activity, the resulting classification and (ontological) characterisation of the described objects is that as well.

I do not intend to reject ontological analysis in general. I rather emphasise the importance of a reflection on the epistemic activities (*why* questions) involved in describing objects of interest, which can also provide an understanding on how objects are getting characterised as being of certain (*what*) kind. In other words, our understanding that not all features, which we use in our descriptions of a target, have the same level of generality, informs us about which features describe the target in a more general way than some others (Smith and Ceusters 2010). Likewise, the understanding that not all features describe classificatory targets in the same way helps us in deciding *what* kind of features to select in distinguishing different types of things. For example, *making a distinction* between classificatory categories of *material things* and certain *human artifacts* such as measurement methods, actually means that *we recognise* that some of the objects of knowledge are perceived and described in terms of physical (material) features of spatio-temporal objects that are not produced by human activity, unlike some others, which are produced by humans and cannot be described in terms of physical (material) features. In explaining *why* we distinguish these two categories as diverse types, we provide an answer that the latter category, unlike the former one, does not describe material objects, but it describes things such as measurements, standards, and values, which are intentionally produced by humans to capture some aspects of the world and to intervene on it. Whether or not we believe that our description captures some metaphysical features that *belong* to the described things, the description of types of things is always our description of how we perceive and distinguish those things. Since, the descriptions involved in *classifying* things result from an epistemic activity, the produced categories should also be considered as the results of that epistemic activity.

An alternative to *The Ontological Grounding* principle can emerge from the discussions that consider epistemic and pragmatic interests as those that drive a classification and distinction of 'objects' as being of certain kind (Dupré 1981, 2002; Boniolo 2012; Hacking 2002). In particular, an acknowledgement that the objects are classified into categories according to the epistemic and pragmatic aims of a community or a research group (Dupré 1981, 2002) can inform an ontological analysis (Sojic and Kutz 2012). Because the ways of describing and grouping the objects may vary, the categories and classifications are also numerous. For example, the criteria of a gourmand in classifying animals may result in a different taxonomy than the one used by evolutionary biologist (Dupré 1981, 2002). The classificatory criteria of a group of gourmand experts, will indeed describe and classify animals according to the taste and softness of their meat etc., i.e. according to that how the animals 'look to us' when we consume them. For sure, the feature that meat of an animal tastes good to humans does not belong to the animals. Yet, animals can be classified according to our 'taste'.

A similar approach to classification is accepted among ontology engineers, who emphasise the pragmatic aspects of the ontology design while representing types and instances (Lord and Stevens 2010). So, the gourmand's ontology is not so far from the ontologies that are actually used in agronomy and animal farms, where human values direct classification. The features that are captured in these ontologies are based on the value criteria of food industry. Moreover, representing, for example, the thickness of animal fat layer needs to include measurements and evaluation methods designed by human standards (Trißl and Reinsch 2011). Likewise, in classification of cancer, a standardised description of tumours includes description of methods (artifacts) by which a diagnostic class gets assigned (Sobin, Gospodarowicz, and

Wittekind) (see also 2.2.1 and 2.2.2). Such a description constitutes meta-data, which specify closer the represented types.

The last objection to *The Ontological Grounding* principle regards its limited applicability. If we agree that things should be classified according to traits that belong to those things, most likely we must exclude fictitious entities and objects of beliefs as classificatory targets. Arguably, the artifacts such as measurements might be questioned as well. Still, a classification of fictitious characters described in the mythology, fables, fairy tales, and other fictional stories is a reasonable classificatory task. Indeed, such a classification can also be useful to psychologists, anthropologists, or the film industry. However, it is not obvious what kind of traits *belong* to Little Prince or to a Unicorn. It might be more reasonable to claim that a description of Unicorn belongs to our beliefs, while our beliefs belong to us. Then, the Unicorn's traits do not belong to him, but they belong to us. Therefore, a classification of fictitious entities would need to classify objects of human beliefs. However, the problem is that our beliefs cannot be described as white and they do not have even a single horn. While, Unicorn can be described as a horse-like fictive animal with a horn, it is unclear to whom those traits indeed belong. Hence, following the demand for *traits that belong to things*, it seems that according to *The Ontological Grounding* principle there cannot be a good classification of fictitious entities. In scientific classifications the similar reasoning might question a classification of the measurement records, which hardly can belong to anything if taken outside of the measurement apparatus context and the human standards on how the measurements are used.

In the end, instead of completely rejecting *The Ontological Grounding* principle, I rather accept its important points that criticise mixing of types and meta-types, objects of classification and the classificatory criteria. However, another (meta) ontological position, which would be more pragmatic regarding the question on 'what there is' (e.g. (Quine 1948; Lord and Stevens 2010; Boniolo 2012)) and more tolerant towards a variety of classificatory targets, respecting the epistemic context in specifying 'objects', can complement *The Ontological Grounding* principle in order to support a broader spectrum of pragmatic interests involved in scientific classification.

b) ***Structure*** is another criterion that a good taxonomy needs to have. That is to say, a good classification, unlike CAT (see (5) and (6)), should be subdivided in types and sub-types , such as genera and species in biology (Jansen 2009). Indeed, the Structure principle is important for representing scientific knowledge about the relations among types (classes) as well as individuals (instances). While simple listing of categories cannot support almost any inference, a hierarchical structure provides information on how scientists understand a domain and how they reason about the domain. For instance, the Disease ontology (DO) is structured in a hierarchy where 'breast cancer' (DOID:1612)[79] is a subclass of 'cancer' (DOID:162), and 'female breast cancer'(DOID:0050671) is subtype of 'breast cancer'. This is a simple example, which illustrates that a female diagnosed with breast cancer will be an instance of a class 'female breast cancer', and accordingly, the instance will belong to all other super-classes of the 'female breast cancer' class, i.e. 'breast cancer', 'cancer' etc.. Note that hierarchical structure reflects scientific knowledge, which sometimes diverges in conceptualising a problem. Therefore, different disciplines might disagree on the choice of the classificatory structure. For example, whether mitochondria are classified as organisms or as cellular components depends on the background theory that is accepted, i.e. whether mitochondria are prokaryotes living within eukaryotic cells. Likewise, Ludger's suggestion that NCIT should subsume poikilotherms under the type of vertebrates, will not be the most useful hierarchical structure for a domain that focuses on the relation between the environment and physiology, which may prioritise body temperature as the main classificatory criterion (Block 1982). So, within such a domain it would be more useful to divide the organisms into supertypes, poikilotherms and homeotherms, while vertebrates would be just a subtype of these two supertypes. Therefore, interoperability between two or more domain classifications, and related ontologies, is not a simple task. Looking at the seemingly same terms across the classifications is not sufficient for the ontology alignment, but the context of the domain criteria, which influences hierarchical structure, has to be carefully analysed. The

[79] DOID stands for the Disease Ontology Identifier, which provides a unique label for each term in DO. The identifiers are then assigned to the classes, when an ontology is implemented.

normalisation methods are often recommended when two ontologies are getting aligned (see e.g. (Rector 2003)).

> [...] the resultant normalized ontology modules should as far as possible reflect the existing disciplinary division of labor in the relevant domain of science. Relevant *inferred* polyhierarchies can then be created according to need, for example, when providing support for information retrieval or for the representation of multidisciplinary scientific content or of the results of a particular set of experiments (Smith and Ceusters 2010).

c) ***Disjointness*** is a criterion, which also contributes structural organisation of a classification. Within a hierarchy of types and sub-types , anything that instantiates a subtype also instantiates the type of which it is a subtype (Jansen 2009). So, the types on the same level of classification should be disjoint. That is to say, no breast cancer can be an instance of both male and female breast cancer. However, both male and female breast cancer are a type of cancer. As Jansen rightly notices, CAT's (1), i.e. the animals that belong to the Emperor, most likely includes trained animals (3). However, subsuming (3) under (1), as Jansen suggests, can induce problems in reasoning about the instances of the classes, because in this case not all trained animals (3) belong to the Emperor (1). For the same reason biomedical taxonomies often keep just a form of listing, without facing the problems in reasoning about the instances that partition in multiple classes.

d) **Exhaustiveness** is a demand that a classification (ideally) needs to subsume all the entities that the classification aims to capture. Obviously, the classificatory exhaustiveness is uneasy to achieve in any empirical domain. For example, it is uneasy to capture extension of the classes, such as all persons on Earth who have breast cancer. However, by limiting the scope of classification to the patients in a hospital or a geographical region, the approximation of an exhaustive class might be achieved. An intensional exhaustiveness of classification (i.e. inclusion of all types, specified according to their meaning), is also just a desirable ideal, because most classifications at some point need to face changes in scientific knowledge that force an inclusion of new categories. For example, new data can provide evidence for the classificatory limits and shortcomings of previously established classifications (Arnone et al. 2009).

e) **No ambiguity** is the demand to use classificatory terms as precisely as possible in order to avoid multiple meanings. The ambiguous use of the term ‘animal’ in CAT mixes biological taxa with painted animals, dead animals, and fictive creatures. Nonetheless, if we refine *The Ontological* principle as I proposed, then a re-structuring of CAT, with a specification of the types and sub-types of classified ‘animals’ that we think of as ‘natural’, ‘fictive’, and ‘painted’, might have improved the classificatory status of CAT, making it less ‘comical’. Thus, ambiguity in a classification can be reduced by specifying the context in which a term is used. Similarly, Lundgers’ objection to NCIT that ‘the artificial type laboratory animal stands out inappropriately when listed alongside the three natural classes [i.e. vertebrates, invertebrates, and poikilotherms], since laboratory animals do not comprise a natural kind’ (Jansen 2009, p.163), can be reassessed by a specification of the criteria that drive this particular classificatory division. Laboratory animals are indeed a special kind of animals (Ankeny and Leonelli 2011). Still, that does not mean animals, which are used as model organisms, should be excluded from NCIT as they play an important role in cancer research. Rather, a specification of the term ‘animal’ and the related sub-types of animals will reduce ambiguity, making distinction between ‘laboratory’ animals and ‘natural kinds’, while keeping usefulness of both classificatory branches of the NCIT.

f) **The uniformity** principle asks for a well-defined classificatory domain. A classificatory criterion should persist in the selection of what is getting classified as types and sub-types throughout the classification. Once again, CAT exemplifies an inconsistent use of the classificatory criteria.

> CAT, however, draws on the distinguishing traits of several different kinds at once. Heading (1) sorts animals according to their owners, (4) according to species membership, among other things, (7) according to species membership plus the lack of an owner, (9) according to behavior, (13) according to the effects of behavior, and (14) according to an animal’s appearance to a remote observer (Jansen 2009).

Keeping uniformity of the classificatory criteria is important for a well structured classification. However, avoiding multiple classificatory criteria within a classification would also reduce the utility of a classification, which aims to capture various aspects of a domain. Thus, various criteria are often combined as, for instance, in the branching of the Gene Ontology[80] into the classification of cell components, biological processes and molecular functions, where three different criteria are used to classify genes and gene products. In the Disease Ontology[81], the types of cancer are classified according to the organ of the cancer origin, the cancer cell type, as well as additional specifications of tumour phenotypes such as hormonal status. Thus, within the classifications of complex domains, the Uniformity principle cannot be consistently applied to the whole classification. Still, the classificatory criterion should be uniformly applied at least to the particular classificatory branches.

g) **Explicitness and precision** poses a demand to classification to be as explicit and precise as possible. So, CAT's (12), i.e. 'others', exemplifies a category that is neither explicit nor precise. Indeed, the 'others' class, consisting of genes or types of disease, signs, and symptoms that have not yet been assigned to any other class, appears also in biomedical classifications and ontologies (Jansen 2009). However, it would not be appropriate to consider members of such a class as classified objects. Such a class rather explicates the weaknesses of a classificatory system that considers as desired some objects which yet cannot fit in.

h) **No meta-types** principle, according to Jansen, says that meta-types that come about through the classification process itself should be excluded from the classification. Obviously, CAT's (8) is a peculiar meta-type (those animals included in the present classification), which includes all other types that are already represented on the same level as this meta-type. Such a category is redundant and confusing. However, I would not agree that any other meta-type is redundant in the same way. As I have previously argued, the classificatory process and classification as the classificatory product are mutually dependent. Thus, specification of certain

[80] http://www.geneontology.org/

[81] http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page

meta-types, e.g. reasoning procedures that have influenced the selection of the classificatory categories (e.g. laboratory animals), may contribute explicitness of the classification and the classificatory concepts. Of course, if the specification is represented as meta-types, those categories should not stay on the same representational level as the types, as stated in (a), (b), and (f). Alternatively, a specification of a classificatory procedure can be provided as meta-data.

In this section I have presented some of demands that a well structured classification should have, i.e. a) Ontological Grounding, b) Structure, c) Disjointness, d) Exhaustiveness, e) No ambiguity, f) Uniformity, g) Explicitness and precision, h) No meta-types. In addition, I discussed the limits and the problems that arise in designing a classificatory system, showing that the presented demands cannot fully capture a variety of practical needs in a biomedical domain. In particular, I proposed an alternative interpretation of the Ontological Grounding principle, emphasising the need for (meta-data) specification that would explicate how the classificatory needs of a domain influence *what* is going to be classified and *how* it will be done.

The following sections focus on the existing classificatory systems for breast cancer, where I analyse similarities, differences, and interdependences of particular sub-domain classifications.

3.5. Clinical classification

Clinical classification of breast cancer employs, explicitly or implicitly, various classificatory systems. Section 3.5., throughout its subsections, addresses an interaction between the classificatory aims and knowledge captured within the existing classifications. I will show how clinical classifications depict clinical reasoning through the specific uses of classificatory terms and relevant kinds of generalisation. The particular choices of labels and classes that are utilised in the clinical domain, but not in some other domains, should also reveal certain normative and interest driven aspects of clinical knowledge.

3.5.1. TNM classificatory system

TNM system is an internationally accepted guide for clinical classification of tumours. This classificatory system is designed to help clinicians by providing them with standards for the assessment of the extent to which cancer has spread in a patient. Moreover, a careful clinical description and classification of tumours serves a number of objectives, such as:

> (1) aiding the clinician in the planning of treatment; (2) giving some indication of prognosis; (3) assisting in the evaluation of the results of treatment; (4) facilitating the exchange of information between treatment centers (5) assisting in the continuing investigation of human cancer (Copeland 1961).

The TNM provides such a standardised reference that supports a description of a cancer stage and its progression throughout three main *anatomical* categories: Tumour size (T), Nodal status (N), and Metastasis (M). So, in the case of a particular patient, clinical findings represented as data records will be matched with a particular TNM class. For example, clinical findings acquired by extraction and measurement of a tumour size will be assigned to one of the established (T) classes.

Historical origins of the TNM system lead back to 1944, when Pierre Denoix (Denoix 1944) published his criteria for staging of colorectal tumours. The importance of systematic collection of data, which can also be statistically analysed, was recognised at the very beginning of a standardised tumour classification. In 1950 the Committee on Tumour Nomenclature and Statistics was appointed (Copeland 1961). Soon after, in 1953, The Committee held a joint meeting with the International Commission on Stage-Grouping in Cancer and Presentation of the Results of Treatment of Cancer (ibid.). The meeting resulted in *an agreement on a general criterion for the TNM classification as based on the characterisation of anatomical extent of tumours*. In 1958 the Committee published the first recommendations for the clinical stage classification of breast and larynx cancers (UICC 1958). Ten years later, the first edition of internationally established TNM classificatory standards included various cancer types (UICC 1968). Every of the following editions used to include certain updates as knowledge in the field was changing. For instance, the sixth version of the TNM classification included a number of prognostic factors that are organ specific (Sobin, Gospodarowicz, and Wittekind 2009).

Basically, the sixth version of TNM system consists of 24 categories that are divided in four main degrees of tumour size (T1-T4), three main degrees of nodal status (N0-N3), and two degrees of metastatic status (M0-M1) (ibid.). For the purposes of easier manipulation and analysis, these groupings are eventually reduced to five stages (S 0 – S IV), from zero to four. Stage zero (S0) stands for carcinoma in situ, while the fourth stage (S IV) is assigned to the most advanced cases with distant metastasis. The second and the third stage correspond to the tumors' progression respectively. *The criteria for the stage* grouping consider each group as '*more or less homogeneous in respect of survival*' (Sobin, Gospodarowicz, and Wittekind 2009). So, in the case of breast cancer the staging from 0 to IV would correspond to 5-year Relative Survival Rate (RSR) respectively: **0** 100%, **I** 100%, **II** 86%, **III** 57%, **IV** 20% (Ries 2007).

The following example, labelled as *TNM (P_0)*, illustrates the reasoning employed in the breast cancer staging, according to TNM system:

TNM (P_0)

a) a patient P_0 has breast cancer with the features (X_0)
   - *Tumour size ≤2 mm*: thus it belongs to the category T1
   - *P_0 has one lymph node involved*: thus N1
   - *There are no metastasis in the body of P_0*: thus M0

b) a set (T1, N1,M0) is classified as stage two (S II) (TNM breast cancer staging system)

c) *P_0 has the stage two of breast cancer* (inferred from (a) and (b))

d) There is 86% probability that P_0 will survive next five years (inferred from the breast cancer statistics for (S II))

Generally speaking, a higher RSR fits a lower disease (tumour) stage, while the lowest RSR is associated with the highest disease (tumour) stage. Still, I shall point out certain aspects that often stay implicit in the tumour classification. The survival rate as well as tumour classification depend on various contingent factors such as available therapy, diagnostic technology and patients' behaviour following diagnosis. A claim that these dependences dramatically change the view on the utmost cases such as the survival (outcome) of the patients with an advanced tumour stage would be an overstatement. However, if we consider how a classification depends on other related factors, we can also understand better the limits and potentials of a classification. For example, in a number of cases that are classified as lying between stages I and IV, those 'contingent' factors might actually influence survival rate and so the accuracy of the assessed classification. For example, it has been shown that physical activity after diagnosis of breast cancer significantly influences survival (Holmes et al. 2005; Holick et al. 2008). Thus, women who were physically more active after diagnosis had shown a lower rate of death from breast cancer than women who were less physically engaged (Holick et al. 2008).

Obviously, information such as *postdiagnostic activity* of patients is not considered when the diagnosis is assigned and the cancer staging assessed. Rather, the staging is based on an abstraction from such information and an 'averaging' of the tumour captured as (X_0) at the moment in time t_0 of the patient's (P_0) diagnosis and prognosis assessment. The particular tumour features described in (a) are assigned to the general classificatory categories (T1, N1,M0). Further, the prognosis in (d) is derived from (c), whereby the tumour description (a) is assigned to the average outcomes (e.g. survival, relapse etc.) related to the similar types of tumours ($X_{1,n}$) that resemble the captured one, primarily by means of similar outcome (b). That is to say, the characterisation a particular tumour (X_0) might be considered as an idealisation (X_{type}) that fits to the stage two (S II).

Therefore, by looking at the TNM classificatory reasoning, we can distinguish that the staging of a particular patient, indexed as P_0 depends on knowledge about

1) *which instances of the tumours,* described as ($X_{1,n}$), can be considered as *the resemblances* with (X_0) so that a kind of identity between P_0's tumour and the resembling tumours, captured as ($X_{1,n}$), can be asserted; The assertion of such a resembling identity (X_0) ≈ ∃($X_{1,n}$) is possible as it is recognised that (X_0) and some of ($X_{1,n}$) are all of the specific type (X_{type});

2) *which features of the tumour are most relevant* for capturing and describing tumour as ($X_{0,n}$); Concurrently, the choice of the relevant features grounds on knowledge of *criteria for establishing the resemblance relation* that will enable the assignment of a particular description ($X_{0,n}$) to a type (X_{type}).

3) *the outcomes of the patients*; i.e. in the case of tumour staging (Sobin, Gospodarowicz, and Wittekind 2009), the main criterion for resemblance among the tumours, captured in TNM classes, is based on knowledge about *the outcomes of the patients* with the tumours that resemble the P_0's tumour, which is characterised as (X_0) and classified as the stage two (S II).

Thus, *TNM (P_0)*, which classifies P_0's tumour, applies knowledge outlined in (2), while selecting the relevant classificatory features such as the tumour size lower then 2mm. That feature will limit the scope of potentially resembling tumours, ($X_{1,n}$) in (1), to those that are less then 2mm in diameter. Another feature, the lymph node involvement will shrink further the scope of the resembling class ($X_{1,n}$). The information on metastatic involvement will eventually define the resembling reference class as the one that fits the tumour of P_0. Accordingly, the tumour will be classified as the stage two (S II), giving P_0 a good prognosis with a high rate of survival in the next five years (86%), because the people with the resembling tumour features usually survive at that rate (3).

However, if we look again at the reasoning employed in the staging of P_0's tumour we can notice a circularity in knowledge as the steps involved in the classification are interdependent. Knowledge of (1) depends on knowledge of (2), while knowledge of (2) depends on knowledge of

(1). Moreover, (3) will depend on (1) and (2), while knowledge of (3) will influence (1) and (2). That is to say, (1) and (2) are mutually dependent because the choice of the relevant classificatory features (2) and the selection of instances that are described as having those features (1) depend on each other. Only by recognising that some instances have similar features (1), the features will be selected as the criteria of similarity (2) for the eventual grouping of instances (1). Likewise, depending on the kind of feature that is selected in (1) and (2), the average measure such as survival rate will change, because diverse features will group the patients in various ways. On the other hand, the information on patients' outcome (3) will help in deciding what kind of features (1) and (2) should be grouped together, whereby (3) becomes the classificatory criteria (2).

This apparent circularity that arises from the interdependences among the various levels of classificatory knowledge can be seen as a conflict between a *static* capturing of tumour type in a *class* and a *dynamic* character of knowledge that constitutes *classification as a process*. At first, I will address the static capturing in the clinical classifications and some advantages that standards for scientific labelling of classes can have. Subsequently, in sections that follow I discuss the dynamics in knowledge that induces modifications in the classificatory labelling.

The description of tumour features that fit into the TNM classificatory system facilitates clinical decisions, because standardised clinical data (e.g. about therapeutic efficacy) can be compared across the health care centres on an international scale, eventually resulting in specific recommendations for the clinical assessment of the patients (Singletary and Connolly 2006). Clinical records, as an organised data collection, facilitate improvements in clinical knowledge. By using standards for tumour characterisation, clinicians and researchers have a common reference in recording and interpreting the data. In other words, the standardised records of tumour staging and TNM categories support knowledge sharing and data analysis. The resulting representation of disease that follows commonly accepted standards, such as TNM, not only facilitates clinical decisions, but it also drives towards a better understanding of disease.

> The TNM system for cancer staging is not perfect, but it represents our current best effort to provide a method that is clinically useful and reflective of the available data. Refinements and amendments of the TNM system have been aimed at improving its ability to estimate prognosis. An important aspect of the new staging system is the definition of a nomenclature and coding system that will standardize the collection of important data that may affect treatment in the future (Singletary and Connolly 2006).

Thus, even if it is almost certain that every classification will have its shortcomings (see 3.4.), only by having some accepted reference class and a standardised nomenclature, the information about tumours can be captured and compared, while the resulting classificatory knowledge can be further tested, accepted and also corrected. For instance, if we look at the case of P_0's tumour, the classification could have been done in another way, using age and ethnicity as the main classificatory criteria. So, P_0 would have been assigned to another tumour type (Y_{type}), e.g. breast cancer in elderly people. Indeed, such a classification is relevant for some clinical questions, because age and ethnicity do influence clinical outcome and therapy response (Shavers, Harlan, and Stevens 2003; Anderson, Jatoi, and Sherman 2009; Jemal et al. 2009). However, unlike some other classifications, the aim of TNM is to capture tumour size, lymph node status, and metastasis as the difference makers, which are highly informative in distinguishing between the bulk of patients who will get recovered easier from the disease and those who are not having a good prognosis (Sobin, Gospodarowicz, and Wittekind 2009). Every other kind of information about patients and the tumour features is considered as an additive value that complements clinical knowledge captured in TNM classification.

3.5.2. Clinical trials shifting the classes

Historically speaking, tumour size, lymph node status, and metastasis were not immediately chosen as classificatory criteria and labelled as TNM categories. Acknowledgment of similarity, which makes a not-yet-distinguished group to be a distinguished class, results from discrimination of some instances as more similar to some and less similar to others, with respect to the selected feature. Accordingly, the recognition of similarity follows the acknowledgement of heterogeneity among the instances and the variety of criteria that make them diverse. Indeed, heterogeneity among tumour specimens was observed and acknowledged in clinical practice at the very beginnings of cancer research (Fisher, Redmond, and Fisher 2008). However, at first, diverse outcomes had been attributed to the diverse responses to surgical treatment. Thus, a particular response to the surgical treatment was considered as the main reason for the diversity of clinical outcomes, while other features of patients and their tumours were not given distinct significance. At least, the information about those other features that might have influenced the clinical outcomes were not systematically distinguished, collected and analysed. However, an accumulation of ‘unexpected observations’ gradually gained its significance among the clinicians, even if only in an anecdotal form (Fisher, Redmond, and Fisher 2008). The repeatable patterns of unexplained associations indicated that older patients respond better to therapy than younger patients, larger tumours were associated with a worse outcome than the smaller ones, and involvement of auxiliary nodes indicated a worse prognosis (ibid.).

Only after a series of controlled clinical studies, which started in the 1960’s, breast cancer heterogeneity also acquired important scientific acknowledgements. Some of the first results of clinical trials reported that involvement of auxiliary nodes is associated with clinical outcome (Fisher 1970). So, involvement of auxiliary nods was one of the first features labelled as a difference maker in breast cancer prognosis, based on the large scale studies that provided statistical evidence. Accordingly, nodal status acquired the status of a standardised parameter in clinical TNM classification of tumours.

It is not by chance that the history of TNM classification overlaps with, and slightly precedes the history of the clinical trials. In the light of the presented historical example, we can see that the tumour size (T), nodal status (N), and metastasis (M) came into clinical use as the labels to capture tumour features before their significance was statistically associated with the outcomes (Denoix 1944). The initial introduction of the TNM classificatory labelling was just a starting point on the way to the expansion of clinical knowledge acquired through the clinical trials (Fisher, Redmond, and Fisher 2008). Consider, once again, the three aforementioned aspects of knowledge involved in the classification of P_0's tumour (Section 3.5.1) as (1) the tumours grouping, (2) the classificatory features fixing, and (3) the references to the outcome. In the clinical observations, before the TNM was established, the relevant TNM features had been selected (2), thus enabling and fixing the classificatory focus that defines group (1). The features are *selected as relevant* even if their relevance has been at first just locally confirmed by a clinician or a restricted community of researchers (Denoix 1944). Concurrently, *the outcome* (3) was recognised in the clinical community as a crucial criterion for deciding which tumour features should be tracked down (Fisher, Redmond, and Fisher 2008). However, before (3) was officially acknowledged by the clinical community (which conducted the trials) as criterion (2) for classification of the instances (1) associated with certain outcome, (3) was just an implicit category utilised by clinicians in the reasoning about the instances and the tumour features. But, in the long run, the controlled clinical studies gained the advantage of the explicit distinction of a number of classificatory criteria, which previously used to have just an implicit (anecdotic) form. Thereby, the systematic recordings of the selected criteria, which were tracing the fixed categories under the study, brought in clinical knowledge a corrective value of collected information.

Knowledge acquired from clinical studies has shifted understanding, treatment and classification of cancer by extending the time period of observations in clinical trials. The time period of observation has shown to be crucial for the interpretation of categories. Namely, it has been shown that long term studies can lead to new conclusions, which might be even opposite to those that are inferred from a short term follow up. For example, randomised clinical trial (RCT) that followed-up 701 women with breast cancer relatively small in diameter (less than 2 cm) after a period of eight years (1973-1980) showed statistically significant difference ($P<0.001$) in local recurrence of tumour between two groups of patients (Veronesi et al. 2002). In the group of patients who underwent radical-mastectomy local recurrence of tumour was much lower than in the patients who had a breast-conserving treatment. However, the long term survival rate demonstrated no difference among the two groups. The rate of death from all causes after 20 years of follow up was identical in the two groups (P=1.0), while the rate of death from breast cancer was almost identical among the groups (P=0.8). That led to a conclusion which changed both the understanding of the disease as well as the therapeutic approach to breast cancer treatment. Namely, the long term studies showed that radical-mastectomy, an established practice that lasted eighty years, does not reduce the risk of tumour recurrence in the long run. In that way a previous belief that a radical surgical intervention is the best treatment for breast cancer was ruled out. The conclusion was applicable to the group of patients with relatively small breast cancer as RCT design covered that particular domain of patients. Thus, the category of tumours smaller than 2cm was re-classified in the prognostic and therapeutic context.

3.5.3. TNM as a dual system

The TNM classification comprises two sub-classificatory systems, clinical and pathological. Therefore, TNM is characterised as a dual system (Sobin, Gospodarowicz, and Wittekind 2009). Clinical TNM classification (cTNM) consists of preoperative information retrieved from physical exam, and imaging data. Pathological TNM (pTNM) is based on evidence retrieved before the treatment that is complemented with information acquired through the methods specific for the pathological classification. The pTNM classification also includes histopathological information acquired after a surgical intervention, which enables pathological assessment of the grade of

- the primary tumour (pT), retrieved through the histopathological analysis of the removed tumour (or biopsy extract);
- the lymph nodes (pN), retrieved through the histopathological analysis, and
- distant metastasis pM, retrieved through the microscopic examination.

Making a distinction between cTNM and pTNM classification is important as these two systems represent information retrieved through different methods and serve different purposes. The main purpose of cTNM is to support choice and evaluation of therapy, while pTNM serves in guiding adjuvant therapy and providing additional data for prognosis assessment (Sobin, Gospodarowicz, and Wittekind 2009). Thus, information about pTNM can also support modifications in the treatment recommendations, such as recommendation of adjuvant therapy. After assignment of cTNM and pTNM, tumour stage can be assessed according to the rules of the site specific staging. The assessment of stage is the final step. It defines the therapy and remains unchanged in the medical records.

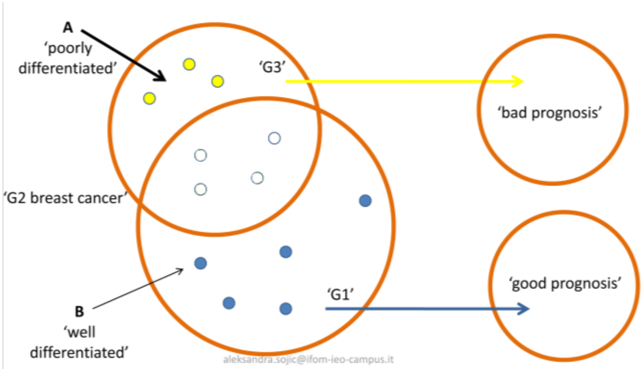
Apparently, numerous advancements of biomedicine, especially molecular aspects of disease, are not incorporated into the TNM system. The most significant input that plays an important role in the cTNM system comes from the advancements of diagnostic technology such as sophisticated imaging tools (Pinder, Harris, and Elston 2008). On the other hand, histopathological data that are used in pTNM modify the therapy choice by providing information on pathological grading (ibid.) (see also Section 3.6.). Yet, pathological grading of tumours usually stays on a morphological observation. For instance, IHC staining of tumours helps in distinguishing 'poorly differentiated' and 'well differentiated' tumour structures (Section 3.5.4.). Even if such an observation looks at a fine-grained morphology of the cancerous tissue and presence of some pathological elements (e.g. HER2 over-presence), it is still an anatomical characterisation of tumour, because it looks for the presence or absence of some parts, their size and structure. Thus, clinical and pathological classifications of tumours might be rather considered as different just in degree and the level of observation, because both sub-systems, cTNM and pTNM, rely on the same basic principle that looks at the anatomical features. Eventually, the methods employed in the tumours comparison and classification, within TNM system, remain in line with the tradition that captures the anatomical rather than the functional characterisation of tumours. Certain exceptions can be seen in parameters relevant for pTNM, which measure tumour markers such as Ki-67, while capturing proliferation rate and tumour behaviour.

One of the main challenges to TNM classification nowadays is the integration of recently acquired knowledge about biological and genetic prognostic factors into the previously established classificatory categories (Zurrida and Veronesi 2011; Arnone et al. 2009; Veronesi et al. 2009). An approach in facing this problem have resulted in adopting anatomical stage grouping and prognostic grouping as two separate systems. The separation in groupings allows an easier extension of prognostic factors within the prognostic grouping (Gospodarowicz 2001), while the anatomic classification remains unaltered. At the moment, *histologic grade* has been incorporated into the soft tissue sarcoma, bone, and prostate tumours classification; a*ge* and *histology* are included into the thyroid tumours classification; while *serum markers* are incorporated into the classification of testis and gestational trophoblastic tumours (Sobin, Gospodarowicz, and Wittekind 2009).

3.5.4. Ambiguity in classification of the grade two tumours

The traditional methods for classifying tumours based on the anatomical features (TNM) are still predominantly used and they are accurate in a significant number of cases (Elston and Ellis 1991, 2002). However, the TNM classification invokes a 'grey area' (Dalton et al. 2000) while dealing with the tumours that might be classified in both ways, as 'poorly differentiated' and 'well differentiated'. Those tumours are classified as histologic grade two (pT2) and, according to some authors, constitute 30-60% of the breast cancer cases (Sotiriou et al. 2006).

Figure 23 represents the sets of biomedical claims about the breast cancer classification and prognosis, as based on the histopathological analysis (pT). The blue, yellow and white spots represent the sets of the claims, where each of the sets is associated with only one tumour specimen. The intersection area represents a significant number of cases (white spots that are labelled as G2) in which the claims are conflicting[82].



Contradictory biomedical claims in the breast cancer classification

Figure 23 'Grey area' in reasoning about grade two tumours

So, in the case of a clinician A that is looking at the histopathological data, within his background knowledge the G2 members, i.e. pathological grade 2 tumours (pT2) will be associated with the bad prognosis. For another clinician, the G2 members will be mapped onto 'the good prognosis' set. So, the members of the intersection area will have assigned the status associated with both G3 and G1 a good prognosis and a bad prognosis. Concordance between two pathologists in reasoning about the grade two tumours has been investigated and found to range from 50% to 85% (Sotiriou et al. 2006).

Even if we accept that a rational subject can have conflicting beliefs at the same time, as an agent in the decision making he has to weigh and decide between the conflicting arguments. The TNM classification does not provide a clear clue in some cases such as the prognosis of the grade two tumours.

[82] The information about G2 reasoning is gathered from the publications and personal contact with the oncologists at the European Institute of Oncology in Milan. See also (Keating and Cambrosio 2003).

This grade is associated with an intermediate risk of recurrence and is thus not informative for clinical decision making. (Sotiriou et al. 2006)

While illustrating uncertainty in clinical observation, a professor of clinical histopathology Peter P. Anthony applies the duck-rabbit dilemma (Figure 24): 'Is it a duck? Is it a rabbit? If they can tell me that it swims on water then I can say confidently that it is a duck. On the other hand, if it runs on grass then surely it is a rabbit.' (Anthony 1998). The reasoning about cellularity of a fine-needle aspirate from a mobile breast lump, according the author, is an analogous example of the context dependent observation, which illustrates the insufficiency of certain types of information. In a young girl that means something quite different than when the aspirate was obtained from a mammographic abnormality in a postmenopausal woman (Anthony 1998). Thus Anthony concludes with a cautious advice '[...] In the absence of necessary clinical information, obtain it yourself and never act alone when an uncertainty exists' (Ibid.). Although, in some cases important information is not available at the moment and a demand for action is urgent, it is important to develop an explicit representation of the contextual dependences in the clinical and biomedical observations.

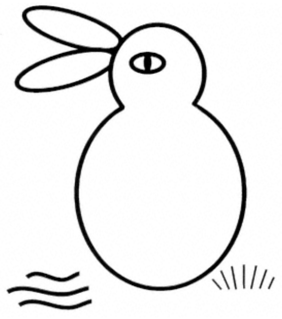


Figure 24 - Duck-rabbit

The breast cancer classification, especially the grade two tumours class, shows to which extent the classificatory categories are flexible and context dependent. In one case the category 'G2 breast cancer' will mean one thing that is 'good prognosis', while in another case 'G2 breast cancer' will be associated with the opposite meaning, i.e. 'bad prognosis'. And, new additional information will reshape again the meaning of this concept.

However, the additional information that can help a non-contradictory prognosis often goes beyond the TNM classification. It includes information on biomarkers[83], such as the degree of a particular hormone presence. Consider 'G2 breast cancer', in the case of an estrogen receptor

[83] Note that histopathological grading sometimes includes tumour markers, for instance Ki-67.

negative patient (ER-), most likely ‘G2 breast cancer’ will mean ‘bad prognosis’, while in the case of estrogen receptor positive (ER+) patient, ‘G2 breast cancer’ could also mean a ‘good prognosis’ (Gasparini and Hayes 2006).

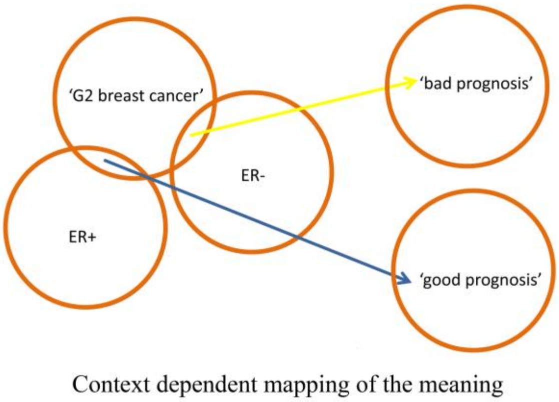


Figure 25 Additional information

However, ‘most likely’ and this level of information will not be sufficient in the case of the clinical decision, which is relevant for a real person. A complex and specific informational context has to be considered in order to support an accurate diagnosis, prognosis and the therapy recommendation (Piccart 2006). As the TNM system considers just a limited number of classificatory parameters, it does not seem to be a sufficient tool for such a complex task (Sotiriou et al. 2006; Veronesi et al. 2009). It rather offers a basic guideline for the clinical decisions and an important framework for the data collection and future comparison.

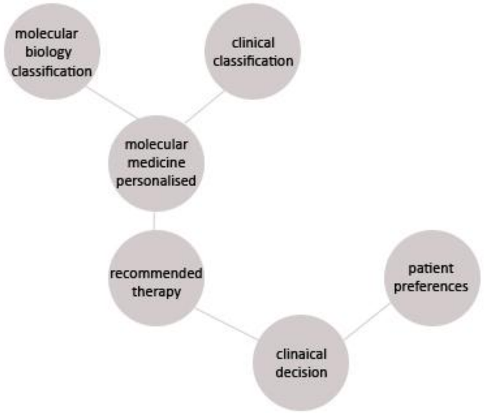


Figure 26 Including preferences

Through the G2 example, the Figures 24 and 25 illustrate how the term meaning and related mapping are context sensitive. Although estrogen receptor as a classifier is not sufficient in covering a wide range of contexts, it represents explicit contextual reasoning. The meaning of a complex term such as ‘breast cancer’ gets a specific meaning through the process of shrinking down the set of potential meanings to the specific one. The process of the meaning specification depends on the pragmatic context, which is in our case a pathological setting of prognosis assessment. Additional restriction comes from the empirical knowledge and statistical analysis that associate ER- tumours with the bad prognosis and ER+ tumours with the good prognosis. Further restrictions of meaning may depend on the information about gene expression, or (in the case of the decision for a therapy) the patients' preference. So, a good prognosis does not need to mean the same thing for every patient. A therapy recommendation proposed by the clinician assigns to the patient a 'good prognosis' under the condition that the recommended therapy is received. The meaning of 'good prognosis' is obviously the clinician cantered interpretation where prolonged survival is considered as good and desirable irrespective of collateral problems that can be critical for the patient's decision (Stockler 2006). Even if the role of values and preferences in classification makes for a topic on its own (Kincaid, Dupré, and Wylie 2007), thanks to the ontological modelling applied to the reasoning with heterogeneous kinds of data, the clinical and molecular information might be, in principle, explicitly integrated with the patients' preferences. Figure 26 outlines a general idea of inclusion of preferences in clinical decision making, where informed patients take an active role.

The following sections examine how biological and molecular parameters support a more specific characterisation of breast cancer in order to avoid ambiguities as those that are seen in the grade two tumour' morphological descriptions.

3.6. Biological classification, histological and molecular sub-types

> Breast cancer is a clinically heterogeneous disease, and existing histological classifications do not fully capture the varied clinical course of this disease. (Pusztai et al. 2006)

Histopathological characterisation of tumours and TNM classification cover just partly the heterogeneity in tumours that manifest different behaviour and response to therapy. In the search for parameters that can support a more accurate prediction, prognosis, and diagnosis, the molecular classifications of breast cancer have been developed (Pusztai et al. 2006; Perou et al. 2000). However, incorporating these new molecular classifiers into clinical practice and the existing clinical classifications is not a simple task. I will address some of the issues in the integration of biological, molecular and clinical classifications. In particular, I identify certain difficulties that originate in the use of domain dependent methods and a classificatory focus, which is always domain specific. At the same time, I stress the advantage of keeping various classifications within separate systems.

We have seen (Sections 3.5.2-3.5.4) that the clinical classification (cTNM) considers tumour size, tumour location, nodal and metastatic status as the most relevant data for the tumour classification, diagnosis, prognosis, and a therapy choice (Sobin, Gospodarowicz, and Wittekind 2009). *The biological classification* of tumours, on the other hand, looks at the deregulated biological processes such as the mitotic activity, nuclear pleomorphism or tubule formation. Such a description of tumour belongs to the pathological assessment of tumour, and effects pathological grading (pT). A pathological assessment of tumours largely relies on the observation of the morphological structures. While trying to capture tumour grade, pathologists evaluate abnormal phenotypes of the tissue samples by assessing degree to which they diverge from what is considered as a normal phenotype (Walker and Thompson 2008). So, unlike normal cells, cancer cells are described as having small cytoplasm, multiple nuclei, multiple and large

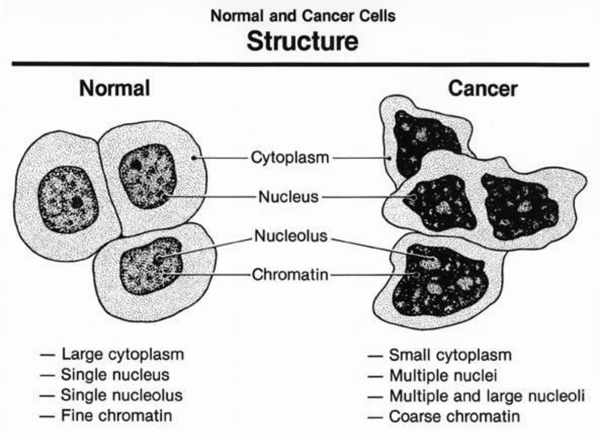


Figure 27 Cancer vs. normal cells - Illustration by National Cancer Institute (Pat Kenny)

nucleoli, and coarse chromatin (Figure 27). Three microscopic features 1) degree of tumour tubule formation (percentage of cancer composed of tubular structures), 2) tumour mitotic activity (rate of cell division), and 3) nuclear grade (cell size and uniformity) are scored on a scale of 1 – 3. This pathological grading system is sometimes labelled as the Bloom and Richardson system, but its significant improvements are contributed to the Nottingham Group (Galea et al. 1992; Walker 2003). The use of this modified *Nottingham grading system* is recommended by the UK and the European Breast Screening Pathology Groups, the US Directors of Anatomic and Surgical Pathology, as well as the UICC and WHO (Walker and Thompson 2008).

Thus, pathological description on the tissue and cell levels belongs to a *biological classification*, because it associates abnormal morphology captured in a microscopy image, with the underlying biological processes that are typical for cancer behaviour. In other words, biological classification of tumours provides information that is supported with a biological explanation (Walker and Thompson 2008). Besides having a predictive power, the detection of nuclear pleomorphism, tubule formation, and an increased mitotic activity can also explain particular tumour behaviour, while distinguishing more and less aggressive tumours. So, the three histological markers provide information on carcinogenic processes that result in a cancer phenotype, thus allowing an estimation of tumour behaviour. On the other hand, clinical classifiers cT and cN do not provide an explicit explanation of why larger tumours and tumours with a higher cN grade have worse prognosis than the smaller tumours and those associated with enlarged lymph nodes. Though statistical data confirm that larger tumours are frequently associated with the lymph nodes involvement and a worse outcome, the cases of very aggressive tumours that are small in size undermine the tumour size as a viable parameter in the explanation for aggressive tumour behaviour.

> Smaller tumors are less frequently node positive than larger lesions, but there is a risk of metastatic lymph node disease even for lesions <10 mm in size. [...] Similarly, the clinical determination of lymph node size does not necessarily reflect the underlying pathology: nodes reactive to, for example,

> previous breast biopsy, may be enlarged without evidence of metastatic disease, whilst nodes containing metastatic carcinoma may be impalpable. Thus, the clinical TNM system is unreliable (Pinder, Harris, and Elston 2008).

Concurrently with questioning the reliability of cTNM, those small and aggressive tumours also lead to dismissing cT and cN as parameters that can provide a comprehensive explanation for the aggressive behaviour of tumours. That is to say, tumour size cannot explain tumour aggressiveness, because there are cases of large and less aggressive and small but very aggressive tumours (Walker and Thompson 2008). Thus, the explanatory power of tumour size in explaining tumour behaviour and predicting the outcome is limited. On the other hand, biological classifiers, which are used in pathological assessment, distinguish specific tumour types as based on their biological features independently on the tumour size, thus covering the 'exceptions', which cTNM is not able to capture. Moreover, cancer screening campaigns additionally influence significance of tumour size as a classifier.

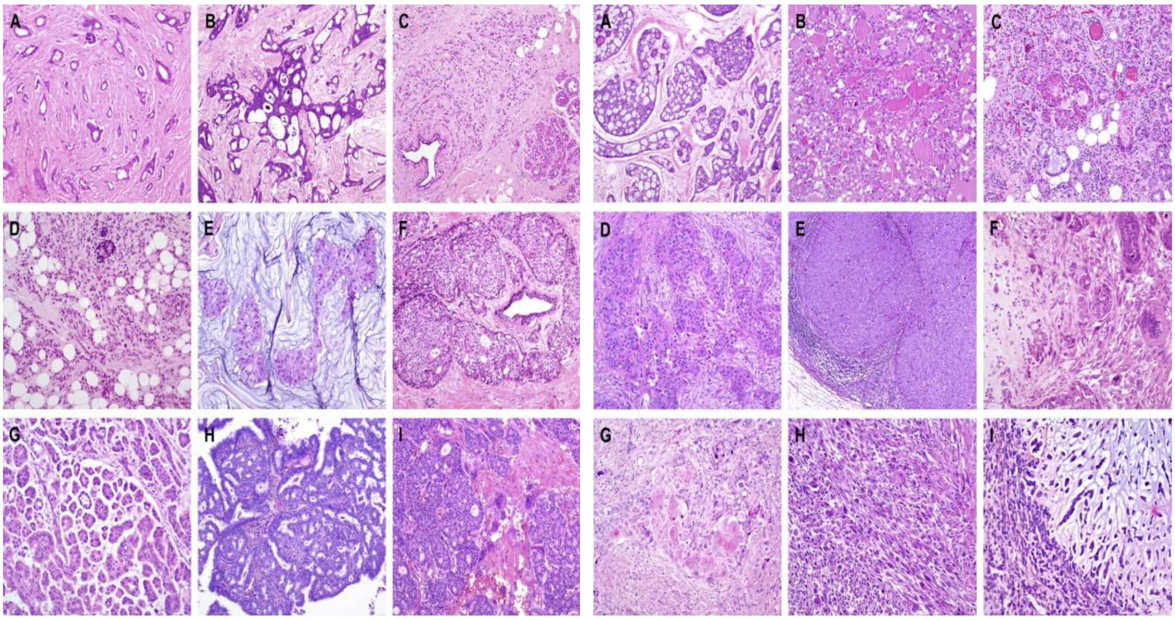
> Specifically, it is recommended that half of the invasive carcinomas detected by mammographic screening in the UK National Health Service Breast Screening Programme should be ≤15 mm in size. (Pinder, Harris, and Elston 2008).

Thereby, biological characterisation of tumours gains an even higher importance, while the existing cT categories go towards an additional revision and a closer specification of the TNM reference values (Arnone et al. 2009; Veronesi et al. 2009).

Histological characterisation of tumours has resulted in distinguishing 17 sub-types of breast cancer, which are grouped into two main groups, according to the status of estrogen receptor (i.e. ER+ and ER-) (Weigelt, Geyer, and Reis-Filho 2010). 'Histological type' refers to the growth pattern of the tumours, i.e. morphological and cytological patterns that were associated with distinct clinical presentations and outcomes (Ibid.). Figure28 presents histological types of breast cancer with distinct phenotypic patterns. The phenotypes presented in group 1 are estrogen positive breast cancer phenotypes, while group 2 presents estrogen receptor negative phenotypes.

A huge heterogeneity in observed phenotypic patterns that are associated with heterogeneous clinical behaviours of tumours, diverse response to therapy and a variety of outcomes, supports the claim that breast cancer is not a single disease but a set of diseases.

> Breast cancer is not a single disease, but is instead a collection of diseases that have distinct histopathological features, genetic and genomic variability, and diverse prognostic outcomes. Thus, no individual model would be expected to completely recapitulate this complex disease (Vargo-Gogola and Rosen).



1) Histological special types of breast cancer preferentially **oestrogen receptor positive**. (A) Tubular carcinoma, (B) cribriform carcinoma, (C) classic invasive lobular carcinoma, (D) pleomorphic invasive lobular carcinoma, (E) mucinous carcinoma, (F) neuroendocrine carcinoma, (G) micropapillary carcinoma, (H) papillary carcinoma, (I) low grade invasive ductal carcinoma with osteoclast-like giant cells.

2) Histological special types of breast cancer preferentially **oestrogen receptor negative**. (A) Adenoid cystic carcinoma, (B) secretory carcinoma, (C) acinic-cell carcinoma, (D) apocrine carcinoma, (E) medullary carcinoma, (F) metaplastic carcinoma with heterologous elements, (G) metaplastic carcinoma with squamous metaplasia, (H) metaplastic spindle cell carcinoma, (I) metaplastic matrix-producing carcinoma.

Figure 28 Histological special types of breast cancer, from Weigelt et. al, 2010, Molecular Oncology 4 (3):192-208.

The features that the members of this heterogeneous set of breast cancers share are: 1) a *common anatomical site*, i.e. mammary gland; 2) the features common to all cancer kinds, thus the breast cancers as well: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (Hanahan and Weinberg 2000).

An important task for a breast cancer ontology is to specify some particular features that more specifically describe the members of the breast cancer kinds, but it is hard to distinguish any particular feature that all of the breast cancer sub-types share (Vargo-Gogola and Rosen). Describing the diverse kinds of breast cancer as biologically heterogeneous implies that a functional characterisation of genetic and phenotypic behavioural patterns cannot cover this heterogeneity in a single model (Ibid).

However, while modelling this heterogeneous set of breast cancer types within an ontology model it is possible to specify meta-data as the types of data and classificatory criteria that play an important role in describing each of the breast cancer types, independently on how the data actually depict a particular breast cancer type[84]. More precisely, independently of the role that, for example, estrogen receptor (ER) plays in a particular cancer subtype, ER stays an important marker that describes tumour behaviour and response to therapy. This fact was the main reason why the histological sub-types were classified into two main subgroups, ER+ and ER- breast cancers (see Figure 28). Thus, specification of the tumour markers, such as ER, provides an important criterion for the classification of breast cancer types.

Tumour markers are molecular classifiers, which endorse the understanding of biological behaviour of tumours and a specific response to therapy (Gasparini and Hayes 2006). However, as discussed in the previous sections (3.1 and 3.5.4), the information on the tumour markers is not sufficient to capture the heterogeneous tumour features. Since, the heterogeneity of a breast

[84] The risk factors and age constitute a special kind of classificatory criteria, which I discuss in Section 3.7.

cancer phenotype cannot be fully covered within a single classification the parameters of several classificatory systems need to be combined (Viale, Ghioni, and Mastropasqua 2010). For example, information about presence of HER2 positive status changes significance of the generalisation, which characterises ER+ category as the one that indicates good prognosis and prolonged survival (Section 3.1). In addition, the information on the age and risk factors modifies interpretation of other classificatory parameters (Section 3.7). Thus, every breast cancer classificatory system provides an important piece of information that can maximise its utility in explaining and predicting cancer behaviour only when it is accompanied with the information recorded and structured within various classificatory systems.

In this section I discuss certain features of bio-molecular classificatory systems, which employ different classificatory criteria to describe breast cancer. I have presented a distinction of histological types as a kind of biological classification that looks at the morphological features of tumours on the tissue and cellular levels. The histological classification acquires additional information about the tumour phenotypes when tumours are classified according to the tumour markers criteria. However, for a variety of reasons the classifications of tumours that employ

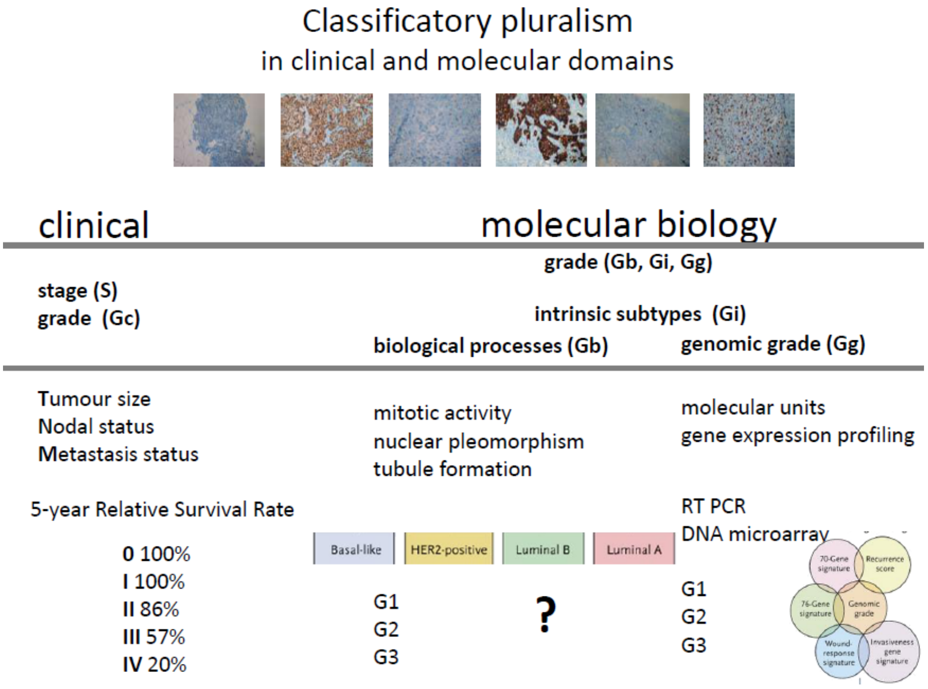


Figure 29 A plurality of breast cancer classifications

diverse classificatory criteria need to stay structured within separate classificatory system (Section 3.4). Mixing together the anatomical and functional characterisation of tumours that considers biological processes would break down the demand for the classificatory uniformity (Section 3.4). On the other hand, since the heterogeneous phenotypic features can be accurately captured only when these diverse classificatory parameters are combined, it is important to sustain the interoperability and reuse of the classificatory parameters. Keeping the classificatory categories within the distinct classificatory systems seems as a more productive approach in capturing the heterogeneous phenotypes, because in that way the classificatory parameters can be more easily combined according to the needs of a clinician or a biologist engaged in a diagnostic or an exploratory endeavour. Such an approach to classification employs distinct classificatory systems as the reference systems, which segments are reused and combined, according to the problem specific demands, e.g. clinical decision making.

Figure 29 summarises the idea of a *classificatory pluralism*, which comprises various clinical and molecular classifications that characterise breast cancer sub-types through the specification of 1) TNM staging and grading, 2) grading of biological processes captured in a histopathological image, 3) biomarkers, 4) 'intrinsic molecular sub-types ', and 5) genomic classifiers.

The biomarkers label biological units (e.g. ER), which have a diagnostic utility within the clinical domain. In addition, knowledge about the biological processes in which the marked biological unit is involved also supports an understanding of the cancer behaviour, e.g. hormone dependent response to therapy (Gasparini and Hayes 2006). Thereby, biomarkers support biological classification of tumours utilised in pathological grading and in a final tumour staging. In addition, biomarkers also provide information on gene expression, because they usually label proteins the expression of which is regulated on the genomic level. Therefore, a classification of tumours that looks at the protein expression also influences specification of the tumours' genomic grade.

The genomic classification is mainly focused on the molecular units or rather the sets of units acquired through the gene expression profiling (Sørlie et al. 2001; Pusztai et al. 2006). To some respect genomic classification has shown to be more similar to the clinical classification than to the biological one. For, the biological classification usually provides a biological explanation that connects the cancer grade with the biological processes, such as tubule formation. On the other hand, the genomic classifiers, analogously to the clinical ones, may lack a straightforward explanatory relation between the classificatory units (a gene set) and the disease specification. Gene expression patterns, which are recognised in breast cancer, distinguish tumour subclasses with clinical implications (Sørlie et al. 2001). Since the set of genes can be associated with the cancer through the multiple processes, diverse gene ontology (GO) categories[85] can be assigned to it. For instance, the 70 genes set[86] is associated with the hundreds of biological processes such as the cell cycle, apoptosis, androgen receptor signalling pathway etc. Yet, in the contexts where the gene set performs the role of a parameter within a predictive model, it functions as a classifier even if an explicit explanation of how the genes from the set are involved in the disease development is missing. In the genomic classification statistical evidence for the correlation between a gene set and the outcomes, as well as the clinical utility of the selected gene set, are often sufficient reasons that justify the classificatory choices.

On the other hand, the specification of *molecular sub-types* provides a link between the biological and genomic classification. It characterises the tumour sub-types by connecting the genomic data with biological knowledge about the tumour phenotypes (Sotiriou et al. 2006). Classification of breast cancers into the molecular sub-types combines classificatory parameters of biological and genomic classificatory systems, while describing the tumour sub-types.

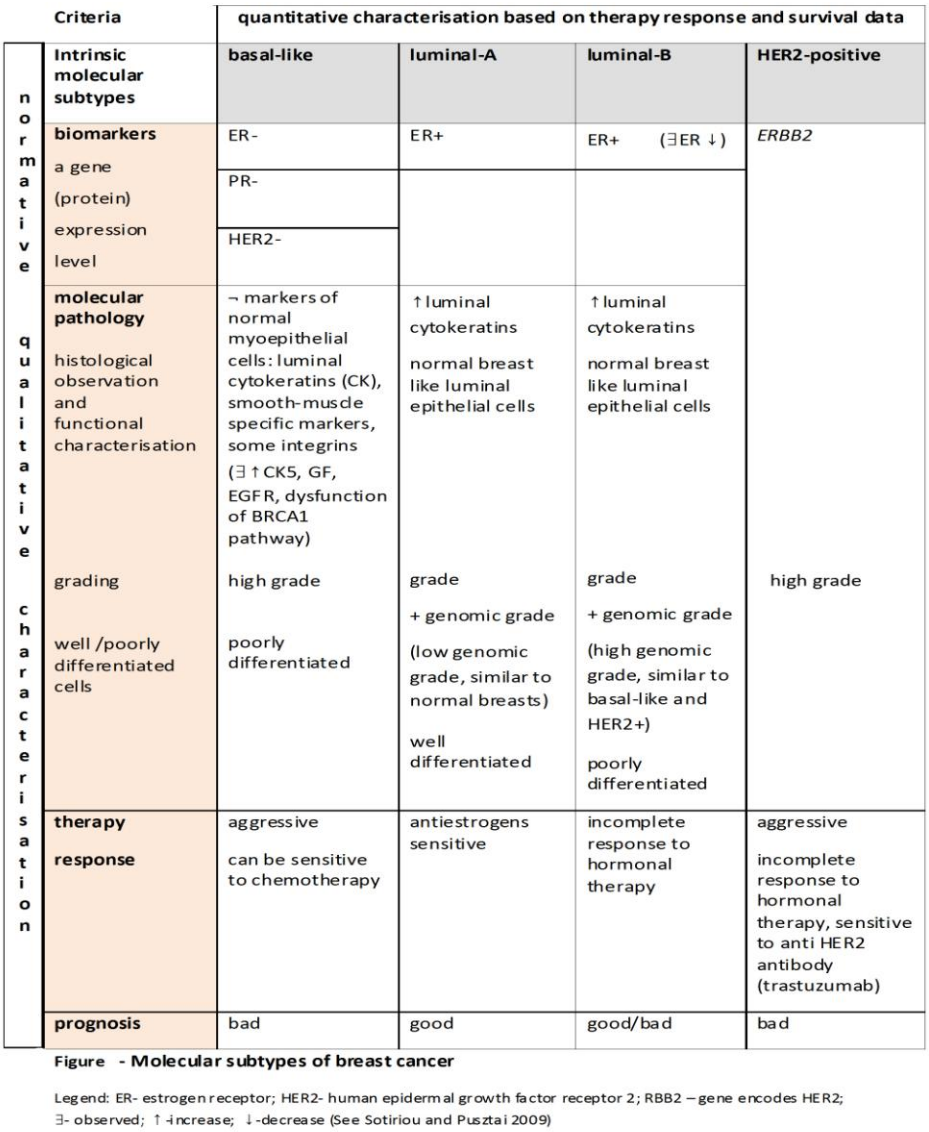
The use of knowledge acquired by the gene expression profiling has resulted in a specification of four main molecular sub-types of breast cancer. Named by Perou (Perou et al.

[85] The Gene Ontology (GO) is an important source, where knowledge about genes and gene products is organised. The aim of GO is to systematise knowledge about genes and gene products, grouping them into the categories of the biological functions, processes and components. Every association of a gene or gene product with a category is supported with evidence (e.g. a publication).

[86] www.adjuvantonline.com

2000) as ‘intrinsic’ these classes are 1) basal-like breast cancers, 2) luminal-A cancers, 3) luminal-B cancers and 3) HER2-positive cancers (Figure 30).

The basal-like type is often called triple negative, because it corresponds to ER-negative, progesterone-receptor negative (PR-) and HER2-negative tumours (Sotiriou et al. 2006). The luminal-A tumours are characterised by ER-positive markers and histologically low grade, while luminal-B tumours show mostly ER-positive and high-grade character (Ibid.). In some cases luminal-B cancer can have a low expression of ER. HER2-positive cancers show amplification high expression of the ERBB2 gene and the genes of the ERBB2 amplicon (Ibid.).



	Criteria	quantitative characterisation based on therapy response and survival data			
normative	**Intrinsic molecular subtypes**	**basal-like**	**luminal-A**	**luminal-B**	**HER2-positive**
	biomarkers a gene (protein) expression level	ER- PR- HER2-	ER+	ER+ (∃ER ↓)	*ERBB2*
qualitative characterisation	**molecular pathology** histological observation and functional characterisation	¬ markers of normal myoepithelial cells: luminal cytokeratins (CK), smooth-muscle specific markers, some integrins (∃ ↑CK5, GF, EGFR, dysfunction of BRCA1 pathway)	↑luminal cytokeratins normal breast like luminal epithelial cells	↑luminal cytokeratins normal breast like luminal epithelial cells	
	grading	high grade	grade + genomic grade (low genomic grade, similar to normal breasts)	grade + genomic grade (high genomic grade, similar to basal-like and HER2+)	high grade
	well /poorly differentiated cells	poorly differentiated	well differentiated	poorly differentiated	
	therapy **response**	aggressive can be sensitive to chemotherapy	antiestrogens sensitive	incomplete response to hormonal therapy	aggressive incomplete response to hormonal therapy, sensitive to anti HER2 antibody (trastuzumab)
	prognosis	bad	good	good/bad	bad

Figure - Molecular subtypes of breast cancer

Legend: ER- estrogen receptor; HER2- human epidermal growth factor receptor 2; RBB2 – gene encodes HER2; ∃- observed; ↑ -increase; ↓-decrease (See Sotiriou and Pusztai 2009)

Figure 30 A classification of the parameters, which describe the molecular sub-types

These classificatory groups correspond to the clinical characterisation that is based on ER and HER2 status, proliferation markers and the histological grade. While supporting the accuracy of the molecular classification, this correspondence, also limits its implementation in clinics. Since, a simpler and more feasible method often takes an advantage in clinical practice, histopathological grading is recommended as a valuable approximation to the genomic grading (Goldhirsch, Ingle, Gelber, Coates, Thurlimann et al. 2009; Goldhirsch et al. 2011).

Figure 30 illustrates how prognostic, therapeutic, and diagnostic criteria are combined together with the genomic and biological descriptions in the characterisation of breast cancer molecular sub-types. Although the quantitative measurements are not specified in the table, the qualitative descriptions (biomarkers, genomic grade etc.) have assigned numerical values, which are acquired through a comparison of ‘normal’ and ‘abnormal’ samples, as well as the information on the related therapeutic response and clinical outcomes (Perou et al. 2000; Sørlie et al. 2001; van de Vijver et al. 2002; Sotiriou et al. 2006).

An advantage of having a closer molecular classification of breast cancer is that it endorses the possibility to tailor a targeted therapy, which relies on a better understanding of the patient’s specific cancer biology (Sotiriou and Pusztai 2009; Sotiriou et al. 2006). Indeed, in some cases the genomic classification of phenotypes when connected with the clinical outcome does give a better prediction than the traditional classification. A comparison[87] of the estimated risk of the cancer reoccurrence and metastasis calculated by conventional criteria (tumour size, nodal status, grade, ER status) and the criteria of the 70 gene signature[88] has shown that in 29% of cases (87 of 302 included patients) the risk assignments were not in agreement (Sotiriou and Pusztai 2009). Among these 87 patients, 59 (68%) of them had assigned high-risk through the conventional methods, while the gene signature ascribed to them low-risk.

On the other hand, 28 patients (32%) were clinically classified as low-risk, while the genomic signature rated them as the high-risk group (Ibid.). This mismatch between the results of

[87] www.adjuvantonline.com

[88] MammaPrint, Agendia, developed in the Netherlands Cancer Institute

the analysis by the conventional clinical criteria and the genomic signatures, demonstrates how a comparison of the domain specific methods influences the advancements of biomedical knowledge. When calculated the 10-year overall survival, the genomic assignment showed as more accurate than the conventional clinical parameters. The survival rate in the group of patients with low-risk according to the genomic classification was 89%, while the group with the high-risk genomic signature had 69% survival rate (Ibid.). A combined approach that uses MammaPrint and clinical guidelines has led to an altered adjuvant treatment recommendations in 26% of cases from a group of 427 patients in the Netherlands hospitals (Bueno-de-Mesquita et al. 2007).

Hence, although diverse, molecular and clinical classificatory knowledge are complementary as well. Knowledge of the genomic grade can improve the accuracy of prognosis and modify decision for a therapy, while the pathological classification of tumours based on histological types often serves as a valuable approximation to the genomic grading (Goldhirsch, Ingle, Gelber, Coates, Thürlimann et al. 2009; Goldhirsch et al. 2011). In addition, it has been shown that the molecular markers have a limited value as clinical parameters, especially when they are evaluated separately (Viale, Ghioni, and Mastropasqua 2010). Likewise, genomic knowledge combined with clinical parameters brings in clinical practice an additive predictive value, while the molecular subtypes particularly contribute to the understanding of the underlining cancer biology. Such an integrative approach to the disease that considers clinical and molecular knowledge, while integrating diverse classificatory systems, is designed to support a personalised medicine (Piccart 2006; van 't Veer and Bernards 2008).

3.7. Risk factors as the criteria for classification

The previous sections have presented various aspects of breast cancer classifications within the clinical and molecular domains. However, an important area of knowledge about breast cancer has been left out of the previously discussed classificatory systems.

The aim of sections 3.7.1-3.7.4 is to clarify how the categories such as age, gender or social status, which are usually understood as risk factors and not as classificatory categories, influence breast cancer classification. I will argue that the (categorical) terms corresponding to the risk factors actually play an important role of the classificatory units. Therefore, I carry on a discussion on why knowledge about the risk factors should be explicitly integrated as a classificatory system into the breast cancer knowledge representation. Moreover, I argue for a necessary inclusion of the risk factors into the breast cancer ontology. I base my argument on an interdependence of different types of classificatory knowledge. Accordingly, I demonstrate that knowledge on a single level, captured within a domain specific classification, will be incomplete and sometimes mistaken, if the inter-domain dependences are marginalised. The following sections identify how risk factors influence breast cancer classification.

3.7.1. Identifying the risk factors

It is already a part of common knowledge that certain factors closely relate with the occurrence of breast cancer[89]. These factors, which are associated with an increased risk of cancer development, are labelled as the risk factors. They might be specified through the categorical and numerical values, thus including a high array of information about genetics, behaviour, life style, the environmental conditions etc. (McPherson, Steel, and Dixon 2000). Even so, the common feature that makes all these factors a unique group is an association with cancer development. Information on the risk factors highly overlaps with the information on the cancer incidence. [90] Only by knowing that something happens at a certain rate is it possible to gain knowledge of how likely it is that it will happen.

For example, it is recognized that information about *age* and *gender* is relevant for breast cancer risk assessment. Statistical data show that the risk for development of breast cancer

[89] http://info.cancerresearchuk.org/cancerstats/types/breast/riskfactors/

[90] http://info.cancerresearchuk.org/cancerstats/types/breast/incidence/

increases with age. That is to say the incidence of breast cancer is higher in the population of women older than 50 (Althuis et al. 2005). Gender is relevant as well, since it does make a difference in the risk assessment. The women are almost 100 times more likely to get breast cancer than men. Only around 1% of all breast cancer cases are registered in the men population (Giordano, Buzdar, and Hortobagyi 2002). For the same reason it is difficult to conduct research and perform large clinical trials on male population, as the male breast cancer counts as a rare disease. *Family history* of breast cancer counts as a risk factor because of a higher incidence of breast cancer in women whose close relatives have had breast (Pharoah et al. 1997; Couto and Hemminki 2007), uterine, ovarian or colon cancer (Broca 1866; Narod et al. 1991; Nelson et al. 1993; van der Groep, van der Wall, and van Diest 2009). *The genetic changes* such as mutated BRCA1 and BRCA2 genes significantly increase the risk of getting breast cancer, up to 80% (King et al. 2003). Other relevant information is about the *menstrual cycle*. It is observed that women whose period starts early, before age 12, or those who go through menopause late, after age 55, have an increased risk of getting the cancer (Brinton et al. 1988; Gao et al. 2000). Information about *parity* is included, since it is recognised that there is a higher risk of breast cancer in women who never had a child or who had first child after age 30, while an early pregnancy and having more than one child reduces the risk (MacMahon et al. 1970; Lambe et al. 1996). Further factors that are associated with the breast cancer are related to the external inputs such as use of alcohol (Longnecker et al. 1988) or use of certain *drugs*, e.g. diethylstilbestrol -DES, which was a prescribed drug as a prevention from miscarriage, between 1940’s and 1960’s (Palmer et al. 2006), and hormone replacement therapy (Schairer et al. 2000). An additional factor is an exposure to *radiation*, especially in childhood (Preston et al. 2002; Kleinerman 2006). *Obesity* is a risk factor associated with an increased production of estrogen, which increases likelihood of breast cancer development (Huang et al. 1997).

3.7.2. What do the risk factors explain?

The research about the relation of the risk factors and the disease development has resulted in various epidemiological and statistical models (Key, Verkasalo, and Banks 2001; Tyrer, Duffy, and Cuzick 2004), which represent knowledge about breast cancer in a way that differs from the clinical and molecular representations. Models that are focused on the risk factors aim at representing the changes in the cancer incidence by looking at the differences in its distribution across the geographical regions and time scales (Key, Verkasalo, and Banks 2001). These models often include the information registered by tracking down co-presence of chosen parameters that are of particular relevance for the cancer incidence (Ibid.). Even if molecular parameters are sometimes present in those models, they are treated as the risk factors on the same line with other risks such as the environmental factors (Key, Verkasalo, and Banks 2001; Tyrer, Duffy, and Cuzick 2004). In that respect, the risk factors simultaneously play the role of the descriptive and predictive units, while representing knowledge that something is the case. For instance, certain epidemiological studies (Althuis et al. 2005) represent *that* the incidence of breast cancer is higher in the population of women, especially those older than 50. An aggregation of the risk factors will further increase the likelihood of a particular woman's risk. So, it will be also possible to predict that within the group of women with family history of breast cancer, a woman of age 60, who was an alcohol consumer at an early age, and who has been on hormonal replacement therapy for 10 years, has a high risk of developing breast cancer (Key, Verkasalo, and Banks 2001; Tyrer, Duffy, and Cuzick 2004). However, the captured information does not go beyond a representation *that* the factors are related. Consequentially, the explanation that comes with the epidemiological and statistical models of breast cancer consists in an explication of *what is the case*, and not why it is the case. Such a population based model that takes into account a multi-factorial risks, genetic and environmental, acts as more accurate predictor than a model that considers the genetic risks only (*ibid*.). Moreover, the interpretation of the results captured by the model explains *that* the genetic risk cannot be sufficient for the cancer development, because some other factors need to be co-present in order to trigger the genetic predispositions to develop cancer (Ibid.). However, an explanation of *why* it is the case stays out of the scope of this particular model, because the identification of the processes and mechanisms that lead towards the cancer development is out of the scope of epidemiological studies.

In a certain respect, the outlined reasoning typical for epidemiology shows similarities with clinical reasoning. That is to say, the relation of TNM classes and the patients' outcome mainly represent *that*, for example, lymph node involvement is positively associated with a bad prognosis, but not how the two are related. Thus, from an explanatory perspective, epidemiological classes are alike the clinical ones.

How is it possible then that the information about the risk factors helps in a better understanding of the disease and how it happens? At this point we can consider again an *asymmetry* and interaction between clinical and biomedical knowledge (see Section 3.2). Information in the form of the observed correlations, while having a *direct* impact on a health policy, which influences clinical practice, actually seems to have only an *indirect* role in the production of biomedical knowledge. For example, most of the listed risk factors are associated with the hormonal changes and metabolic deregulations (Boyle and Leake 1988; Pike et al. 1993). The statistical evidence on the correlation between the information on hormonal changes (e.g. age at the first and/or last menstruation, pregnancy etc.) and cancer incidence directs the focus of biomedical research to understand and explain hormonal interactions and deregulations of metabolic pathways (Ibid.). Therefore, the statistically confirmed association among the risk factors and the development of disease play a mediatory role in the process of explaining *why and how* the cancer occurs. The asserted associative claims, which are epistemically more cautious than the causal ones, serve as a guide in the search for a more specific causal explanation within biomedical knowledge.

If the risk factors are included in the explanation only in this indirect way, do they lose significance in the representation of the breast cancer heterogeneity? In case it is possible to give an accurate representation of the heterogeneity through the molecular characterisation of cancer, without adding the information about the risk factors, the representation of breast cancer heterogeneity could stay on the molecular level only. Consequentially, the breast cancer classification might have been unified into a unique and accurate classification that is the molecular one. However, I will show that this is not the case.

3.7.3. The role of 'Age' across the classificatory domains

In the previous sections (2.1, 3.4, 3.6) I have presented several reasons that support a plurality of the disease representations and classification. I have based my arguments on the identification of specificities and interdependences among the knowledge domains, which are always capturing the problem of interest in a pragmatic and perspective driven manner. Here I defend the position that the similar interdependences hold among the risk factors and other breast cancer classificatory categories. In particular, I consider the role that *age* plays in epidemiological, clinical, and molecular characterisation of breast cancer. However, I demonstrate that the classificatory systems consider the relevance of age on different levels and with different degrees of explicitness. Thus, I will show that these diverse domains conceptualise *age* in different ways, while classifying and organising knowledge about breast cancer.

Information about age plays an important role in clinical reasoning about breast cancer. Because an apparently similar tumour may exhibit quite distinct behavioural patterns across different age groups, clinical decisions about prognosis and therapy choice all depend on information about the age of patients. Therefore, breast cancer in young and elderly people are often considered as distinct, age dependent classes (Anderson, Jatoi, and Sherman 2009; Jatoi, Anderson, and Rosenberg 2008). In general terms, the age dependent differences can be explained by the age related physiological processes that drive an age specific behaviour of the tumour (Thomas and Leonard 2009). On the other hand, a molecular characterisation of cancer aims at capturing typical molecular features of cancer subtypes whereby the age differences are often intentionally neglected. Such an abstraction from age dependent variations in molecular characterisations of cancer subtypes is a useful tool, which supports an understanding of those molecular features that are common to a cancer subtype independently of patients' age.

However, clinical and epidemiological studies show a difference in frequency of breast cancer subtypes, hormonal status, or tumour grade among the age groups (Thomas and Leonard 2009). The question I address considers how molecular classifications and models for breast cancer can embrace clinical and epidemiological knowledge about the varieties associated with age dependent cancer statistics that is explained by age-specific physiological processes involved in carcinogenesis and therapy response. In section 3.7.4 I outline a framework for a model that supports an integration of clinical and molecular classifications. Such an integrative model should avoid a proliferation and repetition of existing molecular classes across age groups. Instead of having separate classes of breast cancer in young and elderly people, we can specify 'age' as a category that can be re-used in various classificatory contexts, while integrating cancer classes across age groups. Of course, the integrative model also needs to capture the heterogeneity of age specific physiology and diverse cancer behaviours. However, a specification of *age* that covers epidemiological, biological, and clinical contexts needs to face many challenges.

Understanding *the concept of age* of an organism in terms of physiological processes and their relation with carcinogenic events is not a simple task. Although age can be specified as a property of an organism it is not obvious what kind of property it is. Apparently, it does not seem suitable to characterise age as a biological function of an organism. Age, rather, is a measure of the process of ageing. However, the process of ageing can be measured in different ways while specific questions are being addressed. Information about age in the molecular domain is specified in terms that are difficult to use in clinical practice where age is being specified as a relative measure calculated from the date of birth of an individual to a certain point in its life-span. On the other hand, the biological process of ageing and its relation to the carcinogenic processes is explained, *inter alia*, by processes such as telomere shortening (Campisi et al. 2001). Thus, telomere, which consists of TTAGGG sequence repeats at the ends of chromosomes and which plays a key role in the maintenance of chromosomal stability, has a function in the ageing of an organism that is sometimes co-occurrent with the disease development (Liu et al. 2003). Stated in simple terms, through every division of a somatic cell a bit of telomere is being ‘lost’. Thus, the telomere shortening which naturally occurs in time provides a molecular explanation of ageing (Campisi et al. 2001; Olovnikov 1996).

However, I argue that information about age will lose important aspects of its explanatory power when getting into details of molecular processes, e.g. representing functions of telomere and its shortening. A detailed representation of molecular events involved in ageing not only likely entails an overload in the information content, making complex events even less comprehensible, but there is also a risk of producing red herrings, misrepresenting the role of *age* in a disease.

Namely, the age of an organism does not need to coincide with the molecular age of a tissue within an organism. Rather, *these different concepts of age are addressing different questions*, while measuring diverse aspects and temporal features on connected, but still distinct levels, focusing on bio-chemical, cellular, metabolic, tissue, organ or organism related processes.

> Breast cancer might be the best example of how a single concept of age cannot cover every research question. For most cancer sites there is a linear loglog relationship between incidence and age. This relationship does not hold for breast cancer, and certain ‘key’ breast cancer risk factors suggest that breast tissue does not ‘age’ in step with calendar time (Pike et al. 1983).

A mismatch between age and cancer incidence has led to the hypothesis that hormonal changes, such as those that occur in pregnancy, result in diverse tissue ageing rates (Ibid.). Thus, an early pregnancy slows down the breast tissue ageing, thereby reducing the risk of cancer development. According to this hypothesis, hormone induced differences in tissue ageing can also explain why the first childbirth at a late age is not as protective as it is at an earlier age, but actually increases the cancer risk. The breast tissue of women who were not pregnant at an early age has aged faster than the tissue of women who have had an early age pregnancy.

If the tissue is aging at a different rate than the organism, this process should be detectable and explainable *on the molecular level* as well. Such a molecular explanation of aging should also explicate the connection of aging and carcinogenesis. However, the theory of aging that looks at the molecular activities does not seem to be compatible with certain cancer theories that attribute 'immortality' to the cancer cells[91]. Instead of going to apoptosis cancer cells continuously divide. The telomere-theory of aging claims that the cells senesce through every cell division (Campisi et al. 2001; Olovnikov 1996). The telomere shortening results in chromosomal instability and sequential accumulation of the mutations that result in cancer (Campisi et al. 2001; Liu et al. 2003). Behind this claim there is a traditional model of carcinogenesis as a gradual progression of events that result in cancer. This view, which relates telomere shortening (aging) with carcinogenesis, seems to extrapolate knowledge from epidemiological studies about the age related cancer incidence to the molecular level. However, such an extrapolation might not be quite appropriate, because knowledge about the role of telomerase in the processes that counteract the aging by the telomere repairing mechanisms does not fit completely into the linear picture of aging as a gradual process related to carcinogenesis. Telomerases are RNA-dependent polymerases that catalyse the synthesis of the telomeric DNA at the tips of eukaryotic chromosomes (O'Reilly, Teichmann, and Rhodes 1999). In addition, telomerase is recognised as crucial for carcinogenesis by counteracting the aging, and contributing to the 'immortality' of the cancer cells (Shay and Bacchetti 1997). As differentiated somatic cells rarely present telomerase activity, an inactive telomerase is attributed to 'normal' somatic cells in contrast to the cancer cells whose telomerases are typically active. Therefore, telomerase is also proposed as the diagnostic and prognostic marker (Shay and Bacchetti 1997). However, when taken together the molecular theory of aging and knowledge about telomerase activity rather seem contradictory than complementary. Since the cancer cells do not senescence as normal cells, coming up with the concept of age *on the molecular level* is a particularly complex issue, which asks for a precise

[91] E.g. a theory that attempts to explain the origin and mechanisms by which cells become 'immortal' (cancer) cells is the cancer stem cell theory (Pece et al. 2010).

distinction of the units and levels on which *age* is measured, e.g. age of cells, cellular components, tissue, and organism. Moreover, this controversy raises the question if 'age' is actually an ambiguous term, which is inappropriately used in some contexts.

Independently of the soundness of the above presented hypotheses of hormone induced differences in the breast tissue ageing and the controversy in the conceptualisation of molecular aging, an observation of interdependences between hormonal changes and breast cancer goes back to the early history of breast cancer research. Indeed, we have only acquired scientific evidence on the association between age, pregnancy, and breast cancer incidence through clinical

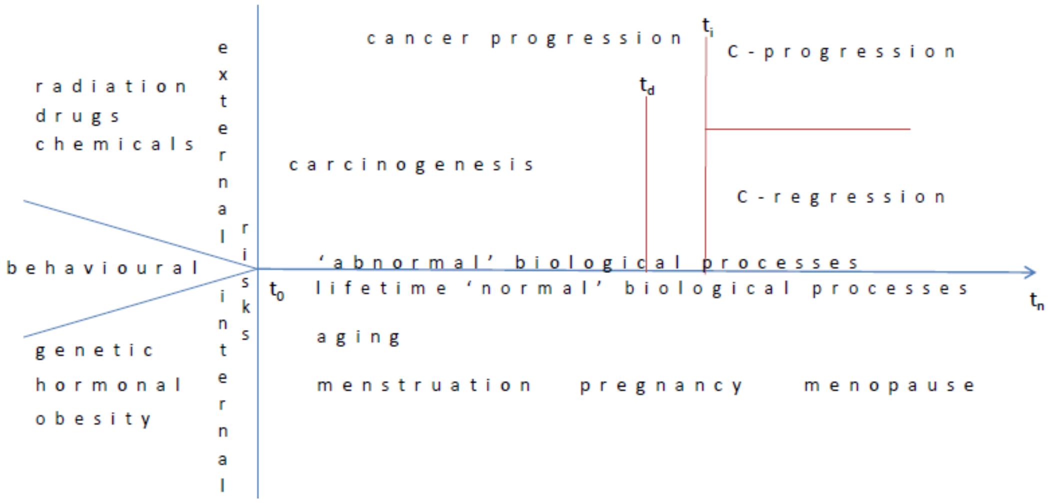


Figure 32 The age related domains: risks, 'normal', and 'abnormal' biological processes

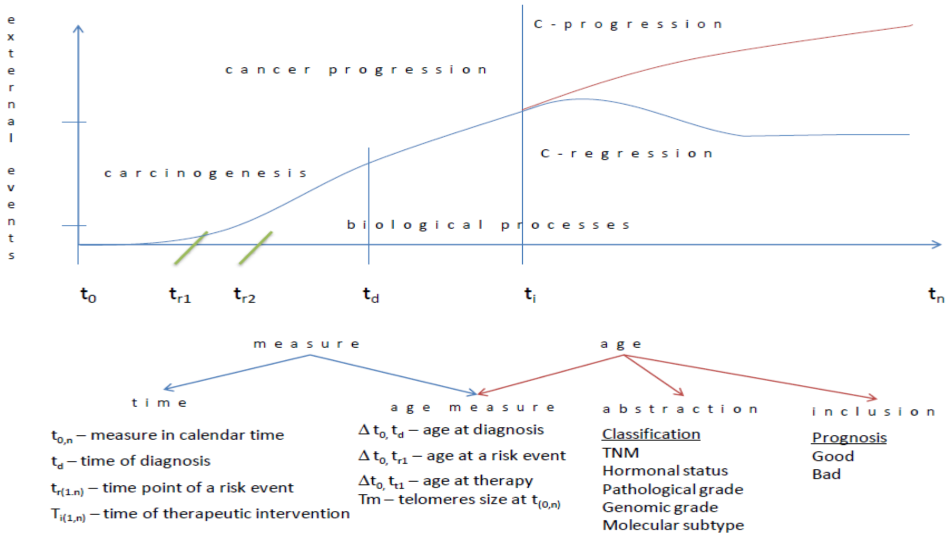

external events
C-progression
cancer progression
C-regression
carcinogenesis
biological processes
t_0
t_{r1}
t_{r2}
t_d
t_i
t_n
measure
age
time
age measure
abstraction
inclusion
$t_{0,n}$ – measure in calendar time
t_d – time of diagnosis
$t_{r(1.n)}$ – time point of a risk event
$T_{i(1,n)}$ – time of therapeutic intervention
$\Delta t_0, t_d$ – age at diagnosis
$\Delta t_0, t_{r1}$ – age at a risk event
$\Delta t_0, t_{t1}$ – age at therapy
Tm – telomeres size at $t_{(0,n)}$
Classification
TNM
Hormonal status
Pathological grade
Genomic grade
Molecular subtype
Prognosis
Good
Bad

Figure 32 Age as a time dependent measure

trials and controlled population studies.

Figures 31 and 32 present a time line, which maps certain parameters (the risk factors and cancer related events), which are outside of the cancer context considered as normal biological processes (e.g. aging, menstruation, pregnancy etc.).

Clinical and epidemiological studies report that the risk factors, tumour characteristics and clinical trial results significantly vary across age groups (Wang et al. 2007). The association of age and cancer characteristics is known as *age interaction*. Here, the age-dependent differences are not only in magnitude, but, more importantly, *information about age changes the evaluation and interpretation of what the risk factors, cancer traits and therapy responses are*. Thus, the relation between age and cancer specification or classification is not just a quantitative one, but also includes a qualitative difference. Moreover, there is not a single cut-off point such as postmenopausal age that is sufficiently relevant for the understanding of the differences. Rather, as Anderson et al. (Anderson, Jatoi, and Sherman 2009) analyse, there are a few age-thresholds where the qualitative age interactions take place. The qualitative age interaction as defined by Peto (Peto 1982) refers to the cases where an inverse or crossover effect for different ages occurs. I here present the summary of results from (Anderson, Jatoi, and Sherman 2009), which need to be considered in an ontology model that specifies breast cancer classes. For example, parity increases the risk of breast cancer before age 33–40 but is protective thenceforth (Pike et al. 1983). While obesity is a risk factor for postmenopausal women, age 50, it is protective for premenopausal women (Cleary and Maihle 1997). Black women have a higher risk of breast cancer than white ones, but only if they are younger than age 40. After this age, the groups swap in the risk assignment (Anderson et al. 2008). Age-specific incidence rates are higher for blacks than whites prior to 40 years of age, but higher for whites after 40 years of age. Also, the tumour characteristics associated with a poor prognosis, such as tumour sizes greater than 2 centimetres, positive axillary lymph nodes, high tumour grade, estrogen and progesterone receptor negative expression, are not equally distributed in all age groups (Anderson, Jatoi, and Devesa 2005). Poor prognostic factors are more common among women before the age of 50, whereas good prognostic factors are more common after the age of 50 (Ibid.). Younger women, before the age of 35, have a much greater risk of developing metastasis, having higher recurrence rates (RR) and a much worse outcome than older women (Nixon et al. 1994). In addition, women before the age of 35, with ER+ breast cancers, have higher rates of relapse after the adjuvant systemic treatment than older women (Goldhirsch et al. 2001). So, ER positivity as a 'good' prognostic factor does not have the same weight in all age groups. This means that the 'poor' prognosis cannot be based just on the characterisation of a tumour type, and a tumour type does not provide a sufficient explanation for breast cancer behaviour and the patient's outcome (Nixon et al. 1994). Age is relevant as well.

Information about age seems to play a crucial role in explaining 1) different responses to a therapy and thus characterisation of disease in cases when the only known difference is age difference; 2) why the breast tumours in old and young groups of patients are characterised as biologically diverse.

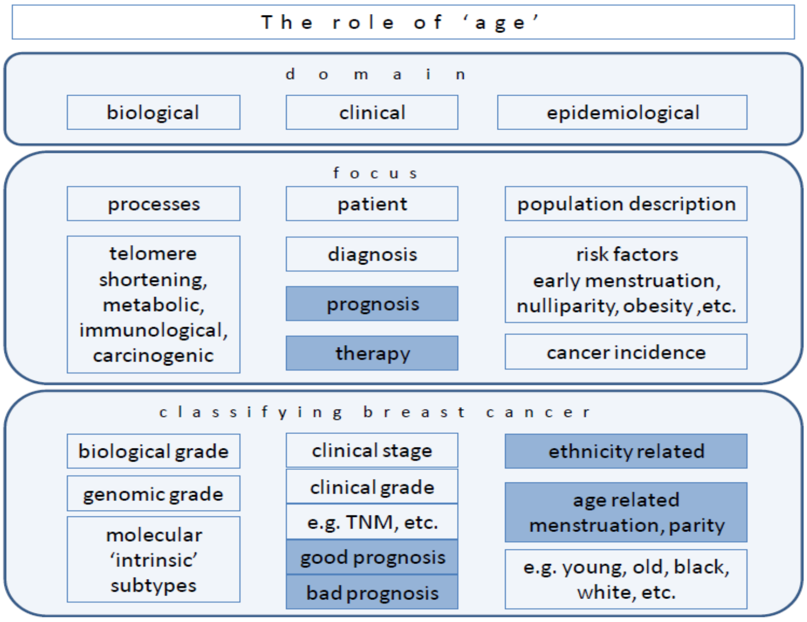


Figure 33 The role of age across domains

However, outside of the epidemiological domain, the risk factors are rarely explicitly recognised as classificatory categories, yet they are unavoidably included in various segments of clinical and molecular knowledge, which drives classifications. Therefore, I here argue that an explicit inclusion of the risk factors as classificatory categories must take place in a breast cancer knowledge representation, because the risk classifiers may contribute to a better understanding of the classificatory systems, and so prevent a possible confusion and misinterpretation of certain classificatory categories.

Figure 33 presents how a specific domain focus, explicitly or implicitly, directs description of age related parameters, which I have previously discussed. Epidemiological descriptions capture age parameters on a population level, which considers risk factors and cancer incidence. Clinical domain focuses on classification of patients, assessment of diagnosis and prognosis, whereby age is used as an important parameter which complements clinical classificatory (TNM) categories. On the other hand, biological parameters capture age by looking at the biological processes such as telomere activity, while the related biological classifications of breast cancer rarely include explicitly age classifiers.

Regarding molecular classifications, a search for capturing age dependent differences has resulted into a proliferation of the molecular groups of breast cancer. So, breast cancer is used to be grouped into the categories: 1) breast cancer in young, 2) breast cancer in very young women, and 3) breast cancer in elderly, potentially also including sub-groupings arranged by the tumour stage into 4) early breast cancer in young woman, or 5) metastatic breast cancer in elderly patients (Piccart 2006). However, these separation and molecular groupings appear only in the context that considers clinical applications of molecular knowledge. The age grouping of molecularly distinct classes mainly concerns the diversity in response to a particular therapy. In the context of molecular research the cancer models do not represent the age differences. It is hard to find any model of breast cancer that employs only molecular representations of the carcinogenic processes that are based on the age differences, i.e. the age specific molecular models. That is to say, the molecular biology research often abstracts the age component and analyses tumours independently of age. Accordingly, the molecular breast cancer sub-types are distinguished and classified by looking at the tumour characteristics, which do not include the age parameters (Figure 32). Only when it comes to the therapy decision, the age depended responses are considered, based on the existing survival data and the related therapy response. Moreover, the data on the elderly patients' therapy response are poor, because the elderly patients are not widely included in the clinical trials. The survival data of elderly patients are also bias for the reason that elderly patients often do not receive the therapy that would prolong their survival at the most (Jones 2006; Stockler 2006).

Eventually, the molecular sub-types are represented on the molecular basis only, while the age of patient plays a classificatory role only in a clinical context. As a result, there is a proliferation of the clinical models of breast cancer as an age dependent disease, while the molecular representations of the disease outside of the clinical and epidemiological context do not consider explicitly the age dependent differences.

3.7.4. Representing 'Age' as an epistemic modifier

In order to resolve the conceptual problems in understanding and modelling the role of *age* across research contexts that specify breast cancer classes in different ways, I propose a pragmatic approach to disease representation while performing an abstraction of domain specific uses of the term 'age'. I treat 'age' as a general classificatory module that plays various roles in distinct knowledge contexts. Thus, although age is neither a biological function of an organism nor of a cell component, *age* still plays an explanatory role in our understanding of disease within each of the contexts. This ambiguity in the meaning of 'age' might be resolved when the context of its use is explicitly specified.

As *age* is perceived in various ways within the domains of epidemiological, clinical and molecular biology research, a plurality of its functions (e.g. predictive, explanatory) can be integrated by an explicit recognition of the diversities and similarities between the pragmatic and epistemic interests within and across the domains. I therefore argue that in the representation of breast cancer, the concept of *age* functions as a cognitive effect modifier.

Anderson et al. (Anderson, Jatoi, and Sherman 2009) introduced *age* as an effect modifier in order to explain why age-dependent variations in breast cancer outcomes and therapy responses should be seriously taken into account in the design of clinical trials. Since I here consider the domain of modelling and the disease representation, I consider *age* as a cognitive effect modifier. That is to say, our contextual understanding of *age* modifies our decision on how we should represent the relations among relevant concepts and particular breast cancer classes. For example, several age points (12, 30, 40, 55), represented as nodes within the breast cancer representation, may be mapped onto other relevant concepts, e.g. breast cancer risk factors. The nodes representing age-points function as reference points in our explanation of the diverse courses of disease. This explanatory function of age as an effect-modifier can be further specified according to the demands of a particular domain. In molecular and cell biology, ‘age’ will be mapped onto fine-grained specifications of molecular age by looking at telomere shortening and telomerase activities that are involved in cancer development. Consequently, an explanatory function of ‘age’ within the molecular context is to connect explanandum, i.e. a specific tumour behaviour, with its explanans, e.g. a description of the temporal molecular and metabolic processes. Likewise, ‘age’ can be mapped onto another set of concepts represented within the clinical and epidemiological domains, while supporting domain specific explanations and predictions. The function of ‘age’ within the clinical domain will be to predict a prognosis for a patient by looking at the differences in the clinical outcomes of the patients belonging to different age groups. On the other hand, ‘age’ in epidemiology functions in the mapping of the disease incidence across the regions, ethnicities, and the diverse social groups. The age-incidence map can be further enriched with the information about co-present risk factors, eventually resulting in a representation of how significant the relation of age is to other risk factors. Such a representation can also make explicit an explanation that the genetic factors and ageing are not sufficient for the development of cancer (Key, Verkasalo, and Banks 2001). The importance of the interaction between the genetic, environmental, and behavioural factors (such as life style) for the explanation of carcinogenic events additionally supports the claim that a molecular characterisation of ‘age’ cannot provide a sufficient explanation for how specific breast cancer classes depend on age.

Figure 34 presents an initial idea of the framework for an ontology model of breast cancer, which integrates various conceptualisations of *age* on the molecular, tissue, organ and organism levels. Biological age is divided into sub-types molecular age, tissue age, organ age, and organism age. However, as *age* is conceptualised in different ways on each of the levels, the model in particular distinguishes measurements of calendar and biological age, which does not need to correspond linearly to the calendar age (Pike et al. 1983). While age stays implicit in the specification of breast cancer subtypes, it explicitly influences the clinical diagnostic and prognostic classes through the assignment of the patient specific parameters about age and age dependent risk factors.

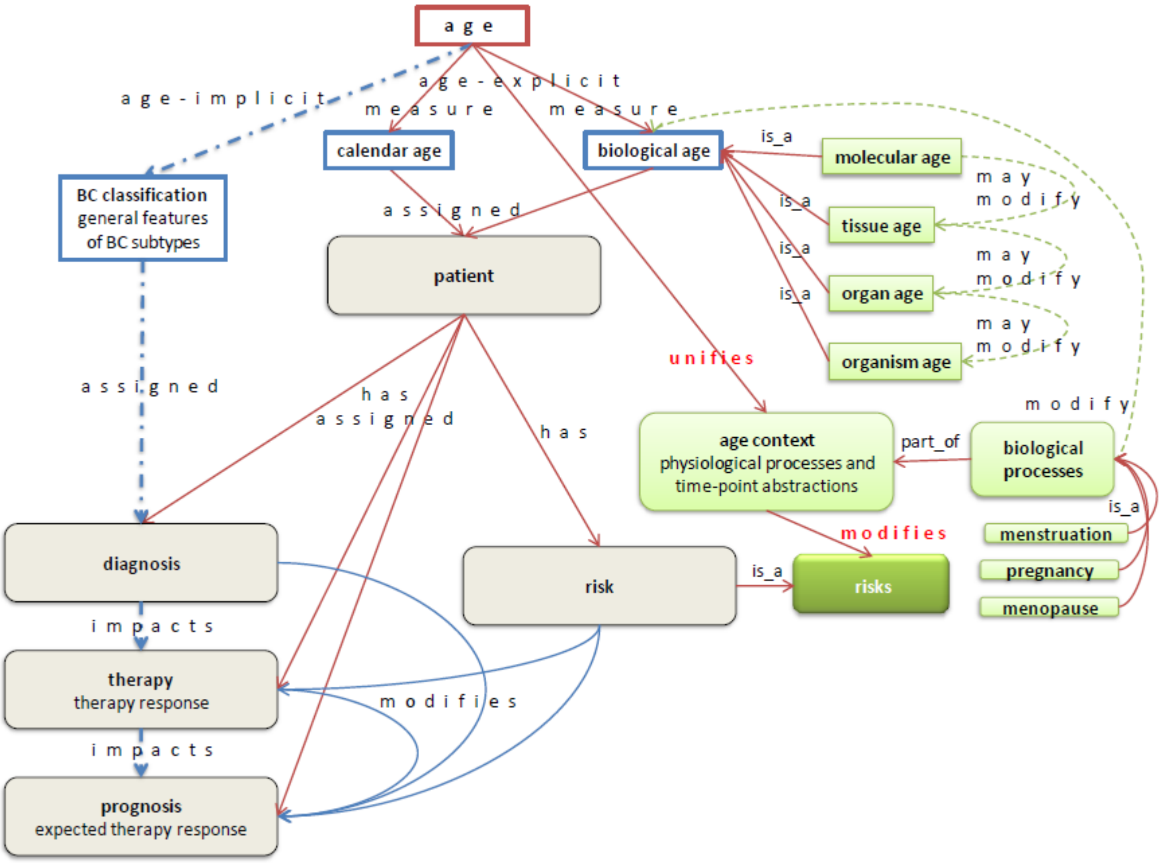


Figure 34 'Age' as an epistemic modifier

Section 3.7.3 clarifies the idea that the risks cannot be interpreted in a linear way, but the age context significantly modifies the understanding of certain parameters (e.g. parity) as beneficial or maleficent in every particular (age) case. Menstruation, pregnancy, and menopause are the examples of age dependent biological processes, which in the age context get a specific interpretation of either the risk factors or normal processes (Section 3.7.3). The distinction between a patient specific risk and the set of potential risk factors is represented by separating the two classes (the grey vs. green box in Figure 34). The patient risk (grey) is a special subclass of the risks (grey), because the class of risks represented in grey includes general knowledge about age dependent risks, whereby what is a risk for one patient (grey) need not be risk for another one. That is to say, formal modelling of the risks may employ non-monotonic reasoning with circumscription (see Section 2.2.2), so that the concept of risk (green) is minimised (and considered as normal) unless the information about the patient's age at an event (e.g. an early age at diagnosis) is present. As it is known that younger women, before the age of 35, have a much *greater risk of developing metastasis*, having higher recurrence rates (RR) and a much worse outcome than older women (Nixon et al. 1994), as well as that women diagnosed with ER+ breast cancers before the age of 35, have higher rates of relapse after the adjuvant systemic treatment than older women (Goldhirsch et al. 2001), the information about *age related risks* will modify the prognostic and diagnostic classes.

Although the information about age related risks is normally considered in clinical domain, a formal representation that explicitly integrates information about the role of age in carcinogenic processes brings an advantage as various pieces of knowledge about cancer biology and epidemiology may complement each other.

Since the concept of age plays an important role in clinical, biological and epidemiological reasoning about cancer, I have explicated a number of problems that need to be resolved for an adequate conceptual modelling of the role of age across research domains (Section 3.7.3). I have shown that the relations between the classificatory parameters might not be always reducible to the representations of molecular mechanisms and processes (e.g. telomerase activity) by which they are involved in disease development. At least, representing details of age related processes is sometimes less useful than an abstraction that considers just the specific time points in which a relevant physiological change takes place. Even if some of the qualitative age interactions in breast cancer can in principle be explained and represented in terms of age-related metabolic processes and hormonal changes, other factors such as parity, obesity, and ethnicity also influence the direction in interpretation (positive/negative) of the 'switching' age points (Section 3.7.3). Therefore, 'age' as having assigned a numerical value that labels a time point in the life-span of an organism provides a more useful representational tool that integrates an array of directions in which the interpretation of the tumour behaviour can take place. Thus, an interpretation of 'age' that I propose plays the role of an epistemic modifier in the representation of many possible outcomes. I have also shown that only by making the classificatory connections explicit, even by 'packing' them into abstract and general categories such as 'age', which I define as an epistemic (cognitive) modifier, further precise connections among the classificatory parameters can emerge. For, rather than being detached from each other, the classificatory categories are closely intertwined. A recognition of similarities and differences across the knowledge domains facilitates the explication of the inter-domain relations, and therefore the integration of the classificatory systems.

From a practical point, such an explicit representation of the classificatory parameters and their interconnections can support clinical reasoning by reducing complexity of physiological details through an abstraction, while preserving accuracy in performing complex reasoning tasks related to the qualitative age interactions. On the other hand, an explicit inclusion of epidemiological parameters, such as risk factors, into a representation that integrates the clinical and molecular breast cancer classificatory categories enhances an understanding of biological mechanisms through which the disease takes course in a particular case. My proposal for the interpretation of 'age' as an epistemic modifier in an integrative breast cancer ontology is open for an inclusion of other domain specific ontologies such as a physiological processes ontology. Indeed, the details do matter, in particular when a personalised medical treatment needs to be tailored. However, before going into the details, we have to abstract from them in order to provide an integrative ontological framework which can embrace a broad array of the particular and numerous physiological patterns.

In order to conclude section 3.7 and the third chapter, I shall briefly summarise the main points about the way in which I specified the clinical and molecular biology approaches to classification as distinct but mutually related.

We have seen how the distinctions among the classificatory systems are sometimes blurred. Still, the methodological differences among the classificatory systems are significant. So within the classificatory systems we can distinguish different interests that direct the classifications. In the first part of the chapter I explained how some of the differences originate in the specific kind of knowledge and reasoning that is typical for the clinical and biomedical domains. I have also presented various systems for the classification of breast cancer, e.g. TNM system, including the clinical staging and grading, and pathological grading. Following the expansion of molecular biology and genomic research, biological grading and molecular grading have been developed in order to support more accurate prognostic and diagnostic strategies. However, I have shown that the resulting clinical and molecular classifications are still represented in separate systems, while knowledge from the cancer biology has not yet been fully incorporated into the clinical practice. An apparent separation of the classificatory systems is in accordance with the demands for a well structured classificatory organisation, presented in section 3.4, whereby classificatory units can be combined according to the needs of the task at hand.

Furthermore, I extended my analysis to the breast cancer risk factors. I have argued that knowledge about the risk factors should be incorporated into the breast cancer knowledge representation in an explicit way, so that the risk factors can also be employed as the means for the breast cancer classification. When connected with other classificatory systems, the risk factors can enrich the understanding of other classificatory categories with knowledge from epidemiology, which is usually incorporated into the classifications in a very indirect way.

The moral of my descriptive and comparative analysis of the classificatory systems is that the classificatory terms designate the dynamic and interdependent categories, and therefore their interdependence should be made explicit.

The explication of the interdependences among the categories has numerous advantages. Consider just the asymmetry between the predictive and the explanatory roles of the classificatory units within the clinical and molecular domains. A clinical or a genomic classification may have a predictive power without having a strong explanatory power (Sections 3.6 and 3.7.2). The associations among the tumour features and the clinical outcomes are often based on the statistical analysis that is lacking an explicit explanation of how and why the association holds (Sections 3.6 and 3.7.2). Therefore, for an evidence based medicine, which goes beyond the statistical evidence, clinical classification has to be accompanied with various histological, molecular and epidemiological classifications into an explicit representation of the scientifically justified connections that explain relations among the classificatory units. I presented the molecular subtypes and histological types of breast cancer (Section 3.6) as an example of the classificatory categories that also provide an explanation for the specific tumour behaviour. In this way the classificatory units play the role of explanatory units. Having explicated the classificatory relations, which hold among the classificatory units, the predictive and explanatory functions of the classificatory units coming from diverse domain classifications may be mutually enriched. Moreover, the analysis of similarities and differences that hold among the classificatory systems, including an explication of the relations among the classificatory categories, provides a conceptual framework for the breast cancer ontology.

The discussion of the risk factors and *age,* in particular, has shown how certain categories and relations are understood across the breast cancer domains, whereby having different degrees of explicitness. Accordingly, I treated 'age' as a general classificatory module that plays various roles in distinct knowledge contexts which can be captured explicitly within an ontology model. I presented how the contextual understanding of age modifies our decision on how we should represent the relations among relevant concepts and particular breast cancer classes. Similarly to the HER2 ontology, formal modelling of the risks may employ non-monotonic reasoning with circumscription (see Section 2.2.2), so that the concept of risk is minimised and considered as normal, unless the information about the patient's age at an event (e.g. an early age at diagnosis) is present. Accordingly, the information about age related risks will modify the prognostic and diagnostic classes.

The discussion of the ontology modelling strategies, specification of the ontology-classes as well as the breast cancer classificatory systems have provided the arguments in favour of a pragmatic and pluralistic view on classification and biomedical ontologies. Classificatory systems are ordering and organising knowledge in a manner that is most useful to the chosen, domain-relevant questions, whereby the classificatory categories capture the most relevant aspects of the phenomena of interest. The ontologies are presented as human engineering products that represent conceptualisation of a domain, as needed and when needed, in order to fulfil a particular purpose. Accordingly, the ontology classes and the relations are specified in a domain-specific manner, while particular modules and classes can be mutually linked and re-used, if needed and when needed.

Chapter IV

The Biomedical Ontologies Facing the Disunity of Science

Having discussed the basic concepts and issues related to biomedical ontologies, knowledge representation, clinical and molecular breast cancer classification, this chapter aims at summarising the discussion in the light of the unity of science debate. I address some of the traditional philosophical problems related to the conflict between two apparently counteracting forces: the unity of science and disciplinary division. In particular, I examine the impact of information technologies and databases for knowledge integration on the unity of science debate.

The development of information technologies and formal tools such as ontologies for knowledge representation has positioned the integration of the heterogeneous knowledge domains and types of representation as an important goal (Chapters I and II). However, aiming towards the integration is not sufficient for claiming a unity of science. Therefore, I shall examine in which sense and to what extent the resulting integration might be considered as a unification of scientific knowledge. While discussing several approaches to the unity and disunity, I indicate a new setting of the debate that emerges with the expansion of the information technologies. Eventually, I argue for a form of collaborative unity, and for a weak form of representational integration. The chapter concludes by proving a view as to how clinical and molecular knowledge (Chapter III) get mutually enriched while being connected by means of biomedical ontologies.

4.1. The problems with merging domain knowledge

Knowledgebases (KBs) and applied ontologies have been recognised as a convenient tool to bridge heterogeneous knowledge domains (Hunter and Summerton 2006; Leonelli 2008; Kutz 2011; Leonelli and Ankeny 2012; Sojic and Kutz 2012). In the following sections l examine what this merging of different knowledge domains actually means. In order to address the issue of knowledge integration, I will first analyse the domain knowledge problems through the examples from molecular oncology and clinical practice.

The computer scientists who deal with interoperability of diverse knowledge domains have explicated certain obstacles and problems for merging knowledge formally into knowledgebases (Hunter and Summerton 2006). While discussing the problems outlined by Hunter and Summerton (for short, H&S), I point out the implications that the presented problems impose to the unity-disunity debate, which I address in the reminder of the chapter.

H&S distinguish two basic *kinds of knowledge used in a knowledgebase*: domain knowledge and generic (general) knowledge (Hunter and Summerton 2006). While *generic knowledge* is defined as a type of knowledge that can be used in multiple applications, *domain knowledge* is characterised as being specific for a particular domain, thereby it has a limited applicability and cannot be easily re-used in other domains (Ibid.). We can illustrate this view by the example of the claims that do hold across multiple domains. For instance, reasoning about *events* always includes some kind of temporal relations. Thus, independently of the domain, the kind of event taken into consideration, and the time-measurements, we perceive an event as composed of basic segments (sub-events) which are related in that way that one sub-event (a state or a process, e.g. transcription factor binding) precedes another sub-event in time (e.g. translation, protein aggregation etc.). Hence, knowledge that an event is composed of one or more ordered ‘segmented events’, i.e. *predecessor* and *successor*, is an example of generic knowledge applicable to various domains. It can be applied to the modelling of business action plans as well as to the modelling of molecular processes. Broadly speaking, top-level ontologies (e.g. DOLCE, BFO) are designed to capture such generic knowledge.

Regarding generality of ontologies that are capturing intensional knowledge (Section 1.2.1, 1.2.2), the distinction between generic and domain knowledge should not be confused with the distinction between intensional and extensional knowledge (Hunter and Summerton 2006). Intensional knowledge is often described as general knowledge (e.g. in description logic and ontology specification implied by Guarino's definition in (2009)), because it communicates a shared meaning and understanding of a represented concept, and it is contrasted to the extension understood as an individual that is an instance of the concept represented within a knowledgebase. However, since intensional knowledge mostly concerns a particular subject, it cannot be qualified as general knowledge in H&S sense, because such knowledge is not applicable to the domains that deal with different subjects. In other words, intensional knowledge is mostly knowledge about a particular subject within a context, and the context of a domain specifies the universe of discourse within which the terms acquire particular meaning thus expressing specific knowledge claims.

Moreover, H&S distinguish several sub-types of domain knowledge: ontological knowledge, relational knowledge, meta-level knowledge, coherence knowledge, and knowledge about allowable ranges of various values (Hunter and Summerton 2006). *Ontological* knowledge considers the terms as belonging to the same equivalence class, and certain hierarchical ordering of the terms according to their generality. For example, the term 'protein' is a more general term than 'protein 53', which belongs to a sub-class of the term 'protein'. *Relational* knowledge structures data into a relational form. A simple example of relational knowledge illustrates the subsumption relation, which connects two terms by *is_a* relation into the claim 'protein 53 *is_a* protein'. Another kind of relation frequently used in biomedical ontologies is the parthood relation. So, a cell is represented as being in a parthood relation to its components, i.e. 'cell *has_part* a cell component' (see 2.2.1). *Meta-level* knowledge includes, for example, preferences over the sources and reliability of the presented information. The evidence codes used in the Gene Ontology are an example of meta-level knowledge. Another example consists of the preferences captured as the classificatory criteria. For instance, animals might be classified according the criteria that prioritise differences in body temperature or they might be classified according to a particular purpose, e.g. model organisms in laboratory research (see discussion about NCIT in 3.4). Meta-level knowledge explicates the criteria used in classification. *Coherence* knowledge considers consistency among the terms. By avoiding statements that contain contradictory terms the consistency of the reasoning within a knowledgebase is preserved. For example, while the terms 'malignant' and 'proliferation' are consistent, the terms 'normal' and 'abnormal' are not, because they cannot be consistently assigned to the same class, within the same context. In other words, something is considered as abnormal only if it is not normal, i.e. 'normal' and 'abnormal' classes are mutually exclusive. Knowledge about *allowable ranges* of values deals with information that can have different values. Thus, specification of the value ranges aims at increasing precision, while reducing vagueness. An explicit specification of allowable ranges is particularly important for the tumours classification. Having explicitly represented the value ranges for the tumour markers such as HER2 it will be possible to classify a tumour as HER2-positive or HER2-negative sub-type (see 2.2.1, 2.2.2).

The presented distinction of the knowledge sub-types is closely linked to the classificatory demands[92] discussed in section 3.4. For instance, the design of a well structured classification asks for relational knowledge. Disjointness of the classes contributes classificatory consistency, whereby coherence knowledge provides a distinction of the terms and related classes that are mutually exclusive, being either inconsistent or coherent. Knowledge about allowable ranges of values supports explicitness, precision, and disambiguation of represented concepts. Regarding meta-level knowledge, H&S seem to defend a position that is in line with the position I argued for

[92] According to Ludger Jansen, the classificatory demands are a) Ontological Grounding, b) Structure, c) Disjointness, d) Exhaustiveness, e) No ambiguity, f) Uniformity, g) Explicitness and precision, h) No meta-types (Jansen 2009). For the critical analysis of these criteria see Section 3.4.

in 3.4. Namely, contrary to Jansen, H&S emphasise importance of meta-level knowledge to representation of domain knowledge. Likewise, ontological knowledge, as presented by H&S, remains in line with the pragmatic approach (see 3.4), which focuses on language use rather than on some metaphysical properties that might describe a domain (Hunter and Summerton 2006). Similarly to classificatory demands, the subtypes of domain knowledge are complementary and interdependent. Relational knowledge informs ontological knowledge on how equivalence classes need to be organised into a hierarchical structure. Coherence knowledge contributes consistency of the represented relational hierarchy. Specification of the allowable ranges and meta-level knowledge on the preferences about reliable sources additionally contributes precision and exhaustiveness of the represented knowledge.

Nevertheless, the distinction of domain knowledge sub-types is operative in pointing to the variety of aspects that domain knowledge captures. Indeed, by looking at a specific knowledge sub-type we can identify the sources of the problems that emerge in the process of merging knowledge from different domains.

For instance, clinical, epidemiological, and molecular domains need not associate the same relations with a term (e.g. 'HER2'). As illustrated in the previous chapters (Sections 3.3, 3.4, 3.7.3), each of the domains prioritises certain relations above the others. Clinical studies focus on the patients, diagnostic, prognostic, and therapeutic terms that can support clinical decisions (3.2, 3.3). Epidemiological studies focus on populations and frequency of cancer incidence, which might be related either to genetic or to the environmental factors (3.7-3.7.4). Molecular studies, on the other hand, look for the relations among the molecular components and mechanisms that provide an insight into molecular characterisation of disease (3.6). The question is if these diverse kinds of relations that represent knowledge about cancer in various domains can be indeed consistently merged together. If so, we shall examine how thus is possible, and what kind of integration this could be. In other words, we should examine how the diversities across the disciplinary domains influence interoperability and integration throughout the ontological knowledge, relational knowledge, meta-level knowledge, coherence knowledge, and knowledge about allowable ranges of values.

Undoubtedly, the first thing to consider when discussing knowledge integration is a distinction between *complementary* and *conflicting* information. If different sources provide complementary information about different aspects of the same phenomenon, the process of knowledge integration seems less questionable than in those cases in which conflicting information is present. The complementary information can be merged by the standardised tools for integration of new information to the existing sources (Hunter and Summerton 2006). However, deciding on what is complementary and what is conflicting is not always a straightforward task. H&S distinguish three basic types of conflict that can appear in the process of merging information: 1) dispositional conflicts; 2) epistemic conflicts; 3) ontological conflicts.

Dispositional conflicts arise as a result of the conflicts of interests and values. An example of dispositional conflict, provided by H&S, refers to the reports such as war reports about the casualties, whereby each side in the war conflict communicates different information. Similarly, the results of political elections often present an interest driven reporting. In the case of weather forecasts a subjective view in reporting can influence a more or less optimistic qualitative description of meteorological conditions. Of course, neither biomedical reporting is free of interest and values. Socio-philosophical studies have already emphasised an important role that values play in scientific reporting (Kincaid, Dupré, and Wylie 2007; Kitcher 2001), whereby certain (dispositional) interests drive selection of a problem, data analysis, selection of particular graphs for the data visualisation, as well as statistical and experimental tools that can support scientific hypothesis.

Epistemic conflicts arise from epistemic reasons such as differences in beliefs. Unlike dispositional conflicts, which also might contain some underlying differences in beliefs, the conflicting beliefs in epistemic conflicts need not have an obvious non-epistemic interest involved. The example of epistemic conflict is illustrated by the conflicting reports about future events such as forecasts (Hunter and Summerton 2006). Indeed, even subtle differences in the experimental methods and measurements can result in different predictions (Ibid.). In scientific practice, the distinction between dispositional and epistemic conflicts is often blurred, because the belief in the justifiability of a scientific hypothesis can induce a dispositional interest in production of the most suitable experimental evidence. Accordingly, the 'unsuccessful' experiments are withdrawn, while only those which seem most valuable in convincing validity of the hypothesis are published and disseminated. Analysis of microarray data is an example that shows how data interpretation can be biased (Ambroise and McLachlan 2002; Pusztai et al. 2006). Depending on the research question, the same data can be interpreted in many different ways, and the selection of a particular method for data analysis involves both epistemic and dispositional interests (Pusztai et al. 2006).

> "If you torture your data long enough, they will tell you whatever you want to hear" has become a popular observation in our office. In plain English, this means that study data, if manipulated in enough different ways, can be made to prove whatever the investigator wants to prove. Unfortunately, this is generally true. Because every investigator wants to present results in the most exciting way, we all look for the most dramatic, positive findings in our data. When this process goes beyond reasonable interpretation of the facts, it becomes data torturing. (Mills 1993)

Of course, the rigor of scientific methodology contributes balancing between the dispositional and epistemic interests through the demands such as replicability of the experiments and cross-validation of the results (Allison 2006). In the formal knowledge representation, epistemic and dispositional conflicts can be, at least partly, reduced by specification of meta-data (meta-level knowledge), which explicate and distinguish the context of the underlying beliefs, presented as the scientific hypotheses (Soldatova, Rzhetsky, and King 2011), classificatory criteria, preferences over the experimental methods, and procedures used in data production (Taylor et al. 2008; Colombo et al. 2009). Accordingly, one of the main tasks for the database-curators is to evaluate confidence of the asserted claims and to relate them to the available data sets, experimental methods and tools (Leonelli 2008; Shimoyama et al. 2009).

Ontological conflicts, according to H&S, arise from differences in terminology and different uses of terminology across various domains and sources. So, different terms are often used for the same concept (synonymy), and a term can also be associated with different concepts (polysemy). H&S argue that terminological disambiguation can ameliorate ontological conflicts by specifying relations of synonymy, polysemy, homonymy, antonymy, meronymy, and hyponymy (Hunter and Summerton 2006). Indeed, a significant effort of the ontological community, which has been working on interoperability issues and knowledge integration, was focused on terminological alignments, development of controlled vocabularies and reference terminologies (Rector 2003; Hartung et al. 2012; Milian et al. 2010). However, as previously discussed, the use of a reference terminology can resolve just one side of the problem (Rector 1999).

The ontological conflicts are, at first, captured as terminological conflicts, because conceptualisation of a domain is communicated in language. Still, the origin of terminological conflict goes beyond a terminological disagreement (Ibid.). In other words, differences in conceptualisation, within and across domains, become explicit in the process of disambiguation of the terms in use. While in many cases the agreement about the terms' meaning can be achieved, there are cases where the term's specification cannot cover the variety of practical (epistemic and pragmatic) interests across the fields (Sojic and Kutz 2012). Various domains necessarily employ specific conceptualisation that fits best the needs of a domain. As discussed on the example of 'HER2', clinicians use 'HER2' in assessing diagnosis, while biomedical researchers find particular interest in molecular structures and functions of the protein.

Accordingly, the conceptualisation of the term 'HER2' as 1) a tumour marker *and* as 2) a protein with specific biochemical features might be considered as an example of polysemy. Polysemy is defined as the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings. Though complementary, these two meanings of ‘HER2’ are not identical, because the same term ‘HER2’ does not have the same semantic and cognitive value in the context when it is used to denote either a tumour marker or a biochemical component. This ambiguity usually does not produce problems in practice, when the context of use directs the language users to select the most appropriate meaning of the term (Section 2.4.).

However, when the information about ‘HER2’ comes to formalisation, the ambiguity is not allowed. Formal specification involves an explicit definition, a formal representation of definiens, and its constitutive parts. So, ‘HER2’ as a tumour marker plays a function of a diagnostic label, having assigned a measurement value, which may be provided through various methods. In the case of FISH method, the signal of HER2 gene expression in nucleus provides a measurement value, while in IHC method the measurement captures expression of HER2 intramembrane protein (Sections 2.2.1, 2.2.2). Still, the meaning of ‘HER2’ as a tumour marker does not change when the method of measurement changes, i.e. in both cases ‘HER2’ has meaning of a diagnostic marker. On the other hand, the meaning of ‘HER2’ as a protein can be formally represented either way, e.g. by capturing the sequence of the coding gene, 3D structure of the protein, or its functional features. Apparently, each of the cases chosen to represent the specific meaning of ‘HER2’ asks for different formalisation. Therefore, the term’s meaning needs to be described in a precise and specific meaner, which fits best particular needs[93]. Terminological disambiguation in the field of applied ontology and knowledge representation involves a collaborative work of various experts who are making explicit conceptualisation of a domain (Ekins et al. 2011; Leonelli and Ankeny 2012; Sojic and Kutz 2012; Tudorache 2010).

Regarding the role of terminology in the process of knowledge integration it is important to make a distinction between terminological agreements as means and ends. If the

[93] Section 4.2.2 presents examples that illustrate how balancing between generality and specificity of the terms supports integration across the fields. While keeping the meaning as general as possible supports the inter-domains connectedness, the context of a given domain supports the expansion of the informational content by providing the meaning with specificity.

establishment of a common reference terminology and controlled vocabulary were the only targets of the integrative endeavour, the terminological disambiguation could have been a sufficient means. However, it is not the case. Terminology is just a tool that needs to serve its purpose to communicate information and represent knowledge.

A standardised and disambiguated reference terminology facilitates communication and the information retrieval across the domains. By linking synonymous terms across distributed databases and the digital information systems, various domains can be accessed in an easy and quick manner (Leonelli 2008). However, this kind of connection across the domains that *links* distributed databases by means of shared terminology mostly contributes a faster exchange of information that can, in principle, take place by any other communication channel, e.g. publications. In the case of dissemination of information through publications, the task of interpretation, understanding and integration of the represented information is left to the reader (see groups 2 and 3, Section 1.2.6, Fig. 14). Likewise, the mere linkage[94] of terms by digital means leaves the process of knowledge integration to the interpreter, e.g. database user. Of course, the user friendly interfaces for the information retrieval allow a faster access to the stored information. However, arguably[95], the very act of information retrieval is not much different from the analogue process, which takes place in well organised museums and libraries that provide access to the stored material objects, data, and documents. The data and documents are classified and stored, whereby the web interfaces just quicken the access. Accordingly, a commonly accepted reference terminology, which *links* information sources, should be considered as a necessary but not sufficient condition for claiming a kind of knowledge

[94] The mere linkage here denotes connection of information by means of technology such as HTML in World Wide Web, which does not include an explicit semantic specification that is used in other platforms such as OWL. For further clarification and the distinction between HTML, OWL, RDF etc. see Uschold (2004).

[95] Bruno Strasser, for example, argues for a more significant difference between traditional and digital databanks (Strasser 2012), while Stefan Müller-Wille (Müller-Wille and Charmantier 2012) defends the position that downplays this difference. Regarding databases as the information storage, my position is in line with Müller-Wille's view, while I argue that digital media and information technology brings a substantial novelty in the field of automated reasoning and KR. In addition, the use of databases seems to play an important role in shifting the way of thinking about a research field, e.g. model organisms (Leonelli and Ankeny 2012).

integration. The ‘digitalisation’ of an agreed terminology is just an initial step, which can, in principle, support knowledge integration.

The remainder of the chapter examines how biomedical ontologies, while dealing with the problems such as merging knowledge from different domains, can inform the philosophical debate about the unity of science.

4.2. A revival of the unity vs. disunity debate

The division of scientific labour and the perspectives on the unity and disunity of science have been thoroughly debated among philosophers (Neurath 1937; Oppenheim and Putnam 1958; Galison and Stump 1996, Bechtel and Hamilton 2007). While disciplinary division has been shown to be inevitable (Dupré 1983; Kitcher 1990, 2001; Rosenberg 1994; Cartwright 1999), a need for the integration of the dispersed pieces of knowledge acquired in particular disciplines has been recognised as well (Kitcher 2001; Bechtel and Hamilton 2007; Bechtel 1984; Darden and Maull 1977; Potochnik 2010). This section gives an overview of the debate, which in the light of the ontology expansion gets a new turn that revives the old ideas both ways, for and against the unity.

The idea of the unity of science goes back to Aristotle. Bechtel and Hamilton recognise five historical attempts to unify knowledge: 1) Aristotle’s metaphysical and hierarchical unity; 2) the Enlightenment project of the French Encyclopedists; 3) the systematic unity of Naturphilosophen Lorenz Oken; 4) the methodological unity of the Vienna School’s Encyclopedia of Unified Science; 5) the organizational unity of cybernetics and general systems theory (Bechtel and Hamilton 2007). I argue that the recent expansion of the information technologies marks a new era in which the idea of the unification in science plays an important though testing role. The development of standardised reference terminologies, the data exchange platforms, the ontology languages, and models bring onto the scene a number of the old ideas of 'unified science' that are blended together in a particular way.

Aristotle portrayed science as a unified hierarchy, which is subdivided into the theoretical sciences (metaphysics, mathematics, and physics), the practical sciences (ethics and politics), and the productive sciences (poetry, rhetoric, etc.). According to Aristotle, the theoretical sciences and metaphysics in particular lie on the top of the hierarchy whereby investing the first causes provides the universal understanding of "all the underlying subjects" (Aristotle 2008). Aristotle's influence on the contemporary approaches to the ontology modelling and its application is more than apparent. In particular, top-level ontologies such as BFO (Grenon 2003; Jansen 2009) strongly rest on such an Aristotelian tradition (Pisanelli 2004; Jansen 2009; Smith 2004). The following passage illustrates the view on ontology as a basic and unifying science.

> The task of ontology is to represent reality or, rather, to support the sciences in their representation of reality. In the last chapter, the reader became acquainted with an important means of doing so, namely: the technique of classification. But, in any classification, what are the very first kinds? What should the top level look like? [...] From the point of view of the philosophical tradition of ontology, the question of a top-level ontology is tantamount to the question of the most basic categories. In order to develop some alternative suggestions, the nature of categories must first be addressed. To this end, I appeal to the philosopher whose ideas are pivotal in influencing our current understanding of ontology: Aristotle [...] (Jansen 2009).

Thus, Ontology is described as a unifying science that can provide general knowledge, which connects the sciences on a top-level by means of the most basic categories that distinguish dependent versus independent entities, continuents versus occurrents, and universals versus particulars. A need for clear understanding and specification of classificatory categories that are used to represent knowledge seems undisputable. However, the question is how these categories should be specified. As already discussed through numerous examples (Chapter III), the agreement on one and the best way to classify things might never be achieved, neither in principle nor in practice (Chapter III). The context of a domain, particular modelling and problem solving tasks are directing decision on the most suitable classificatory criteria. Thereby, the domain-specific demands for the classificatoty choices undermine the viability of such a unifying classificatory system. In other words, the classificatory pluralism in the sciences brings up the issue of what these most general categories can unify. The question is if the BFO categories can indeed be *consistently*, *unexceptionally*, and *simultaneously* mapped to the various classifications across the scientific domains, thus providing the unity of science. The examples of classificatory pluralism (Chapter III) illustrate a crucial problem for BFO in the attempt to capture *unexceptionally*, and *simultaneously* the distributed domain classifications. The classificatory heterogeneity persists as a useful tool in addressing particular tasks (Dupré 1983; Mai 2004; Lord and Stevens 2010), thus making the sciences reluctant to such a uniform treatment. Nevertheless, the integration and interoperability of the domains are still an important task for the ontologists (Leonelli 2008; Hoehndorf, Dumontier, Gennari et al. 2011; Hoehndorf, Dumontier, Oellrich et al. 2011). Accordingly, the *consistency* of biomedical ontologies is often preserved in practice by taking into account the contextual limitations without the ambition to produce an *unexceptionally*, and *simultaneously* unified classification. In order to achieve such an interoperability across the domains, not only ontology languages are heterogeneous (Kutz, Mossakowski, and Lücke 2010; Lange et al. 2012; Lange, Mossakowski, and Kutz 2012), but also several upper-level ontologies are currently in use so as to fit best the practical needs (Hoehndorf, Dumontier, Gennari et al. 2011; Lord and Stevens 2010). Such a practice supports both directions of thought about science: 1) it supports the idea that science is unified, because various research groups do collaborate, exchange information, and represent it formally; 2) it supports the idea that science is disunified, because any attempt to build a uniform framework that fits all scientific domains fails to grapple with pluralism. Accordingly, the debate on the disunity of science endures.

Another well known undertaking to unify knowledge is the Enlightenment project of the French Encyclopedists. The ontological efforts in knowledge integration reveal many similarities with this Enlightenment project. Unlike the Aristotelian and Scholastic approaches, which consisted of metaphysical speculations, the main guides for the Encyclopedists were reason and empirical knowledge, which had been fostered by the scientific revolution. *The Encyclopedia, or Reasoned Dictionary of the Sciences, Arts, and Trades* (1751–1772), edited by Denis Diderot and Jean Le Rond d'Alembert, was published in 17 volumes. The project was an expansion of *The Universal Dictionary of Arts and Sciences* (1743-1745), which under Diderot turned into 'a monumental effort to outline the present state of knowledge in the sciences, arts, and practical crafts and to make this knowledge widely accessible' (Bechtel and Hamilton 2007).

Publication of the comprehensive knowledge, produced by experts and disseminated openly shows an obvious similarity with the aims and noticeable results of the 'digital Encyclopedists'. Through the World Wide Web, information technologies and ontology tools, the new Encyclopedists also employ 'digital dictionaries' such as SNOMED and other reference terminologies that connect various areas of knowledge, making it open and easily accessible, mostly available to everyone who needs information. This aspect of 'unification', unlike the Aristotelian view presented above, seems undisputable and widely supported in society of experts and non-experts. However, as Bechtel and Hamilton put it, 'Encyclopédie represents a compilation of knowledge rather than an integration of it. In many respects, this reflects our contemporary situation' (Bechtel and Hamilton 2007). The 'unification' provided by means of publicly available knowledge is rather to be addressed in sociological terms, which I will get back to in section 4.2.1.

Other historically relevant undertakings that consider science as unified are Cybernetics[96] (Norbert Wiener - 1948), General Systems Theory (Ludwig von Bertalanffy - 1951) and Dynamical Systems Theory (DST). Although I will not go into detail in presenting these views, it is important

[96] The term is coined by Wiener in 1948 and applied to systems that could steer themselves. The serial of conferences organised by Wiener was referred to as 'Cybernetics'. See (Bechtel and Hamilton 2007).

to note their common feature and influence, in particular on the fields such as science and society studies (STS). Rosenblueth, Wiener, and Bigelow in 1943 introduced the idea of *feedback*, which considers biological and artificial systems to be goal-directed and mutually interconnected. Such a feedback organization (Cybernetics) is seen as a *unifying* force that connects biological and social systems. In STS terms, science and technology, while being focused on biological systems, are also an inseparable part of the social systems. So, biological and social systems are interconnected in a dynamic way, thus continuously *reshaping* each other through the feedback loops (Jasanoff 2004; Jasanoff et al. 2001). Regarding computational ontologies, the interaction between science and society in shaping directions of knowledge collection, labelling, representation, and integration is recognised and addressed in a number of papers, e.g. (Leonelli 2009b; Leonelli and Ankeny 2012; Mai 2004)[97]. While discussing the epistemic groups involved in knowledge organisation, I have outlined just a few interdependences among the scientific communities that are integral part of society (Sojic and Kutz 2012). In section 4.2.1 I argue that social (collaborative) aspects of science make the best candidate for seeing science as unified.

The last to discuss, but nonetheless important, is the influence of the logical empiricism, which is revived in our debate mostly because of the role that logic, language, and empirical knowledge play in the new unifying endeavour. The scientist and philosophers focused around the Vienna[98] and Berlin[99] circles, in line with August Comte (19th C), strongly rejected metaphysics while accepting (positive) knowledge as grounded on observation and experimentation. At the beginning of the 20thC the influence of Comte's view and Ernst Mach's radical empiricism

[97] Note that these papers do not refer explicitly to the Cybernetics and General Systems Theory. The analogy is rather a matter of my observation aimed to attract attention to the relation between contemporary approaches that analyse collaborative efforts of the ontological community and its historical predecessors.

[98] While the leading people were Hans Hahn, Moritz Schlick, Rudolf Carnap, and Otto Neurath, the number of the participants in the movement, in a direct or indirect way, was far more numerous. (e.g. Kurt Gödel, A.J. Ayer, Herbert Feigl, Philipp Frank, Hans Hahn, Carl Hempel, Karl Menger, Richard von Mises, Ernest Nagel, Karl Popper, W.V. Quine, Frank Ramsay, Hans Reichenbach, Alfred Tarski, Friedrich Waismann, and Ludwig Wittgenstein.)

[99] Led by Hans Reichenbach, the circle included Kurt Grelling, Walter Dubislav, Kurt Lewin, Richard von Mises, and Paul Oppenheim, and many others.

expanded into various directions and interpretations of empiricism. The history of philosophy marks this period as a struggle to found philosophy as a discipline that can give a general account of knowledge, which can go in hand with other sciences that by that time already diverged into specialised disciplines.

On the way to found philosophy on stable grounds, logical analysis acquired a prominent role as the means to capture general scientific claims that go beyond particular empirical observations. Apparently, the rejection of metaphysics led towards the acceptance of modern logic (Frege, Peano, Russell, Whitehead) rather than Aristotelian 'underlying' categories. Logical empiricism as a movement was heterogeneous in the ways in which the role of logic, language, and empirical experience were perceived (see e.g. (Creath 2011; Uebel 2011; Bechtel and Hamilton 2007)). Even so, the joint efforts to provide a common methodological ground for all sciences resulted in *The International Encyclopedia of Unified Science*, edited by Otto Neurath, Rudolf Carnap, and Charles Morris. The mission behind The Encyclopedia was to provide a methodological support to sciences in order to achieve the cross-disciplinary connections and exchange of knowledge, so that advances in one field can stimulate advances in other fields. Contrary to the ambitious goal of The Encyclopedia as a long term project, only two volumes were published.

Neurath's vision of science as a pragmatic unity of inter-connected sciences, which in a dynamic way interact, mutually enriching each other by exchanging knowledge, seems to be relevant nowadays more than ever before. The information technologies, in general, and the ontologies, in particular, foster this process of knowledge exchange. However, the issue of 'unity' and the kind of integration that results from this process still need to be examined.

The quest for the methodological unity undertaken by the logical empiricists, *inter alia*, produced a controversial view according to which sciences can be unified by means of theoretical reduction. *The reductive unity* position defends the view that the diverse scientific theories are either in principle or in practice reducible to a common origin, which is represented in common terms, mostly those of physics. Although there are different approaches to how such a reductive unity can be achieved[100], a general idea that is present in its every version is the idea of getting to the common origin that unifies the sciences on that reduced level.

Those who hold that theory reduction is not possible often argued against establishing the appropriate bridge principles among the theories. For instance, Paul Feyerabend (Feyerabend 1970) claimed that terms in different theories, even if they have the same lexical form, may actually have different meanings which depend on different theoretical contexts that they are part of. Therefore, terms that are seemingly the same may in fact be incommensurable. Since it seems impossible to construct bridge principles that would adequately relate the term ‘temperature’ as it is used in classical thermodynamics and in statistical thermodynamics, these two theories are to be considered as incommensurable. Following the same line of reasoning, Thomas Kuhn (Kuhn 1970) argued that the meaning of the term ‘mass’ as used in the Newtonian mechanics is incommensurable with the meaning of the same term as used in Einsteinian mechanics.

Apparently, the debate about the appropriateness of mapping among the ‘synonymous’ terms coming from different domains revives the view of the disunity through the examples that the ontologists deal with in a daily practice (Tordai et al. 2010). In Section 4.2.2 I argue that, even if not incommensurable, the apparently synonymous terms, when explicated formally, rather seem ambiguous.

In line with Feyerabend and Kuhn, David Hull (Hull 1972) also challenged a reductive view in the case of Mendelian and molecular accounts of genetics. While molecular genetics characterizes traits in molecular terms, the Mendelian approach relates the concept of gene as related to phenotypic traits, such as ‘tall’. Thus, ‘Mendelian genes’, when described in molecular

[100] For instance, some of the approaches proposed the bridging role principles that would connect the theories, e.g. Nagel (1961, 1974) proposed rules of correspondence as the links among the theoretical terms; some others, like Robert Causey (1977), employed an ontologically committed interpretation where the theories at the lower level describe the operation of parts of the structured wholes (see Bechtel and Hamilton (2007)).

terms, result in specification of a number of molecular mechanisms which could produce the same phenotypic trait (i.e. multiple realizability). On the other hand, Hull argued, such a reductive approach is not satisfactory also because the same molecular mechanism can produce different phenotypic effects (Ibid.) (For the discussion see (Lewontin 2004; Rheinberger and Müller-Wille 2004)).

Since it has also been argued that the set of molecular mechanisms which result in certain traits might be provided by empirical investigation (Rosenberg 1994), the various descriptions of molecular pathways could be in principle related to the traits described in terms of Mendelian genetics (Ibid.). However, Rosenberg claimed, even if that might be possible in principle, providing such a detailed set of disjunctions that would describe all possible directions of genotype-phenotype relation would not have contributed the comprehensibility, but rather a confusion with the overwhelming information.

In practice, databases and the information technologies nowadays provide a platform to organise and store a huge amount of information. Thus, the information technologies, which enable practically tractable knowledge about billions of 'stored' relations, seem to endorse the reductionists' arguments, at least in principle. For instance, a number of model organism databases include descriptions of the phenotypes that are associated with specific genotypes (Hoehndorf, Schofield, and Gkoutos 2011). Regarding biomedical ontologies, the explicit representation of genotype-phenotype relations has been particularly investigated in its relevance to the understanding of genetic basis of the human diseases (Ibid.).

Nonetheless, a problem that seems to remain unresolved by capturing these numerous genotype-phenotype relations into the organised databases that store information in a tractable way concerns the focus of explanation. As Rosenberg (1994) argued, a reduction of the phenotypic characterisation, as described in terms of functional (Mendelian) traits, to the description of the structural traits (i.e. molecular genotypes), does not contribute a comprehensible explanation in the case of natural selection. Since selection might take place on various levels, particular disciplines will always have specific kinds of questions, being focused either at the Mendelian or at the molecular traits. In order to provide the most appropriate answer, the disciplinary interest directs the interpretation of 'genes' in a domain specific way, thus supporting the disunity rather than the unity of science.

Rosenberg's argument involves a particularly influential version of the reductive unity, which focuses on its methodological aspects related to explanation. According to *explanatory reductionism*, the unity of science is achievable through a reduction of the diverse scientific explanations to the explanations provided in terms of some basic principles (e.g. microphysical laws (Oppenheim 1991; Hempel and Oppenheim 1948; Nagel 1984)). The proponents of the explanatory unity were aiming to provide the principles such as *connectability* and *derivability*, which would hold between the derivable and derived theories. In order to avoid the acknowledged problems with the covering low (D-N)[101] model of explanation (e.g. (Hempel and Oppenheim 1948)), Phillip Kitcher's early[102] account of the explanatory unity outlines the explanatory system as a whole set of explanations that unifies science (Kitcher 1981). Accordingly, Kitcher considers science as a unity, because no single scientific explanation can be derived independently of other explanations that are part of the explanatory system, which as such gets derived from the best available and well systematised knowledge in science at a given time.

The aim to represent scientific knowledge as composed of well systematised 'explanatory'[103] patterns seems to be a quite topical strategy in the practice of biomedical ontologists. Ontology engineers, in collaboration with domain experts, use the best available

[101] Kitcher (1981) points out that the D-N does not provide an account of 1) how scientific explanation advances the understanding, 2) how to measure and compare the explanatory power of theories, 3) how to justify the basic distinction between laws and the accidental generalisations, which is in itself problematic.
[102] Kitcher in his later works adoptees a weaker version of unity, i.e. 'modest unificationism' (Kitcher 1999).
[103] Note that the explanatory patterns in formal ontology denote the represented axiomatic relations and inferences. However, unlike the axioms of logical empiricists, the axioms in an ontology are substantially enriched with the semantic values in such a way that the validity of the inferential process between the premises and their entailments need to conform not only to the formal logical rules of reasoning but also to the empirical content of the best available scientific knowledge at a time. The analysis of the relation between logical empiricism and ontology engineering surely deserves much more attention, which exceeds the scope of my discussion here.

scientific knowledge in order to derive the patterns which can cover a number of phenomena of a similar kind. Alike Kitcher's view of the unity as 'best systematisation', which minimises the number of argument patterns and maximises the set of conclusions, the ontological patterns are designed to serve such purpose (Alizadeh et al. 2001; Hoehndorf et al. 2010; Hoehndorf, Ngonga Ngomo, and Kelso 2010). The ontological patterns aim at capturing formally conceptualisation of a domain through the relational schemes (Hoehndorf et al. 2010), so that the resulting model can be applicable to various situations (Guarino, Oberle, and Staab 2009). Even so, the pattern modelling and the ontological axiomatisation rather seem to fit a pluralistic than a unifying view of science (see e.g. (Kutz, Mossakowski, and Lücke 2010; Lange, Mossakowski, and Kutz 2012; Sojic and Kutz 2012)). Since various domains are often conceptualised and formalised in different ways, the pattern design is heterogeneous as well (Ibid). Thus, instead of unifying the patterns, the ontologists are developing methods that can make the ontological patterns and models interoperable and re-usable across the domains (Kutz, Mossakowski, and Lücke 2010; Hoehndorf, Ngonga Ngomo, and Kelso 2010; Hoehndorf et al. 2010; Lange, Mossakowski, and Kutz 2012).

The idea of the reductive unity has been attacked many times with different arguments (Dupré 1983; 1993; Bechtel 2007; Cartwright 1983). In general terms, the opponents of reductionism agreed that 'What is left is a kind of pluralism of scientific language, practice, and subject matter. These, Suppes argues, are diverging rather than converging, and this is as it should be' (Bechtel and Hamilton 2007). While Dupré severely discredited classificatory unity (Dupré 1981, 1983, 1993, 2002, 2004), Cartwright demonstrated that scientific practice employs the abstractions in a way that cannot by any means fit into the D-N model, which has been proposed by the empirical positivists (Cartwright 1983; Dupré 1981, 1993, 2002).

Since classification plays a central role in the ontological domain, Dupré's arguments that favour pluralism are well placed in the debate about the classification and representation by means of biomedical ontologies (see 3.6, and e.g. (Mai 2004; Sojic and Kutz 2012; Leonelli 2012b; Lord and Stevens 2010)). Likewise, Cartwright's view on scientific models (i.e. human engineering products; theoretical descriptions which are not capturing the messy world in an absolute manner), the accounts proposed by ontology engineers consider ontologies as engineering products designed to serve a particular purpose, while grappling with real world problems (Mai 2004; Lord and Stevens 2010). Accordingly, the ontology models, alike any other model, might have a limited applicability outside of the scope the model has been designed for (see the domain knowledge problems, Section 4.1).

On the other hand, the integration of the 'patchwork of theories and models' (Cartwright 1983) seems to play an important task in the ontological community. In order to make various ontology models interoperable, there have been developed various approaches that use the ontology modularisation and the ontology matching methods (Shvaiko and Euzenat 2013; de Bono et al. 2011; Sheth et al. 2008; Lange, Mossakowski, and Kutz 2012; Normann and Kutz 2010). Section 4.2.2 addresses the issue of ontology mapping, while examining the kind of 'unity' that is achieved through the emergence of the interlinked knowledgebases.

4.2.1. Examining the collaborative unity

The shortcomings of those approaches that defend reductive unity have motivated several other approaches[104], which attempt to provide an account of how it is possible that, despite the disciplinary divisions, separate scientific disciplines still seem connected, functioning together, mutually exchanging knowledge claims, data, methods, and evidence. In this section I focus on one particular approach proposed by Angela Potochnik. By examining Potochnik's arguments, developed in (Potochnik 2010), I will draw a number of conclusions that may clarify certain unifying aspects of biomedical ontologies, which emerge as a novel kind of unity in

[104] The alternatives to the disunity nowadays are mostly framed in terms of integration in science. See e.g. Bechtel and Hamilton (2007); Sandra Mitchell (2003) proposes a version of integrative pluralism; Darden and Maull (1977) define interfield connections; Kitcher (2001) argues for a well organised science as balanced within the society; See also Galison and Stump (1996).

science. This critical approach to Potochnik's paper serves as a heuristic tool in distinguishing the promising from the misleading paths of thinking about ontologies and unity of science.

Potochnik presents the idea of *Coordinated unity* as a solution that, according to the author, provides an argument for the unity of science, while avoiding the problems that reductive unity fails to face (Potochnik 2010). The coordinated unity, supposedly, provides the most satisfactory solution to the debate, because it depicts scientific practice more accurately than the approaches of the disunity proponents.

I show that Potochnik's arguments fail to dismiss the arguments in favour of the disunity of science for several reasons. First, Potochnik misinterprets some of the arguments she attacks. Then, she gets into a number of fallacies by mixing epistemological and Ontological levels of discussion. Moreover, the paper does not provide the conclusive arguments for unity either. On the other hand, I argue in favour of Potochnik's view that particular scientific disciplines indeed function in co-ordination. Yet, I stress that Potochnik's argument for the unity which refers to the scientific practices has a limited scope. The validity of her argument from scientific practices applies only to the practices that, as such, are a particular sociological phenomenon.

In other words, the question of unity in *science* requires specification of what one exactly means by science. Certain attacks to the unity position criticise precisely the lack of demarcation criteria, which would distinguish science from non-science (Dupré 1993). The evaluation of Potochnik's argument requires a clear distinction between two interpretations of the term 'science': if *science* is understood in sociological terms as collaboration among diverse research groups, there is an acceptable way in which science can be considered as unified[105]. However, if *science* is understood in terms of scientific 'objects' (e.g. theoretical products, problems, theories, explanations, and representations), the idea of unity stays undermined. At least, it seems that Potochnik fails to show that science is unified in the latter sense.

[105] Note that Dupré in (1995), while arguing against the unity, admits that science might be considered as the well functioning collaboration in terms of social activity. So, Potochnik misinterprets Dupré's view.

In her paper, Potochnik reframes the Neurathian view of science understood as a pragmatic unity, which is reflected through scientific practice. The paper attempts to defend the claim that science is unified and not disunified because the diverse scientific disciplines do collaborate by sharing the same evidence. The reason for the coordinated behaviour of science, according to the author, is the causal complexity of the world, which is so complex that it needs to be addressed by diverse disciplines.

At this point, Potochnik slips into a fallacy, wherein she remains throughout the paper. She introduces causal complexity of the world as a fact of the world. In philosophy, such an approach is considered as a matter of an Ontological statement. Thereby, the complexity presented in this way is to be understood as an Ontological state of the world. By noting this, Potochnik's moves to the level of criticising causal, representational, and classificatory disunity, without providing any account of how these numerous causes, which act in all the directions out-there in the world, actually succeed in keeping sciences together. Among others, Nancy Cartwright (Cartwright 1983; Dupré 1993; Cartwright 1999) has shown that scientific capturing of the causes can go any direction. What we know about the causes in the world is mediated by scientific representations, justifications, and explanations. As Jon Williamson put it,

> [...] certain causal beliefs are appropriate or rational on the basis of observed evidence; [...] Causality, then, is a feature of our epistemic representation of the world, rather than of the world itself. (Williamson 2006)

Moreover, irrespectively of the way in which we treat causal complexity, either in ontological or in epistemological terms, causal complexity can be a strong reason for arguing in favour of the disciplinary division rather than unity. Since a problem such as cancer cannot be grasped by only one research group and method, various sciences focus on different aspects of it, thus fostering a division of scientific labour. Accordingly, the divided research groups *collaborate* when they need some additional information, which can be useful for their own research[106] (Potochnik 2010).

Although related, collaboration and co-ordination are two distinct concepts. While a common problem drives scientists to *collaborate* in exchanging knowledge and data, the *co-ordination* takes place on the level of society. Such a co-ordination might be exercised on various scales of society, from its micro scales (e.g. a lab leader) to the macro scales such as national founding policy or the World Health Organisation programs. Potochnik's argument, however, seems to mistake collaboration for co-ordination. Thus, her examples in favour of the *coordinated* unity actually go in favour of a sort of collaborative unity, while failing to address the co-ordinated unity.

Surely, we shall also examine the effects that a *collaborative* unity can have on the debate of the unity of science. As the main tool for having such a unity Potochnik refers to *evidence* that by being *circulated* across the communities justifies the unity. Since the shared evidence plays the crucial role in the argument for unity, I will call this unity the *evidential unity*. In spite of being intuitively appealing, the argument for the evidential unity is not pernicious for the proponents of the disunity. The reason is simple. Circulation of evidence does not unify science.

The circulation of evidence cannot be used as a universal criterion that unifies science. For, even if evidence can be re-used in some cases it does not mean that evidence is generally re-usable on all scales of scientific practice. Given that there are the exceptions showing that evidence is not re-usable, evidence cannot ground the unity of science[107]. At least, the evidential unity cannot be a universally valid criterion for claiming the unity. If the criterion is not universally applicable to all sciences, than such a unity is leastwise a partial unity. Certainly, a partial unity is open to the objections from the proponents of the disunity.

[106] Potochnik's example from evolutionary biology which presents scientific collaboration in acquiring information of their own interest thus illustrates the division of labour.

[107] Besides, by establishing evidence as a unifying criteria a lot of science will be excluded, e.g. mathematics. It is not clear how theoretical research in mathematics participates in sharing evidence with other sciences.

In the following sections I outline some problems for treating evidence as a unifying criterion in science.

Though evidence has a crucial place in Potochnik's argument, she does not specify how we should understand evidence in the first place. Instead of going into a controversial discussion about the concept of evidence, I accept the following minimalistic characterisation of evidence:

> In any event, the concept of evidence is inseparable from that of justification. When we talk of 'evidence' in an epistemological sense we are talking about justification: one thing is 'evidence' for another just in case the first tends to enhance the reasonableness or justification of the second. (Kim 1988)

Presumably, Potochnik would agree that we can consider scientific claims as a special kind of 'things' which a) must be supported by evidence and b) can be used as evidence.

However, while Potochnik claims that

> [...] evidential relationships do not respect field boundaries. (Potochnik 2010, p.3)

I argue that this claim has a limited scope and cannot cross certain field boundaries. Since Potochnik's claim is neither supported by evidence which can be valid irrespectively of the field, nor can be used as evidence irrespectively of the field, this claim cannot justify the claim that science is unified.

In order to support my own claim, I distinguish two basic kinds of 'things' that are commonly used as evidence and circulated in scientific community: *data* and *knowledge claims*.

I show next that 1) shared *data* used as evidence do not unify knowledge across the domains; 2) shared *knowledge claims* do not unify domains by means of evidence. Thus, if evidence is supposed to support the unity of sciences, that must be due to some other reason rather than by means of 'things' that are circulated as evidence.

1) **Shared data do not unify domains**

Namely, even if two research domains use the same data as evidence to support their claims, this does not imply that *the shared data* play the same evidential role in these domains, i.e. they are not the same evidence. The same data may play diverse evidential roles when they have been given different interpretations in different contexts. The unification by the evidential connection experiences serious problems in biological data interpretation, even within a single discipline. For instance, the same data can support opposite evidential claims even within the same group of scientists when they apply different methods to analyse data (Pusztai et al. 2006). James Mills' view on the ambiguity in data interpretation applies to evidence when those data are used for evidential purposes.

> "If you torture your data long enough, they will tell you [justify][108] whatever you want to hear" (Mills 1993)

So, the sharing of data as "naked evidence" that is free of the experimental context, which would associate the data with certain meaning, is not sufficient to support the unity on the grounds of the shared evidence.

Given that the data used as evidence are context sensitive, it becomes questionable what is shared and unified by data sharing. In other words, it is not obvious how mere data sharing contributes the evidential connections of the disciplines.

A proper positioning of the evidential vs. unifying role is also relevant for the understanding of the role that data sharing plays in the domain of biomedical ontologies. For instance, Sabina Leonelli has shown how data and claims about phenomena, when circulated by means of bio-ontologies, can have various degrees of locality and generality across scientific communities (Leonelli 2009a). Leonelli claims that these features, *inter alia*, also expand the evidential scope of data (Ibid.). Surely, the fact that the 'digitalised' data are circulated across

[108] I add '[justification]' to the original quote as it can apply to the cases when the 'tortured' data are used as justification and evidence for the related claims.

communities much faster than ‘non-digitalised’ data explains an accelerated dissemination of information acquired within a research field. However, the evidential scope[109] of the ‘digitalised’ data is not expanded in any other way than the one that is the result of the accelerated distribution of presented information. As in the case of ‘non-digitalised’ data, only the interpretation of the ‘digitalised’ data will determine which evidential role the data actually acquire when they are used. Thus, an expansion of the evidential scope that takes place by means of ontologies is not due to the emergence of some special evidential features that digitalised data have, but it is rather due to the technological advansments that contribute to the enchantment in the information accessibility. Accordingly, sharing data for evidential purposes makes science neither more nor less unified than it was before. The same data can be used in many different ways and the interpretation of data results in a huge variety of claims, some of which are also conflicting (Mills 1993; Pusztai et al. 2006; Anderson, Jatoi, and Sherman 2009). Thus, it is not the shared data that unifies science.

Even if the evidential role of data across the research contexts appears to be an unconvincing reason for unity, the success of science to circulate and re-cycle information might be the strongest reason for urging that there is some kind of unity. So, let us consider an alternative view on the unity achieved through shared evidence.

Suppose that the coordinated unity allows the enrichment in knowledge of one domain through sharing evidence produced in another domain - not by means of sharing “naked” data but rather the interpreted data. Presumably, the commonest way to represent and share interpretations is to make propositional claims, thus I focus on the claims that are used as evidence across the domains. While examining this alternative, I show that the “naked data problem” shifts to the problem with diversity of evidential contexts which provide the claims.

[109] I thank Matteo Mameli for bringing to my attention the problems with the evidential scope.

2) **Shared claims do not unify domains by means of shared evidence**

In the following argument against the view that evidence is the unifying means, I will use examples from clinical and biomedical knowledge domains. I first introduce a basic epistemological distinction of different kinds of knowledge. Next, I develop my argument that the "naked data problem" holds also in the case in which claims are used as evidence.

In general terms, epistemologists distinguish three types of knowledge: 1) Ability knowledge (knowing how), 2) Propositional knowledge (knowing that), 3) Interrogative knowledge (knowing why, what, where, when, which, whether, who) (Pritchard 2006).

Accordingly, if someone possesses the ability knowledge (For short, *A*) that means he is able to perform a certain action, e.g. talking, riding a bicycle or operating a patient. Propositional knowledge (For short, *P*) is knowledge that something is the case. So, propositional knowledge can be expressed in the form of a proposition. Interrogative knowledge (For short, *I*) is ascribed to someone who is able to answer the *wh- questions*, i.e. why, what, where, when, which, whether, who. On the other hand, *why-answers* constitute a special kind of 'because-[so and so]' propositions.

Suppose now that *P* about a phenomenon and *I* on the same phenomenon have been represented explicitly as the scientific claims, e.g. written on paper. We shall examine now if language as a common medium to represent *P* and *I*, supports the unity of knowledge about the phenomenon by using the claims as shared evidence. In particular, we shall see if such a connection indeed constitutes the evidential unity of the disciplines. Namely, I will test the idea of the unity in science by an exercise that *explicates propositional knowledge* (*P*), by expressing it through its *interrogative form*, i.e. because-propositions (*I*). The reason for doing this is the following.

The claims in question are scientific claims. A claim is a scientific claim only if it is supported by evidence that can convince acceptance of the claim. Thus, for any scientific claim it should be possible to provide another claim that will support its justification, thereby communicating the reasons *why* we should accept it. For the sake of my argument I am not concerned with possible objections of the radical skeptics. I assume that any material and observational evidence acquired through scientific research procedures is viable evidence, and *because-propositions* are *relevant answers to why-questions*. Thus, the claims in question are the same kind of claims that we see in scientific publications.

So, a scientific claim that something is the case (*P*), in its explicit (*I*) form will represent a set of *because* claims. In this way, knowledge stated in a not-yet-explicated propositional form (*P*) will be represented in an explicit form (*I*). Doing so, the justification of the initial propositional claim will be represented explicitly. Such an explicit representation of justification will be extremely relevant in the case of clinical and biomedical knowledge, which demands the evidence based medicine (EBM).

> […] EBM is not designed to be a comprehensive account of medical knowledge, but only an account of that part of medical knowledge which is propositional. […] there is the apparent possibility of testing propositional claims one at a time […]. (Ashcroft 2004)

In our case, the propositions explicated through the (*I*) form serve to test the (*P*) form propositions. However, as we will see, the attempt to explicate propositional clinical and biomedical knowledge in its (*I*) form, will result in several problems due to contextual sensitivity of the domain knowledge. In Chapter III I have discussed the distinctions between clinical and biomedical knowledge. The domains were presented as having different focus in the problems they address, the terminology they use, the epistemic and practical aims, e.g. explanatory, predictive, curative etc. Here I rather use the drawn distinctions in order to examine responsiveness of the two domains to an attempt of their unification. I present two possible courses of the attempt to unify the clinical and biomedical domains by means of evidence. I phrase my exercise in terms of a dilemma which questions possibility of unified knowledge.

The dilemma about the possibility of unified knowledge

Assume that Potochnik is right in claiming that science is unified and evidence for that are, just as she says, the evidential relations which circulate freely across the scientific fields. Since science is unified and not disunified, the 'evidential connections do not respect disciplinary boundaries' (Potochnik 2010). We have already excluded the possibility that shared data can play this role of evidence by means of which science gets unified. Consider now the other option, which we left unexamined. It is not data, but the claims as to what is used to circulate as evidence across the domains to make science unified.

Suppose also that there is such *a propositional claim (a)* which exemplifies the case of unity between clinical and biomedical knowledge exactly because it can be used as evidence interchangeably between the domains. So, the initial assumptions are:

- (a) exemplifies the unity of clinical and biomedical domain by evidential relations.
- (a) exemplifies the unity of science.
- In the unified science evidential relations do not respect disciplinary boundaries.
- Since science is unified, (*I*) forms of (a) obviously support the unity and not disunity.

Let us see now what happens with (a) when it comes to its natural inhabitant, i.e. clinical and biomedical context. Being a scientific claim, (a) can also be explicated and represented in its interrogative form (*I*). Given that science is unified, (*I*) of (a) should not threaten unity by any means. After all, (a) is just an explicit saying of the evidential content that supported acceptance of (a) in the first place. Note that (a) was initially specified as the propositional claim which exemplifies the unity between clinical and biomedical knowledge because it can be used as evidence interchangeably between the domains. Given that science is unified and the evidential relations do not respect the disciplinary boundaries, the interrogative form of (a) in clinical domain (for short *Ic*) and the interrogative form of (a) in biomedical domain (for short *Ib*) should confirm the unity, which is achieved by means of evidential relations.

Consider now the case of reasoning that can take place within the context of clinical and biomedical domains. In fact, Chapter III presented numerous examples of a similar kind.

So, we will substitute the abstract claim (a) with the claim that *stem cell marker*[110] *predicts aggressiveness of tumour*:

(a)⇔ 'The stem cell marker predicts aggressiveness of tumour.'

The interrogative form of this claim is provided by a clinician and by a biologist. Both of them agree that this is a scientific claim. However, as we will see, they do so for different reasons[111]. The best way to find out the reasons why they consider it a scientific claim is by asking them *why* questions. In other words, we will have two sets of (*I*) claims:

1) claims provided by the clinician (*Ic*)	**2) claims provided by the biologist (*Ib*)**
Ic1)'The correlation between marker positivity and clinical outcome is statistically significant.'	Ib1) 'Cancers with **stem cell origin** are aggressive and the marker efficiently targets them.'
Ic2)'Clinical studies have shown the marker efficacy.'	Ib2) 'Molecular pathways [...] targeted by the marker result in accelerated cancer proliferation and metastasis.'
Ic3)'Most of the patients with the positive stem cell marker have an aggressive form of tumour.'	Ib3) 'The marker targets **cancer stem cell** efficiently, thus predicting cancer behaviour.'
Ic4)'The marker is efficient tool in clinics for the patients' assessment.'	Ib4) 'The correlation between the marker positivity and clinical outcome is statistically significant.'
Ic5) 'The marker targets efficiently the cells which result in an aggressive form of tumour.'	Ib5) 'Clinical studies have shown marker efficacy.'

When we compare these two sets of claims most of them will be coherent and even complementary. However, when the clinician and the biologist exchange their lists with the written claims, it comes out that the lists contain either too much or too little in terms of evidential statement. What the biologist means by accepting (a) as the scientific claim has a different meaning in the context of the clinician's domain knowledge. The biologist also doubts

[110] At present, there is a lively debate over the issue of what is the best stem cell marker. One of the most promising candidates is CD24$^-$/CD44$^+$ status (Frank, Schatton, and Frank 2010; Al-Hajj et al. 2003). For the sake of my argument, I assume the agreement about the stem cell marker has been already achieved.

[111] For the distinctions between clinical and biomedical reasoning see the discussion in Chapter III. Here I illustrate it through the example, which I designed after discussing it with a few clinicians and biologists. Surely, a more extensive empirical study would make my distinction stronger, but even this way it serves its basic purpose.

that the clinician really knows *why* the marker is an efficient predictor, since he cannot answer by *which* molecular mechanisms the aggressiveness of the targeted tumours develops. On the other hand, the clinician will disagree about the biologist's claims (Ib1) and (Ib3), because he does not accept that there is sufficient evidence for the claim that the marker indeed targets some kind of stem cells. He even does not believe that there is such an entity as cancer stem cell. Since (*Ib*) also contains the claims which denote an non-existing entity, according to the clinician, the set (*Ib*) is not an acceptable set of claims.

The example shows that particular disciplines disagree in the context of reasoning about evidence. This is the case which a biomedical expert rejects (*Ic*) as evidence for (a). Similarly, the clinician rejects that (*Ib*) is evidence for (a), while accepting (a) nonetheless, he does so *not* by means of the consistently shared evidential relation. It follows that either **(a) is not justified**, so cannot serve as evidence that crosses domains and unifies science, or **(a)** accepted by the clinician **is not the same (a)** that the biomedical expert accepts. It also follows that either (i) (*I*) cannot be circulated as shared evidence, or (ii) the claim (a) having a different meaning, should be considered as an ambiguous claim in the two domains, which cannot be represented consistently when the domains are merged together. Hence, since either (*I*) or (a) cannot cross disciplinary boundaries by means of the shared evidential relation and unify the domains, **(a) exemplifies the disunity rather than the unity of science.** In simple terms, the interrogative forms of (a) in clinical and biomedical domains lead to a disagreement that (*I*) can be used as evidence across the domains of discourse. Thus, the claim does not unify the two domains by means of shared evidence.

The example illustrates a problem which emerges when we consider that a shared claim plays the role of the shared evidence which unifies the domains. Of course, the example is designed to serve its heuristic purpose. It shows that 1) sharing the claims do not unify science by means of shared evidence; 2) certain claims are shared among the disciplines even if the criteria for accepting them differ across the fields. It also illustrates that the clinical and biomedical communities support two different methodologies of justification.

The diverse treatment of evidence across various disciplinary contexts undermines the view that science is unified because the evidential relations do not respect disciplinary boundaries. I have shown that 1) the shared *data* used as evidence do not unify science 2) the shared *claims* do not unify domains by means of shared evidence. Thus, if science is unified that must be due to some other reason rather than by means of sharing 'things' that are circulated as evidence. Since both the data and claims are context sensitive, we should examine next what is shared and unified by the sharing of data and claims across the scientific domains.

In the light of this, I will show next what does connect the disciplines and in which terms science might be indeed considered as unified. In order to do so, I will respect the distinction of the two aforementioned aspects of science. Namely, I distinguish 1) science understood in terms of a co-ordinated social activity and 2) science understood in terms of scientific 'objects' such as representations, explanations etc., traditionally examined in philosophy of science. Obviously, the two aspects of science are interdependent and these interdependences are a frequent topic in sociology, ethnology, and history of science. However, instead of following the historical and sociological approaches that examine these interdependences, I find the distinction between (1) and (2) a helpful analytic tool in positioning the debate about the unity in science.

Actually, the co-ordination of science, which takes place on the various levels of society, enables a fruitful collaboration of scientific communities. While in some cases collaboration emerges from the practical needs such as problem[112] solving (Potochnik 2010), the collaboration is possible only because of the standards which are established in the community. A high array of standards in the form of technological platforms, research methods, terminology, and the criteria for treating evidence (Keating and Cambrosio 2003), are the tools that circulate across the scientific communities, thus enabling collaboration among the scientific disciplines.

Another important concept that enables and explains collaboration in science is the concept of trust (Rousseau et al. 1998; Hardwig 1991). Without trust it would be hard to imagine

[112] Potochnik (2010) gives an example of evolutionary biologists who use the results of research from molecular biology.

any kind of collaboration, in particular in science that diverges into multiple specialised research domains.

> Modern knowers cannot be independent and self-reliant, not even in their own fields of specialization. In most disciplines, those who do not trust cannot know; those who do not trust cannot have the best evidence for their beliefs. In an important sense, then, trust is often epistemologically even more basic than empirical data or logical arguments: the data and the argument are available only through trust. If the metaphor of foundation is still useful, the trustworthiness of members of epistemic communities is the ultimate foundation for much of our knowledge. (Hardwig 1991)

The complex role that trust plays in science and society has been addressed through numerous multidisciplinary perspectives that examine the ways in which trust enables communication and collaboration among the fields (Rousseau et al. 1998). From the perspective of trust that holds among the research communities, it can be explained how evidence actually can be shared and circulated, despite the disciplinary divisions. The scientific claims do not circulate across the fields because the science is unified in sharing the 'things' that are circulated as evidence. The claims can circulate and get shared across the scientific domains exactly because science is divided into the communities which mutually *share trust* in each other's independency and peculiar ways in addressing the problems, while using particular research methods and procedures.

In order to examine certain aspects of unity in the latter sense, e.g. representations, in the following section I introduce the idea of representational interoperability that holds across the scientific domains. In particular, I argue that ontologies emerge as a novel kind of 'unity' by supporting the co-ordinated and explicit representation of the domain conceptualisations, thus resulting in *the extended-meaning* 'unity'. I will show how ontology enables knowledge exchange and enrichment across various contexts, while preserving disciplinary diversities and specificities. Eventually, various scientific domains can be considered as constituents of a well organised socio-technological structure, which makes domain knowledge consistently interconnected rather than unified.

4.2.2. The extended-meaning ‘unity’

I have addressed certain aspects of the contrastive forces that act towards scientific specialisation, on the one hand, and towards knowledge integration, on the other. In particular, I have acknowledged the co-ordinated behaviour of science as an enabling condition for the collaboration and exchange of knowledge across the divided scientific communities (Section 4.2.1). However, science is more than a social interaction that relies on trust among the agents involved in the cross-domain communication. There are also ‘objects’[113] that are exchanged in scientific communication thus informing particular sciences about their research topic.

In this section I discuss a special kind of ‘objects’ that are exchanged in scientific communication. I focus on *terms* that serve as interoperable *labels* in the implemented ontology models[114]. The validation of terms, which are associated with the same meaning, as synonymous, conduces the linking of various scientific domains (Shvaiko and Euzenat 2013; de Bono et al. 2011). In order to understand how biomedical ontologies, while using terms, contribute to the integration of the divided scientific disciplines, I will examine the roles that definition, synonymy and polysemy play in the process of knowledge integration.

Apparently, scientific communities rely on established and well specified terminologies. In this way, the transmission of knowledge within and across communities is already established through the circulation of common protocols, experimental documentation and publications.

[113] Sections 1.2.3. and 1.2.4 examine how *terms* play the role of ‘objects’ in the ontology representations, so that particular conceptualisation, meaning ,and reasoning acquires an explicit form, e.g. labelling terms to the reference classes explicates relational statements etc..

[114] Section 1.2.3. discusses how terms are used in labelling of the classes in the ontology vs. information models. While ontology models do not need to have assigned specific values for the individual members of the classes, the ontologies implementation provides the value assignment to the classes through the mapping of the information models onto the datasets.

However, a closer analysis of scientific practice, whereby the terms are used to label classes and represent reasoning about the domain problems (Chapter III), demonstrates that it is neither feasible nor advisable to unify scientific terminology in a rigid way. Having different domain interests, different disciplines use terms in a specific way as to fit best particular needs.

The resulting terminological discordance becomes a central problem when the terms need to face an explicit formal representation. On the one hand, a common reference terminology, which aims at covering a huge array of domain interests, often results in very general definitions that are sometimes also ambiguous[115]. On the other hand, formal representation of the terms' and their meaning attempts to capture accurately and consistently these varieties. Thus, the debate on the unity and disunity in science here emerges as a struggle between 1) a unifying terminology that can be used independently of the domain and 2) a specific domain terminology which is useful and informative in representing specific problems.

The acknowledged resistance of terms to capture the variety of meanings into a single specification weakens the viability of a rigid synonymy in practice (Section 3.4, 3.7.3). Seemingly synonymous terms, when employed in context, are used to denote various kinds of things and to communicate different cognitive content (see 'HER2' example, Section 4.2.1). Therefore, I examine the role of synonymous terms in the integration of biomedical knowledge into knowledge bases (KBs). At the same time I interpret KBs as an emerging battlefield for testing opposite views on the unity of science. I will argue that the implementation of ontologies exceeds the mere linkage of domains, not only because the terms are synonymous on a general level (lexical definition), but rather because the applied ontologies also allow the extension of the terms' meaning across the representational contexts. Thus, ontologies provide a certain kind of representational integration.

In the reminder of the chapter I demonstrate how ontologies enable balancing between generality and specificity of the scientific knowledge captured by the terms in use. I first discuss

[115] The ambiguity of 'HER2' in two different contexts is an example of ambiguity. The examples that follow in the text additionally illustrate this point.

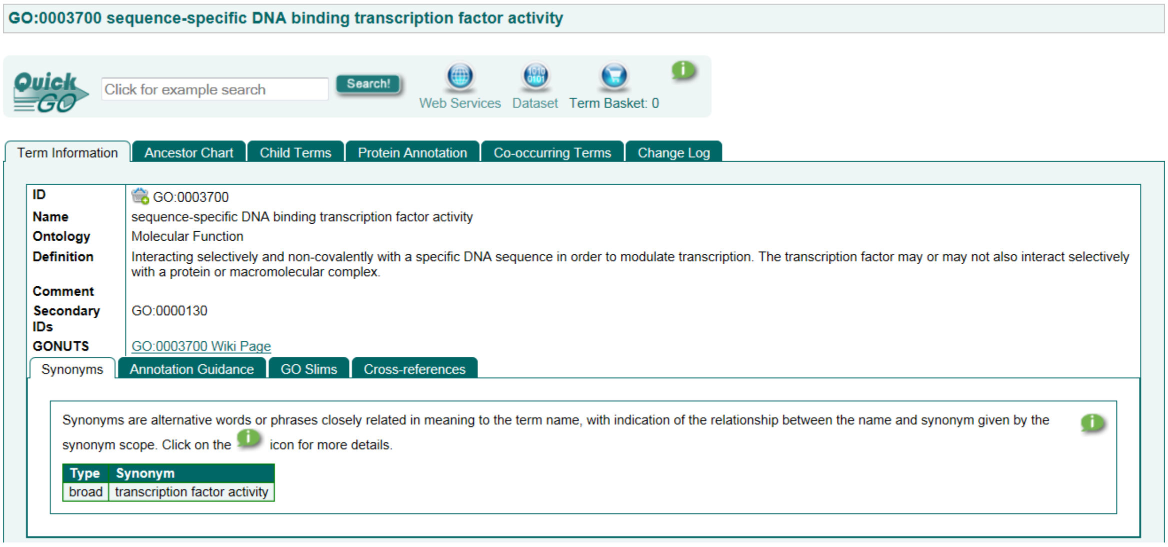
how the context of communication influences transmission of knowledge by means of language. I then examine the examples of the terms in use, showing that ontologies (using terms as labels) enable 1) the inter-domain connections, 2) enrichment in the term's meaning. Thus, I will argue that ontologies do provide a novel kind of representational integration.

In several accounts that aim at explaining the conventional and social character of language, language use has been characterised as a game where the successful moves have been made if the same meaning is assigned to an expression 'a' by the expression utterer and by the hearer (Gärdenfors 2005; Hardwig 1991; Wittgenstein et al. 2009; Hurford 2007). So, the cases such as Humpty Dumpty's arbitrary assignment of the meaning is acknowledged as a counter-example for a successful game player (Wittgenstein et al. 2009). Namely, a (knowledge[116]) message is transmitted only if the two-way process is accomplished (see also section 2.4). First, the term is getting assigned a meaning within a context specified by the knowledge sender (input). Second, the receiver (output) interprets the message accepting the context and the meaning specified by the sender. The informative content of a term is in that way transmitted when both directions (input and output) of the transmission fit together. That is to say knowledge transmission is achieved when both the knowledge sender and the receiver could agree about the meaning. Obviously, such an ideal case does not exist in practice.

Scientific language avoids many ambiguities in representing knowledge through the explicit definitions of terms (Neurath 1983). However, scientific terms are frequently used in various ways in different disciplines (Dupré 2004). As Neurath notices (See also (Hull 1988)), ambiguity is an important feature of language and, I will argue, essential for an enrichment of scientific knowledge.

[116] The messages that transmit knowledge claims are especially complex and problematic as a number of conditions need to be satisfied, e.g. justificatory criteria and expectations of the agents involved might be in disagreement. For the sake of simplicity I assume an ideal case where the agreement has been achieved through social conventions.

The following examples demonstrate that such an ambiguous and flexible nature of language, which is co-ordinated on several levels (Kalra 2006; Leonelli 2009b; Perry et al. 2008; de Bono et al. 2011; Ekins et al. 2011), actually enables the ontologies to be interoperable and informative in representing knowledge. The interoperability is achieved, *inter alia*, through the process which I will characterise as a *co-ordinated de-contextualisation*[117] of the term's meaning.



GO:0003700 sequence-specific DNA binding transcription factor activity

ID	GO:0003700
Name	sequence-specific DNA binding transcription factor activity
Ontology	Molecular Function
Definition	Interacting selectively and non-covalently with a specific DNA sequence in order to modulate transcription. The transcription factor may or may not also interact selectively with a protein or macromolecular complex.
Comment	
Secondary IDs	GO:0000130
GONUTS	GO:0003700 Wiki Page

Synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope. Click on the i icon for more details.

Type	Synonym
broad	transcription factor activity

Figure 35 GO definition for GO:0003700

For example, consider a definition[118] of a molecular function provided by the Gene Ontology curators that is labelled with the term '*sequence-specific DNA binding transcription factor activity*', having assigned identifier GO:0003700.

DEF. '*sequence-specific DNA binding transcription factor activity*' (definiendum)

[117] Note that the idea of co-ordinated 'packaging' of data by removing the context in order to allow the data to travel across different research contexts has been addressed in (Leonelli 2010). I here examine how this 'packaging' contributes knowledge integration by means of representing the meaning of the labels explicitly.

[118] There is no a unique definition of what *definition* is. There are different kinds of *definition* such as a dictionary definition, an ostensive, stipulative, descriptive, explicative, or a definition can be given informally as a kind of remark. Exactly having that in mind I do not focus on one definition type since the definitions are often used in a flexible manner complementing each other. Add (Michael, Mejino Jr, and Rosse 2001)

> Interacting selectively and non-covalently with a specific DNA sequence in order to modulate transcription. The transcription factor may or may not also interact selectively with a protein or macromolecular complex. [source: GOC:curators, GOC:txnOH] [119] (definiens)

Although the definiens contains new information compared to the definiendum (i.e. the involvement of a non-covalent selective activity that modulates transcription), in a certain way it is a trivial definition. It is trivial in the same way in which it would be trivial to say: 'The raining activity is an activity that modulates the weather conditions so that it is either raining or not raining.' Namely, if a definition contains a disjunctive form, as many of the GO definitions do, 1) it leaves room for the numerous potential connotations, 2) it increases generality, and 3) it reduces informativeness[120] of the definition. However, ambiguity of this kind, exactly by leaving room for various interpretations, actually brings an advantage in practice.

The contextual dependencies, which describe biological processes, are uneasy to capture within a rigid definition. Apparently, transcription factor activity is not a process that occurs as a reaction between two isolated biological components. It has been shown that there are many other components and interactions that condition the activity of the transcription factors. On the other hand, by providing a precise contextual definition, the term would have significantly decreased its applicability across variety of contexts. So, through the process which Leonelli describes as 'packaging' of data and claims (Leonelli 2009a, 2010), the curators through achieving an agreement about general definitions (Shimoyama et al. 2009) actually strip off the contextual meanings of the defined terms in order to cover a variety of possible meanings that are associated with the same term. The presented GO definition is an example that shows how by balancing

[119] http://amigo.geneontology.org/cgi-bin/amigo/term-details.cgi?term=GO:0006915&session_id=1432amigo1291905253&

[120] According to Popper, the amount of empirical information conveyed by a theory or its empirical content increases with its degree of falsifiability (Popper 1959) . So, claims that have a disjunctive form (as the aformentioned one) are often unfalsifiable and uninformative. In computational linguistics it has been recognised that the term's informativeness correlates with its specificity and it is reciprocal to the term's generality. However, there have been proposed different methods for the measuring of informativeness, e.g. Kireyev 2009 .

between generality and specificity of description the database curators attempt to provide a useful characterization of a concept, which will be 1) as specific as useful and 2) as general as possible.

Consider next how terms get enriched with a specific meaning within a particular context. The following example should illustrate how, similarly to ordinary language, information captured by the definition of a scientific term will indeed be informative only when it gets connected with other terms, which specify its meaning in that particular context. In the case of our definiendum, the particular meaning of '*sequence-specific DNA binding transcription factor activity'* is specified within the context in which this term is *used* to describe the binding activity. For instance, an *in vitro* experiment has demonstrated that the ubiquitin specific protease, USP7 or HAUSP, regulate the sequence-specific DNA binding mediated by the core domain of p53 (Sarkari, Sheng, and Frappier 2010). The description of the experiment provides information about the precise experimental setting, thus allowing one to claim that 'p53 has sequence-specific DNA binding transcription factor activity' in that particular context. Sarkari et. al. have characterised 'p53' as a transcription regulator not because p53 satisfies the definition of '*sequence-specific DNA binding transcription factor activity'* in a trivial, but rather in a very particular way. The entire experimental setting that is described in the published paper (Sarkari, Sheng, and Frappier 2010), is at the same time an exemplar and an explication of the term '*sequence-specific DNA binding transcription factor activity'* within the context.

Accordingly, a restriction provided by the context which specifies the meaning, directs the interpretation of the term from the set of multiple potential meanings towards a particular meaning. Thus, when a term is interpreted in the context its meaning shifts from a general and less informative to a specific and more informative meaning. That is to say, the term is *enriched in informativity* when connected into the network of the related terms that are used to describe the experimental context.

On the other hand, though Sarkari et al. have demonstrated that it is justified to assign this molecular function to p53 due to its activation triggered by USP7 and HAUSP, their claim is to be considered as a context dependent claim, thus not sufficient evidence for a general claim that 'p53 has sequence-specific DNA binding transcription factor activity'. At this point, the role of curators in validation of generality of the claims comes into practice as a *co-ordinated* activity (Shimoyama et al. 2009). The curators validate the methods used in experiments, as well as mutual support of evidence coming from numerous experimental systems (Leonelli 2009a). Therefore, the curators contribute co-ordination of a dynamic process that drifts between contextualization and de-contextualisation in representing knowledge.

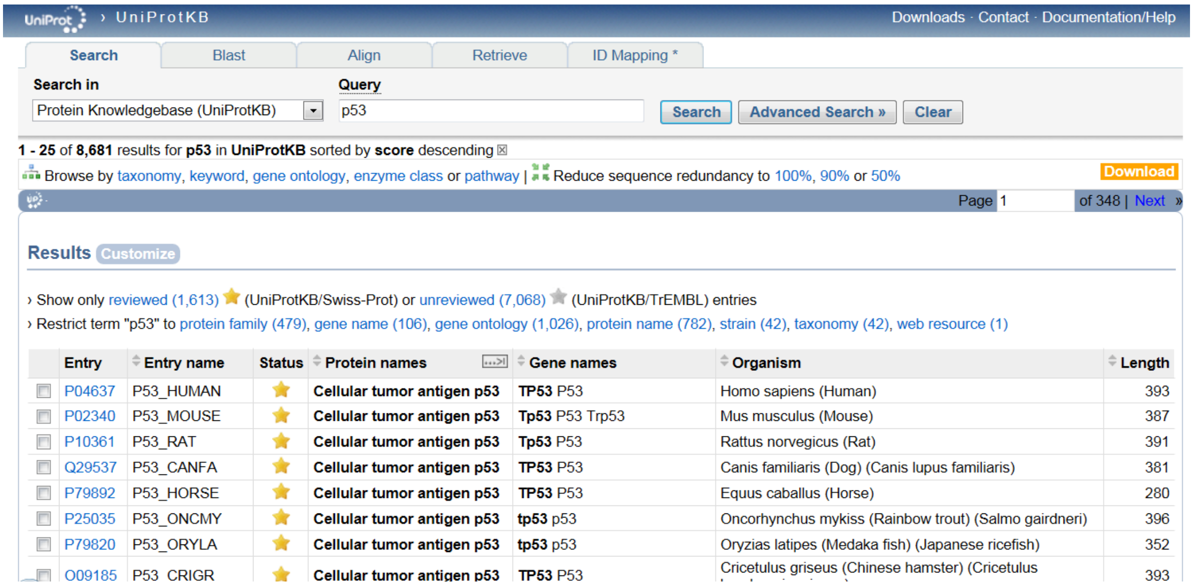
We should examine next the implications that this contextualization and de-contextualisation of meaning has for the unity of science in the case of knowledge bases.

The computer scientists working on knowledge representation and interoperability among information systems intensely investigate the methods for 'the integration of knowledge in meaningful chunks'. These meaningful chunks have been captured by the ontologies that support knowledge sharing and knowledge re-use (Chapter I). We have seen that the possibility of knowledge sharing and re-use highly depends on a well defined terminology (Section 1.1., 3.4.). So, a defined term such as 'p53' has assigned an identifier (e.g. P04637 for human p53 in UniProtKB[121]), which is mapped onto the synonymous terms in various contexts. Note that the process of alignment and mapping also takes place as a *co-ordinated* activity through numerous standards (Leonelli 2009b; Shvaiko and Euzenat 2013).

In the information models the same term 'p53' has been assigned a unique resource identifier (URI) that is mapped onto the synonymous terms, the ontology identifiers (e.g. UniProt: P04637, P02340, P10361), as well as onto other more general classes that are associated with the term (e.g. protein, cell component etc.). So, the same URI is associated with the specific set of descriptive terms that capture various contexts in which the protein p53 can be described. Likewise, 'p53' is purposely defined in a way that is general enough to be shared among the research contexts, still being as unambiguous as possible.

[121] http://www.uniprot.org/uniprot/P04637

Thus, a term can be shared across knowledge bases only if it has been de-contextualised through an agreed general definition. Moreover, instead of a specification in a form of nominal definition, ‘p53’ is getting mapped onto the string values (e.g. full name, symbol), synonyms terms, representations of the peptide sequence or primary sequence, synonymous terms characterised in the related biological species, etc. For instance, Figure 36 is a preview of UniProt Knowledgebase that shows various ID numbers of ‘p53’ that are mutually aligned across biological species.



Protein Knowledgebase (UniProtKB)

p53

1 - 25 of 8,681 results for p53 in UniProtKB sorted by score descending ☒

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50%

Page 1 of 348 | Next »

› Show only reviewed (1,613) ★ (UniProtKB/Swiss-Prot) or unreviewed (7,068) ★ (UniProtKB/TrEMBL) entries

› Restrict term "p53" to protein family (479), gene name (106), gene ontology (1,026), protein name (782), strain (42), taxonomy (42), web resource (1)

	Entry	Entry name	Status	Protein names	Gene names	Organism	Length
□	P04637	P53_HUMAN	★	**Cellular tumor antigen p53**	**TP53** P53	Homo sapiens (Human)	393
□	P02340	P53_MOUSE	★	**Cellular tumor antigen p53**	**Tp53** P53 Trp53	Mus musculus (Mouse)	387
□	P10361	P53_RAT	★	**Cellular tumor antigen p53**	**Tp53** P53	Rattus norvegicus (Rat)	391
□	Q29537	P53_CANFA	★	**Cellular tumor antigen p53**	**TP53** P53	Canis familiaris (Dog) (Canis lupus familiaris)	381
□	P79892	P53_HORSE	★	**Cellular tumor antigen p53**	**TP53** P53	Equus caballus (Horse)	280
□	P25035	P53_ONCMY	★	**Cellular tumor antigen p53**	**tp53** p53	Oncorhynchus mykiss (Rainbow trout) (Salmo gairdneri)	396
□	P79820	P53_ORYLA	★	**Cellular tumor antigen p53**	**tp53** p53	Oryzias latipes (Medaka fish) (Japanese ricefish)	352
□	O09185	P53_CRIGR	★	**Cellular tumor antigen p53**	**TP53** P53	Cricetulus griseus (Chinese hamster) (Cricetulus	393

Figure 36 UniProt IDs for 'p53' across the species

In this way, the ‘digitalised’ term ‘p53’ is able to play very diverse roles in the scientific claims. First, it can be assigned to diverse types of biological entities such as *gene* or *protein*. It can also be characterised as a cell component. The representational scope will include numerous forms and modifications that characterise the family of ‘p53’ proteins. Further, it is associated with numerous molecular functions, biological processes such as apoptosis, cell cycle arrest, and carcinogenesis.

Seemingly, the plurality in ‘p53’ characterisation across knowledgebases is not problematic since many of the mentioned meanings are not mutually exclusive. However, contradictory claims arise when some of these different meanings are used at the same time. It is obvious that ‘p53’ will not play all the aforementioned roles at the same time. The moment ‘p53’ is used to represent a claim, it will be used to denote exclusively a gene, a protein, or certain molecular functions, but not every possible function; it will be involved either in representation of a 'normal' or of an 'abnormal' function etc.. Having at the same time multiple contextual specifications makes a problem for the alignment across the representational domains, especially in the case of its formal characterisation. However, even in this case, instead of considering this promiscuous permissiveness of meanings as impermissible for the sciences, its acceptance might actually be the only way in which knowledge can be integrated. Using terms as labels enables the inter-domain connections, but the enrichment in the term's meaning is achieved when different contexts are brought together.

The distributed knowledge bases, such as UniProt[122], Swissprot[123], The Gene Ontology[124] and KEGG[125], and 329 registered biomedical ontologies[126], although focussed on different aspects of biological and medical knowledge, are interconnected due to the semantic generality and flexibility of the represented terms such as 'p53'. In that way the numerous knowledge bases are interlinked by a common label, while representing it in different knowledge contexts.

The mapping of terms and related ontologies is a coordinated process, which relies on collaboration of researches that work in different fields. Due to the collaborative efforts in the biomedical domain, the annotation and mappings of 5,437,452 terms, coming from 329 ontologies, are available at BioPortal, which is continuously getting updated with new ontologies, mappings, and annotations (http://bioportal.bioontology.org/mappings).

Regarding synonymy, it is important to note that interoperability across the domain representations relies on several kinds of synonymy among the terms. The aforementioned definition of 'p53' is synonymy in a *broad* sense, while for some purposes, e.g. a formal representation and the ontology alignment, an exact and narrow synonymy can suit better.

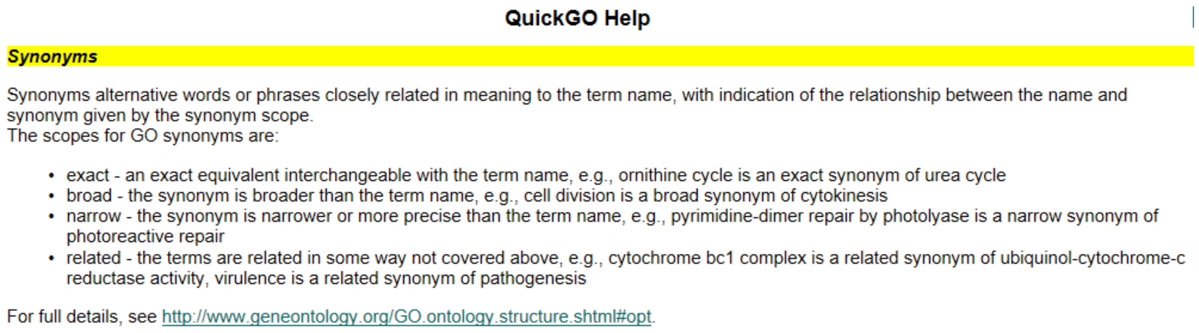
[122] http://www.uniprot.org/

[123] http://www.uniprot.org/help/uniprotkb

[124] http://www.geneontology.org/

[125] http://www.genome.jp/kegg/

[126] http://bioportal.bioontology.org/ontologies



QuickGO Help

Synonyms

Synonyms alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope.
The scopes for GO synonyms are:

- exact - an exact equivalent interchangeable with the term name, e.g., ornithine cycle is an exact synonym of urea cycle
- broad - the synonym is broader than the term name, e.g., cell division is a broad synonym of cytokinesis
- narrow - the synonym is narrower or more precise than the term name, e.g., pyrimidine-dimer repair by photolyase is a narrow synonym of photoreactive repair
- related - the terms are related in some way not covered above, e.g., cytochrome bc1 complex is a related synonym of ubiquinol-cytochrome-c reductase activity, virulence is a related synonym of pathogenesis

For full details, see http://www.geneontology.org/GO.ontology.structure.shtml#opt.

In particular, an exact synonymy significantly facilitates ontology matching and integration. Ontology matching uses standardised methods (Shvaiko and Euzenat 2013) to establish ‘correspondences between semantically related entities of ontologies’ (ibid.) in order to perform a particular tasks (e.g. ontology merging, query answering, or data translation). Since the matching is often performed on large ontologies and datasets, special algorithms have been developed for various purposes (for a review of methods see Shvaiko (2013)). The matching can be performed on the strings (terminological), structure (structural), data instances (extensional) or models (semantics) (Ibid.). Thus, an integration of knowledge that is represented in knowledgebases can take place on various levels.

Nevertheless, every attempt to merge knowledge domains is designed to fit a particular task, be this an integration of thesaurus (Rector 2003; Hartung et al. 2012; Milian et al. 2010) or the breast cancer related ontologies. Accordingly, the most suitable method for finding the relation of similarity, the source ontology to be merged together, and the relations to represent depend on the particular purpose that motivates the integration.

In Section 2.1.1. I have discussed several ontologies (e.g. FMA, PATO, GO, ChEBI etc.), each of which is designed for a particular purpose. Accordingly, the represented concepts within these ontologies are selected and related among each other in a way that best fits the aims of the particular ontology. However, when a particular task, such as the representation of knowledge about breast cancer has to be achieved, various modules from domain ontologies need to be combined (Kutz, Mossakowski, and Lücke 2010; Hoehndorf, Oellrich, and Rebholz-Schuhmann 2010; Kutz 2011). For instance, in the case of breast cancer ontology, The International

Classification of Disease (ICD)[127] needs to be combined with OBO foundry ontologies and other formal ontologies such as General Formal Ontology GFO (Herre 2010), GALEN (Rector and Nowlan 1994), and the Dolce ontology (Masolo et al. 2003). In addition, several OBO ontologies include the terms and relations that represent clinical and biomedical knowledge about breast cancer, e.g. the Human Disease Ontology (DO)[128] , The Foundational Model of Anatomy (FMA), The Phenotypic Quality Ontology (PATO)[129] , The Units of Measurement Ontology (UO), The Ontology for Biomedical Investigation (OBI)[130], etc. On the other hand, the way in which particular terms and modules from the existing ontologies are going to be selected depends on the modelling task, be that a representation of a breast cancer phenotype, or a clinical decision making. Thus, the goal oriented representations already co-exist together and support unity of science in terms of co-ordinated behaviour that fosters not unification but interoperability across the representational domains in order to take advantage of knowledge that is distributed crosswise.

As a final thought, let us consider one of the advantages that the use of ontology can bring into the representation of knowledge about breast cancer. For example, the classificatory terms, which are represented through different levels of granularity, can be now integrated on a formal level. Thus, 'HER2' with the assigned meaning of a tumour marker (clinical knowledge that HER2 positive patient has an aggressive tumour) can be integrated with the fine grained knowledge about biological processes (biomedical knowledge about HER2 expression represented through the signalling pathways).

The representation of knowledge about molecular processes in this way enriches clinical knowledge with a justification that also explains why the high expression of HER2 results in an aggressive tumour. For example, tumour aggressiveness is explained, *inter alia*, through the

[127] Respectively, http://ncit.nci.nih.gov/, http://www.ihtsdo.org/snomed-ct/, http://www.nlm.nih.gov/research/umls/, http://www.who.int/classifications/icd/en/

[128] http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology

[129] http://www.bioontology.org/wiki/index.php/PATO:Main_Page

[130] http://obi-ontology.org/

process of HER2 phosphorylation, which triggers the transduction cascades resulting in alterations of gene expression, transcription, translation, and protein stability (Nahta et al. 2006). These processes eventually affect cell growth, proliferation, migration, adhesion, and survival resulting in cancer. Of course, the molecular representation of carcinogenesis can also have diverse levels of granularity, capturing the main pathways as divided into the sub pathways. A detailed representation of the HER2 related carcinogenic processes can focus on Akt pathway that intermediates between HER2 receptor activity and HER2 gene expression (Nahta et al. 2006). If HER2 receptor has reached a certain threshold in terms of HER2 transmembrane presence, its quantitative difference will make a qualitative change among others through the Akt activation. That is because Akt can enhance a carcinogenic loop effect by a suppression of the tumour suppressor p27 as well as by an activation of the oncogene Nf-kB (Ibid.). For, both possible directions will result in the cell proliferation and the cell growth, including an additional HER2 amplification that will enhance the carcinogenic effects even more.

Modelling formally an ontology that represents these detailed molecular processes as incorporated into clinical knowledge, so that clinical and molecular knowledge are practically used in a synergy for the diagnosis assessment, clinical decisions, and the drug design, is no longer a matter of remote future but rather a matter of the day.

Conclusion

The advancements in knowledge about breast cancer achieved through the employment of novel research technologies, methods of data analysis, as well as cancer-related social concerns, have resulted in an intensification of the research, an accumulation of data, and the production of numerous models and classificatory systems. At the same time, the expansion of information technologies has provided the biomedical domain with an infrastructure for organising and compiling the growing body of knowledge and data relevant to the understanding and dealing with the rather dispersed pieces of knowledge.

I have addressed certain problems, limits, and potentials of the integration of knowledge across heterogeneous domains, each of which captures knowledge about breast cancer from a specific perspective. In particular, I focused on the novelty that information technologies bring to the biomedical domain.

Reasoning *about* health records and reasoning *with* health records can be considered two closely related, but separate processes. Reasoning about health records includes a broad set of activities such as the cognitive processes of the individuals or groups who think about the records while designing, using, or evaluating them (Section 1.2., 2.2.1, 3.7.). The process of the record design includes reasoning about a domain and about information that needs to be structured and organised within repositories. As in the case of non-electronic records, reasoning about electronic ones is grounded on reasoning about information which might not yet been recorded into an electronic health system.

Obviously, from this perspective, reasoning about health information is not a novelty. Health-related data were recorded and organised many centuries before electronic systems were invented. For instance, a historically important document that presents systematically collected data on mortality is 'Early Bills of Mortality' (Walford 1878). However, the development of information technologies has brought at least two significant changes to the practice of biomedical reasoning. The first change relates to established standards such as *how data should be structured* in order to fulfil the requirements of an electronic health system (Kalra 2006). Thus, the design of an electronic health system influences the way its users think about information that needs to be recorded. Of course, the process goes both ways, because the needs of users influence the design of the system. But, once the electronic health system is established, the users need to organise the data according to the technical requirements of the system. So, the standards and technical constraints pose certain restrictions on the way information is thought of (Patel et al. 2000). For instance, empirical research has demonstrated the profound influence of technology in shaping the cognitive behaviour of clinicians.

> Results indicate that exposure to the computer-based patient record was associated with changes in physicians' information gathering and reasoning strategies. Differences were found in the content and organization of information, with paper records having a narrative structure, while the computer-based records were organized into discrete items of information. The differences in knowledge organization had an effect on data gathering strategies, where the nature of doctor-patient dialogue was influenced by the structure of the computer-based patient record system (Patel et al. 2000).

Accordingly, I focused on classificatory knowledge in order to clarify how the collected information is organised into *discrete units*, which are labelled by key *terms* and recorded into an electronic health system (Chapter I, Chapter III).

The second change that comes with the development of information technologies concerns reasoning *with* health records, which can be processed automatically by machines. So, when we speak about reasoning *with* electronic health records, we also enter the field of automated reasoning, where formal representation of knowledge plays a central role.

Sections 1.1 and 1.2.2 acknowledged certain dependencies between classification, data collection, and digital data recording processes. Since reasoning *about* health information is frequently communicated in natural language, I discussed the role language plays in capturing the chosen aspects of knowledge. I examined how 'discrete items of information', labelled by specific lexical terms, is structured into an ontology model in order to represent explicitly how the domain experts conceptualise the targeted domain for a particular purpose (1.2.2-1.2.6, 2.2.1-2.2.3).

From a technical point of view, reasoning about health information is captured in a model that structures health information accordingly. The introduced distinction between information models and ontology models (Section 1.2.2.) shows that reasoning about health information into an electronic system can take place at various levels. Information models include models of codes and models of data structures (Section 1.2.4, Figure 6). Thereby, a relational database captures reasoning about health records as reasoning over extensional structures. The extensional structures of an information model are data records organised and stored in a database. On the other hand, ontology models organise concepts into conceptual structures, which represent explicitly how the domain experts understand the concepts within the targeted domain. As such, ontology models are understood as 'upper level' models in a knowledgebase, which represent reasoning as the intensional structures. An ontology model represents what the concepts and classes mean in a particular domain and how they relate to each other. Even so, information models and ontology models can be mutually aligned, so that the ontology models can play the role of reasoning schemes for information models, while the structures of information models can be extrapolated from a database scheme and represented in the form of ontology (Martinez-Cruz, Blanco, and Vila 2011).

I have shown that ontology models and information models share a number of features with the reasoning of biomedical experts because the models are designed to capture human reasoning about a domain, e.g. reasoning about the objects of interest, parthood relations, subordination of concepts, i.e. organisation of the concepts into hierarchical structures. This analogy between human reasoning about a biomedical domain and related ontology and information models, which represent reasoning in a machine readable form, has directed the focus of my analysis to the basic 'units', represented as terms, codes, and classes that are selected and organised according to the aims of a particular sub-domain.

In order to clarify how various domain experts select the most relevant 'units' to represent knowledge in practice, I undertook an analysis of: 1) diversity in breast cancer phenotype representation (Chapter II); and 2) diversity captured through various breast cancer classificatory systems (Chapter III).

I demonstrated that the plurality of biomedical representations within and across the research domains in both these cases, the breast cancer phenotypes and breast cancer classificatory systems, depend on the research questions and the aims for which the representations and classifications are designed. Moreover, I showed that addressing problems such as breast cancer requires a combination of various representational and classificatory perspectives. Only by combining knowledge acquired in several domains can a problem such as breast cancer be addressed in its complexity so that specialised knowledge from one domain can inform, correct, and enrich knowledge from other domains (Chapter II and III).

Regarding a practical application that combines perspectives, I proposed a model, developed in collaboration with Oliver Kutz, which represents 'normal' and 'abnormal' phenotypes by integrating certain clinical and molecular criteria which are used to describe the HER2+ tumours (Section 2.2.1-2.2.3). The most relevant feature of the phenotype that this model captures is overexpression of the HER2 protein. I explained how a detection system (IHC method stains a tissue sample with a coloured dye) produces an image that represents HER2 overexpression as a ‘brown colour’ of the membrane. Since the colour *marks* the presence of the protein within the membrane (as mediated by the detection system), the *quality* of the ‘brown’ membrane within the HER2+ breast tissue sample is used to represent a *quantity*, i.e. the amount of the HER2 proteins (marked by the dye) within the cell membrane. The detected amount of the HER2+ cell membranes corresponds with the number of HER2+ cells, while a tissue sample with more than 30% HER2+ cells acquires a ‘HER2+ cancer’ phenotype that is also assigned to the mammary gland. While capturing the reasoning of histopathologists about HER2 positivity, such an approach allowed us to describe HER2 ‘overexpression’ as *mereology* of the HER2+ phenotype.

Furthermore, the discussion of the HER2 model has demonstrated that the classification of organ parts etc. into *normal* vs. *abnormal* (Section 2.2, Figure 15) is not a classical dichotomy. Since HER2 protein is present in both normal and abnormal phenotypes, attaching ‘abnormality’ to a particular tissue depends non-monotonically on the presence of information concerning overexpression of HER2. That is to say, unless such information is explicitly known, the default tissue is labelled as ‘normal’. Therefore, adding information to the formal modelling (e.g. more instances of HER2 are part of some tissue) results in retracting the property of ‘normality’. Moreover, the precise information about the ranges of values that influence decisions on the assignment of ‘abnormality’ to some tissues has been presented as a matter of convention and the best available knowledge at a current time. Thus, the example shows that the classification and modelling of a disease phenotype is not fixed once and forever by some intrinsic Ontological features, but rather is the result of a pragmatic approach to modelling that concerns defeasible reasoning, while leaving room for new information and scientific revisions of previously established classificatory knowledge. Accordingly, section 3.4 particularly examined Ludger Jansen’s view on what a good classification is and how *the Ontological Grounding principle* supports classification.

The HER2 ontology has also demonstrated that formal modelling of ‘abnormality’ requires the non-monotonic extension of an ontology, as well as a combination of several ontologies. Since the majority of currently available ontology repositories (e.g. BioPortal) lack the ability to host heterogeneous ontologies, the HER2 ontology is hosted on OntoHub, a recently launched ontology repository, which supports a multi-logic modelling environment and the pluralistic approaches to ontology design (Sojic and Kutz 2012).

In addition, I outlined a framework for a model that would, in a similar manner, combine clinical, molecular, and epidemiological knowledge about the role that age plays in breast cancer (Section 3.7.). I argued that the classificatory systems consider the relevance of age on different levels and with different degrees of explicitness. Epidemiological descriptions capture age parameters on a population level, which considers risk factors and cancer incidence. Clinical domain focuses on the classification of patients, the assessment of diagnosis and prognosis, while age of individual patients is measured in a calendar time. Biological classifications capture age by looking at biological processes such as telomere activity, while the related biological classifications of breast cancer rarely explicitly include age classifiers. Since diverse domains conceptualise age in different ways, while classifying and organising knowledge, I specified ‘age’ as an *epistemic modifier* that directs the interpretation of the classificatory categories according to the practical needs of a domain. Therefore, I proposed to treat ‘age’ as a general classificatory module that plays various roles in distinct knowledge contexts. Accordingly, the contextual conceptualisation of age modifies the decision as to how to represent the relations among particular breast cancer classes. On a formal level, modelling of the risks may employ non-monotonic reasoning with circumscription, specified in section 2.2.2, so that the concept of risk is minimised and considered normal, unless the information about the patient’s age at an event (e.g. an early age at diagnosis) is present. Thus, information about age-related risks will modify the prognostic and diagnostic classes.

Nonetheless, I acknowledged that the distinctions between clinical and biomedical knowledge induces a sort of epistemic gap (Section 3.2.). The discussions in the third chapter showed that clinical and biomedical experts need not prioritise the same features while representing disease. Consequently, the domain of discourse, the selection of the most relevant classifications, explanations, and justifications fit a variety of practical needs, e.g. diagnosis assessment, classifications of patients, molecular components, populations, risk factors, etc. Instead of being sceptical about the gap, I stressed how personalised medicine and formal knowledge representation can contribute to the understanding of the gap as a rather fruitful tool that explicates the diversity of the epistemic needs across various clinical and biomedical domain representations and classifications (Section 3.1).

The process of collaboration among the various epistemic groups, which address the problems in domain-specific ways, yet aim to bring this knowledge together, were outlined in Section 1.2.6. Semantic and concept maps (1.2.3.) were presented as mediators for a formal representation of knowledge that explicitly captures a shared conceptualisation. An advantage of semantic and concept maps, as a representational tool, is their ability to provide an explicit representation of conceptual structure, which reflects reasoning about the targeted domain. So, the maps are often used as an intermediary step in designing ontology models and automated reasoning systems (Section 1.2.3, 1.2.6). The constitutive units of semantic and conceptual maps are the *terms* used to represent concepts and label 'objects' of interest in a particular domain. The terms are connected into a structure by chosen relations so that the relations that connect the represented units also represent reasoning about the units, e.g. parthood, hierarchy etc.

From the perspective of scientific representation and the representational means, I distinguished ontology representations from other kinds of scientific representation. I argued that ontologies which target the way domain experts conceptualise a problem constitute a special kind of scientific representation because they *represent reasoning explicitly* (Section 1.2.4.). The ontology model for the HRE2 breast cancer phenotype, developed and discussed in sections 2.2-2.2.3, is an example of how reasoning about phenotypes is represented in a formal way.

As a heuristic tool for understanding how ontologies and knowledgebases capture knowledge through collaborative efforts, I distinguished the epistemic agents and groups involved in ontology building (Section 1.2.6). Ontology building as an interdisciplinary endeavour necessarily deals with diverse epistemic and pragmatic interests. So, an ontology model which aims at capturing a shared conceptualisation might not be indeed shared across the particular knowledge domains. Therefore, I decided to address the implications that these attempts to integrate knowledge by means of an ontology have on the debate on unity and disunity of science.

In the fourth chapter, I examined the ways in which ontologies and knowledge bases can be considered a novel phenomenon that unifies science. I distinguished several cases that demonstrate the revival of the debate on the unity and disunity of science, which now takes place in a digitalised medium that interconnects the distributed knowledgebases.

I argued that even in this new medium, the domain problems ask for a plurality of classifications, representations, and ontological formalisms. Accordingly, any unification of knowledge that would dismiss the plurality of methodologies and perspectives that depend on the heterogeneity of the domain interests would detract from the practical advantages that domain-specific approaches bring to the enrichment of scientific knowledge. So, I characterised this novel unity as an extended-meaning unity, which relies concurrently on the precision and ambiguity of scientific language. Ontology modelling through the co-ordinated processes of contextualisation and de-contextualisation of the term's meaning contributes to the interconnectedness among the distributed knowledge domains, while enriching them with informative content according the particular needs. Thus, I conclude that the only way science can be literally considered as unified is through a co-ordination of scientific endeavours on the level of society that provides research policy, funds, socio-technological infrastructure, and various kinds of standards.

Surely, the effects of new technologies that foster co-ordination and collaboration across the biomedical domains bring significant improvements to molecular and clinical oncology, supporting clinical decisions and better understanding of dispersed pieces of knowledge.

Bibliography

(UICC), International Union Against Cancer. 1958. Committee on Clinical Stage Classification and Applied Statistics: Clinical stage classification and presentation of results, malignant tumours of the breast and larynx. Paris.

———. 1968. TNM Classification of malignant tumours. Geneva.

Al-Hajj, M., M. S. Wicha, A. Benito-Hernandez, S. J. Morrison, and M. F. Clarke. 2003. Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences* 100 (7):3983-3988.

Alizadeh, A. A., D. T. Ross, C. M. Perou, and M. van de Rijn. 2001. Towards a novel classification of human malignancies based on gene expression patterns. *The Journal of pathology* 195 (1):41-52.

Allison, D.B, Page, P.G, Beasley, T.M., Edwards, J.W., ed. 2006. *DNA Microarrays and Related Genomic Techniques: Design, Analysis, and Interpretation of Experiments*: Chapman & Hall.

Althuis, Michelle D., Jaclyn M. Dozier, William F. Anderson, Susan S. Devesa, and Louise A. Brinton. 2005. Global trends in breast cancer incidence and mortality 1973-1997. *International Journal of Epidemiology* 34 (2):405-412.

Ambroise, C., and G. J. McLachlan. 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences* 99 (10):6562-6566.

Aminoff, M. J. 2010. *Brown-Séquard: An Improbable Genius Who Transformed Medicine*: Oxford University Press.

Anderson, W. F., I. Jatoi, and S. S. Devesa. 2005. Distinct breast cancer incidence and prognostic patterns in the NCI's SEER program: suggesting a possible link between etiology and outcome. *Breast cancer research and treatment* 90 (2):127-137.

Anderson, W. F., P. S. Rosenberg, I. Menashe, A. Mitani, and R. M. Pfeiffer. 2008. Age-related crossover in breast cancer incidence rates between black and white ethnic groups. *Journal of the National Cancer Institute* 100 (24):1804-1814.

Anderson, William F., Ismail Jatoi, and Mark E. Sherman. 2009. Qualitative Age Interactions in Breast Cancer Studies: Mind the Gap. *J Clin Oncol* 27 (32):5308-5311.

Ankeny, Rachel A. 2000. Fashioning descriptive models in biology: Of Worms and wiring diagrams. *Philosophy of Science* 67 (3):272.

Ankeny, Rachel A., and Sabina Leonelli. 2011. What's so special about model organisms? *Studies In History and Philosophy of Science Part A* 42 (2):313-323.

Anthony, Peter. P. 1998. *Diagnostic pitfalls in histopathology and cytopathology practice*: Cambridge University Press.

Aristotle. 2008. *The Metaphysics*. Translated by J. H. McMahon: Cosimo.

Arnone, Paolo, Stefano Zurrida, Giuseppe Viale, Silvia Dellapasqua, Emilia Montagna, Paola Arnaboldi, Mattia Intra, and Umberto Veronesi. 2009. The TNM classification of breast cancer: need for change. *Updates in Surgery* 62 (2):75-81.

Aronson, J. L., R. Harré, and E. C. Way. 1995. *Realism rescued: how scientific progress is possible*: Open Court Pub Co.

Ashcroft, R. E. 2004. Current epistemological problems in evidence based medicine. *Journal of Medical Ethics* 30 (2):131-135.

Beatson, G. T., L. Francis, and M. Eng. 1896. On the treatment of inoperable cases ofcarcinoma of the mamma: suggestions for a new method of treatment, with illustrative cases. *Lancet* 148:104-7.

Bechtel, William. 1984. Reconceptualizations and Interfield Connections: The Discovery of the Link between Vitamins and Coenzymes. *Philosophy of Science* 51 (2):265-292.

Bechtel, William P., and Andrew Hamilton. 2007. Reduction, integration, and the unity of science: Natural, behavioral, and social sciences and the humanities. In *Philosophy of Science: Focal Issues (Volume 1 of the Handbook of the Philosophy of Science)*, edited by T. Kuipers: Elsevier.

Bernard, C. 1999. *Experimental Medicine*: Transaction.

Bernard, C., H. C. Greene, and L. J. Henderson. 1957. *An Introduction to the Study of Experimental Medicine*: Dover Publications.

Blasimme, A., and P. Maugeri. 2011. Humanised models of cancer in molecular medicine: the experimental control of disanalogy. *History and philosophy of the life sciences* 33 (4):603-622.

Block, W. 1982. Cold hardiness in invertebrate poikilotherms. *Comparative Biochemistry and Physiology Part A: Physiology* 73 (4):581-593.

Bodenreider, O. 2008. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform* 67:79.

Bonatti, P., C. Lutz, and F. Wolter. 2006. Description logics with circumscription. *Proc. KR* 6:400-410.

Boniolo, G. 2001. Concepts as representations and as rules. *Revista de filosofía* 25:93-115.

———. 2005. A Contextualized Approach to Biological Explanation. *Philosophy* 80 (02):219-247.

———. 2007. *On scientific representations: from Kant to a new philosophy of science*: Palgrave Macmillan.

Boniolo, G., M. D'Agostino, and P. P. Di Fiore. 2010. Zsyntax: a formal language for molecular biology with projected applications in text mining and biological prediction. *PLoS ONE* 5 (3):e9511.

Boniolo, G., M. D'Agostino, M. Piazza, and G. Pulcini. 2012. A logic of non-monotonic interactions. *Journal of Applied Logic*.

Boniolo, G., Valentini, S. 2012. Objects. A Study in Kantian Formal Epistemology. *Notre Dame Journal of Formal logic*.

Borst, W. N. 1997. Construction of engineering ontologies for knowledge sharing and reuse, PhD thesis, Institute for Telematica and Information Technology, Universiteit Twente, Enschede, The Netherlands.

Boshuizen, HPA, and HG Schmidt. 1992. On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science: A Multidisciplinary Journal* 16 (2):153-184.

Boyle, Peter, and Robin Leake. 1988. Progress in understanding breast cancer: Epidemiological and biological interactions. *Breast cancer research and treatment* 11 (2):91-112.

Brachman, R. J. 1979. On the Epistemological Status of Semantic Networks In *Associative networks: Representation and use of knowledge by computers*, edited by N. V. Findler: Academic Press New York.

Brachman, R., and H. Levesque. 2004. *Knowledge Representation and Reasoning*: Elsevier Science.

Brinkman, R. R., M. Courtot, D. Derom, J. M. Fostel, Y. He, P. Lord, J. Malone, H. Parkinson, B. Peters, and P. Rocca-Serra. 2011. Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 1 (Suppl 1):S7.

Brinton, Louise A., Catherine Schairer, Robert N. Hoover, and Joseph F. Fraumeni. 1988. Menstrual Factors and Risk of Breast Cancer. *Cancer Investigation* 6 (3):245-254.

Broca, P. 1866. *Traité des tumeurs*. Vol. 1: P. Asselin.

Brooks, L. R., G. R. Norman, and S. W. Allen. 1991. Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General* 120 (3):278-287.

Bueno-de-Mesquita, Jolien M., Wim H. van Harten, Valesca P. Retel, Laura J. van 't Veer, Frits S. A. M. van Dam, Kim Karsenberg, Kirsten F. L. Douma, Harm van Tinteren, Johannes L. Peterse, Jelle Wesseling, Tin S. Wu, Douwe Atsma, Emiel J. T. Rutgers, Guido Brink, Arno N. Floore, Annuska M. Glas, Rudi M. H. Roumen, Frank E. Bellot, Cees van Krimpen, Sjoerd Rodenhuis, Marc J. van de Vijver, and Sabine C. Linn. 2007. Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). *The Lancet Oncology* 8 (12):1079-1087.

Callebaut, Werner. 2012. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1):69-80.

Campisi, J., S. Kim, C. S. Lim, and M. Rubio. 2001. Cellular senescence, cancer and aging: the telomere connection. *Experimental gerontology* 36 (10):1619-1638.

Carlson, R. W., S. J. Moench, M. E. Hammond, E. A. Perez, H. J. Burstein, D. C. Allred, C. L. Vogel, L. J. Goldstein, G. Somlo, and W. J. Gradishar. 2006. HER2 testing in breast cancer: NCCN Task Force report and recommendations. *Journal of the National Comprehensive Cancer Network: JNCCN* 4:S1.

Carnap, Rudolf. 1950. Empiricism, semantics, and ontology. *Revue Internationale De Philosophie* 4 (2):20--40.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Cartwright, Nancy. 1999. *The dappled world: a study of the boundaries of science*: Cambridge University Press.

Chakravartty, A. 2007. *A Metaphysics for Scientific Realism: Knowing the Unobservable*: Cambridge University Press.

———. 2010. Informational versus functional theories of scientific representation. *Synthese* 172 (2):197-213.

Chalmers, D. J., D. Manley, and R. Wasserman. 2009. *Metametaphysics: new essays on the foundations of ontology*: Oxford University Press.

Charlin, Bernard, Jacques Tardif, and Henny P. A. Boshuizen. 2000. Scripts and Medical Diagnostic Knowledge: Theory and Applications for Clinical Reasoning Instruction and Research. *Academic Medicine* 75 (2):182-190.

Chen, Jake, and Amandeep S. Sidhu. 2007. *Biological Database Modeling*: Artech House, Inc.

Chuang, Han-Yu, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. 2007. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3 (1).

Cimino, J. J. 1998. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine-Methodik der Information in der Medizin* 37 (4):394-403.

Clark, Gary M. 1994. Do we really need prognostic factors for breast cancer? *Breast Cancer Research and Treatment* 30 (2):117-126.

Cleary, M. P., and N. J. Maihle. 1997. The role of body mass index in the relative risk of developing premenopausal versus postmenopausal breast cancer.

Colombo, G., D. Merico, Z. Nagy, F. De Paoli, M. Antoniotti, and G. Mauri. 2009. Ontological modeling at a domain interface: bridging clinical and biomolecular knowledge. *Knowledge Engineering Review* 24 (3):205-224.

Committee on Improving the Patient Record. 1991. *The computer-based patient record: an essential technology for health care*. Edited by R. S. Dick, Steen, E. B.: Institute of Medicine. National Academy Press.

Contessa, Gabriele. 2007. Scientific Representation, Interpretation, and Surrogative Reasoning. *Philosophy of Science* 74 (1):48-68.

Copeland, M. M. 1961. Clinical staging system for cancer and end results reporting. *CA: A Cancer Journal for Clinicians* 11 (2):42-47.

Correa Geyer, F., and J. S. Reis-Filho. 2009. Microarray-based gene expression profiling as a clinical tool for breast cancer management: are we there yet? *Int J Surg Pathol* 17:285 - 302.

Couto, E., and K. Hemminki. 2007. Estimates of heritable and environmental components of familial breast cancer using family history information. *Br J Cancer* 96 (11):1740-1742.

Creath, Richard. *"Logical Empiricism" in The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.)* (Winter 2011 Edition) 2011 [cited. Available from URL = <http://plato.stanford.edu/archives/win2011/entries/logical-empiricism/>.

Dalton, L. W., S. E. Pinder, C. E. Elston, I. O. Ellis, D. L. Page, W. D. Dupont, and R. W. Blamey. 2000. Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement. *Modern Pathology* 13 (7):730-735.

Darden, Lindley, and Nancy Maull. 1977. Interfield Theories. *Philosophy of Science* 44 (1):43-64.

David, P. A. *Public dimensions of the knowledge-driven economy. A brief introduction to the OECD/CERI project* 2002 [cited. Available from http://www.oecd.org/dataoecd/47/37/2074404.pdf.

Davis, Jerel C., Laura Furstenthal, Amar A. Desai, Troy Norris, Saumya Sutaria, Edd Fleming, and Philip Ma. 2009. The microeconomics of personalized medicine: today's challenge and tomorrow's promise. *Nat Rev Drug Discov* 8 (4):279-286.

Davis, R., H. Shrobe, and P. Szolovits. 1993. What is a Knowledge Representation? *AI Magazine* 14 (1):17-33.

de Bono, Bernard, Robert Hoehndorf, Sarala Wimalaratne, George Gkoutos, and Pierre Grenon. 2011. The RICORDO approach to semantic interoperability for biomedical data and models: strategy, standards and solutions. *BMC Research Notes* 4 (1):313.

De Matos, P., R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, and C. Steinbeck. 2010. Chemical entities of biological interest: an update. *Nucleic acids research* 38 (suppl 1):D249-D254.

Denoix, PF. 1944. Tumor, node and metastasis (TNM). *Bull. Inst. Nat. Hyg (Paris)* 1:69.

Di Cataldo, S., E. Ficarra, A. Acquaviva, and E. Macii. 2010. Automated segmentation of tissue images for computerized IHC analysis. *Computer methods and programs in biomedicine* 100 (1):1-15.

Dietze, H., and M. Schroeder. 2009. GoWeb: a semantic search engine for the life science web. *BMC Bioinformatics* 10.

Doms, A., and M. Schroeder. 2005. GoPubMed: exploring PubMed with the gene ontology. *Nucleic acids research* 33 (suppl 2):W783-W786.

Donegan, W. L., and J. S. Spratt. 2002. *Cancer of the breast*. 5th ed. Philadelphia: Saunders.

Donnellan, K. S. 1966. Reference and definite descriptions. *The Philosophical Review* 75 (3):281-304.

Donner, R. S., and H. Bickley. 1993. Problem-based learning in American medical education: an overview. *Bulletin of the Medical Library association* 81 (3):294.

Dumontier, M., B. Andersson, C. Batchelor, C. Denney, C. Domarew, A. Jentzsch, J. Luciano, E. Pichler, E. Prud’hommeaux, and P. L. Whetzel. 2010. The Translational Medicine Ontology: Driving personalized medicine by bridging the gap from bedside to bench. *In Proceedings of The 13th Annual Bio-Ontologies Meeting*:120-123.

Dupré, John. 1981. Natural kinds and biological taxa. *Philosophical Review* 90 (1):66-90.

———. 1983. The Disunity of Science. *Mind* (367):321-346.

———. 1993. *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*: Harvard University Press.

———. 2002. *Humans and Other Animals*: Clarendon Press.

———. 2002. Is 'Natural Kind' a Natural Kind Term? *The Monist* 85 (1):29-49.

———. 2004. Human Kinds and Biological Kinds: Some Similarities and Differences. *Philosophy of Science* 71 (5):892-900.

Ekins, S., M. A. Z. Hupcey, A. J. Williams, and A. Bingham. 2011. *Collaborative Computational Technologies for Biomedical Research*: John Wiley & Sons.

El Emam, K., L. L. B. Michael Power, and D. Willison. 2007. Privacy Guidelines Workshop Report: CHEO Research Institute, Ottawa, Canada.

Elsenbroich, C., O. Kutz, and U. Sattler. 2006. A case for abductive reasoning over ontologies. *OWL: Experiences and Directions* 67:81-82.

Elston, C. W., and I. O. Ellis. 1991. Pathological prognostic factors in breast cancer I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 19:403 - 410.

———. 2002. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. C. W. Elston & I. O. Ellis. Histopathology 1991; 19; 403–410. *Histopathology* 41 (3a):151-151.

Enderling, Heiko, Mark A. J. Chaplain, Alexander R. A. Anderson, and Jayant S. Vaidya. 2007. A mathematical model of breast cancer development, local treatment and recurrence. *Journal of Theoretical Biology* 246 (2):245-259.

Evans, A. David , and Vilma Patel, eds. 1989. *Cognitive Science in Medicine: Biomedical Modeling*. Cambridge, MA: MIT Press.

Fackenthal, J. D., and O. I. Olopade. 2007. Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nature Reviews Cancer* 7 (12):937-948.

Faratian, Dana, Robert G. Clyde, John W. Crawford, and David J. Harrison. 2009. Systems pathology - taking molecular pathology into a new dimension. *Nat Rev Clin Oncol* 6 (8):455-464.

———. 2009. Systems pathology[mdash]taking molecular pathology into a new dimension. *Nat Rev Clin Oncol* 6 (8):455-464.

Farfán, F., Varadarajan, R., and Hristidis, V. . 2009. Electronic Health Records. In *Information Discovery on Electronic Health Records*, edited by V. Hristidis: Chapman & Hall/CRC.

Feyerabend, P. 1970. Against method: outline of an anarchistic theory of knowledge. In *Minnesota Studies in the Philosophy of Science Vol*, edited by R. A. S. Winokur: University of Minnesota Press.

Fisher, B., Slack, N.H. 1970. Number of lymph nodes examined and the prognosis of breast carcinoma. *Surg Gynecol Obstet* 131 (1):79-88.

Fisher, Bernard, Carol K. Redmond, and Edwin R. Fisher. 2008. Evolution of Knowledge Related to Breast Cancer Heterogeneity: A 25-Year Retrospective. *Journal of Clinical Oncology* 26 (13):2068-2071.

Frank, N. Y., T. Schatton, and M. H. Frank. 2010. The therapeutic promise of the cancer stem cell concept. *The Journal of clinical investigation* 120 (1):41.

French, S. 2003. A Model-Theoretic Account of Representation (Or, I Don't Know Much About Art... But I Know It Involves Isomorphism). *Philosophy of Science*:1472-1483.

Frigg, R. 2006. Scientific representation and the semantic view of theories.

———, ed. 2010. *Beyond mimesis and convention: representation in art and science*. Vol. 262: Springer Verlag.

Frigg, R., and S. Hartmann. 2006. Models in science. *Stanford Encyclopedia of Philosophy*, http://stanford.library.usyd.edu.au/entries/models-science/.

Fritz, A. G., Percy, C., Jack, A., Shanmugaratnam, K., Sobin, L., Parkin, D.M., Whelan, S., ed. 2000. *International classification of diseases for oncology: ICD-O*. Geneva: World Health Organization.

Galea, M. H., R. W. Blamey, C. E. Elston, and I. O. Ellis. 1992. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat* 22:207 - 219.

Gangem, A., C. Catenacc, and M. Battaglia. 2004. lnflammation Ontology Design Pattern: An Exercise in Building a Core Biomedical Ontology With Descriptions and Situations. *Ontologies in medicine* 102:64.

Gao, Yu-Tang, Xiao-Ou Shu, Qi Dai, John D. Potter, Louise A. Brinton, Wanqing Wen, Thomas A. Sellers, Lawrence H. Kushi, Zhixian Ruan, Roberd M. Bostick, Fan Jin, and Wei Zheng. 2000. Association of menstrual and reproductive factors with breast cancer risk: Results from the Shanghai breast cancer study. *International Journal of Cancer* 87 (2):295-300.

Gärdenfors, Peter. 2005. *The dynamics of thought*: Springer.

Gasparini, G., and D. Hayes. 2006. *Biomarkers in breast cancer: molecular diagnostics for predicting and monitoring therapeutic effect*: Humana Press.

Gaudillière, Jean-Paul, and Hans-Jörg Rheinberger, eds. 2004. *From Molecular Genetics to Genomics : The Mapping Cultures of Twentieth-Century Genetics*. London: Routledge.

Giere, R. N. 2006. Perspectival pluralism. In *Scientific pluralism*, edited by S. H. Kellert, H. E. Longino, and C. K. Waters: University of Minnesota Press.

Giere, Ronald. 2010. An agent-based conception of models and scientific representation. *Synthese* 172 (2):269-281.

Giordano, S. H., A. U. Buzdar, and G. N. Hortobagyi. 2002. Breast cancer in men. *Annals of internal medicine* 137 (8):678-687.

Goldhirsch, A., R. D. Gelber, G. Yothers, R. J. Gray, S. Green, J. Bryant, S. Gelber, M. Castiglione-Gertsch, and A. S. Coates. 2001. Adjuvant therapy for very young women with breast cancer: need for tailored treatments. *JNCI Monographs* 2001 (30):44-51.

Goldhirsch, A., J. N. Ingle, R. D. Gelber, A. S. Coates, B. Thürlimann, and H. J. Senn. 2009. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. *Annals of Oncology* 20 (8):1319-1329.

Goldhirsch, A., J. N. Ingle, R. D. Gelber, A. S. Coates, B. Thurlimann, and H. J. Senn. 2009. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. *Ann Oncol* 20:1319 - 1329.

Goldhirsch, A., W. C. Wood, A. S. Coates, R. D. Gelber, B. Thürlimann, and H. J. Senn. 2011. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Annals of Oncology*:1736-1747.

Gonzalez-Angulo, Ana Maria, Bryan T. J. Hennessy, and Gordon B. Mills. 2010. Future of Personalized Medicine in Oncology: A Systems Biology Approach. *Journal of Clinical Oncology* 28 (16):2777-2783.

Goodman, N. 1968. *Languages of art: An approach to a theory of symbols*: Indianapolis: The Bobbs-Merrill Company.

Goodman, Nelson. 1978. *Ways of worldmaking*: Hackett.

Gospodarowicz, M. K. et al., ed. 2001. *Prognostic factors in cancer*: International Union against Cancer. Wiley-Liss.

Gospodarowicz, M. K., B. O'Sullivan, and L. H. Sobin, eds. 2006. *Prognostic factors in cancer*: International Union against Cancer. Wiley-Liss.

Grenon, P. 2003. BFO in a Nutshell: A Bi-categorial Axiomatization of BFO and Comparison with DOLCE. IFOMIS Report 06/2003. *Institute for Formal Ontology and Medical Information Science (IFOMIS), University of Leipzig, Leipzig, Germany*.

Griffiths, F. 2009. *Research Methods for Health Care Practice*: SAGE.

Gruber, Thomas R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5 (2):199-220.

———. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies* 43 (5-6):907-928.

Gruhn, J. G., and R. R. Kazer. 1989. *Hormonal regulation of the menstrual cycle: the evolution of concepts*: Plenum Medical Book Co.

Gruninger, M. 2004. Ontology of the process specification language. *Handbook on Ontologies*:575-592.

Guarda, P. 2011. *Fascicolo sanitario elettronico e protezione dei dati personali*. Trento: Università degli Studi di Trento - Dipartimento di Scienze Giuridiche.

Guarino, N. 1994. The ontological level. In *Philosophy and the Cognitive Science*, edited by R. Casati, Smith, B, and White, G. Vienna: Hölder-Pichler-Tempsky.

———. 2009. The ontological level: Revisiting 30 years of knowledge representation. In *Conceptual Modeling: Foundations and Applications*, edited by A. Borgida, Chaudhri, V., Giorgini, P., Yu, E.: Springer.

Guarino, N., D. Oberle, and S. Staab. 2009. What is an Ontology? *Handbook on Ontologies*:1-17.

Gurwitz, David, Jeantine E. Lunshof, and Russ B. Altman. 2006. A call for the creation of personalized medicine databases. *Nat Rev Drug Discov* 5 (1):23-26.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*: Cambridge University Press.

———. 2002. *Historical Ontology*: Harvard University Press.

Hadzic, M., and E. Chang. 2005. Ontology-based support for human disease study. *Proceedings of the 38th Hawaii International Conference on System Sciences - 2005. IEEE*:143a-143a.

Hamburg, M. A., and F. S. Collins. 2010. The path to personalized medicine. *New England Journal of Medicine* 363 (4):301-304.

Hanahan, Douglas, and Robert A. Weinberg. 2000. The Hallmarks of Cancer. *Cell* 100 (1):57-70.

Hanson, N. R. 1958. *Patterns of discovery: an inquiry into the conceptual foundations of science*: University Press.

Hardwig, J. 1991. The role of trust in knowledge. *The journal of philosophy* 88 (12):693-708.

Harnad, S. R. 1990. Category induction and representation. In *Categorical perception: the groundwork of cognition*, edited by S. R. Harnad: Cambridge University Press.

Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, and C. Mungall. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* 32 (Database issue):D258.

Hartung, M., A. Gross, T. Kirsten, and E. Rahm. 2012. Effective Mapping Composition for Biomedical Ontologies. *In Proc. of Semantic Interoperability in Medical Informatics (SIMI-12), Workshop at ESWC-2012*.

Harvey, Jennet M., Gary M. Clark, C. Kent Osborne, and D. Craig Allred. 1999. Estrogen Receptor Status by Immunohistochemistry Is Superior to the Ligand-Binding Assay for Predicting Response to Adjuvant Endocrine Therapy in Breast Cancer. *Journal of Clinical Oncology* 17 (5):1474.

Hastings, J., O. Kutz, and T. Mossakowski. 2011. How to model the shapes of molecules? Combining topology and ontology using heterogeneous specifications. In *Proc. of the Deep Knowledge Representation Challenge Workshop (DKR-11), K-CAP-11*. Banff, Alberta, Canada.

Hempel, C. G. 1966. *Philosophy of natural Science*. Oxford, England: Prentice-Hall.

Hempel, Carl G., and Paul Oppenheim. 1948. Studies in the logic of explanation. *Philosophy of Science* 15 (2):135-175.

Henderson, John. 2005. Ernest Starling and 'Hormones': an historical commentary. *Journal of Endocrinology* 184 (1):5-10.

Herre, H. 2010. General Formal Ontology (GFO): A foundational ontology for conceptual modelling. In *Theory and Applications of Ontology: Computer Applications*, edited by R. Poli, and Obrst, L. Berlin: Springer.

Higginbotham, James 2006. Truth and Reference as the Basis of Meaning. In *The Blackwell guide to the philosophy of language*, edited by M. Devitt and R. Hanley: Blackwell Pub.

Hoehndorf, Robert, Michel Dumontier, John Gennari, Sarala Wimalaratne, Bernard de Bono, Daniel Cook, and Georgios Gkoutos. 2011. Integrating systems biology models and biomedical ontologies. *BMC Systems Biology* 5 (1):124.

Hoehndorf, Robert, Michel Dumontier, Anika Oellrich, Dietrich Rebholz-Schuhmann, Paul N. Schofield, and Georgios V. Gkoutos. 2011. Interoperability between Biomedical Ontologies through Relation Expansion, Upper-Level Ontologies and Automatic Reasoning. *PLoS ONE* 6 (7):e22006.

Hoehndorf, Robert, Frank Loebe, Janet Kelso, and Heinrich Herre. 2007. Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. *BMC Bioinformatics* 8 (1):1-12.

Hoehndorf, Robert, Axel-Cyrille Ngonga Ngomo, and Janet Kelso. 2010. Applying the functional abnormality ontology pattern to anatomical functions. *Journal of Biomedical Semantics* 1 (1):4.

Hoehndorf, Robert, Anika Oellrich, Michel Dumontier, Janet Kelso, Dietrich Rebholz-Schuhmann, and Heinrich Herre. 2010. Relations as patterns: bridging the gap between OBO and OWL. *BMC Bioinformatics* 11 (1):441.

Hoehndorf, Robert, Anika Oellrich, and Dietrich Rebholz-Schuhmann. 2010. Interoperability between phenotype and anatomy ontologies. *Bioinformatics*.

Hoehndorf, Robert, Paul N. Schofield, and Georgios V. Gkoutos. 2011. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic acids research*.

Hofweber, Thomas. 2011. Logic and Ontology. *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/logic-ontology/#Ont.

Holick, Crystal N., Polly A. Newcomb, Amy Trentham-Dietz, Linda Titus-Ernstoff, Andrew J. Bersch, Meir J. Stampfer, John A. Baron, Kathleen M. Egan, and Walter C. Willett. 2008. Physical Activity and Survival after Diagnosis of Invasive Breast Cancer. *Cancer Epidemiology Biomarkers & Prevention* 17 (2):379-386.

Hollnagel, H. 1999. Explaining risk factors to patients during a general practice consultation: conveying group-based epidemiological knowledge to individual patients. *Scandinavian journal of primary health care* 17 (1):3-5.

Holmes, Michelle D., Wendy Y. Chen, Diane Feskanich, Candyce H. Kroenke, and Graham A. Colditz. 2005. Physical Activity and Survival After Breast Cancer Diagnosis. *JAMA: The Journal of the American Medical Association* 293 (20):2479-2486.

Horwich, P. 1998. *Truth*: Clarendon Press.

Hristidis, V., ed. 2009. *Information Discovery on Electronic Health Records*: Chapman & Hall/CRC.

Huang, Zhiping, Susan E. Hankinson, Graham A. Colditz, Meir J. Stampfer, David J. Hunter, JoAnn E. Manson, Charles H. Hennekens, Bernard Rosner, Frank E. Speizer, and Walter C. Willett. 1997. Dual Effects of Weight and Weight Gain on Breast Cancer Risk. *JAMA: The Journal of the American Medical Association* 278 (17):1407-1411.

Hull, D. L. 1972. Reduction in genetics--biology or philosophy? *Philosophy of Science*:491-499.

———. 1988. Science as a Process. *University of Chicago Press*.

Hunter, Anthony, and Rupert Summerton. 2006. A knowledge-based approach to merging information. *Know.-Based Syst.* 19 (8):647-674.

Hurford, J. R. 2007. *The origins of meaning*: Oxford University Press.

ISO/DTR 20514. 2004. Health informatics – Electronic health record – Definition, scope, and context.

Ivshina, Anna V., Joshy George, Oleg Senko, Benjamin Mow, Thomas C. Putti, Johanna Smeds, Thomas Lindahl, Yudi Pawitan, Per Hall, Hans Nordgren, John E. L. Wong, Edison T. Liu, Jonas Bergh, Vladimir A. Kuznetsov, and Lance D. Miller. 2006. Genetic Reclassification of Histologic Grade Delineates New Clinical Subtypes of Breast Cancer. *Cancer Research* 66 (21):10292-10301.

Jansen, L. 2009. Categories: The top-level ontology. In *Applied Ontology: An Introduction*, edited by K. Munn, and B. Smith: Ontos Verlag.

———. 2009. Classifications. In *Applied Ontology: An Introduction*, edited by K. Munn, and B. Smith: Ontos Verlag.

Jansen, L., Schulz , S. 2011. The Ten Commandments of Ontological Engineering. In *OBML 2011 Workshop Proceedings*, edited by H. Herre, Hoehndorf, R, Loebe F: Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE), Universität Leipzig.

Jasanoff, S. 2004. *States of knowledge: the co-production of science and social order*: Routledge.

———. 2005. Civic Epistemology. In *Designs On Nature: Science And Democracy In Europe And The United States*: Princeton University Press.

Jasanoff, S., T. Pinch, G. E. Markle, and J. C. Peterson. 2001. *Handbook of science and technology studies*: Sage Publications, Incorporated.

Jatoi, I., W. F. Anderson, and P. S. Rosenberg. 2008. Qualitative age-interactions in breast cancer: a tale of two diseases? *American journal of clinical oncology* 31 (5):504-506.

Jemal, Ahmedin, Rebecca Siegel, Elizabeth Ward, Yongping Hao, Jiaquan Xu, and Michael J. Thun. 2009. Cancer Statistics, 2009. *CA Cancer J Clin* 59 (4):225-249.

Jones, R. and Leonard, R.C.F. . 2006. Metastatic Breast Cancer: Tailored Chemotherapy for the Elderly Woman. In *Breast Cancer and Molecular Medicine: Towards Tailored Approaches*, edited by M. J. Piccart, Hung, Mien-Chie, Solin, Lawrence J., Cardoso, Fatima und Wood, William C. : Springer-Verlag Berlin Heidelberg.

Kalra, D. 2006. Electronic health record standards. *IMIA yearbook of medical informatics* 2006:136-144.

Karp, P. D., S. Paley, C. J. Krieger, and P. Zhang. 2004. An evidence ontology for use in pathway/genome databases. *In Pac. Symp. Biocomput* 9:190-201.

Keating, Peter, and Alberto Cambrosio. 2003. *Biomedical platforms : realigning the normal and the pathological in late-twentieth-century medicine*. Cambridge, Mass.: MIT Press.

Kellert, S. H., H. E. Longino, and C. K. Waters, eds. 2006. *Scientific pluralism*: University of Minnesota Press.

Kern, J., K. Fister, and O. Polasek. 2009. Active Patient Role in Recording Health Data. In *Encyclopedia of Information Science and Technology*, edited by M. Khosrow-Pour. Hershey. New York: Information Science Reference.

Key, Timothy J., Pia K. Verkasalo, and Emily Banks. 2001. Epidemiology of breast cancer. *The Lancet Oncology* 2 (3):133-140.

Kim, J. 1988. What is" naturalized epistemology?". *Philosophical perspectives* 2:381-405.

Kincaid, H., J. Dupré, and A. Wylie. 2007. *Value-free science?: ideals and illusions*: Oxford University Press.

King, Mary-Claire, Joan H. Marks, Jessica B. Mandell, and Group The New York Breast Cancer Study. 2003. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. *Science* 302 (5645):643-646.

Kireyev, Kirill. 2009. Semantic-based estimation of term informativeness. *In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*:530–538.

Kitcher, Philip. 1981. Explanatory unification. *Philosophy of Science* 48 (4):507-531.

———. 1990. The division of cognitive labor. *Journal of Philosophy* 87 (1):5-22.

———. 1999. Unification as a regulative ideal. *Perspectives on Science* 7 (3):337-348.

———. 2001. *Science, Truth, and Democracy*: Oxford University Press.

Kleinerman, R. A. 2006. Cancer risks following diagnostic and therapeutic radiation exposure in children. *Pediatric radiology* 36:121-125.

Kljakovic, M. 2001. The use of single case studies by academic general practitioners. *Australian family physician* 30 (12):1195.

Knuuttila, T. 2010. Some Consequences of the Pragmatist Approach to Representation. *EPSA Epistemology and Methodology of Science*:139-148.

Knuuttila, Tarja. 2011. Modelling and representing: An artefactual approach to model-based representation. *Studies In History and Philosophy of Science Part A* 42 (2):262-271.

Kripke, S. 1977. Speaker's Reference and Semantic Reference. *Midwest studies in philosophy* 2 (1):255-276.

Kuhn, T. S. 1970. *The structure of scientific revolutions*: University of Chicago Press.

Kumar, G. L., Badve, S. S. 2008. Milestones in the Discovery of HER2 Proto-Oncogene and Trastuzumab (Herceptin™). *Connection*:9.

Kutz, O., T. Mossakowski, and D. Lücke. 2010. Carnap, Goguen, and the Hyperontologies: Logical Pluralism and Heterogeneous Structuring in Ontology Design. *Logica Universalis* 4 (2):255-333.

Kutz, Oliver, Mossakowski Till, Hastings Janna, Garcia Castro Alexander, and Sojic Aleksandra. 2011. Hyperontology for the Biomedical Ontologist: A sketch and some examples. *Working with Multiple Biomedical Ontologies Workshop, ICBO 2011*:399-408.

La Caze, Adam. 2010. The role of basic science in evidence-based medicine. *Biology and Philosophy* 26 (1):1-18.

Lacroix, M., and G. Leclercq. 2004. Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast cancer research and treatment* 83 (3):249-289.

Lambe, Mats, Chung-cheng Hsieh, Hsiao-wei Chan, Anders Ekbom, Dimitrios Trichopoulos, and Hans-Olov Adami. 1996. Parity, age at first and last birth, and risk of breast cancer: A population-based study in Sweden. *Breast cancer research and treatment* 38 (3):305-311.

Lange, C., T. Mossakowski, and O. Kutz. 2012. LoLa: A Modular Ontology of Logics, Languages, and Translations. In *In Proceedings of the 6th International Workshop on Modular Ontologies (WoMO) 2012*, edited by T. Schneider, Dirk, Walther. Graz, Austria: ceur-ws.

Lange, C., T. Mossakowski, O. Kutz, C. Galinski, M. Grüninger, and D. C. Vale. 2012. The Distributed Ontology Language (DOL): Use cases, syntax, and extensibility. *Arxiv preprint arXiv:1208.0293*.

Le, X. F., F. Pruefer, and R. C. Bast. 2005. HER2-targeting antibodies modulate the cyclin-dependent kinase inhibitor p27Kip1 via multiple signaling pathways. *Cell Cycle* 4 (1):87-95.

Leonelli, S., A. Diehl, K. Christie, M. Harris, and J. Lomax. 2011. How the gene ontology evolves. *BMC Bioinformatics* 12 (1):325.

Leonelli, Sabina. 2008. Bio-ontologies as Tools for Integration in Biology. *Biological Theory* 3 (1):7-11.

———. 2009a. On the locality of data and claims about phenomena. *Philosophy of Science* 76 (5).

———. 2009b. Centralising labels to distribute data: The regulatory role of genomic consortia. In *The Handbook for Genetics and Society: Mapping the New Genomic Era. London: Routledge*, edited by P. Atkinson, Glasner, P, and Lock, M.

———. 2010. Packaging small facts for re-use: databases in model organism biology. In *How well do facts travel? The dissemination of reliable knowledge*, edited by P. Howlett and M. S. Morgan.

———. 2012a. Classificatory Theory in Data-intensive Science: The Case of Open Biomedical Ontologies. *International Studies in the Philosophy of Science* 26 (1):47-65.

———. 2012b. Classificatory Theory in Biology. *Biological Theory*:1-8.

———. 2012c. When Humans Are the Exception: Cross-Species Databases at the Interface of Clinical and Biological Research. *Social Studies of Science* 42 (2):214-236.

Leonelli, Sabina, and Rachel A. Ankeny. 2012. Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1):29-36.

Lewontin, Richard. 2012. *The Genotype/Phenotype Distinction*. SEP, Apr 26, 2011, 2004 [cited 2012]. Available from http://plato.stanford.edu/entries/genotype-phenotype/.

Lifschitz, V., F. van Harmelen, and B. Porter, eds. 2008. *Handbook of knowledge representation*: Elsevier Science Limited.

Lipton, Allan, S. M. Ali, K. Leitzel, L. Demers, V. Chinchilli, L. Engle, Harold A. Harvey, C. Brady, C. M. Nalin, M. Dugan, W. Carney, and J. Allard. 2002. Elevated Serum HER-2/neu Level Predicts Decreased Response to Hormone Therapy in Metastatic Breast Cancer. *Journal of Clinical Oncology* 20 (6):1467-1472.

Liu, L., R. C. Wylie, L. G. Andrews, and T. O. Tollefsbol. 2003. Aging, cancer and nutrition: the DNA methylation connection. *Mechanisms of ageing and development* 124 (10):989-998.

Longnecker, Matthew P., Jesse A. Berlin, Michele J. Orza, and Thomas C. Chalmers. 1988. A Meta-analysis of Alcohol Consumption in Relation to Risk of Breast Cancer. *JAMA: The Journal of the American Medical Association* 260 (5):652-656.

Lord, Phillip, and Robert Stevens. 2010. Adding a Little Reality to Building Ontologies for Biology. *PLoS ONE* 5 (9):e12258.

Love, Richard R., and John Philips. 2002. Oophorectomy for Breast Cancer: History Revisited. *Journal of the National Cancer Institute* 94 (19):1433-1434.

Ludwig, J. A., and J. N. Weinstein. 2005. Biomarkers in cancer staging, prognosis and treatment selection. *Nature Reviews Cancer* 5 (11):845-856.

MacMahon, B., P. Cole, T. M. Lin, C. R. Lowe, A. P. Mirra, B. Ravnihar, E. J. Salber, V. G. Valaoras, and S. Yuasa. 1970. Age at first birth and breast cancer risk. *Bulletin of the World Health Organization* 43 (2):209.

Mai, J. E. 2004. Classification in context: relativity, reality, and representation. *Knowledge organization* 31 (1):39-48.

Mäki, U. 2010. Models and truth. *EPSA Epistemology and Methodology of Science*:177-187.

Malterud, K. 2000. Symptoms as a source of medical knowledge: understanding medically unexplained disorders in women. *FAMILY MEDICINE-KANSAS CITY-* 32 (9):603-611.

———. 2001. The art and science of clinical knowledge: evidence beyond measures and numbers. *The Lancet* 358 (9279):397-400.

Martinez-Cruz, C., I. J. Blanco, and M. A. Vila. 2011. Ontologies versus relational databases: are they so different? A comparison. *Artificial Intelligence Review*:1-20.

Masolo, C., S. Borgo, A. Gangemini, N. Guarino, A. Oltramari, and L. Schneider. 2003. The WonderWeb Library of Fundational Ontologies and the DOLCE ontology, D18, Final Report (vr. 1.0. 31-12-2003). In *WonderWeb Deliverable*: ISTC-CNR.

McCarthy, J. 1980. Circumscription--a form of non-monotonic reasoning. *Artificial intelligence* 13 (1-2):27-39.

McPherson, K., C. M. Steel, and J. M. Dixon. 2000. Breast cancer—epidemiology, risk factors, and genetics. *BMJ: British Medical Journal* 321 (7261):624–628.

McWhinney, I. R. 1989. 'An acquaintance with particulars...'. *Family medicine* 21 (4):296-8.

Medin, D. L., and L. J. Rips. 2005. Concepts and categories: Memory, meaning, and metaphysics. In *The Cambridge handbook of thinking and reasoning*, edited by K. J. Holyoak, Morrison, R.G. (Jr.).

Merrill, Gary H. 2010. Ontological realism: Methodology or misdirection? *Applied Ontology* 5 (2):79-108.

Metcalfe, K. A., D. Birenbaum-Carmeli, J. Lubinski, J. Gronwald, H. Lynch, P. Moller, P. Ghadirian, W. D. Foulkes, J. Klijn, and E. Friedman. 2008. International variation in rates of uptake of preventive options in BRCA1 and BRCA2 mutation carriers. *International Journal of Cancer* 122 (9):2017-2022.

Michael, J., J. L. Mejino Jr, and C. Rosse. 2001. The role of definitions in biomedical concept representation. *Proceedings of the AMIA Symposium*:463–467.

Milian, K., Z. Aleksovski, R. Vdovjak, A. ten Teije, and F. van Harmelen. 2010. Identifying Disease-Centric Subdomains in Very Large Medical Ontologies: A Case-Study on Breast Cancer Concepts in SNOMED CT. Or: Finding 2500 Out of 300.000. *Knowledge Representation for Health-Care. Data, Processes and Guidelines*:50-63.

Miller, W. L. 1992. Routine, ceremony, or drama: an exploratory field study of the primary care clinical encounter. *The Journal of family practice* 34 (3):289.

Mills, James L. 1993. Data Torturing. *New England Journal of Medicine* 329 (16):1196-1199.

Mitchell, S. D. 2003. *Biological Complexity and Integrative Pluralism*: Cambridge University Press.

Montazemi, A. R., Esfahanipour, A. 2009. Application of Cognitive Map in Knowledge Management. In *Encyclopedia of Information Science and Technology*, edited by M. Khosrow-Pour. Hershey. New York: Information Science Reference.

Moore, R. C. 1985. Semantical considerations on nonmonotonic logic. *Artificial intelligence* 25 (1):75-94.

Morgan, M. S., and M. Morrison, eds. 1999. *Models as mediators: Perspectives on natural and social science*. Vol. 52: Cambridge University Press.

Mossakowski, T., C. Maeder, and K. Lüttich. 2007. The heterogeneous tool set, HETS. In *In Proceedings of the 13th international conference on Tools and algorithms for the construction and analysis of systems*: Springer-Verlag.

Müller-Wille, Staffan, and Isabelle Charmantier. 2012. Natural history and information overload: The case of Linnaeus. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1):4-15.

Mungall, C., G. Gkoutos, N. Washington, and S. Lewis. 2007. Representing phenotypes in OWL. *OWL: Experiences and Directions (OWLED 2007), Innsbruk, Austria*.

Mungall, C. J., G. V. Gkoutos, C. L. Smith, M. A. Haendel, S. E. Lewis, and M. Ashburner. 2010. Integrating phenotype ontologies across multiple species. *Genome biology* 11.

Nagel, Ernest. 1984. *The structure of science problems in the logic of scientific explanation*. Indianapolis; Cambridge: Hackett.

Nahta, R., D. Yu, M. C. Hung, G. N. Hortobagyi, and F. J. Esteva. 2006. Mechanisms of disease: understanding resistance to HER2-targeted therapy in human breast cancer. *Nature clinical practice. Oncology* 3 (5):269-80.

Nahta, Rita, Dihua Yu, Mien-Chie Hung, Gabriel N. Hortobagyi, and Francisco J. Esteva. 2006. Mechanisms of Disease: understanding resistance to HER2-targeted therapy in human breast cancer. *Nat Clin Prac Oncol* 3 (5):269-280.

Narod, S. A., J. Feunteun, H. T. Lynch, P. Watson, T. Conway, J. Lynch, and G. M. Lenoir. 1991. Familial breast-ovarian cancer locus on chromosome 17q12-q23. *Lancet* 338 (8759):82.

Nelson, Christine L., Thomas A. Sellers, Stephen S. Rich, John D. Potter, Paul G. McGovern, and Lawrence H. Kushi. 1993. Familial clustering of colon, breast, uterine, and ovarian cancers as assessed by family history. *Genetic Epidemiology* 10 (4):235-244.

Neurath, Otto. 1983. Protocol Statements. In *Otto Neurath - Philosophical Papers 1913-1946*, edited by R. S. a. N. M. Cohen. Dordrecht: D. Reidel.

Newell, A. 1982. The knowledge level. *Artificial intelligence* 18 (1):87-127.

Nixon, A. J., D. Neuberg, D. F. Hayes, R. Gelman, J. L. Connolly, S. Schnitt, A. Abner, A. Recht, F. Vicini, and J. R. Harris. 1994. Relationship of patient age to pathologic features of the tumor and prognosis for patients with stage I or II breast cancer. *Journal of Clinical Oncology* 12 (5):888-894.

Normann, I., and O. Kutz. 2010. Ontology Reuse and Exploration via Interactive Graph Manipulation. *In Proc. of the ISWC Workshop on Ontology Repositories for the Web (SERES-2010)(ISWC-2010)*.

O'Reilly, Marc, Sarah A. Teichmann, and Daniela Rhodes. 1999. Telomerases. *Current Opinion in Structural Biology* 9 (1):56-65.

Olovnikov, A. M. 1996. Telomeres, telomerase, and aging: origin of the theory. *Experimental gerontology* 31 (4):443-448.

Oppenheim, H. Putnam and P. 1991. Unity of Science as a Working Hypothesis. In *The philosophy of science*, edited by R. Boyd, P. Gasper and J. D. Trout. Cambridge, Mass.: The MIT Press.

Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. 1997. CATH-a hierarchic classification of protein domain structures. *Structure* 5 (8):1093-1109.

Overton, J. A., C. Romagnoli, and R. Chhem. 2011. Open Biomedical Ontologies applied to prostate cancer. *Applied Ontology* 6 (1):35-51.

Palmer, Julie R., Lauren A. Wise, Elizabeth E. Hatch, Rebecca Troisi, Linda Titus-Ernstoff, William Strohsnitter, Raymond Kaufman, Arthur L. Herbst, Kenneth L. Noller, Marianne Hyer, and Robert N. Hoover. 2006. Prenatal Diethylstilbestrol Exposure and Risk of Breast Cancer. *Cancer Epidemiology Biomarkers & Prevention* 15 (8):1509-1514.

Patel, V. L., D. R. Kaufman, and S. Magder. 1991. Causal explanation of complex physiological concepts by medical students. *International journal of science education* 13 (2):171-185.

Patel, V. L., A. W. Kushniruk, S. Yang, and J. F. Yale. 2000. Impact of a computer-based patient record system on data collection, knowledge organization, and reasoning. *Journal of the American Medical Informatics Association* 7 (6):569-585.

Patel, V.L., A.E. Evans and G.J. Groen. 1989. Biomedical knowledge and clinical reasoning. In *Cognitive Science in Medicine: Biomedical Modeling*, edited by A. E. David and P. Vilma: MIT Press.

Pece, S., D. Tosoni, S. Confalonieri, G. Mazzarol, M. Vecchi, S. Ronzoni, L. Bernard, G. Viale, P. G. Pelicci, and P. P. Di Fiore. 2010. Biological and molecular heterogeneity of breast cancers correlates with their cancer stem cell content. *Cell* 140 (1):62-73.

Perou, Charles M., Therese Sorlie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Christian A. Rees, Jonathan R. Pollack, Douglas T. Ross, Hilde Johnsen, Lars A. Akslen, Oystein Fluge, Alexander Pergamenschikov, Cheryl Williams, Shirley X. Zhu, Per E. Lonning, Anne-Lise Borresen-Dale, Patrick O. Brown, and David Botstein. 2000. Molecular portraits of human breast tumours. *Nature* 406 (6797):747-752.

Perry, N., M. Broeders, C. de Wolf, S. Törnberg, R. Holland, and L. von Karsa. 2008. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition-summary document. *Annals of Oncology* 19 (4):614-622.

Peto, R. 1982. Statistical aspects of cancer trials. In *The treatment of cancer* edited by H. K. E. London: Chapman & Hall.

Pharoah, Paul D. P., Nicholas E. Day, Stephen Duffy, Douglas F. Easton, and Bruce A. J. Ponder. 1997. Family history and the risk of breast cancer: A systematic review and meta-analysis. *International Journal of Cancer* 71 (5):800-809.

Piccart, Martine J., Hung, Mien-Chie, Solin, Lawrence J., Cardoso, Fatima und Wood, William C. , ed. 2006. *Breast Cancer and Molecular Medicine: Towards Tailored Approaches*: Springer-Verlag Berlin Heidelberg.

Pike, M. C., M. D. Krailo, B. E. Henderson, J. T. Casagrande, and D. G. Hoel. 1983. 'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer. *Nature* 303 (5920):767-770.

Pike, Malcolm C., Darcy V. Spicer, Laila Dahmoush, and Michael F. Press. 1993. Estrogens, Progestogens, Normal Breast Cell Proliferation, and Breast Cancer Risk. *Epidemiologic Reviews* 15 (1):17-30.

Pinder, S. E., G. C. Harris, and C. W. Elston. 2008. The role of the pathologist in assessing prognostic factors for breast cancer. In *Prognostic and Predictive Factors in Breast Cancer*, edited by R. A. Walker and A. M. Thompson.

Pisanelli, D. M. 2004. Biodynamic ontology: applying BFO in the biomedical domain. *Ontologies in medicine* 102:20.

Popper, K. R. 1959. The logic of scientic discovery. *London: Hutchinson*.

Potochnik, Angela. 2010. A Neurathian Conception of the Unity of Science. *Erkenntnis* 74 (3):305-319.

Preston, D. L., A. Mattsson, E. Holmberg, R. Shore, N. G. Hildreth, and J. D. Boice Jr. 2002. Radiation effects on breast cancer risk: a pooled analysis of eight cohorts. *Radiation research* 158 (2):220-235.

Pritchard, Duncan. 2006. *What is This Thing Called Knowledge?*: Routledge.

Pusztai, Lajos, Chafika Mazouni, Keith Anderson, Yun Wu, and W. Fraser Symmans. 2006. Molecular Classification of Breast Cancer: Limitations and Potential. *Oncologist* 11 (8):868-877.

Putnam, Hilary. 1973. Meaning and reference. *Journal of Philosophy* 70 (19):699-711.

———. 1975. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7:131-193.

Quine, W. V. 1948. On what there is. *Review of Metaphysics* 2:21--38.

———. 1951. Main trends in recent philosophy: Two dogmas of empiricism. *The Philosophical Review*:20-43.

Rector, A. 2006. Model of Meaning & Model of Data structure, edited by http://www.cs.man.ac.uk/~rector/presentations/KRMED-2006/KRMED-2006-rector.pdf.

Rector, A. L. 1999. Clinical Terminology: Why is it so hard? *Methods of information in medicine* 38 (4/5):239-252.

———. 2003. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. *Proceedings of the 2nd international conference on Knowledge capture*:121-128.

Rector, A. L., and W. A. Nowlan. 1994. The GALEN project. *Computer methods and programs in biomedicine* 45 (1-2):75-78.

Rector, A. L., W. A. Nowlan, and S. Kay. 1991. Foundations for an electronic medical record. *Methods of information in medicine* 30 (3):179-86.

Reichenbach, H. 1951. The rise of scientific philosophy: Berkeley: University of California Press.

Reiter, R. 1980. A logic for default reasoning. *Artificial intelligence* 13 (1-2):81-132.

Rheinberger, H.J., and S. Müller-Wille. 2012. *Gene*. SEP, Apr 26, 2011, 2004 [cited 2012]. Available from http://plato.stanford.edu/entries/gene/.

Ries, L.A.G., Eisner, M.P. 2007. Cancer of the female breast. In *SEER Survival Monograph: Cancer Survival Among Adults: U.S. SEER Program, 1988-2001, Patient and Tumor Characteristics*, edited by L. A. G. Ries, Keel, G.E, Eisner M.P, et al. : National Cancer Institute, Bethesda, MD. NIH publication 07-6215.

Röhl, J. Mechanisms in biomedical ontology. *Journal of Biomedical Semantics* 3 (Suppl 2):S9.

Rosenberg, A. 1994. *Instrumental biology, or, The disunity of science*: University of Chicago Press.

Rosenbloom, S. T., R. A. Miller, K. B. Johnson, P. L. Elkin, and S. H. Brown. 2006. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *Journal of the American Medical Informatics Association* 13 (3):277-288.

Rosse, C., and J. L. V. Mejino. 2003. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics* 36 (6):478-500.

Rousseau, D. M., S. B. Sitkin, R. S. Burt, and C. Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of management review* 23 (3):393-404.

Sadegh-Zadeh, K. 2011. The Logic of Diagnosis. In *Philosophy of Medicine*, edited by F. Gifford: Elsevier.

Sarkari, Feroz, Yi Sheng, and Lori Frappier. 2010. USP7/HAUSP Promotes the Sequence-Specific DNA Binding Activity of p53. *PLoS ONE* 5 (9):e13040.

Sassower, R., and M. A. Grodin. 1987. Scientific uncertainty and medical responsibility. *Theoretical Medicine and Bioethics* 8 (2):221-234.

Schaffner, Kenneth F. 1993. *Discovery and Explanation in Biology and Medicine*: University of Chicago Press.

Schairer, Catherine, Jay Lubin, Rebecca Troisi, Susan Sturgeon, Louise Brinton, and Robert Hoover. 2000. Menopausal Estrogen and Estrogen-Progestin Replacement Therapy and Breast Cancer Risk. *JAMA: The Journal of the American Medical Association* 283 (4):485-491.

Scharfenberg, F. J., F. X. Bogner, and S. Klautke. 2007. Learning in a gene technology laboratory with educational focus: Results of a teaching unit with authentic experiments. *Biochemistry and Molecular Biology Education* 35 (1):28-39.

Schön, D. A. 1991. *The reflective practitioner: How professionals think in action*. London: Avebury.

Schulz, S., R. Cornet, and K. Spackman. 2011. Consolidating SNOMED CT's ontological commitment. *Applied Ontology* 6 (1):1-11.

Sengupta, K., A. Krisnadhi, and P. Hitzler. 2011. Local closed world semantics: Grounded circumscription for OWL. *The Semantic Web-ISWC 2011* 7031:617-632.

Shavers, Vickie L., Linda C. Harlan, and Jennifer L. Stevens. 2003. Racial/ethnic variation in clinical presentation, treatment, and survival among breast cancer patients under age 35. *Cancer* 97 (1):134-147.

Shay, J. W., and S. Bacchetti. 1997. A survey of telomerase activity in human cancer. *European Journal of Cancer* 33 (5):787-791.

Sheth, Amit, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy Finin, Krishnaprasad Thirunarayan, Faezeh Ensan, and Weichang Du. 2008. An Interface-Based Ontology Modularization Framework for Knowledge Encapsulation. In *The Semantic Web - ISWC 2008*: Springer Berlin / Heidelberg.

Shi, S. R., C. Liu, B. M. Balgley, C. Lee, and C. R. Taylor. 2006. Protein extraction from formalin-fixed, paraffin-embedded tissue sections: quality evaluation by mass spectrometry. *Journal of Histochemistry & Cytochemistry* 54 (6):739.

Shimoyama, M., G. T. Hayman, S. J. F. Laulederkind, R. Nigam, T. F. Lowry, V. Petri, J. R. Smith, S. J. Wang, D. H. Munzenmaier, and M. R. Dwinell. 2009. The rat genome database curators: who, what, where, why. *PLoS computational biology* 5 (11):e1000582.

Shvaiko, P., and J. Euzenat. 2013. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering* 25 (1):158 - 176.

Singletary, S. E., and J. L. Connolly. 2006. Breast cancer staging: working with the sixth edition of the AJCC Cancer Staging Manual. *CA: A Cancer Journal for Clinicians* 56 (1):37-47.

Sintonen, Matti. 2005. Scientific Explanation: Conclusiveness Conditions on Explanation-Seeking Questions. *Synthese* 143 (1):179-205.

Sioutos, N., S. Coronado, M. W. Haber, F. W. Hartel, W. L. Shaiu, and L. W. Wright. 2007. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics* 40 (1):30-43.

Slamon, D. J., G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire. 1987. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235 (4785):177.

Smith, B. 2004. Beyond concepts: ontology as reality representation. Paper read at Proceedings of the third international conference on formal ontology in information systems (FOIS 2004).

Smith, B., and W. Ceusters. 2010. Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Applied Ontology* 5 (3):139-188.

Smith, C. L., C. A. W. Goldsmith, and J. T. Eppig. 2004. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology* 6 (1):R7.

Sobin, L. H., M. K. Gospodarowicz, and C. Wittekind. 2009. *TNM Classification of Malignant Tumours*: John Wiley & Sons.

Sojic, A., and O. Kutz. 2012. Open biomedical pluralism: formalising knowledge about breast cancer phenotypes. *Journal of Biomedical Semantics* 3 (Suppl 2):S3.

Soldatova, L. N., A. Rzhetsky, and R. D. King. 2011. Representation of research hypotheses. *J Biomed Semantics* 2 (Suppl 2):S9.

Sørlie, Therese, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Thor Thorsen, Hanne Quist, John C. Matese, Patrick O. Brown, David Botstein, Per Eystein Lønning, and Anne-Lise Børresen-Dale. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* 98 (19):10869-10874.

Sotiriou, Christos, and Lajos Pusztai. 2009. Gene-Expression Signatures in Breast Cancer. *New England Journal of Medicine* 360 (8):790-800.

Sotiriou, Christos, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, Christine Desmedt, Denis Larsimont, Fatima Cardoso, Hans Peterse, Dimitry Nuyten, Marc Buyse, Marc J. Van de Vijver, Jonas Bergh, Martine Piccart, and Mauro Delorenzi. 2006. Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. *J. Natl. Cancer Inst.* 98 (4):262-272.

Sowa, J. F. 2000. *Knowledge representation: logical, philosophical, and computational foundations*. Vol. 13: MIT Press.

Spyns, P., R. Meersman, and M. Jarrar. 2002. Data modelling versus ontology engineering. *ACM SIGMod Record* 31 (4):12-17.

Stanley, J. 2005. *Knowledge and Practical Interests*: OUP Oxford.

———. 2007. *Language in context: selected essays*: Oxford University Press.

Sterelny, K., and P. E. Griffiths. 1999. *Sex and death: An introduction to philosophy of biology*: University of Chicago Press.

Stingl, John, and Carlos Caldas. 2007. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nat Rev Cancer* 7 (10):791-799.

Stockler, Martin , Duric, Vlatka , and Coates, Alan S. 2006. Patients' Preferences: What Makes Treatments Worthwhile? In *Breast Cancer and Molecular Medicine: Towards Tailored Approaches*, edited by M. J. Piccart, Hung, Mien-Chie, Solin, Lawrence J., Cardoso, Fatima und Wood, William C. : Springer-Verlag Berlin Heidelberg.

Strasser, Bruno J. 2012. Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1):85-87.

Strawson, P. F. 1964. Identifying reference and truth-values. *Theoria* 30 (2):96-118.

Suárez, M. 2010. Scientific representation. In *Blackwell's Philosophy Compass*.

Suárez, Mauricio. 2003. Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science* 17 (3):225-244.

———. 2004. An inferential conception of scientific representation. *Philosophy of Science* 71 (5):767-779.

Swoyer, C. 1991. Structural representation and surrogative reasoning. *Synthese* 87 (3):449-508.

Symons, John. 2010. Ontology and Methodology in Analytic Philosophy. In *Theory and applications of ontology: Philosophical perspectives*, edited by R. Poli, Seibt, Johanna Springer Verlag.

Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, and P. Bork. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* 39 (suppl 1):D561.

Tata, J. R. 2005. One hundred years of hormones. *EMBO reports* 6 (6):490.

Taylor, Chris F., Dawn Field, Susanna-Assunta Sansone, Jan Aerts, Rolf Apweiler, Michael Ashburner, Catherine A. Ball, Pierre-Alain Binz, Molly Bogue, Tim Booth, Alvis Brazma, Ryan R. Brinkman, Adam Michael Clark, Eric W. Deutsch, Oliver Fiehn, Jennifer Fostel, Peter Ghazal, Frank Gibson, Tanya Gray, Graeme Grimes, John M. Hancock, Nigel W. Hardy, Henning Hermjakob, Randall K. Julian, Matthew Kane, Carsten Kettner, Christopher Kinsinger, Eugene Kolker, Martin Kuiper, Nicolas Le Novere, Jim Leebens-Mack, Suzanna E. Lewis, Phillip Lord, Ann-Marie Mallon, Nishanth Marthandan, Hiroshi Masuya, Ruth McNally, Alexander Mehrle, Norman Morrison, Sandra Orchard, John Quackenbush, James M. Reecy, Donald G. Robertson, Philippe Rocca-Serra, Henry Rodriguez, Heiko Rosenfelder, Javier Santoyo-Lopez, Richard H. Scheuermann, Daniel Schober, Barry Smith, Jason Snape, Christian J. Stoeckert, Keith Tipton, Peter Sterk, Andreas Untergasser, Jo Vandesompele, and Stefan Wiemann. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotech* 26 (8):889-896.

Thagard, Paul. 1998. Explaining Disease: Correlations, Causes, and Mechanisms. *Minds and Machines* 8 (1):61-78.

Thomas, G. A., and R. C. F. Leonard. 2009. How age affects the biology of breast cancer. *Clinical oncology* 21 (2):81-85.

Tordai, A., A. Ghazvinian, J. van Ossenbruggen, M. Musen, and N. and Natasha. 2010. Ontology Reuse and Exploration via Interactive Graph Manipulation. In *In Proc. of the ISWC Workshop on Ontology Repositories for the Web (SERES-2010)(ISWC-2010)*.

Trißl, S., and N. Reinsch. 2011. Developing an Animal Trait Ontology - Why Phenotype Ontologies are not enough. In *In Proceedings of the 3rd Workshop of GI Workgroup 'Ontologies in Biomedicine and Life Sciences' (OBML)*, edited by H. Herre, Hoehndorf, R, Loebe, F.: Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE), Universität Leipzig.

Tu, S. W., O. Bodenreider, C. Çelik, C. G. Chute, S. Heard, R. Jakob, G. Jiang, S. Kim, E. Miller, and M. M. Musen. 2010. A content model for the ICD-11 revision: Technical Report BMIR-2010-1405, Stanford Center for Biomedical Informatics Research.

Tudorache, T., S. Falconer, N. Noy, C. Nyulas, T. Üstün, M. A. Storey, and M. Musen. 2010. Ontology Development for the Masses: Creating ICD-11 in WebProtégé. In *Knowledge Engineering and Management by the Masses*, edited by P. Cimiano, H. Pinto, G. Burel, A. Cano, M. Rowe and A. Sosa: Springer Berlin Heidelberg.

Tyrer, J., S. W. Duffy, and J. Cuzick. 2004. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine* 23 (7):1111-1130.

Uebel, Thomas. *"Vienna Circle" in The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.)* (Summer 2012 Edition) 2011 [cited. Available from URL = <http://plato.stanford.edu/archives/sum2012/entries/vienna-circle/>.

Uschold, M., and M. Gruninger. 2004. Ontologies and semantics for seamless connectivity. *ACM SIGMod Record* 33 (4):58-64.

van't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, and A. T. Witteveen. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415 (6871):530-536.

van 't Veer, Laura J., and Rene Bernards. 2008. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 452 (7187):564-570.

van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. 2002. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347:1999 - 2009.

van der Groep, P., E. van der Wall, and P. J. van Diest. 2009. Pathology of hereditary breast cancer. *Cellular Oncology*:1-18.

Van Der Linden, H., D. Kalra, A. Hasman, and J. Talmon. 2009. Inter-organizational future proof EHR systems: A review of the security and privacy related issues. *International journal of medical informatics* 78 (3):141-160.

van Fraassen, B. C. V. 1980. *The Scientific Image*. Oxford: Oxford University Press.

———. 2008. *Scientific representation: paradoxes of perspective*: Clarendon Press.

Vargo-Gogola, Tracy, and Jeffrey M. Rosen. 2007. Modelling breast cancer: one size does not fit all. *Nat Rev Cancer* 7 (9):659-672.

Veronesi, Umberto, Stefano Zurrida, Aron Goldhirsch, Nicole Rotmensz, and Giuseppe Viale. 2009. Breast Cancer Classification: Time for a Change. *Journal of Clinical Oncology* 27 (15):2427-2428.

Viale, Giuseppe, Mariacristina Ghioni, and Mauro G. Mastropasqua. 2010. Traditional molecular markers and response to adjuvant endocrine or trastuzumab-based therapies. *Current Opinion in Oncology* 22 (6):541-546.

Walford, Cornelius. 1878. Early Bills of Mortality. *Transactions of the Royal Historical Society* 7:212-248.

Walker, R. A. 2003. Assessment of breast cancer grading using the Nottingham combined histological grading system. *Prognostic and Predictive Factors in Breast Cancer*.

Walker, R. A., and A. M. Thompson, eds. 2008. *Prognostic and Predictive Factors in Breast Cancer, Second Edition*: Taylor & Francis.

Wang, R., S. W. Lagakos, J. H. Ware, D. J. Hunter, and J. M. Drazen. 2007. Statistics in medicine-reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* 357 (21):2189-2194.

Wears, R. L., and M. Berg. 2005. Computer technology and clinical work. *JAMA: The Journal of the American Medical Association* 293 (10):1261.

Weigel, Marion T., and Mitch Dowsett. 2010. Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocr Relat Cancer* 17 (4):R245-262.

Weigelt, B., F. C. Geyer, and J. S. Reis-Filho. 2010. Histological types of breast cancer: how special are they? *Molecular Oncology* 4 (3):192-208.

WHOFIC Network. 2007. *Production of ICD-11: The overall revision process*. WHO 2007 [cited 13.02.2012 2007]. Available from http://www.who.int/classifications/icd/ICDRevision.pdf.

Wiggins, David. 1980. *Sameness and Substance*: Oxford, Blackwell.

Williams, M. 1996. *Unnatural doubts: Epistemological realism and the basis of scepticism*: Princeton Univ Pr.

Williamson, J. 2006. Causal pluralism versus epistemic causality. *Philosophica-Gent-* 77:69.

Wimsatt, W. C. 2007. *Re-engineering philosophy for limited beings: piecewise approximations to reality*: Harvard University Press.

Wittgenstein, L., G. E. M. Anscombe, P. M. S. Hacker, and J. Schulte. 2009. *Philosophical investigations*: Wiley-Blackwell.

Woodfield, A. 2007. Public Words Considered as Vehicles of Thinking. In *Explaining the mental: naturalist and non-naturalist approaches to mental acts and processes*, edited by C. Penco, M. Beaney and M. Vignolo.

Woods, J., and A. Rosales. 2010. Virtuous Distortion. *Model-Based Reasoning in Science and Technology* 314:3-30.

World Health Organization. 2005. *ICD-10: International Statistical Classification of Diseases and Related Health Problems*: World Health Organization.

Wright, C. 1992. *Truth and objectivity*. Cambridge, MA: Harvard University Press

Yon Rhee, Seung, Valerie Wood, Kara Dolinski, and Sorin Draghici. 2008. Use and misuse of the gene ontology annotations. *Nat Rev Genet* 9 (7):509-515.

Zurrida, Stefano, and Umberto Veronesi. 2011. A new TNM classification for breast cancer to meet the demands of the present and the challenges of the future. *Women's Health* 7 (1):41-49.