The Knowledge Engineering Review, Vol. 24:3, 205–224. © 2009, Cambridge University Press
doi:10.1017/S0269888909990026 Printed in the United Kingdom

Ontological modeling at a domain interface: bridging clinical and biomolecular knowledge

GIANLUCA COLOMBO[1], DANIELE MERICO[1,2,3], ZOLTÁN NAGY[4], FLAVIO DE PAOLI[1], MARCO ANTONIOTTI[1] and GIANCARLO MAURI[1]

[1]*Department of Computer Science, Systems and Communication (DISCo), University of Milan—Bicocca, viale Sarca, 336/14, 20126 Milan, Italy;*
e-mail: gianluca.colombo@disco.unimib.it, flavio.depaoli@disco.unimib.it, marco.antioniotti@disco.unimib.it, mauri@disco.unimib.it
[2]*Department of Biomolecular Sciences and Biotechnologies (DSBB), University of Milan, Via Celoria, 26, 20133 Milan, Italy;*
e-mail: daniele.merico@gmail.com
[3]*Terrence Donnelly Centre for Cellular and Biomolecular Research (CCBR), Banting and Best Department of Medical Research, University of Toronto, 160, College Street, M5S 3E1 Toronto, Ontario, Canada*
[4]*Department of Vascular Neurology, Semmelweis University, Huvosvolgyi Street 116, 1021 Budapest, Hungary;*
e-mail: nagy@opni.hu

Abstract

In this paper, we discuss the challenges posed by the NEUROWEB project, as a case study of ontological modeling at a knowledge interface between neurovascular medicine and genomics. The aim of the project is the development of a support system for association studies. We identify the notion of clinical phenotypes, that is, the pathological condition of a patient, as the central construct of the knowledge model. Clinical phenotypes are assessed through the diagnostic activity, performed by clinical experts operating within communities of practice; the different communities operate according to specific procedures, but they also conform to the minimal requirements of international guidelines, displayed by the adoption of a common standard for the patient classification. We develop a central model for the clinical phenotypes, able to reconcile the different methodologies into a common classificatory system. To bridge neurovascular medicine and genomics, we identify the general theory of biological function as the common ground between the two disciplines; therefore, we decompose the clinical phenotypes into elementary phenotypes with a homogeneous physiological background, and we connect them to the biological processes, acting as the elementary units of the genomic world.

1 Introduction

Ontologies are a popular research topic in various communities such as knowledge engineering (Kitamura & Mizoguchi, 2003; Gomez-Perez *et al*., 2003), natural language processing (Zaihrayeu *et al*., 2007), cooperative information systems (Benassi *et al*., 2004; Guizzardi, 2007), intelligent information integration (Guarino, 1998; Bouquet *et al*., 2002; Lenzerini, 2002), and knowledge management (Lueg, 2002; Bergamaschi *et al*., 2007). In computer science, ontologies can be defined as computational models which describe the kind of entities, properties of entities, and relations between entities that exist in a specified domain of knowledge; under an artificial intelligence (AI) perspective, they can be regarded as *content theories*, in opposition to *mechanism theories* (Chandrasekaran *et al*., 1999).

One of the most popular end-use of ontologies in information technology is to create an agreed-upon vocabulary for information exchange within a domain. For instance, interoperability among autonomously developed databases (DBs) can be significantly hindered by the absence of uniform semantics; reconciliation can be accomplished through the use of a reference ontology to which the semantics of the individual resources can be related (Benassi *et al*., 2004).

Standard interchange languages (XML, OWL, RDF) were developed to enable the establishment of Web services to retrieve and exchange ontology-encoded information through the Web (W3C, Web Ontology Language). Nonetheless, the choice of specific representation languages for the formal ontology encoding, and the technological issues of database interoperability, do not wear out the whole range of problems concerning ontological modeling (Gruber, 1995; Guarino, 1995; Thomasson, 2004). Despite the effort for defining a standard semantic, experts seem to resist to any attempt of homogenization. Partly, this is due to practical problems, but there are also theoretical reasons why this is not easily accepted and not even desirable. In fact, lots of cognitive and organizational studies show that there is a close relationship between knowledge and identity. Knowledge is not simply a matter of accumulating 'true sentences' about the world, but it is also a matter of interpretation schemes—for example, paradigms (Kuhn, 1970), contexts (Benerecetti *et al.*, 2000), mental models (Johnson-Laird, 1983), perspectives (Boland & Tenkasi, 1995)—which allow people to make sense of what they know. All these mental structures are an essential part of what people know, as each of them provides an alternative lens through which reality can be read.

A simple example of that divergence is the notion of herbivore and carnivore, an apparently intuitive, simplistic, and unambiguous category, which can be understood and successfully used even by children. Considering the standpoint of a farmer, or the zoo personnel responsible for the animal feeding, an animal is categorized as carnivore or herbivore according to its feeding habits, based either on meat and very few vegetables (carnivore), or only vegetables and no meat (herbivore), or a balanced amount of both (omnivorous); extending this criterion to human consumers (their categorizer being the food-retail personnel or the nutritionist), a vegan could be classified as herbivore: albeit unable to digest grass, his feeding is constituted only by vegetables. However, considering the standpoint of the paleontologist, or the veterinary dentist, the anatomy of the teeth set is the crucial criterion. According to that criterion, an animal is permanently carnivore, herbivore or omnivorous from its birth till its death, notwithstanding the feeding behavior exhibited during a certain time span; as a consequence, a vegan, having the teeth set of an omnivorous, is not herbivore. The case is analogous assuming a different anatomic criterion, for example, the structure of the full digestive system[1].

The example displays that herbivore, omnivorous, and carnivore categories can be assigned according to different underlying criteria, defined within different classification systems. We argue that different classification criteria can be explicitly taken into account in the knowledge representation frame, and categories grounded on different criteria can be decomposed into common elements. Indeed, the feeding habits are enabled by the anatomical and physiological structure (digestive system), but co-determined by other factors (e.g. food availability for animals, or ethical choices for humans). Within this enlarged frame, a vegan can be successfully categorized as having an omnivorous digestive system, and herbivore feeding habit due to an ethical choice to avoid meat, overcoming the diverging classificatory solutions. Besides this simple example, and considering more articulated scientific contexts, the divergence of classificatory systems can be associated with different theories operating within a common discipline, in the context of the same research tradition (Laudan, 1977).[2] The notion of *discipline*, though apparently intuitive, was never formalized by prominent Philosophy of Science scholars such as Kuhn (1970), Feyerabend (1975), Laudan (1977), and Lakatos (1978). Here, a discipline is characterized by (a) the domain of reality being the object of study (e.g. neurovascular pathologies, or the human genome), (b) the

[1] This example was discussed during the tutorial of the Bio-ontologies Special Interest Groups (SIG), within the joint bioinformatics conference ISMB/ECCB (Intelligent Systems for Molecular Biology/European Conference on Computational Biology) 2007, in Wien.

[2] Following (Laudan, 1977), we refer to the notion of Research Tradition as a specialization of Kuhn's *paradigm* concept. While a paradigm casts a general light over science and its methods, relying on a historical and cultural background (e.g. Western Medicine versus Eastern Medicine), a Research Tradition better represents a specific corpus of meta-theories (e.g. Atomism, Materialism) and methodologies (e.g. Inductivism, Operativism) acting in the same paradigm.

aims, (c) the general methodology, (d) the general content theory; beside the aims, all the other components can be specified by different schools. Roughly speaking, these criteria underlie the different aggregations of scientific communities proposed by different Philosophy of Science scholars (Kuhn, 1970; Feyerabend, 1975; Laudan, 1977; Lakatos, 1978). Within the general scientific arena, the social group endorsing a certain theory is commonly termed *School*. Under this epistemological perspective, a School gathers groups of research communities, which share the same methodology of inquiry[3] and a common vision over a specific research area, including the adoption of common classificatory systems, as shown in (Thomasson, 2004). Different schools are unified within a discipline by sharing the same objectives: they address the same problems, but they adopt different frames, leading to diverging solutions; concerning the content theory component, this situation is mirrored by the adoption of different classificatory systems. Under an ontological perspective, diverging content theories, belonging to different schools, can be reconciled by decomposing their categories according to the classificatory criteria adopted, and identifying common elementary units. The crossing of disciplines borders, to support multi-disciplinary research projects, poses even greater challenges than managing the existence of different schools. The same decomposition strategy can be applied to integrate the different content theories, but only at the condition of identifying a common ground between the two disciplines, underlying the different objects of study, aims, and methodologies.

In this paper, we discuss the NEUROWEB project as a case study of the modeling challenges arising in the outlined epistemic scenario. The first part concerns the aims of the project, and the way ontological modeling can contribute to their fulfillment; the second part outlines the different scientific communities involved in the project, and the consequent epistemological challenges; the third part presents the general Knowledge Acquisition (KA) and Knowledge Representation (KR) strategy devised to address those challenges; and the fourth part is devoted to a more detailed analysis of the ontological model.

2 The NEUROWEB project: aims

NEUROWEB is a EU (European Union)-funded, health-care oriented project. Its aim is the development of an IT (information technology) infrastructure to support association studies in the neurovascular domain. As their name suggests, association studies consist in the assessment of the statistical significance of association (i.e. co-occurrence) between a phenotype and genetic, environmental or demographic factors; typical examples are the search for genetic markers associated with an increased susceptibility to type-2 diabetes, or the discovery of a relation between the average number of daily-smoked cigarettes and the incidence of lung cancer. In these examples, the phenotype is always a pathology (type-2 diabetes, lung cancer); however, the notion of phenotype is commonly used also in non-pathological contexts (e.g. blood groups, somatic traits): according to a more general definition, a phenotype is an observable state of an individual organism (Bard *et al*., 2004). Phenotypes can consist of a single parameter (e.g. glycemia, the level of glucose in venous blood), or a structured ensemble of inter-related features (e.g. diabetes is a complex pathological state comprising abnormal glycemia). Within the NEUROWEB project, the phenotype is intended as a pathological state, due to the occurrence of a neurovascular lesion, and diagnosed by a clinical expert; this specific declension will be referred as *clinical phenotype* throughout the paper.

[3] By methodology we mean the corpus of methods (i.e. rules and principles) guiding the sensory and cognitive processes required to achieve the goals of a specific practice. Whereas in AI it is possible to clearly distinguish between a content theory and a mechanism theory (Chandrasekaran *et al*., 1999), human knowledge displays circular relations between the two components: for instance, the use of different 'sensory devices', and different mechanisms to elaborate their results, is normative for any ontological categorization. Although ontological modeling addresses the content theory component, we argue that the methodological aspects should be taken into account, especially when dealing with tacit knowledge.

Association studies require large patient cohorts to be effective. For that reason, the NEUROWEB consortium is constituted by four clinical sites from different EU-member countries, which are recognized excellence centers in the field of neurovascular disorders, with a particular focus on ischemic stroke. However, the amount of available data is not a valuable asset itself, unless data quality and methodological coherence are carefully preserved. Considering the specific NEUROWEB context, the critical point is whether phenotypes are defined and assessed homogeneously across different sites. Indeed, although all sites comply with international guidelines, they also have developed specific diagnostic procedures, and in-depth competencies in different areas of stroke diagnosis (such as imaging, biochemical essays, etc.)[4]. This situation is mirrored by the clinical repositories, which were independently developed in each local site to store patient profiles. Clearly, enforcing a top-down standardization of phenotype assessment is neither practicable nor desirable in the applicative context outlined (Van der Vet & Mars, 1998). Therefore, to address the problem of methodological coherence, the phenotypes were encoded in a specific Reference Ontology, which will be referred as the *NEUROWEB Reference Ontology* (or, shortly, the NEUROWEB ontology); the NEUROWEB ontology enables the NEUROWEB system to operate on different clinical repositories, and to retrieve patients characterized by a given (set of) phenotype(s), formulated according to negotiated criteria.

General-purpose medical ontologies (OBO; The Gene Ontology Consortium; The National Center for Biomedical Ontologies; Galen; PATO; SNOMED-CT) could not be used to meet that end, as they proved unsuitable to represent the specific expert knowledge of the NEUROWEB neurovascular communities (Bard *et al.*, 2004; Bodenreider *et al.*, 2007). The diagnostic process has traditionally resisted many efforts to be fully automatized (Miller, 1994; Sicilia *et al.*, 2009); it is important to point out that the aim of the NEUROWEB infrastructure is not to replace the clinical expert in the diagnostic activity, but rather to reconstruct the patients' clinical phenotypes, according to the evidences collected—and interpreted—by the medical experts during the diagnostic process; the rationale of using explicit and analytical phenotype formulations is to integrate data from different sources preserving methodological coherence.

Association studies within the NEUROWEB project have a special commitment to genotype–phenotype relations [A, B]; specifically, the genotypic component assessed is the value of single nucleotide polymorphisms (SNPs) [C, D] in the patients' genome. Beyond the search for statistically relevant SNPs–phenotype associations, there are valuable functionalities requiring to extend the NEUROWEB Reference Ontology to gene functions, and enabling it to treat phenotype formulations as provided by genomic resources. Conjugating phenotypes, as conceived within neurovascular medicine, with entities belonging to the genomic world, clearly requires to bridge two knowledge domains belonging to different disciplines; whereas clinical medicine is committed to the identification of pathological states (diagnosis) and their treatment (or prevention), molecular biology and molecular genetics are committed to the study of biological mechanisms at the molecule level and to the study of the information encoded by the genome.

3 The NEUROWEB project: epistemological challenges

NEUROWEB clinical sites are communities of neurovascular experts, whose daily work is the examination of patients; in each center, the patients have been previously diagnosed with a stroke in an ordinary hospital, and they are admitted to the center in order to undergo a more detailed diagnostic process, as well as to receive the most suitable treatment. The evidences collected during the diagnostic process include both objective/quantitative data (e.g. cholesterol level), and subjective/qualitative data, that is, filtered by the expert's judgment (e.g. presence of a brain lesion); they are stored in locally run repositories and are available to the other clinicians of the center.

[4] International guidelines should not be regarded as exhaustive formulations of diagnostic procedures, but rather as minimal requirements to be met in order to formulate correct diagnoses.

These data can be further exploited for association studies, the task specifically addressed within the NEUROWEB project, as they enable to define a clinical phenotype. Besides the exchange of clinical records, each clinical community is characterized also by the encoding of their own procedural rules for diagnostics into written documents, termed *Diagnostic Protocols*. As mentioned before, local protocols have to comply with general guidelines, negotiated among top experts all over the world; yet, they retain specific elements, characteristic of the local communities' customs and expertise. Considering the outlined features, the NEUROWEB clinical communities can be conveniently regarded as Research Communities, each distinguished from the others by different tradition of problem solving (i.e. diagnostic activity), but gathered together by a common vision over the problems to be solved and the ways of inquiry them.

As remarked by Laudan (1977), the knowledge cycle of a Research Community typically originates a domain conceptualization (a domain vision), usually implicit and not formalized (Davenport & Prusack, 1998; Holsapple & Joshi, 2001). These elements can be clearly identified in the NEUROWEB clinical communities:

- The examination of patients involves collective forms of working (e.g. advises from elder colleagues, collective discussions over anomalous cases);
- The examination of patients, followed by the diagnosis formulation, and the treatment decision, is a problem-solving activity; each problem-solving case undergoes reification into a clinical record;
- The Diagnostic Protocol can be regarded as a procedural manual, and it is the final product of the negotiation phase.

The domain conceptualization of a community may remain implicitly encoded in its routines and written documents, or it may undergo a process of structuring into a *Representational Artifact*. Examples of Representational Artifacts are, on the side of the content theories, classifications of pathologies; and on the side of mechanism theories, diagnostic algorithms; that is, decision trees in which the if-then-else clauses are formulated in Natural Language (NL) (Bowker *et al.*, 1997; Bandini *et al.*, 2003; Smith *et al.*, 2006). The capital difference between a narration, or a manual, and a Representational Artifact is the presence of an explicit structure; the structuring process consists in transferring the knowledge content from the NL-encoding to the structure of the artifact. Though the Representational Artifact is not directly encoded in a formal language, its construction is the necessary condition for knowledge formalization[5]. Representational Artifacts always have residual semantic elements, encoded as NL, and their structure may be partial or unsystematic (Smith *et al.*, 2006). Therefore, it is not practicable to treat them as the sole knowledge asset. This is in agreement with consolidated knowledge engineering (KE) literature, stating that expert knowledge is primarily detained by subjects rather than artifacts or written documents (Nonaka & Takeuchi, 1995; Studer *et al.*, 1998; Guida & Berini, 2000), and consequently postulating the necessity of KA from the experts preliminary to the KR phase (Preece *et al.*, 2001). We argue that Representational Artifacts are valuable assets to guide the KA activity, but, as knowledge sources, they cannot replace the experts' knowledge. In order to support association studies, we are specifically interested in Representational Artifacts (in our case, classificatory systems for clinical phenotypes) able to guide the clustering of patients. In the context of the NEUROWEB consortium, the structuring process does not occur within the local communities, but rather as a negotiation among different research communities. The TOAST (Trial of Org 10172 in Acute Stroke Treatment)

[5] A Research Community can be epistemologically conceived as the crossing over between a *scientific* community that adheres to a hypothetical-deductive paradigm of scientific thinking, and a *Community of Practice* (Brown & Duguid, 1991; Wenger, 1998; Hildreth *et al.*, 2000) that does not strictly adhere to it. Whereas hypothetical-deductive knowledge is often encoded in *formal constructs* (e.g. in physics, Newton's Law of Mechanics; in Chemistry, the Periodic Table), the knowledge of a *Community of Practice* tends to be implicit, that is, *tacit knowledge* (Nonaka & Takeuchi, 1995), or encoded as a semi-formal *Representational Artifact*.

classificatory system (Adams *et al.*, 1993; Lee *et al.*, 2000; Goldstein *et al.*, 2001; Ay *et al.*, 2005) is the Representational Artifact fulfilling, at best, the NEUROWEB needs. It was developed by the international neurovascular community to classify stroke patients on a diagnostic basis, and it was already adopted by most of the clinical sites before the existence of the NEUROWEB consortium. The adoption of the same Representational Artifact (the TOAST) testifies for the presence of a common conceptual and methodological ground across the different NEUROWEB Research Communities; therefore, it is possible to regard them as members of a neurovascular school. As a matter of fact, the TOAST assumes specific diagnostic criteria grounded not only on a common conceptualization of clinical phenotypes (content theory component), but also on shared methodological principles (mechanism theory component). These aspects will be better elucidated in the next session. So far, two major modeling issues have to be managed:

- The TOAST is a classificatory system with a strong NL-component; to be represented with a formal language, it requires a further structuring effort (i.e. refinement of the conceptual model); that activity cannot be carried out without involving the clinical experts from the Research Communities.
- Managing the difference among the Research Communities, which requires to conjugate two different objectives:
  - federating their clinical repositories, and coherently formulating the TOAST clinical phenotypes on that basis;
  - besides the TOAST phenotypes, allowing also the formulation of customized clinical phenotypes, in order to convey the Research Communities' specificities.

As we briefed in the Introduction, to support genotype–phenotype association studies, it is valuable to connect clinical phenotypes to genomic entities; we argue that such an operation requires bridging two knowledge domains belonging to different disciplines. In this case, the two disciplines are clinical medicine, specifically vascular neurology, and molecular biosciences (the convergence of molecular biology, molecular genetics and genomics). The divergence between the two disciplines concerns multiple epistemological dimensions:

- The *object of study* is the same, the human body (though biosciences are not restricted to the study of humans, but also other life forms as well). However, the focus is on different facets, occupying different granularity levels: medicine mainly addresses the level of organs, anatomical parts, and global body parameters, though it also takes into account the tissue, cell, and molecule levels; molecular biosciences addresses the molecular level, with a particular focus on the genetic information encoded by the DNA molecule, and the processes centered on it. Therefore, clinical medicine and molecular biosciences have different reference levels of granularity.
- The two disciplines have different *aims*: clinical medicine aims at determining the pathological states of patients (diagnosis), in order to formulate a prognosis and a suitable treatment—in brief, to save lives; molecular biosciences aim at elucidating the molecular mechanism of biological organisms.
- The *methodology* is different as well (e.g. different experimental techniques), as a consequence of the different facets in the object of study, and the different aims.
- Finally, the *content theory* (the catalogue of entities, their properties and their relations) is different as well. This difference is influenced by the object of study and the methodology.

The genomic information of interest for NEUROWEB (in our case, gene functions and phenotypes) is typically stored within general-purpose repositories, which are not the expression of a research community with a specific methodology; as a consequence, the influence of the methodological component in the genomic setting is weaker than in the clinical setting. However, that does not imply that methodological problems do not occur in the genomic field; it simply means that they are not explicitly managed in the publicly available resources. For instance, HGMD (Human Gene Mutation Database) (Stenson *et al.*, 2003), a repository for genotype–phenotype associations,

includes phenotypes referring to pathologies, thus underpinning a clinical perspective, altogether with phenotypes referring to the alteration of specific molecular functions and processes, thus underpinning a molecular bioscience perspective; the former, however, are not treated with the level of detail adopted in the NEUROWEB project, and the methodological problem of how the pathology is diagnosed is not addressed. Nonetheless, even in this setting, significant discrepancies need to be managed, concerning the different granularity and the different aims. We will address these problems with full details in the next section, altogether with the solutions proposed.

As a whole, we argue that an explicit analysis of the epistemological setting constitutes:

- a valuable contribution to ontological modeling, by elucidating the role of the disciplines aims and methodology in shaping the content theory;
- a fulfillment of the KE methodology for ontological modeling; we specifically argue that a clear picture of the epistemological setting is extremely valuable to efficiently organize the KA and to drive the KR task[6].

4 The ontological modeling strategy adopted

The core concept of the NEUROWEB Reference Ontology is the phenotype. The ontology is primarily committed to the representation of clinical phenotypes, federating the different clinical communities without compromising methodological coherence. Decomposing the clinical phenotypes into general domain components enables to extend the NEUROWEB Reference Ontology to genomic entities, bridging two different knowledge domains.

This section is organized into two parts: in the first part, we address the ontological model for clinical phenotypes; in the second part, we discuss how that model was suitably extended to take into account genomic entities.

4.1 A strategy for clinical phenotypes

The NEUROWEB project is committed to associations' studies, and the primary computational task is the retrieval of patients sharing a common phenotype. As we discussed in the previous sessions, this task is not straightforward. On the one hand, each site independently collects and organizes the clinical evidences required to formulate a phenotype, mirroring the methodological specificities of the local Research Community[7] (see Figure 1). On the other hand, however, the clinical communities accept the federation under a common standard for phenotype classification, the TOAST. All in all, the solution we adopted is an ontological model formulating clinical phenotypes on the basis of clinical evidences stored in the repositories. Given an input phenotype, the ontological model enables the query system to retrieve patients from different repositories; the adoption of an ontological phenotype formulation, opposed to a looser query system, grants the application of coherent criteria for phenotype formulation across different repositories. In the beginning, we devised an initial two-layered model. The lower layer is the Core Data-Set (CDS), a set of common indicators from the local repositories, which clinicians consider essential for stroke

[6] We acknowledge that certain schools of ontological engineering advocate the representation of 'reality', avoiding to mingle with epistemological complications (Smith, 2004). The theoretical debate between *Conceptualism* and *Realism* is a long-standing issue in philosophy; realism advocates the principle of identity as 'objective in its essence'; conceptualism argues for the existence of *conceptualizations* of reality, shared across different subjects and communities. It is out of the scopes for this paper to address that debate, and rule out the ideal approach for ontological modeling. Nonetheless, we argue that the refusal to recognize the differences between conceptualizations, and their underlying rationale, is a major weakness of the realistic approach. Assuming a generally valid reality, as in the case of general-purpose ontologies, may also neglect the specific elements of the expert knowledge accrued within a Research Community.

[7] In the Semantic Web community, this problem would be typically identified as a *Data Integration* problem (Benassi *et al*., 2004). Here we specifically stress the methodological implications of divergence, hence the need of a knowledge-accurate ontological model to support the integration task.

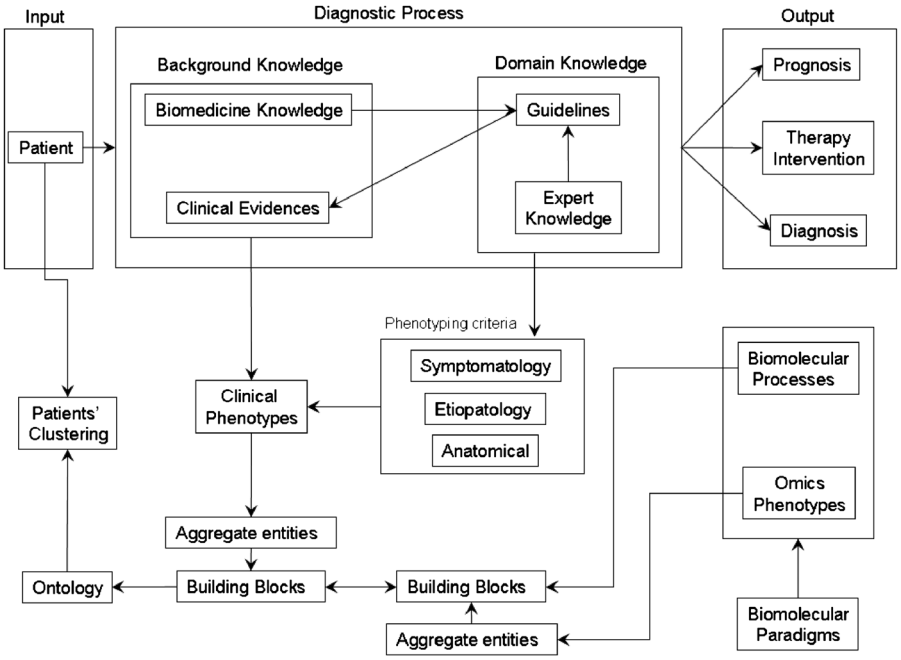

Figure 1 The figure depicts the main epistemological dimensions of knowledge identified within the NEUROWEB project, the TOAST classification criteria exploited for the clinical phenotype decomposition, and the ontological modeling strategy adopted to bridge the genomic entities to the clinical phenotypes

diagnosis (e.g. anterior cerebral artery lesion, blood pressure). The CDS enables to connect clinical evidences having the same semantic across the different communities. The CDS is the result of an iterative negotiation process among the NEUROWEB clinical communities, supervised and aided by the NEUROWEB Knowledge Engineers and IT personnel.

Under an IT perspective, the CDS alone would be enough to support a query system, operating on the integrated ensemble of the four clinical repositories; in that setting, the task of phenotype formulation would be executed by the user by assembling different CDS indicators into a query formula. We can reasonably expect that there is a set of clinical phenotypes used on a regular basis for association studies—the TOAST phenotypes, in the case of the NEUROWEB consortium, as previously discussed; a query system relying only on the CDS would oblige the user to reformulate the phenotypes for every new access to the system, a solution which would be time-consuming and error-prone. For that reason, we collect the TOAST phenotypes in a taxonomy of stroke types and subtypes, termed Top Phenotypes. The Top Phenotypes constitute the second layer of the model, and they are connected to the CDS indicators via a definition formula. The formula is structured as a conjunction/disjunction of criteria on the CDS indicators, expressed as equality/inequality bounds, or quantitative ranges to be satisfied. An example is: (Blood pressure > 50) AND (Anterior cerebral artery lesion = yes). The existence of a negotiated, but centrally encoded, phenotype formulation establishes a methodologically coherent federation of the different communities' resources. Local practices are reconciled within this model, as different sets of CDS indicators can be used to identify a common phenotype. The easiest case occurs when the same clinical evidence can be assessed using different technologies but achieving the same degree of probatory strength (e.g. severe stenosis in the internal carotid artery can be determined by a duplex AND a computed tomography angiography scan, OR by a duplex AND a magnetic resonance angiogram scan). The way phenotypes are encoded strictly resembles the typical formulation of a query, and thus can be easily handled by a clinical expert; for that reason, the two-layered model was successfully used for the KA activity.

It is important to reassert that NEUROWEB is not a system for automatic diagnosis. If the CDS indicators were objective parameters, collecting all the clinical evidences for a patient, and

then applying the phenotype formulation, as encoded by the NEUROWEB Ontology, would constitute a system for automatic diagnosis. That is not the case, as only some of the CDS indicators are objective, for example, the total level of cholesterol. Other CDS indicators, such as the presence or absence of a relevant lesion, require to be assessed (1) the expert's interpretation of a scan image (*subjective component*), (2) to take into account not only on the features of the scan image itself, but also the other features of the patient (*contextual component*), only partially encoded as CDS indicators; in other words, the latter point means that some CDS indicators depend on the value of other indicators. Decomposing the CDS indicators into independent and objective elements is not a practicable solution, as the CDS indicators constitute the minimal elements of agreement among the NEUROWEB neurovascular community. We argue that these complications mirror the well-documented difficulty of diagnosis automatization (Szolovits *et al.*, 1988; Miller, 1994)[8]. Nonetheless, we explicitly model the inter-dependencies among CDS indicators as logical formulas, used for methodological coherence checking: if a patient displays incoherence among the CDS elements required for a certain phenotype, that patient is discarded. An example is the CDS indicator *Relevant Scan Lesion*. According to the diagnostic knowledge, a lesion is relevant only in presence of a *co-axial scan lesion* (i.e. the evidence of some brain tissue damage) and *stenosis* (i.e. the evidence of a partial occlusion in a brain-afferent artery). It follows that Relevant Scan Lesion = *yes* is correct only when Scan Lesion = yes, Stenosis Degree > 50%, Side of the Scan Lesion = DX (SX), and Side of the Stenosis = DX (SX) [9].

The two-layered model is a simple and effective model to encode a set of given phenotypes, organized according to homogeneous classificatory criteria, as in the case of the TOAST. The NEUROWEB Consortium accepts the TOAST as a federation standard, but the ontology should be flexible enough to manage the existence of alternative neurovascular schools, adopting a classification alternative to the TOAST. The introduction of stroke types organized according to different classificatory criteria would be managed in the two-layer model by introducing an additional taxonomy, besides the TOAST taxonomy. However, the exploitation of clinical phenotypes for association studies requires to dynamically redefine the phenotypes, going beyond the categories of a specific classificatory system; in this scenario, the TOAST is a source of robust clinical phenotypes, grounded onto the diagnostic practice, which can be refined or modified to meet, at best, the ends of association studies. We argue that to enable, at best, the manipulation of Top Phenotypes, it is necessary to introduce an additional layer in which the Top Phenotypes are decomposed into their *building blocks*, according to the classification criteria. We will also show that decomposing the Top Phenotypes according to classification criteria enables to identify general domain concepts, supporting the bridging to genomic entities. The CDS indicators are not suitable building blocks, as

- the formulation of Top Phenotypes using the CDS indicators does not explicitly represent the classificatory criteria;
- CDS indicators do not represent general concepts of the medical domain; instead, they are the minimal elements of the diagnostic practice, as it is exerted in the NEUROWEB clinical communities. For example, anatomical parts are often referred in the CDS indicators, but they are not represented as stand-alone entities of the CDS.

The Top Phenotypes taxonomy enables to represent the characterizations of the stroke types with increasing specificity from the root to the leaves (e.g. Ischemic Stroke, Atherosclerotic Ischemic Stroke, Atherosclerotic Ischemic Stroke Evident, Atherosclerotic Stroke Evident with

[8] Specifically, the diagnostic process should not be regarded as a sum of independent tests on single parameters, but rather as the progressive recognition of the patient's state, through mutual reinforcement of different observations; that means that a decision tree is not a suitable model for the cognitive processes underlying the diagnostic activity.

[9] As we will display with more details in the next session, these logical formulas are formally encoded as DL axioms.

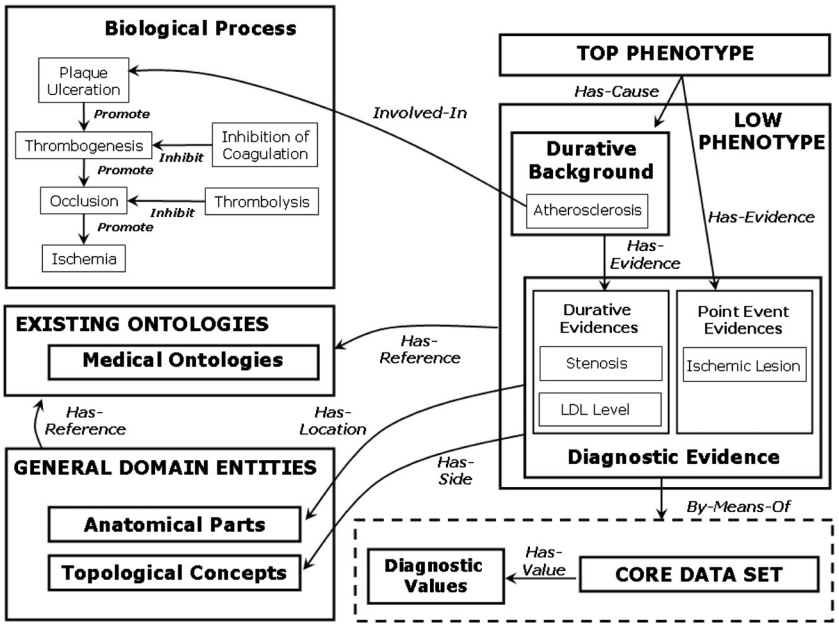

Figure 2 The figure depicts the NEUROWEB Ontology components, and their relations

Intracranial Cerebral Artery Stenosis); however, the taxonomy is intrinsically incapable of representing the notion of part, as required by the identification of the building blocks. To identify the building blocks of the TOAST Top Phenotypes, it is necessary to consider what are the TOAST classification criteria:

- etiology (i.e. Atherosclerotic, Cardioembolic, Lacunar Stroke);
- confidence of the etiological assessment (i.e. Evident, Probable, Possible), depending on the probatory strength of the diagnostic evidence;
- anatomy (i.e. the location of the lesion, e.g. left carotid artery).

The TOAST classification criteria suggest the use of specialized relations[10] to deconstruct the Top Phenotypes into building blocks (see Figure 2):

- Has-Cause (pointing to the specific etiology);
- Has-Evidence (pointing to diagnostic evidences);
- Has-Location (pointing to the anatomical site of the neurovascular lesion);
- Has-Side (pointing to the side of the neurovascular lesion).

To account for the anatomical criterion, we add to the ontology the Anatomical Parts; the Anatomical Parts of interest for the NEUROWEB project are typically vascular territories, such as the intracranial cerebral artery (ICA). These concepts are not phenotypes, since phenotypes are generally defined as observable properties of an organism; on the contrary, Anatomical Parts are physical entities bearing observable properties. We also add to the ontology a set of Topological Concepts, to manage the side of the lesions (including the notions of spatial coaxiality and symmetry). Whereas the CDS indicators derive from the community-level diagnostic knowledge, and the TOAST Top Phenotypes derive from the school-level diagnostic knowledge, the Anatomical Parts and the Topological Concepts are general domain concepts (i.e. on the Biomedical Knowledge level).

[10] As these relations express the partonomy of the phenotype (i.e. its constitutive parts), they can be regarded as specialized mereological relations (cf. Simons, 1987; Sattler, 2000).

The etiological criterion implicitly assumes a partition between:

- the Stroke as a neurovascular occlusive event, causing a brain lesion, and
- the pathophysiological factors causing the occlusive event.

This difference implicitly organizes the diagnostic activity, yet it essentially relies on the physiological nature of stroke as conceived in the general biomedical knowledge. Ischemic stroke literally means a sudden alteration of neurological functions due to an ischemic (i.e. absence of blood flow) event in a brain area. With more details, there is a cascade of sequential events: the initial trigger is the generation of the occluding body or structure, which stochastically leads to vessel occlusion; as a consequence, the brain area downstream the occluded vessel suffers a severe restriction in blood supply (i.e. there is a *local ischemia*); that leads to tissue damage, partial functional impairment at organ and system level (brain and nervous system), and finally behavioral and cognitive anomalies. The stroke can be diagnosed by looking for clinical evidences of these events: the brain region affected by ischemia and necrosis can be identified by means of imaging techniques (i.e. Relevant Lesion); behavioral and cognitive anomalies are typical symptoms exhibited by the stroke patient. The initial trigger, that is the generation of the occluding body, has causal roots in a specific pathological background; the TOAST assumes three etiological groups:

- Atherosclerotic,
- Cardioembolic,
- Lacunar.

Each of these groups refers to a *durative* pathological background, eventually leading to the generation of the occlusive body. For instance, *Atherosclerosis* is a systemic blood vessel pathology, affecting multiple sites in the circulatory system; it consists in the deposition of cholesterol on the blood wall, contextual to the establishment of an inflammatory process, eventually leading to the formation of an atherosclerotic plaque; if the inflammatory process overrides the wound healing process, the plaque ruptures, thus triggering the coagulation process and the generation of a clot on the inner side of the blood wall; if the clot detaches from the wall (i.e. embolization), it circulates until it behaves as an occluding body. Atherosclerosis is diagnosed by the observation of stenosis in the brain afferent arteries, altogether with other diagnostic evidences. For these reasons, we introduce the concept of *Durative Background* and *Diagnostic Evidences*; Diagnostic Evidences are further specified into *Durative* and *Point Event*, to distinguish the diagnostic evidences for the events triggered by the occlusion, and the ones enabling to recognize the durative background.

The decomposition according to these criteria leads to the identification of concepts we term Low Phenotypes. Altogether with Anatomical Parts and Topological Concepts, the Low Phenotypes constitute a new middle layer, connected to the Top Phenotypes and the CDS indicators. We argue that the criteria adopted to identify the Low Phenotypes are implicitly guided by the separation of different pathophysiological processes; in other words, Low Phenotypes are characterized by different underlying pathophysiological processes. For instance, Atherosclerotic Ischemic Stroke Evident (AISE) is caused by Atherosclerosis, whereas Lacunar Ischemic Stroke Evident (LISE) is caused by Small Vessel Disease; however, Atherosclerotic Ischemic Stroke Possible (AISPo) is caused by Small Vessel Disease and Atherosclerosis, thus there is a situation of two concurring causes. Therefore, LISE and AISPo share a common pathophysiological process, Small Vessel Disease. That information was not encoded by Top Phenotypes alone, and it would have been cumbersome to reconstruct it from the CDS-based definition formulas. Decomposing clinical phenotypes in terms of processes implies a shift from a clinical understanding—rooted onto diagnostic evidences, methods, and theories—to the general theory of biological function, which belongs to the corpus of biomedical knowledge. That is the common ground between medicine and molecular biosciences, providing the basis for the cross-domain bridging.

4.2 A strategy to bridge clinical phenotypes and genomic entities

Clinical phenotypes were encoded in order to conceive the diagnostic knowledge of the NEUROWEB clinical communities, and then decomposed to extrapolate general domain concepts. In order to specifically support genotype–phenotype association studies, it is necessary to establish relations between the clinical phenotypes and genomic entities. Specifically, two functionalities can be addressed:

- represent relations between gene functions and clinical phenotypes;
- integrate phenotypes from the clinical and genomic domains, managing
  - the different aims and methodologies;
  - the different granularities (spanning single molecules, biomolecular processes, cell types, tissues, organs, systems, body).

The first task was accomplished by introducing new relations and entities to connect the Low Phenotypes to the reference gene functions, as they are provided by the controlled vocabulary Gene Ontology. As we argued in the previous section, the decomposition into Low Phenotypes is implicitly guided by the separation of different pathophysiological processes. We add to the NEUROWEB Reference Ontology specific Biological Processes[11], relevant for the stroke and its causal backgrounds (e.g. Atherosclerotic Plaque Ulceration, Thrombogenesis). These concepts are connected to through the Low Phenotypes via the Involved-In relation (e.g. Atherosclerotic Plaque Ulceration Involved-In Atherosclerosis). NEUROWEB Biological Processes are then connected to their Gene Ontology counterparts via the Has-Reference relation. We argue that decomposing the clinical phenotypes into Low Phenotypes, associated with different processes, enables to overcome the discrepancy between the clinical and genomic phenotype representation, due to the different aims and methodologies of the two disciplines. For instance, the association between a genotype and the equivalent of the Low Phenotype Atherosclerosis is significantly more frequent than the association with the corresponding Top Phenotype, Atherosclerotic Ischemic Stroke. Indeed, the decomposition of Top Phenotypes into Low Phenotypes supports the analytical treatment of the context factor: Atherosclerotic Ischemic Stroke (Top Phenotype) is a stroke event (Low Phenotype: Relevant Scan Lesion) in the causative context of Atherosclerosis (Low Phenotype), but the notion of Top Phenotype alone does not enable to recognize these two separate elements.

As far as the second task is concerned, a granularity discrepancy arises when phenotypes—as represented in genomic resources—refer to entities spanning different level of organization (ranging from single molecules to the whole body); for instance, a phenotype can be expressed as the altered level of a protein or a molecular complex (e.g. increased level of LDL (low-density lipoproteins)). Granularity can be fully managed only in presence of a general-purpose content theory of biological parts and functions; since there is no such reference ontology available for the biological domain, it was not possible to accomplish this task, as out of the project reach. Although the task of managing the different granularities is not presently supported, the ontology can be suitably extended to handle it, just by adding more modules and relations to the existing structure: as we have already shown, the NEUROWEB Low Phenotypes can be put in relation with biomolecular processes; we argue that it is possible to establish relations also with other general domain concepts, such as cell types, tissues, and organs. For instance, a greater resolution in the representation of the Atherosclerotic processes and physical parts can provide the groundwork for the detailed acquisition and elaboration of genomic data. For instance, Atherosclerosis is a pathology with multiple facets, including molecular functions, biomolecular processes, cell types, anatomical sites, parts, and systems. The liability to trigger a stroke cascade depends on the state of all these components, organized at different level of granularity; different evidences from the genomic research (e.g. gene expression patterns, genotypic profiles, phenotypes) can be related to the clinical phenotypes, only in presence of a content theory organizing them accordingly.

[11] We adopt the term 'Biological' Process instead of 'Pathophysiological' in analogy to the Gene Ontology Biological Processes.

4.3 Knowledge acquisition strategy

The KA was carried out with the four clinical partners. A substantial loop occurred between the establishment of the CDS indicators, and the identification of the clinical phenotypes of interest. In particular, the TOAST classification played an important role in shaping a common vision over the classification criteria. After several rounds of meetings and negotiation among the clinical experts, once the CDS reached a sufficiently stable version, the TOAST was encoded according to the two-layer model, and the experts were asked to formulate the clinical phenotypes according to that frame. This process was closely assisted by the KE team, organizing focused meetings with the most experienced clinicians to discuss the TOAST classification criteria, the phenotype formulation being compiled, and the limits of the model. During this process, the specialized relations used for the Low Phenotypes were introduced in a preliminary version of the three-layer model, and validated with the clinicians. In parallel, the genomic resources were analyzed to identify the discrepancies and the common ground with the clinical phenotype model. As a consequence, the three-layer model was further refined, reaching the final version.

5 The NEUROWEB ontology: the formal and the computational model

In this section, we systematically review the structure of the NEUROWEB Reference Ontology knowledge, formally represented using OWL-DL. We also present the architecture of the IT infrastructure that exploits the NEUROWEB ontology to retrieve patients from the local clinical repositories.

5.1 Methods

Description Logics were adopted as the formal language for the NEUROWEB Ontology. Description logics (Horrocks *et al.*, 1999; Sattler, 2000; Baader *et al.*, 2003) are a family of logic-based KR formalisms designed to represent and reason about the knowledge of an application domain in a structured and well-understood way. The basic notions in description logics are atomic concepts and atomic roles (unary and binary predicates in the terminology of first order language, respectively). A relation between two individual concepts (A and B) is represented as a binary predicate; the relation is termed *role*, the *role concept* is the first argument of the predicate (concept A), and the *role filler* is the second argument of the predicate (concept B). For instance, hasPart.Wheel represents the property of cars having wheels, in which the individual objects belonging to the concept Wheel are fillers of the role hasPart. A specific description logic is mainly characterized by the constructors it provides to form complex concepts and roles from the atomic ones. The language we used to formalize the ontological clinical knowledge is $\mathcal{SHOIN}(\mathbf{D})$, which is an extension of the basic description logic; specifically, we have adopted the OWL-DL version (Horrocks & Patel-Schneider, 2003). The editor adopted for the OWL file generation is Protg, a well known tool developed and distributed by the Standford University; in Protg, the NEUROWEB Ontology concepts are represented as T-Box (Terminological Box) entities.

5.2 The formal model of the ontology

The final ontology model is composed of three main layers: Top Phenotypes (the top layer), Low Phenotypes (the middle layer), and CDS (the bottom layer). The layer of the Low Phenotypes includes three additional modules, whose concepts are not phenotypes: Anatomical Parts, Topological Concept, and Biological Processes. Mappings to the external resources (ontologies, controlled vocabularies, genomic databases) are provided via the Has-Reference relation. The layer of the Top Phenotypes is a taxonomy of stroke types and subtypes according to the TOAST classification; alternative classifications can be accommodated as independent taxonomies, whereas user-customized TOAST phenotypes are included in the TOAST taxonomy. These entities represent clinical phenotypes (i.e. pathological states of stroke patients) judged relevant by clinicians, and recognizable through the diagnostic activity. The root of the TOAST taxonomy is

Ischemic Stroke, whose children are Atherosclerotic Stroke, Cardioembolic Stroke and Lacunar Stroke; these stroke types constitute the three etiological groups according to the TOAST classification. Each of them is then divided into Evident, Probable, and Possible. Deeper levels of phenotype specification lead to the addition of children to these stroke subtypes. Although the reference to anatomical parts is important for the definition of the TOAST categories[12], the TOAST has no further phenotype specification explicitly based on the anatomical criterion (e.g. Atherosclerotic Ischemic Stroke Evident with Intracranial Cerebral Artery Stenosis). Some of the clinical experts have suggested the addition of a further layer to the Top Phenotypes TOAST taxonomy, based on the site of the lesion; after a phase of negotiation in the consortium, those phenotypes could be eventually added to the Reference Ontology.

Reflecting the TOAST classification criteria, the Top Phenotypes are decomposed into Low Phenotypes through two main relations:

- Has-Cause points to the log-term pathology underlying the stroke, represented by the Durative Background subclass of the Low Phenotypes (e.g. Atherosclerosis);
- Has-Evidence points to the diagnostic evidences for the stroke event, represented by the Point-Event Diagnostic Evidences subclass of the Low Phenotypes.

The Has-Evidence relation is also used to connect the Durative Background to its diagnostic evidences, represented by the Low Phenotypes subclass Durative Diagnostic Evidences. Please remark that the distinction between *Durative* and *Point-Event* diagnostic evidences is not coupled to any specification of the Has-Evidence relation. The Has-Location and Has-Side relations are used to connect a Low Phenotype to the anatomical site of a lesion (an Anatomical Part), and to the side of the lesion (a Topological Concept).

A Low Phenotype is finally decomposed into the clinical evidences required for its identification. The By-Means-Of relation connects a low Phenotype to a CDS indicator, and to specific value ranges (termed Diagnostic Values). The Has-Value relation connects the CDS indicator to the Diagnostic Values, and it is nested within the By-Means-Of formula. Since CDS indicators refer to anatomical parts, we introduce the relation Has-Location-CDS to explicitly represent it. An Anatomical Part is connected both to a Low Phenotype (Has-Location) and to the CDS indicators assessing that phenotype by referring to that Anatomical Part (Has-Location-CDS)[13].

Biological Processes constitute an additional module of the NEUROWEB Reference Ontology, and they are inter-related by Activates and Inhibits relations. Low phenotypes are connected to Biological Processes via the Involved-In relation.

To provide mappings to external resources, the ontology includes Has-Reference relations connecting NEUROWEB concepts to entities from genomic databases and biomedical ontologies or thesauri.

Metamodelling formulas

The following formula represents at a high level the NEUROWEB Taxonomy and the main axioms ruling the connections among the different NEUROWEB Entities. The meaning of the roles refer to the one informally described in the present section.

(1) $\texttt{NEUROWEBOntology} \sqsubseteq \texttt{CDS} \sqcap \texttt{GeneralDomainEntity} \sqcap \texttt{Phenotype} \sqcap \texttt{OtherOntoResources} \sqcap \texttt{TopologicalEntity}$

(2) $\texttt{Phenotype} \sqsubseteq \texttt{TopPhenotype} \sqcap \texttt{LowPhenotype}$

(3) $\texttt{LowPhenotype} \sqsubseteq \texttt{DurativeBackground} \sqcap \texttt{DiagnosticEvidence}$

[12] The Atherosclerotic Stroke implies an obstructive event in the large brain arteries; the Cardioembolic Stroke implies the generation of the obstructing body (thromboembolism) in the heart; the Lacunar Stroke implies an obstructive event in the small brain arteries.

[13] The only case in which this correspondence does not occur is when the CDS indicator has a coarse anatomical resolution, though more details on the anatomical part can be referred to the Low Phenotype, as in the case of the Small Vessel Disease.

(4) DiagnosticEvidence ⊑ DurativeEvidence⊓PointEventEvidence
(5) GeneralDomainEntity ⊑ AnatomicalEntity⊓DiagnosticValue
(6) GeneralDomainEntity ⊑ ∀hasReference.OtherOntoResources
(7) Phenotype ⊑ ∀hasReference.OtherOntoResources
(8) TopPhenotype ⊑ ∀hasEvidence.DiagnosticEvidence
(9) TopPhenotype ⊑ ∀hasCause.DurativeBackground
(10) DurativeBackground ⊑ ∃hasEvidence.DurativeEvidence
(11) DiagnosticEvidence ⊑ ∃byMeansOf.(CDS⊓(hasValue.DiagnosticValue))
(12) LowPhenotype ⊑ ∀hasLocation.AnatomicalEntity
(13) LowPhenotype ⊑ ∀hasSide.TopologicalEntity
(14) CDS ⊑ ∀hasLocationCDS.AnatomicalEntity
(15) CDS ⊑ ∀hasSide.TopologicalEntity
(16) CDS ⊑ ∀hasValue.DiagnosticValue

The following formulas represent the Top Phenotype Atherosclerotic Stroke Evident, and its decomposition into Low Phenotypes.

- Ax1 encodes the Top Phenotype decomposition into the Durative Background, Atherosclerosis, and the evidence of the ischemic brain lesion proving the occurrence of the stroke event, Relevant Lesion;
- Ax2 encodes the decomposition of Relevant Lesion into Left Relevant Lesion and Right Relevant Lesion: the ischemic lesion can be on either side of the brain;
- Ax3 encodes the Left Relevant Lesion, requiring the co-axiality between the ischemic lesion site and the stenosis site (Stenosis is the evidence for Atherosclerosis); the encoding of the Right Relevant Lesion is omitted, as strictly similar to the left one;
- Ax4 encodes the left co-axility between the ischemic lesion site and the stenosis site; again, the right co-axiality as omitted, as strictly similar to the left one.

The Low Phenotypes Moderate and Severe Lesion are decomposed into CDS indicators and value validity ranges through the by-Means-Of and Has-Value relations; in axioms (Ax5:8) we display only the axioms referring to the Moderate Lesion. In axioms (Ax9:15) the Low Phenotype Severe Stenosis is connected to the anatomical part, and decomposed into CDS indicators.

(Ax1) AthIschStrokeEvident≡∃hasEvidence.RelevantLesion ⊓ ∃hasCause.(Atherosclerosis ⊓ ∃hasEvidence.SevereStenosis)
(Ax2) RelevantLesion≡∃hasEvidence.LeftRelevantLesion ⊔ ∃hasEvidenceRightRelevantLesion
(Ax3) LeftRelevantLesion≡∃hasEvidence.LeftCoaxialityLesionStenosis
(Ax4) RightCoaxialityLesionStenosis≡∃hasEvidence.((ModerateLesion ⊔ SevereLesion) ⊓ ∃hasSide.Right) ⊓ ∃hasEvidence.(SevereStenosis ⊓ ∃hasSide.Right)
(Ax5) ModerateLesion≡Lesion ⊓ ∃byMeansOf(CT-From2.5to5Centimeters Lesion ⊔ MRI-From2.5to5CentimetersLesion ⊔ PET-From2.5to5 CentimetersLesion)
(Ax6) CT-From2.5to5CentimetersLesion≡CT ⊓ ∃hasValue.2.5-5centimeters
(Ax7) MRIFrom2.5to5CentimetersLesion≡MRI ⊓ ∃hasValue.2.5-5centimeters
(Ax8) PET-From2.5to5CentimetersLesion≡PET ⊓ ∃hasValue.2.5-5centimeters
(Ax9) SevereStenosis≡Stenosis ⊓ ∃byMeansOf.((CTA-MoreThan60PercentInICA ⊔ DSA-MoreThan60PercenInICA ⊔ MRA-MoreThan60PercenInICA) ⊔ (CTA-MoreThan60PercentInCarotidArtery ⊔ DSA-MoreThan60PercentInCarotid Artery ⊔ MRA-MoreThan60PercentInCarotidArtery)) ⊓ ∃hasLocation.ICA
(Ax10) CTA-MoreThan60PercentInICA≡CTA ⊓ ∃hasValue.MoreThan60Percent ⊓ ∃ hasLocationCDS.ICA
(Ax11) DSA-MoreThan60PercentInICA≡DSA ⊓ ∃hasValue.MoreThan60Percent ⊓ ∃hasLocationCDS.ICA

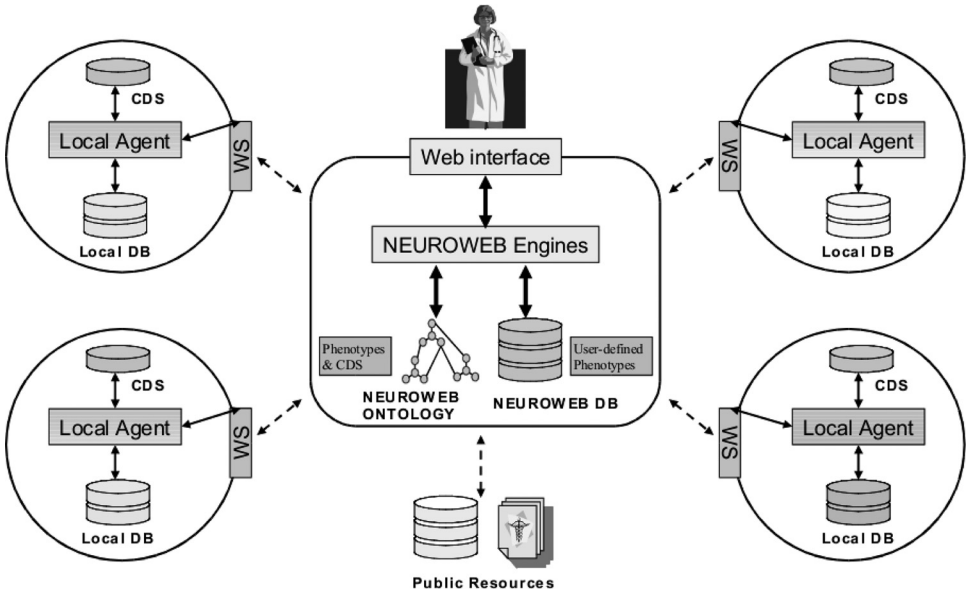

Figure 3 The NEUROWEB architecture

(Ax12) `MRA-MoreThan60PercentInICA` ≡ `MRA` ⊓ ∃`hasValue.MoreThan60Percent` ⊓ ∃`hasLocationCDS.ICA`

(Ax13) `CTA-MoreThan60PercentInCarotidArtery` ≡ `CTA` ⊓ ∃`hasValue.MoreThan60Percent` ⊓ ∃`hasLocationCDS.CarotidArtery`

(Ax14) `DSA-MoreThan60PercentInCarotidArtery` ≡ `CTA` ⊓ ∃`hasValue.MoreThan60Percent` ⊓ ∃`hasLocationCDS.CarotidArtery`

(Ax15) `MRA-MoreThan60PercentInCarotidArtery` ≡ `CTA` ⊓ ∃`hasValue.MoreThan60Percent` ⊓ ∃`hasLocationCDS.CarotidArtery`

5.3 The overall computational architecture

The NEUROWEB architecture has been designed to cope with the heterogeneity of the data sources, that is, the databases that are maintained at the participant institutions, public repositories and scientific literature. Figure 3 reports the overall architecture of the system. Web service technology has been exploited to decouple the NEUROWEB central components from the local sites in order to enable new clinical institutions to join the consortium. A key point is the unified language that every node uses to talk to the center of the system: every conversation is carried on in terms of the Reference Ontology. At every site a translation to the local terminology occurs to retrieve the patient clusters (see Figure 4).

To retrieve a patient cluster given a clinical phenotype encoded in the NEUROWEB ontology, it is necessary to

1. process the axioms of the ontology, converting the phenotypes into a CDS-based formula;
2. exploit the mapping between the CDS and the local repository, which is typically implemented as a relational database.

A NEUROWEB user formulates a query by means of Top Phenotypes, Low Phenotypes, and CDS elements. The queries are expressions that use standard operators (logical and relational operators), and are based on the axioms included in the NEUROWEB ontology[14]. A query can

[14] Other approaches, such as the one adopted in OQAFMA (Mork *et al.*, 2003), define SQL-like languages (StruQL in OQAFMA) to support queries over semantic descriptions. The disadvantage of these approaches is the need to learn a specific query language, which can reduce the usability of the system by medical users, who are not already familiar with writing SQL queries.

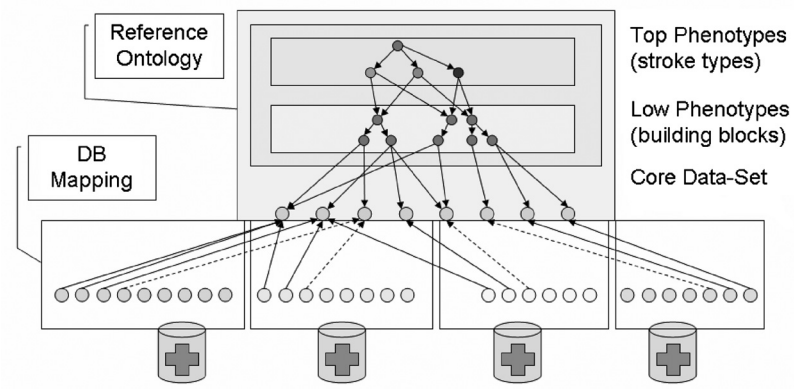

Figure 4 The NEUROWEB Reference Ontology is composed of two parts: Phenotype Ontology and CDS. CDS is the set of clinical indicators concerning the required exams to perform diagnostic activities. It supports interoperability through the mapping established between its entities and the clinical data from local repositories. The Phenotype Ontology is composed of Top Phenotype and Low Phenotype layers. The Top Phenotype layer is an expert-defined taxonomy of strokes connected to the Low Phenotype layer, which provides an explicit representation of the classification criteria, the building blocks of phenotypes. Finally, the Low Phenotypes are connected to the CDS entities via axioms, which encode the value-ranges on CDS clinical variables required for the phenotype occurrence

reflect a built-in phenotype definition (i.e., the associated axiom) as well as user-defined phenotypes to support the research of new associations. Such queries are sent to the Clinical Query Application module of each site to be processed.

In the current version of the system, queries are processed by a software module, the Phenotype Converter, which converts the ontology into a relational schema by navigating the ontology from the top-level phenotype axioms to low phenotypes, and finally to conditions on CDS elements, which are the elementary components of the phenotypes 4. For this navigation, the has-cause, has-evidence and by-means-of relations are used. The results are regular SQL queries containing only CDS indicators. The rewritten queries exploit the mapping between the local data schema and the CDS to query the local database and finally retrieve the matching patient cluster. For the clinical partners currently involved in the NEUROWEB consortium, we implemented the mapping by building a view over the existing databases; this work was done in collaboration with local medical experts that were familiar with the local coding, use and meaning of the local database fields.

The described solution enables for fast integration of new institutions: what is required is just a mapping table between the CDS elements and the local DB elements. This solution is feasible since the CDS was designed to accommodate the most important elements to fully describe a patient profile. However, this solution has the drawback of loosing the semantic definition of phenotypes included in the ontology. A more sophisticated solution can be supported by creating mappings between the NERUROWEB ontology (not just the CDS part) and the local database schema. Moreover, if an institution has already adopted an ontology to organize its phenotype knowledge, the mapping can occur between the local ontology and the reference ontology to deliver the most powerful solution possible.

To facilitate the entry of new institutions, a complete open-source solution has been developed to integrate local databases. The Clinical Query Application runs as a Web application on a Glassfish server, which is a lightweight and safe technology to manage call wrappings to expose local interfaces as WSDL (Web service definition language) documents (i.e. as Web Services). The Phenotype Converter is implemented in Java, exploiting the Jena programming interface, which is the most popular tool to manage ontologies. The database view is implemented in Postgres (Douglas, 2005), which supports all key features required for efficient database programming.

6 Conclusions and future works

The NEUROWEB project posed relevant challenges to the ontological modeling effort. On the one hand, it was necessary to take into account the specific methodologies adopted by different clinical communities of practice, and reconcile them it into a common classificatory system (the TOAST), developing an ontology for the clinical phenotypes; the resulting ontology supports a query system operating across the different repositories. As an explicit ontology design principle, we represented the classification criteria using specialized mereological relations; we argue that this feature enables to reconcile in a common representation frame different classificatory system, adopted by different neurovascular communities. On the other hand, a greater challenge was posed by the bridging of clinical phenotypes, belonging to the world of diagnostic medicine, onto gene functions and phenotypes belonging to the field of genomics. The strategy adopted was to identify a common ground, the general theory of biological function, and accordingly decompose the concepts from the different disciplines into common elementary units (i.e. building blocks). Specifically, we identified clinical phenotypes having a homogeneous physiological background, the Low Phenotypes, and we established connections to the biomolecular processes. A similar solution strategy was applied in the apparently unrelated field of mechanical engineering, where the representation of the design processes involves a deep analysis of the functional roles carried by the mechanical parts of an artifact (Colombo *et al.*, 2007). In addition, we argue that an explicit analysis of the epistemological challenges is crucial for the ontological modeling activity. We specifically identified the community-specific methodologies as the source of tacit knowledge. The KA activity addressed these sources, and was guided by the aim of formalizing the TOAST system, a semi-formal Representational Artifact generated by the an international process of negotiation among the different neurovascular communities.

The software infrastructure required to exploit the classification services offered by the DL-Reasoners is currently under definition. Assertions at the A-Box level can be generated exploiting the mapping between the Reference Ontology and the local repositories. In addition, we are currently designing a user interface able to provide a conceptual view on the ontology, avoiding the direct access to DL axioms. The goal is to enable the clinical experts to interact with the ontology. The efficacy of the user interface is critical for the updating of the ontology, and its use among clinicians as a basis for the negotiation of new classification categories. We argue that the formalization of the TOAST provides the basis to extend the NEUROWEB framework to a system supporting cooperative work.

Acknowledgements

The work presented in this paper reflects the initial activities in the NEUROWEB project (project number 518513). We wish to thank the clinical partners: Istituto Nazionale Neurologico Carlo Besta (INNCB, Milan, Italy), Orszagos Pszichiatriai esNeurologiai Intezet (AOK-OPNI, Budapest, Hungary), University of Patras (UOP, Patras, Greece), Erasmus Universitair Medisch Centrum Rotterdam (MI-EMC, Rotterdam, Holland). In particular, we wish to thank Dr Yiannis Ellul and Dr Stella Marousi from UOP, Dr Zoltan Nagy and Dr Csaba Ovary from AOK-OPNI, Dr Aad Van Der Lugt and Dr Philip Homburg from MI-EMC, and Dr Giorgio Boncoraglio from INNCB, for the valuable contributions during the knowledge acquisition campaign and model refinement process.

References

Adams, H. P. Jr., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L. & Marsh, E. E. 3rd. 1993. Classification of subtype of acute ischemic stroke, definition for use in a multicenter clinical trial, TOAST. Trial of Org 10172 in acute stroke treatment. *Stroke* **24**, 35–41.

Ay, H., Furie, K. L., Singhal, A., Smith, W. S., Sorensen, A. G. & Koroshetz, W. J. 2005. An evidence-based causative classification system for acute ischemic stroke. *Annals of Neurology* **58**, 688–697.

Baader, F., Calvanese D., McGuinness D. L. *et al.* 2003. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

Bandini, S., Colombo, E., Colombo, G., Sartori, F. & Simone, C. 2003. The role of knowledge artifacts in innovation management: the case of a chemical compound designer CoP. *International Conference on Communities and Technologies*, Kluwer Academic Publishers, 327–345.

Bard, J. B., Rhee, L. & Seung, Y. 2004. Ontology in biology: design, application and future challenges. *Nature Reviews Genetics* **5**(3), 213–222.

Benassi, R., Beneventano, D., Bergamaschi, S., Guerra, F. & Vincini, M. 2004. Synthesizing an integrated ontology with MOMIS. *International Conference on Knowledge Engineering and Decision Support (ICKEDS)*, Porto, Portugal, 21–23 July 2004.

Benerecetti, M., Bouquet, P. & Ghidini, C. 2000. Contextual reasoning distilled. *Journal of Theoretical and Artificial Intelligence (JETAI)* **12**, 279–305.

Bergamaschi, S., Guerra, F., Orsini, M. & Sartori, C. 2007. Extracting relevant attribute values for improved search. *IEEE Internet Computing* **11**(5), 26–35.

Bodenreider, O., Smith, B., Kumar, A. & Burgun, A. 2007. Investigating subsumption in DL-based terminologies: a case study in SNOMED CT. *Artificial Intelligence in Medicine* **39**(3), 183–195.

Boland, J. & Tenkasi, R. V. 1995. Perspective making and perspective taking in communities of knowing. *Organizational Science* **6**(4), 350–372.

Bouquet, P., Don, V. & Serafini, A. 2002. ConTeXtualized local ontology specification via ctxml. In *Proceedings of AAAI workshop on Meaning Negotiation*, Edmonton, Alberta, Canada.

Bowker, G. C., Turner, W., Star, S. L. & Gasser, L. 1997. *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*. Lawrence Erlbaum Associates.

Brown, J. S. & Duguid, P. 1991. Organizational learning and community of practice: toward a unified view of working. *Organization Science* **2**(1), 40–57.

Chandrasekaran, B., Josephson, J. R. & Benjamins, V. R. 1999. What are ontologies, and why do we need them? Intelligent systems and their applications. *IEEE* **14**(1), 20–26.

Colombo, G., Mosca, A. & Sartori, F. 2007. Towards the design of intelligent CAD systems: an ontological approach. *International Journal on Advanced Engineering Informatics – Special Issue on Ontology and Epistemology of Systems and Software Engineering* **22**(2), 153–168.

Davenport, T. & Prusack, L. 1998. *Working Knowledge: How Organizations Manage What They Know*. HBS Press.

Douglas, K. 2005. *PostgreSQL* (2nd edn). Sams.

Feyerabend, P. 1975. *Against Method: Outline of an Anarchistic Theory of Knowledge*. Verso.

Goldstein, L. B., Jones M. R., Matchar D. B. *et al.* 2001. Improving the reliability of stroke subgroup classification using the trial of ORG 10172 in acute stroke treatment (TOAST) criteria. *Stroke* **32**, 1091–1097.

Gomez-Perez, A., Corcho-Garcia, O. & Fernandez-Lopez, M. 2003. *Ontological Engineering*. Springer-Verlag.

Gruber, T. R. 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human–Computer Studies* **43**(4–5), 907–928.

Guarino, N. 1995. Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human–Computer Studies* **43**(5–6), 625–640.

Guarino, N. 1998. Formal ontology in information systems. In *Proceedings of FOIS'98*, Trento, Italy. IOS Press, 3–15.

Guida, G. & Berini, G. 2000. *Ingegneria della Conoscenza: Strumenti per Innovare e per Competere*. EGEA.

Guizzardi, G. 2007. On ontology, ontologies, conceptualizations, modeling languages, and (meta)models. In *Frontiers in Artificial Intelligence and Applications, Databases and Information Systems IV*, Vasilecas, O., Eder, J. & Caplinskas, A. (eds). IOS Press, 18–39.

Hildreth, P., Kimble, C. & Wright, P. 2000. Communities of practice in the distributed international environment. *Journal of Knowledge Management* **4**(1), 27–37.

Holsapple, C. W. & Joshi, K. D. 2001. Organizational knowledge resources. *Decision Support Systems* **31**(1), 39–54.

Horrocks, I. & Patel-Schneider, P. 2003. Reducing OWL entailment to description logic satisfiability. In *Proceedings of the 2nd International Semantic Web Conference (ISWC)*.

Horrocks, I., Sattler, U. & Tobies, S. 1999. *Practical Reasoning for Expressive Description Logics*. Springer-Verlag.

Johnson-Laird, P. N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.

Kitamura, Y. & Mizoguchi, R. 2003. Ontology-based description of functional design knowledge and its use in a functional way server. *International Journal of Expert System with Application* **24**(2), 153–166.

Kuhn, T. 1970. *The Structure of Scientific Revolutions*. University of Chicago Press.

Lakatos, I. 1978. *The Methodology of Scientific Research Programmes*, Philosophical Papers, volume 1, Cambridge University Press.

Laudan, L. 1977. *Progress and its Problems: Toward a Theory of Scientific Growth*. University of California at Berkeley.

Lee, L. J. *et al*. 2000. Impact on stroke subtype diagnosis of early diffusion-weighted magnetic resonance imaging and magnetic resonance angiography. *Stroke* **31**, 1081–1089.

Lenzerini, M. 2002. *Data Integration: A Theoretical Perspective*. ACM PODS, 233–246.

Lueg, C. 2002. Knowledge management and information technology: relationship and perspective. *Upgrade—Knowledge Management and Information Technology*. Introduction to the special issue, III(1).

Miller, R. A. 1994. Medical diagnostic decision support systems. Past, present, and future: a threaded bibliography and brief commentary. *Journal of American Medical Informatics Association* **1**(1), 8–27.

Mork, P., Brinkley, J. F. & Rosse, C. 2003. OQAFMA Querying Agent for the Foundational Model of Anatomy: a prototype for providing flexible and efficient access to large semantic networks. *Journal of Biomedical Informatics* (36), 501–517.

Nonaka, H. & Takeuchi, I. 1995. *The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press.

OBO. 2007. Open Biomedical Ontologies. `http://obo.sourceforge.net/` (November 2007).

Preece, A., Sleeman, D. H. & Flett, A. N. *et al*. 2001. Better knowledge management through knowledge engineering. *IEEE Intelligent Systems* **16**(1), 36–43.

Sattler, U. 2000. Description logics for the representation of aggregated objects. In *Proceedings of the 14th European Conference on Artificial Intelligence*, IOS Press.

Sicilia, J. J., Sicilia M. A., Sánchez-Alonso S. *et al*. 2009. Knowledge representation issues in ontology-based clinical knowledge management systems. *International Journal of Technology Management* **47**, 191–206.

Simons, P. 1987. *Parts: A Study in Ontology*. Oxford University Press.

Smith, B. 2004. Beyond concepts: ontology as reality representation. In *Proceedings of FOIS'04*.

Smith, B., Kusnierczyk, W., Schober, D. & Ceusters, W. 2006. Towards a reference terminology for ontology research and development in the biomedical domain. In *Proceedings of KR-MED*.

Protégé version 3.0 2004. Standford University. `http://protege.stanford.edu/` (13 October 2004).

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., Abeysinghe, S., Krawczak, M. & Cooper, D. N. 2003. Human Gene Mutation Database (HGMD). *Human Mutation* **21**(6), 577–581.

Studer, R., Benjamins, R. V. & Fensel, D. 1998. Knowledge engineering: principles and methods. *Data Knowledge Engineering Journal* **25**(1–2), 161–167.

The Gene Ontology Consortium. `http://www.geneontology.org`

The National Center for Biomedical Ontologies, `http://www.bioontology.org/wiki/index.php/Mainpage`

Thomasson, A. L. 2004. Methods of categorization. In *Proceedings of the 3rd International Conference (FOIS'04)*, IOS Press, 3–16.

Van der Vet, P. E. & Mars, N. J. I. 1998. Bottom-up construction of ontologies. *IEEE Transactions on Knowledge and Data Engineering* **10**(4), 513–526.

Wenger, E. 1998. *Community of Practice: Learning, Meaning and Identity*. Cambridge University Press.

W3C—Web Ontology Language 2004. `http://www.w3.org/TR/owl-guide/` (10 February 2004).

Zaihrayeu, I., Sun L., Giunchiglia F. *et al*. 2007. From Web Directories to Ontologies: Natural Language Processing Challenges. In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference*, Busan, Korea, 623–636.