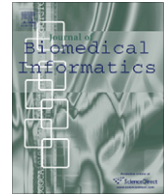




Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Towards an ontology for sharing medical images and regions of interest in neuroimaging

Lynda Temal^{a,b,c}, Michel Dojat^{d,e}, Gilles Kassel^f, Bernard Gibaud^{a,b,c,*}^aINRIA, VisAGes Project-Team, F-35042 Rennes, France^bINSERM, U746, F-35042 Rennes, France^cUniversity of Rennes 1, CNRS, UMR 6074, IRISA, F-35042 Rennes, France^dINSERM, U836, Grenoble, F-38043, France^eJoseph Fourier University, Institute of Neurosciences, Grenoble, F-38043, France^fLaRIA, CNRS (FRE 2733) and Jules Verne University of Picardie, Amiens, F-80039, France

ARTICLE INFO

Article history:

Received 6 August 2007

Available online 17 March 2008

Keywords:

Medical imaging

Data integration

Mediation system

Neuroscience

Biomedical ontologies

Semantic annotation

Data sharing

ABSTRACT

The goal of the NeuroBase project is to facilitate collaborative research in neuroimaging through a federated system based on semantic web technologies. The cornerstone and focus of this paper is the design of a common semantic model providing a unified view on all data and tools to be shared. For this purpose, we built a multi-layered and multi-components formal ontology. This paper presents two major contributions. The first is related to the general methodology we propose for building an application ontology based on consistent conceptualization choices provided by the DOLCE foundational ontology and core ontologies of domains that we reuse; the second concerns the domain ontology we designed for neuroimaging, which encompasses both the objective nature of image data and the subjective nature of image content, through annotations based on regions of interest made by agents (humans or computer programs). We report on realistic domain use-case queries referring to our application ontology.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

Neuroimaging includes a variety of techniques to explore brain structure and function, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET), Single Photon Emission Computed Tomography (SPECT), and Magnetoencephalography (MEG). It has become a major tool for scientists and physicians, in their quest for a better understanding of the mechanisms involved in brain development, brain functions and brain disorders. Moreover, neuroimaging is emerging as a prominent tool to assess the efficacy of new drugs against brain pathologies, such as cancer, neurodegenerative diseases or psychiatric disorders.

In addition to standard image interpretation based on human visual screening, computer-generated imaging biomarkers provide quantitative information useful in medical decision making. The introduction of such computerized markers, precisely defined and automatically extracted from the processed images, supports the use of well-defined protocols or guidelines to optimize imaging processing chains and to standardize image acquisition sequences and scanner calibration procedures.

* Corresponding author. Address: Unité/Projet VisAGes U746, INSERM/INRIA/CNRS/U. de Rennes 1, IRISA, Faculté de médecine, 2 Avenue du Pr Leon Bernard, F-35043 Rennes, Cedex, France. Fax: +33 2 99 84 71 71.

E-mail addresses: lynda.temal@irisa.fr (L. Temal), michel.dojat@ujf-grenoble.fr (M. Dojat), gilles.kassel@u-picardie.fr (G. Kassel), bernard.gibaud@irisa.fr (B. Gibaud).

Today, such developments in data sharing are considered necessary to improve the relevance and efficacy of large scale multi-center clinical trials and are therefore strongly encouraged by stakeholders, such as the NIH (National Institute of Health) and FDA (Food and Drug Administration) in the US, and the framework programs for research in the EU. Moreover, these developments will benefit basic brain research that highlights the relationships between morphology and function in the central nervous system, especially research based on molecular imaging and MR imaging (anatomical and functional MRI, and diffusion-weighted MRI for brain connectivity assessment). This creates a strong need for formal definition of imaging-related information, consisting of acquired data, namely raw data and acquisition conditions, as well as processed and interpreted data, namely data resulting from a specific procedure performed by an agent (human or machine). Thus, data sharing is a challenging topic in biomedical domain and several ongoing efforts are being performed. For instance, the Cancer Biomedical Informatics Grid (CaBIG)¹ [1] is a NCI (National Cancer Institute) initiative to gather in a common cyberarchitecture, a network of cancer centers and research laboratories. To reach this goal, CaBIG is developing standards, policies, guidelines, common applications, open source tools and a middle-ware infrastructure. Similar goals are being pursued by the Biomed-

¹ <https://cabig.nci.nih.gov/>.

ical Informatics Research Network (BIRN)², another initiative supported by the NIH, in the field of neurosciences, in the continuity of the Human Brain Project [2]. Related activities concern both the development of a mediation infrastructure, based on ontologies, and tested applications called Morphology BIRN, Function BIRN and Mouse BIRN, addressing various kinds of needs in neurosciences [3,4]. Similar efforts exist in Europe, e.g. in the context of the Virtual Physiological Human initiative, supported by the European Union [5].

The NeuroBase project, launched in France in 2002, pursues the same general objective: to share images and processing tools in the context of distributed and heterogeneous systems [6]. The goal of the NeuroBase project is twofold: to manage and share the large quantity of data produced (1Gb/subject), and to provide a federated platform for the interoperability of processing tools. Presently, data and tools are disseminated in three French centers, all partners of the NeuroBase project. The objectives can be summarized in three main points: (i) carrying out large scale experiments by sharing heterogeneous distributed data, (ii) combining existing image processing tools to define new data processing pipelines, and (iii) evaluating these heterogeneous pipelines on large datasets produced by the imaging centers.

The cornerstone of this project, and the major focus of this paper, is the design of a common semantic model, according to an ontological approach, which provides a unified view of all data and tools to be shared via the federated system. Our ultimate aim is: (1) to define an easily maintainable and extensible reference ontology for a broad community of neuroscientists. Currently targeted applications concern cognitive science (visual cortex exploration) and neurological pathologies (e.g. neurodegenerative diseases); (2) to integrate conceptualizations from different fields, e.g. neuroanatomy, neurophysiology or neuropathology, into a consistent whole; and (3) to define an ontology that can be mapped with other ontologies, in order to ensure interoperability with external systems.

To define such an ontology, called OntoNeuroBase, we adopted a multi-layer approach and based our ontological commitments on well-known ontologies existing at different levels of abstraction, e.g. top-level ontologies or core domain ontologies [7]. We maintain simultaneously two manifestations of the ontology. The first manifestation is specified in the semi-informal language of the OntoSpec methodology [8]. It is semantically rich as it makes use, in particular, of temporally-indexed relations and meta-properties considered by the OntoClean methodology [9]. As such, this manifestation is intended to facilitate the mapping of OntoNeuroBase with other ontologies. The second manifestation is specified in the formal Web Ontology Language OWL. This manifestation is semantically poorer but enables to use OntoNeuroBase to perform inferences.

Beside its primary role within a federated system to provide a common unified schema for the mapping of the local database schemas, the basic added value of an ontology is that it enables reasoning about shared information. Such reasoning may concern querying by introducing new capabilities based on formal semantics of the concepts and relations expressed in the ontology. Beyond querying, reasoning may also be applied to image annotations with a view to enhance image interpretation, e.g. relating measurements on images having different modalities to the characteristics of the real-world entities being imaged. This may involve representations of space and topological properties of these real-world entities. Composing image processing tools conceived in different contexts to define innovative processing chains requires representing sufficient knowledge about such tools, as well as the data processed, in order to achieve interoperability between the tools.

Hence, this paper mainly focuses on three crucial aspects: (1) how neuroimaging data can be organized in a consistent set of categories to facilitate sharing, (2) what the images actually represent, e.g. an MR signal intensity, a 3D volume, or a time sequence of 3D volumes, (3) how “Regions of Interest” (often abbreviated ROIs) can be represented in the images, and what they mean for agents (humans or programs) involved in their creation, querying and use. This third issue is fundamental for using imaging biomarkers consistently and relating observations of the same reality through several imaging modalities, and for exploring various aspects of brain structure and brain metabolism or function. The literature contains many relevant contributions [10–12], but is still incomplete.

Our work makes two major contributions. The first is a methodology to build a multi-layered application ontology. The second consists of a novel conceptualization of neuroimaging data, encompassing both the objective nature of image data and the subjective nature of annotations made by intelligent agents. This conceptualization is based primarily on the definition of ROIs.

The paper is structured as follows. Section 2 presents our methodology to build a multi-layered application ontology. Section 3 presents the Datasets Ontology, the principal kernel of the OntoNeuroBase ontology which specializes I& DA (Information and Discourse Acts). We focus first on the semantic axes, which allow categorizing the images; we then describe a way of modeling objective image content by mathematical functions; lastly, we introduce ROIs and related annotations, as a means to express subjective information. Section 4 provides an illustrative example, based on our current implementation of the ontology. Section 5 discusses some of our choices regarding both the general design methodology and our novel conceptualization of neuroimaging data. Finally, development perspectives for our project are evoked in Section 6.

2. The ontological reference framework

To fulfill our objectives, we chose a design framework that structures the ontology at different levels of abstraction (Fig. 1) while respecting common conceptualization choices. In this section, we focus on the modular approach [13] adopted in the design of the global structure of the OntoNeuroBase conceptualization, disregarding the specification languages.

At the highest level is a *top-level ontology* that includes abstract concepts and relationships valid across domains. *Foundational ontologies* are such top-level ontologies whose concepts and relations share a common philosophical foundation. We adopted *DOLCE* (Descriptive Ontology for Linguistic and Cognitive Engineering), which serves as a foundational ontology [14].

We then added *Core ontologies*, which provide generic, basic and minimal concepts and relations in a specific domain [7]. By minimal we mean that core ontologies should include only the most reusable and widely applicable categories. These kinds of ontologies are essential for sharing intended meaning between different domains. We adopted I& DA (Information and Discourse Acts), a core ontology initially built for classifying documents as a function of their content [15]. We use it to model medical images, which we consider as types of documents. *Participant Roles* [16] is the core ontology we use to describe the modes of image participation in data processing. I& DA and Participant Roles are built according to DOLCE ontological commitments.

On the basis of these two layers, we constructed our *Domain ontology* dedicated to conceptualizing a specific domain, in this case neuroimaging. Obviously, large domains such as neuroimaging can be divided into sub-domains for the sake of modularization.

² <http://www.nbirn.net/index.shtml>.

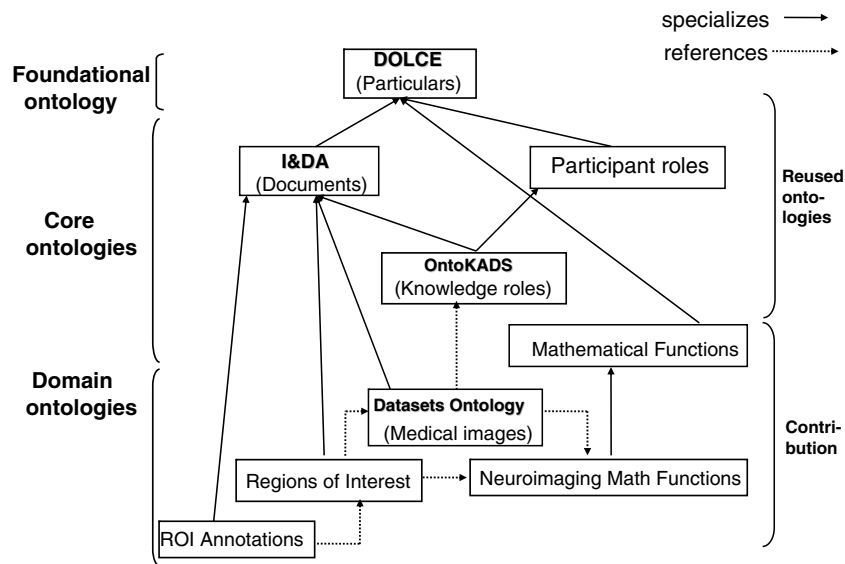


Fig. 1. An overview of the application ontology framework: A solid line going down from sub-ontology O_1 to O_2 means that the entities of O_2 (concepts and relations) specialize the entities of O_1 ; a dashed line between two sub-ontologies, from O_1 to O_2 , means that O_2 concepts reference O_1 concepts without specialization.

This complete framework (Fig. 1) constitutes OntoNeuroBase, our *Application ontology*. Various domain ontologies can be added to extend OntoNeuroBase, but should respect the ontological commitments driving our conceptualization. In the following pages, we detail these main conceptualization choices underlying the foundational and core ontologies we have adopted.

2.1. DOLCE (Particulars)

DOLCE³ [14] (Descriptive Ontology for Linguistic and Cognitive Engineering) is a foundational ontology of *Particulars* which are classified into four separate categories, depending on their modes of existence (Fig. 2).

- *Endurants* are entities that “are wholly present in time” (e.g. you and your parts). Among *Endurants*, and according to whether the entity has direct spatial qualities, *Physical Objects* (e.g. your brain) are distinguished from *Non-Physical Objects* (e.g. your knowledge about neuroimaging), which cover social and cognitive entities. The notion of *Collection* was added recently [17] as a specialization of *Non-physical Object*, in order to represent plural entities (e.g. fiber collection).
- *Perdurants* are entities that “occur in time” (e.g. cerebral blood circulation) in which *Endurants* (e.g. cerebral blood) participate. Among *Perdurants*, *Statives* are distinguished from *Events* according to whether the *Perdurants* are cumulative⁴ or not. *Events* are divided into *Achievements* and *Accomplishments* according to whether they are atomic or not. *Actions* are *Accomplishments* which are intentionally controlled by *Agents*.
- *Endurants* and *Perdurants* are characterized by inherent *Qualities*, which are seen as the basic properties we can perceive or measure (e.g. the density of tissues in a specific anatomical structure). *Qualities* are divided into *Temporal Qualities*, *Physical Qualities*, and *Abstract Qualities*, which are, respectively, inherent to *Perdurants*, *Physical Endurants*, and *Non-Physical Endurants*.
- *Qualities* take “values”, called *Quales* (e.g. a particular grey level encoded by the number 225), within quality region spaces. *Quales* are *Abstract* entities.

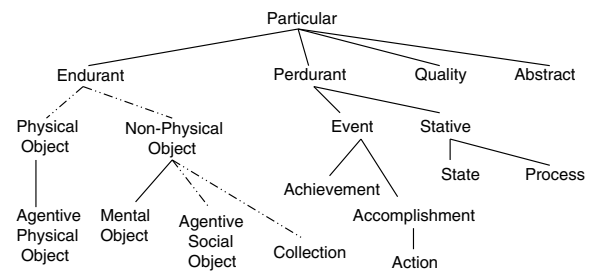


Fig. 2. An excerpt from DOLCE's top-level taxonomy. A solid line between two concepts represents a direct specialization relation. A dashed line reflects the existence of intermediate concepts.

2.2. I& DA (Documents)

I& DA is a core ontology in the domain of semiotics that was initially built for classifying documents by their contents [15]. I& DA extends DOLCE by introducing three main types of entities (Fig. 3):

- *Inscriptions* (e.g. written texts, images) are knowledge forms materialized by a substance (e.g. ink) and inscribed on a physical support (e.g. a sheet of paper, a hard disk). The particularity of these *Physical Endurants* lies in their intentional nature. *Inscriptions* stand for other entities: *Expressions*.
- *Expressions* (e.g. texts, equations) are non-physical knowledge forms ordered by a communication language. *Inscriptions realize Expressions* and, like *Inscriptions*, *Expressions* are intentional entities conveying contents for *Agents*.
- *Conceptualizations* consist of the means by which *Agents* can reason about a world. Functionally, one distinguishes between two kinds of *Conceptualizations*: *Propositions*, as a means of describing a state of affairs; and *Concepts*, as a means of classifying entities. *Messages* are specializations of *Propositions* which result from *Discourse Acts* (e.g. Informing, Defining). *Conceptualizations* can be expressed by *Expressions* and *physicallyRealized*⁵ by *Inscriptions*. *Propositions* can reference *Particulars*, and can have concepts for subject (*hasForSubject Concepts*).

⁵ In the rest of the paper, relation names will be written using a Java-like notation.

³ <http://www.loa-cnr.it/DOLCE.html>.

⁴ *Perdurants* are *stative* or *eventive* according to whether they are the mereological sum of two of their instances [6].

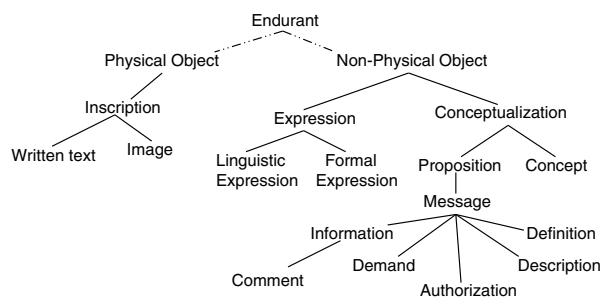


Fig. 3. I&DA's top-level taxonomy.

Note that I&DA made the choice to consider three distinct entities, rather than three different points of view on one entity. We will see that this modeling choice has important consequences for our modeling of medical images and their content.

2.3. Participant roles and knowledge roles

In this section, we introduce the notion of participant roles [16] and their relation with *Perdurants*. This notion is important for representing how image content is processed.

The participant roles inform us about the manner in which an *Endurant* participates in a *Perdurant*. The particularity of roles is they are *anti-rigid*, in the sense that they are non-essential for all their instances [18].

The distinction is made between *Determinants* and *Patients* according to whether they control the *Perdurant* in which they participate, or whether they are affected by it. Among *Patients*, *Data* and *Results* are introduced to model the participation modes of *Non-Physical Objects* in a particular *Action*. These roles (*Data*, *Results*) are knowledge roles [19], which can be played by non-physical objects (*Conceptualizations*, *Expressions*). This is particularly useful for the modeling of our domain, where we essentially deal with image content, which is the *result* of particular *Actions*, e.g. image processing.

3. Neuroimaging domain (OntoNeurobase)

OntoNeuroBase is an application ontology which covers two essential domains: medical images and medical image processing tools [20]. In this paper, we focus on the medical images domain.

3.1. Rationale for image modeling

Querying and processing neuroimaging data in a heterogeneous and distributed environment is based on the assumption that the common characteristics of the data are properly identified and managed. However, this is currently hampered by several factors. One of them is inherent to the ambiguities of the term “image” in user discourse; it sometimes refers to the physical instances of the images, but may also be used to designate the visual rendering of the images or their actual content. Even in terms of image content, an ambiguity may exist between the objective content of an image (e.g. measurement of some physical quantity using imaging equipment) and what users may describe subjectively (e.g. referring to real-world entities such as the subject's anatomy).

3.1.1. The “Tower of Babel” of image formats

Creating an ontology of neuroimaging data implies removing such ambiguities by proposing and organizing meaningful categories of image data in a taxonomy. The basic information to build such categories sometimes exists, and can be found in existing im-

age formats, mainly DICOM, Analyze, NIFTI, VTK, GIS⁶, etc. A problem arises from the multiplicity of these formats, and from the fact that they partly overlap, and that they do not represent the data structure and semantics in an explicit and consistent way. Such formats usually separate metadata (data describing the structure and semantics of image data), from the image data itself, sometimes using different physical files. The DICOM format is primarily used as a native format to represent images created by acquisition equipment. It is not yet well suited to representing processed images, or processes of image processing. It is rather complete in terms of contextual and technical metadata, but it has no ontological foundation, which would facilitate automatic reasoning. DICOM format specifications are primarily organized according to imaging modalities (MRI, PET, CT, etc). In contrast, formats like Analyze, NIFTI, etc., are neutral in terms of modalities and more often used for representing processed data, such as segmentation results or statistical maps. However, their specifications are not rigorous; for example, they do not explicitly distinguish between mandatory and optional information. Generally speaking, they are not conceived to support interoperability of independent applications; they simply support image representation and storage in small communities that share common (non-explicit) ways to represent data.

The extraction of a common, sharable conceptualization of image categories is not facilitated by image formats. For instance, T1 weighted MR images are acquired for exploring brain anatomy. Based on image format alone, the information can only be retrieved indirectly: via tree file organization, the relevant files being located in a special directory called “anatomical”; or via DICOM, by combining various data elements that may or may not be present, since their presence is not required by the standard. Several pieces of information are also not explicitly described and thus not easily retrieved, such as the physical nature of the sampling variables, the pixel or voxel values, and the relation between them. Such information is necessary as soon as one tries to represent semantically the constraints that govern the applicability of a given processing tool for particular sets of images, e.g. to express that a rigid registration assumes that both the source and target images are sampled according to similar X Y Z space variables.

3.1.2. Towards representing imaging biomarkers

Concerning image content, what is at stake is the ability to share consistent representations of imaging biomarkers derived from image processing, as well as the image regions from which they have been specifically derived. This is essential to conducting wide-scale studies with thousands of subjects, on pathologies like Alzheimer's disease or other dementia, in order to quantify brain changes over time, both morphologically (volume of specific cortical or sub-cortical regions) and functionally (using fMRI). This kind of need is addressed in the DICOM standard with the structured reporting paradigm (DICOM SR), based on tree representation of the various observations and facets of each observation (e.g. numeric value of a measurement, code, etc). This paradigm has been successfully used in CAD (Computer Assisted Detection) applications, e.g. for chest or mammography CAD, but so far it has received little attention in neuroimaging. The need to depict regions of interest has been taken into account in DICOM SR, using the notion of SCOOD tree node (spatial coordinates), but in a way that is not fully relevant for neuroimaging applications. Moreover,

⁶ DICOM (Digital Imaging and Communications in Medicine) is a standard for the exchange of medical images and related data developed by the DICOM Standards Committee. The Analyze format is an image format developed by the Mayo Clinic and used in their image processing Package Analyze. The NIFTI image format was proposed in the context of the Neuroimaging Informatics Technology Initiative. VTK is a file format developed by Kitware for its Visualization Toolkit Package. GIS is an image file format introduced in France in the 90s in the context of the Groupe d'Intérêt Scientifique “Sciences de la Cognition”.

the relatively rich semantic capabilities of DICOM SR have no ontological foundation, which in practice seriously compromises reasoning.

In conclusion, image format is a poor representation of image complexity. The following part of this Section introduces a novel conceptualization to solve some of the problems mentioned above, based on the ontological choices we have presented. Section 3.2 addresses the specific question of the relation between content, expression and physical inscriptions of the images. Section 3.3 tackles the difficult question of how to categorize neuroimaging data in meaningful categories. Section 3.4 focuses on the objective content of the images and highlights mathematical structure of this content by introducing a (functional) relation between sampling variables and pixel or voxel values. Section 3.5 addresses the question of ROIs and the related annotations.

3.2. The Datasets Ontology: Relation between content, expression and physical inscriptions

The distinctions made in I&DA (*Inscriptions/Expressions/Conceptualizations*) led us to identify various entities linked to the notion of “dataset”.⁷ First, we distinguished between the following entities:

- A *Dataset Expression* as an *Expression* by means of an encoding format (e.g. a *DICOM Expression*, an *Analyze Expression*, a *GIS Expression*, etc).
- An *Image* as an *Inscription* (on a computer screen, for example). *Images* can be further differentiated according to the image dimension (e.g. *2D Image*, *3D Image*) and the kind of rendering (e.g. *Color Image*, *Black and White Image*). *Datasets* stored in *Files* represent other kinds of *Inscriptions*. *File* is further differentiated according to the kind of encoding format (e.g. *DICOM File*, *Analyze File*, *GIS File*). These *Files* realize corresponding *Dataset Expressions* (e.g. an *Analyze File* realizes an *Analyze Expression*), and *physicallyRealizes* a corresponding *Dataset*.

Each *Dataset* is *expressedBy* at least one *Dataset Expression* and a *Dataset* can be *physicallyRealizedBy* at least one *File*.

With such a model, various aspects of neuroimaging data can be separated using the *quality* category (see Section 2.1). For instance, the *quality format encoding type*, which *hasForQuality* U8 or U16 values, is naturally inherent to the *Dataset Expression* and independent of the content, i.e. the *Dataset*. Qualities such as image intensity and visual aspects are inherent to *Images* displayed on a screen, or printed on paper, independently of the semantics related to the image content. Location quality on a specific support is inherent to *Files*, independent of its format and content.

The organization of the Datasets Ontology thus allows us to perform a number of interesting queries. For example, we can easily find all *Datasets expressedBy* a *DICOM Expression*. We can find *Files* which *physicallyRealize Datasets* and *realize GIS Expressions* that are encoded in U8 *format encoding type*. Qualities associated with each entity represent criteria that can be used to query image representations.

Finally, another dimension of a dataset is captured by considering a *Dataset* as a *Proposition*, which corresponds to the image content, and is detailed in the following Section. The three conceptual dimensions of *Dataset* are reflected in Fig. 4.

⁷ Terms like “image” and “dataset” are widely used by clinicians and scientists in the neuroimaging domain to designate the image and its content as a whole. In our ontology, we use the word “Dataset” in a dedicated way to mean the content only. However, in the rest of the paper, due to the broad utilization of this word, we continue to use it in the two senses: to refer to content, we write *Dataset*; to refer to the whole, we write “dataset”.

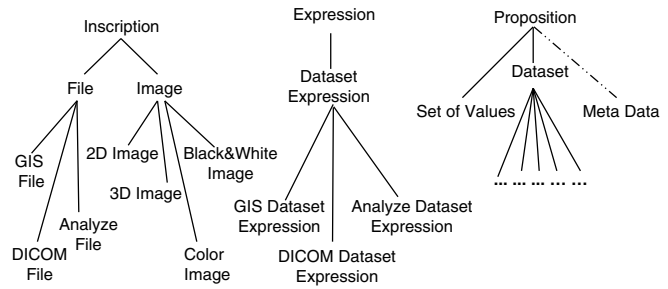


Fig. 4. An excerpt from the Datasets Ontology. This conceptualization considers three dimensions present in the Dataset concept: *Proposition* corresponds to the objective image content, *Expression* defines the encoding format, and *Inscription* defines the way data are materialized.

3.3. Categorization of datasets

A *Dataset* as *Proposition* is a complex, structured entity, which consists of data concerning a subject or a group of subjects. It is considered as a description, with one part corresponding to a structured set of values (*Set of Values*), and another part associating *Meta-Data*. Roughly speaking, *Set of Values* (which is a *Proposition*), stands for the actual kernel of a *Dataset*, independently of any encoding format. It is represented by a function denoting the distribution in space and/or time of physical quantities, such as MR signal intensity, regional cerebral blood flow or a displacement vector (see Section 5.3). *Meta-Data* as *Information* (which is a special kind of *Proposition*) includes information referring to real-world entities, such as: (1) the *Dataset's* acquisition context (acquisition protocol, acquisition equipment) in terms of calibration and parameter settings (e.g. echo time and inversion time for MR images, etc.), (2) the anatomical structure or brain function explored in the scan session (e.g. brain, vision, audition), and (3) the *Dataset's* orientation, i.e. the orientation of the subject with respect to the sampled spatial variables. The detailed modeling of *Meta-Data* is still in progress and will not be presented here.

In practice, *Datasets* can be designated by neuroscientists according to different points of view. For example, a given *Dataset* may be classified as an MRI dataset because it was acquired with MRI acquisition equipment, or because it was derived from an MRI dataset. Likewise, it can be classified as an anatomical dataset because it explores anatomical entities, or as a reconstructed dataset because it results from reconstruction processing. These different points of view serve as criteria to organize *Dataset* categories according to various semantic axes: categories based on modality (Section 3.3.1), categories based on data processing (Section 3.3.2) and categories based on the explored structure or function (Section 3.3.3).

3.3.1. Categories based on modality

Organizing *Datasets* according to the modality (e.g. MRI, PET, CT) is relevant because users intuitively relate *Datasets* to the kind of equipment used for image acquisition. However, this category is determined more by the kind of signal measured (e.g. an MR signal) than by the equipment itself. For example, an *MRI Dataset* processed for bias correction remains an *MRI Dataset*, because the physical property measured is not altered and remains an MR signal.

All categories based on modality (Fig. 5a) are mutually disjunctive: for example, an instance of *MRI Dataset* cannot be an instance of *PET Dataset*.

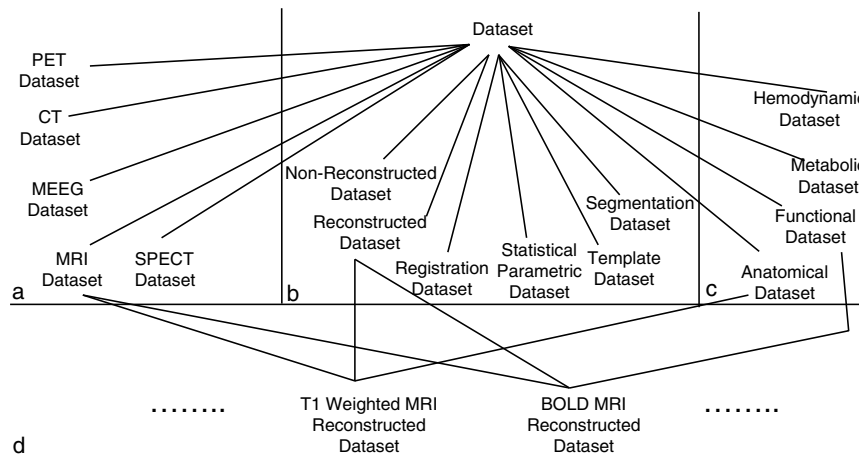


Fig. 5. An excerpt from the top-level Dataset categories based on (a) modality, (b) data processing, (c) entity explored and (d) low-level categories of Datasets inheriting properties related to the three previous semantic axes.

3.3.2. Categories based on data processing

The principle is to determine Dataset categories according to the kind of Data Processing from which they result. Thus, a Reconstructed Dataset is the result of a Reconstruction, whereas a Non-Reconstructed Dataset has not undergone any Reconstruction. A Segmentation Dataset is the result of a Segmentation, whatever the nature of this segmentation (e.g. contour detection, region classification). A Registration Dataset is the result of a Registration; it represents a geometrical transformation (e.g. a 4 × 4 matrix, a displacement field) of one Dataset onto another. A Template Dataset results from calculating the mean of several images (Averaging); it may be used as a reference for multi-subject image registration (Fig. 5b).

3.3.3. Categories based on the explored structure or function

In practice, a Dataset is acquired for a specific goal, i.e. the exploration of brain anatomy or brain physiology, according to a particular experimental protocol. Anatomical Datasets explore brain anatomy, for instance, via a T1 weighted MR Signal that provides good contrast for brain tissues and structures (Fig. 5c). Functional Datasets explore neural correlates following brain stimulation (e.g. BOLD contrast imaging). Hemodynamic Datasets explore brain hemodynamic function (e.g. perfusion imaging for blood volume and flow measurement). Lastly, Metabolic Datasets explore brain metabolism (e.g. MR spectroscopy for metabolite distribution). So, in summary, Anatomical Datasets explore Anatomical Structures (e.g. left hemisphere), whereas Functional Datasets, Hemodynamic Datasets and Metabolic Datasets explore Brain Functions or Physiological Processes (e.g. vision, blood flow or lactate distribution).

This organization provides rich possibilities to formulate queries according to the previous categories, e.g. retrieving all MRI Datasets that are Reconstructed Datasets and which explore anatomy (Anatomical Dataset). The Meta-Data play an essential part by supplementing semantics related to Datasets. They represent additional criteria to refine these queries (e.g. retrieve all MRI Datasets acquired using an MRI Acquisition Protocol).

3.4. Mathematical structure of Dataset content

Because imaging acquisition techniques are based on physical properties, they generally measure a physical quality in space and/or time (e.g. MR signal intensity). Thus, the Set of Values, part of Dataset, can be represented by a function denoting the distribution in space and/or time of such physical qualities.

In this section, we aim at describing what Datasets actually represent, using Mathematical Functions. Explicitly, each Mathematical Function defined in our domain represents exactly one Dataset. The benefit of representing the Dataset at the mathematical level lies in the ability to express the range and domain of the functions explicitly and to document which physical domain they address (e.g. space, time, energy, activity, signal intensity, etc).

Gruber et al. [21] proposed an interesting Ontology for Engineering Mathematics which conceptualizes notions of physical quantities, physical dimensions and functions of quantities. Some of the notions presented below are inspired by this ontology, which we adapted to comply with DOLCE.

We first built a Mathematical Functions Ontology by defining the basic essential concepts and relations required to model Mathematical Functions, taking account of the distinction made in DOLCE between Perdurants/Endurants and Qualities/Abstracts. Because there is some variation between the fields when it comes to intuition about functions, notation, and even the very meaning of the term “function”, we focused on the most abstract sense related to this notion. The characteristic property of a function is that it relates exactly one output to each of its admissible inputs, where the set of inputs represents the domain of the function, and the set of outputs represents its range. In accordance with this definition, we introduced two main concepts (Relation, Mathematical Function) and two relationships (hasForDomain, hasForRange) such that a Mathematical Function is a Relation that relates one input element to exactly one output element. Every Mathematical Function hasForDomain a Set and hasForRange a Set.

Thus, we used this ontology to specialize the mathematical functions relevant to our domain (Fig. 6), and to link them to the physical quality domain they address.

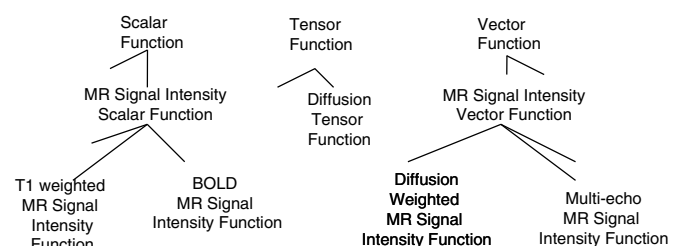


Fig. 6. An excerpt of the taxonomy of mathematical functions.

There are two significant aspects to our proposition. The first is to define a specialization of mathematical functions according to the value type of its range (*Scalar Function*, *Vector Function*, *Tensor Function*, *Probabilistic Function*, etc). The second is to associate the domain and range of each function to the qualities measured.

In neuroimaging, one example of a function representing *Datasets* expresses the mapping of spatial displacements (e.g. displacements according to the three coordinates *X*, *Y* and *Z* of a spatial reference system), which represents the domain of the function, in relation to physical qualities (e.g. MR signal intensity, regional cerebral blood flow), which in turn represent the range of the function. Thus, to describe functions that maps measurements of physical qualities, the first step is to precisely specify the *Qualities* for which measurements are provided, and the *Endurants* or *Perdurants* to which the *Qualities* are inherent.

According to DOLCE, domains and ranges of functions are *Regions*.⁸ Each *Region* is associated with a *Quality* such that elements belonging to *Regions* are “values” taken by these *Qualities*. *Qualities* are inherent in *Endurants* (e.g. an *MR signal intensity* is a quality inherent in an *MR Signal*) or in *Perdurants* (e.g. *Regional Cerebral Blood Flow* is a quality inherent in *Cerebral Blood Circulation*).

We can take the example of modeling a *T1 weighted MR Signal Intensity Function*. The scalar values of its range are *values* of the *MR Signal Intensity Qualities*, which are *inherent* in the *T1 weighted MR Signal*. The values of its domain are all possible triplets of the Cartesian product composed of three intervals, e.g. [0, 127] [0, 127] [0, 48] whose values are *values*, respectively, to a *X displacement*, a *Y displacement*, and a *Z displacement*.

From this representation, one can retrieve all *Datasets* that are represented by the *T1 weighted MR Signal Intensity Function* whose Cartesian product is composed of three intervals, such that the interval on *X displacement* is [0, 127], the interval on *Y displacement* is [0, 127] and the interval on *Z displacement* is [0, 48].

3.5. Regions of interest and annotations

ROIs are defined as a *selected subset of samples within a dataset identified for a particular purpose*. According to this definition and with respect to previous choices made in the *Datasets Ontology* based on I& DA, an *ROI* is a *Proposition* and corresponds to a selected subset of the *Set of Values*. Thus, an *ROI* is a *properPartOf* a *Set of Values*. The definition of an *ROI* as a *proper part* of a *Set of Values* implies the existence of its *ROI expression*, which is a *proper part* of the corresponding *Dataset Expression*, and the *ROI inscription* as a *proper part* of the corresponding *Inscription*. In this paper, we focus solely in *ROI* as *Proposition*. Since a *Set of Values* is *representedBy* a *Mathematical Function* (see Section 3.4), the subset of these values (i.e. the *ROI*) is *representedBy* a *Function Restriction* which is a restriction of the *Mathematical Function*, such that the domain and range of the *Function Restriction* are subsets of the domain and range of the *Mathematical Function*. This representation offers two advantages: (1) the semantics associated with the *Mathematical Function*, which represents the *Dataset*, are the same as for the *Function Restriction*, which represents the *ROI*; and (2) the *Function Restriction* can be determined either by defining its domain or by defining its range. Furthermore, we will see in Section 3.5.1 that the structure of the domain allows us to distinguish between different categories of *ROIs*.

Among the other information for describing *ROIs*, we must include the *Agent* that creates the *ROI* and the date of its conception. For this purpose we use existing relations in I& DA: *isConceivedBy*, which relates a *Proposition* to the *Agentive* that conceived it; and *hasForConceptionDate*, which relates a *Proposition* to its conception Date.

3.5.1. Regions of interest categories

The distinction is made between a *Geometrical ROI* (e.g. a *Parallelepipedic ROI*, a *Spherical ROI*, an *ellipsoidal ROI*) (Fig. 7), which is *representedBy* a *Function Restriction* whose domain is computed from a geometrical primitive (e.g. a parallelepiped, a sphere, an ellipse); and a *Free Form ROI*, which is *representedBy* a *Function Restriction* whose domain cannot be computed from a geometrical primitive. Two major cases of *Free Form ROI* can be mentioned. The first involves contour based *ROIs* that can be defined on serial slices (polylines), e.g. to depict a tumor on a preoperative MRI. Another case involves defining the range of the *Function Restriction* by selecting a particular value from the range of a related *Segmentation Dataset*. For instance, all values 255 in the function representing the *Segmentation Dataset* compose the “White Matter”.

3.5.2. ROI annotations

As underlined above, the aim of defining *ROIs* in *Datasets* is to characterize a specific anatomical, functional or pathological entity. However, associating an *ROI* with a reference to a particular real-world entity (e.g. the hippocampus of the subject) may be subjective and error-prone, so it is important to record sufficient information about observation context, using *Annotations*.

An *Annotation* is a *Comment* (a kind of *Proposition*), which is *anchoredOn* the object it annotates. Then *Annotations* are specialized as *ROI Annotations* which are *anchoredOn ROIs*. An *ROI Annotation* is *anchoredOn* exactly one *ROI*. Hence, we relate an *ROI Annotation* to the real-world entity being referenced by the relation *references*. In practice, an *ROI* may delimit the real-world entity according to three possible relationships, which denote a topological relation between the *ROI* and the region of space occupied by the real-world entity. In the first relation, the *ROI* delimits exactly the anatomical, pathological or physiological entity. In the second, the *ROI* delimits only a sub-part of the real-world entity. In the third, the *ROI* delimits a region that totally includes the real-world entity. To make these different kinds of referencing explicit, we specialize the *references* relation between *ROI annotation* and the real-world entity (anatomical, pathological or physiological) as three relations: *statesThatROIDelimitsExactly*, *statesThatROIDelimitsSubsetOf*, and *statesThatROIDelimitsContains*. For instance, an *ROI Annotation*, which is *anchoredOn* a *Free Form ROI*, *statesThatROIDelimitsExactly* a *Tumor*, which is a *Pathological Entity*. Actually, it is the *ROI* that delimits a real-world entity. However, because this decision is subjective and error-prone,

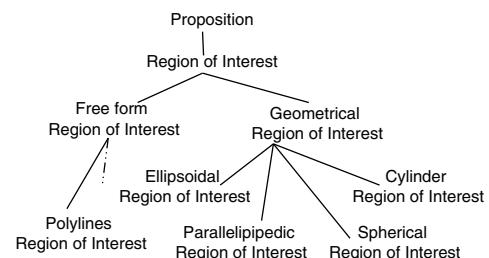


Fig. 7. An excerpt of the taxonomy of regions of interest.

⁸ Here, *Region* means “quality region space”, as defined in DOLCE. Not to be confused with *Region of Interest*.

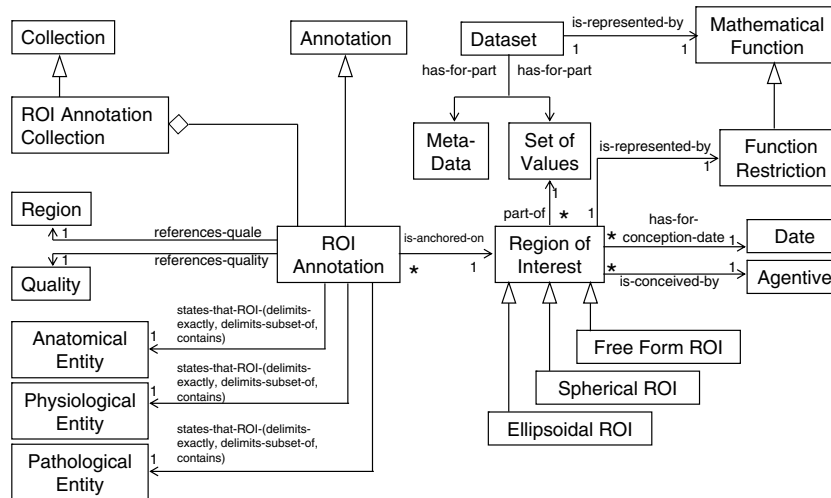


Fig. 8. An overview of regions of interest and annotations modeling.

we do not link an *ROI* directly to the delimited real-world entity, but rather through an *ROI Annotation*, endorsing the observer's subjectivity.

Furthermore, *Annotations* provide the capability to associate qualitative or quantitative information about the referenced entity, such as the volume of the tumor, the depth of the sulcus, the mean activity of a physiological process, etc. Therefore, we relate an *ROI Annotation* to the quality measured, as well as to the value of this quality. Then, we introduce two relationships, *referencesQuality* and *referencesQuale*, which specialize *references* such that an *ROI Annotation referencesQuality* a *Quality* (e.g. a Volume), and *referencesQuale* a *Quale*, which is an atomic part of a *Region* (e.g. 35 mm³).

The last point we want to highlight is the dependency that may exist between *ROIs* from the same *Dataset*, which are annotated during the same interpretation process by the same *Agent*. In most cases, several *ROIs* within the same *Dataset* are commented simultaneously, by associating each *ROI* with the *ROI Annotation* that *references Anatomical, Pathological or Physiological Entities*, e.g. "White Matter", "Grey Matter" and "Cerebrospinal fluid".

When this notion of dependency is significant, a set of *ROI Annotations* can be grouped as an *ROI Annotation Collection*, which is a *Collection* with the constraint that all *ROI Annotations* belonging to the same *ROI Annotation Collection* are conceived by the same *Agent* in the same *Action*.

Fig. 8 summarizes all the general notions described in this section and gives the overview of our model.

4. Illustrative examples

The following realistic use-cases illustrate how our conceptualization model supports various queries.

Several syntactic manifestations of OntoNeuroBase currently exist. As previously mentioned, the initial version, resulting from the modeling process, is semi-informal and structured following the OntoSpec methodology [8], which relies on OntoClean methodology [9]. The ontology expressed with this formalism is semantically rich and language independent, and can be easily translated into OWL. It constitutes our semantic reference and provides documentation for the other formalisms. The OWL version is formal, semantically poorer (e.g. in contrast to the OntoSpec formalism, n-ary relations are not allowed), but it is suitable for automatic reasoning.

The OWL version of the OntoNeuroBase ontology was developed using the PROTEGE⁹ ontology editor. It consists of 445 concepts and 189 relations. The set of instances involved in our illustrative examples has been produced by program, using the Protégé Application Programming Interface (API). Then, the knowledge base (ontology+instances) (Fig. 9) is exploited using the CORESE¹⁰ search engine [22] which internally works on conceptual graphs. It implements RDF, RDFS¹¹ and some statements from OWL-Lite and the query pattern part of SPARQL¹². The query language integrates additional features such as approximate search by computing semantic distance between concepts. Thus, in the initialization phase CORESE loads the files containing the ontologies and the instances and reorganizes them in conceptual graphs. After this initialization phase, CORESE accepts SPARQL queries, processes them, and provides the results in XML format. Clearly, in a final system, the inherent complexity of the SPARQL queries will be hidden to the final end-users, using appropriate search, visualization and manipulation tools.

Use Case 1: This example concerns data obtained in the context of a PET study aimed at highlighting the effect of subthalamic nucleus (STN) stimulation on frontal limbic areas in patients with Parkinson disease. PET studies were performed before and after STN stimulation. After a rigid registration of PET and MR images, PET regional cerebral blood flow was quantified by measuring the mean activity in ROIs, determined manually on the anatomical MRI according to a methodology similar to the one used in [23]. These ROIs concerned nine anatomical regions: orbital frontal cortex (bilateral), right and left anterior cingulate gyri, right and left superior, middle and inferior frontal gyri.

A relevant query is then: "retrieve *Datasets* (1) and *Subjects* (2), such that *Datasets* have *ROIs* (3–6) for parts annotated by *ROI Annotations* (9–10) indicating that *ROIs* delimit exactly, e.g. the *Anterior Cingulate Gyrus* (14–16); retrieve the *Values* associated with the *Qualities* (11–13) related to these ROIs" (the numbers appearing in parenthesis refer to the line numbers within the Q₁ query).

⁹ <http://protege.stanford.edu/>.

¹⁰ <http://www.sop.inria.fr/acacia/soft/corese/>.

¹¹ <http://www.w3.org/TR/rdf-schema/>.

¹² <http://www.w3.org/TR/rdf-sparql-query/>.

Q_1 is the corresponding query expression expressed in SPARQL language.

```

Q1: SELECT ?Patient ?Dataset ?DatasetType ?ROIType ?AnatomicalEntityType ?QualityType ?Value group ?Value
      DISPLAY xml WHERE{
(1) ?Dataset rdf:type ds:dataset
(2) ?Dataset ds:concerns ?Patient
(3) ?Dataset rdf:type ?DatasetType
(4) ?Dataset dol:proper-part ?SetOfValues
(5) ?SetOfValues rdf:type ds:set-of-values
(6) ?SetOfValues dol:proper-part ?ROI
(7) ?ROI rdf:type roi:region-of-interest
(8) ?ROI rdf:type ?ROIType
(9) ?ROI anno:has-for-anchor ?ROIAnnotation
(10) ?ROIAnnotation rdf:type anno:region-of-interest-annotation
(11) ?ROIAnnotation anno:references-quality ?Quality
(12) ?Quality rdf:type ?QualityType
(13) ?ROIAnnotation anno:references-quala ?Value
(14) ?ROIAnnotation anno:states-that-ROI-delimits-exactly ?AnatomicalEntity
(15) ?ROIAnnotation rdf:type ?ROIAnnotationType
(16) ?AnatomicalEntity rdf:type ana:anterior-cingulate-gyrus
(17) ?AnatomicalEntity rdf:type ?AnatomicalEntityType
}

```

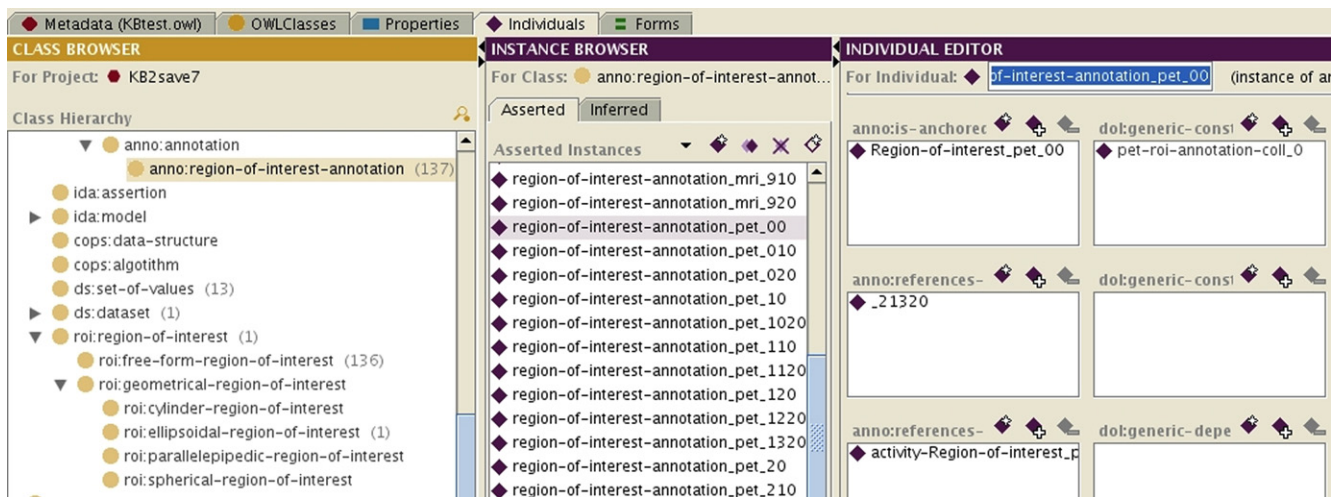


Fig. 9. An excerpt of OntoNeuroBase edited using Protégé.

The XML result generated by CORESE in response to the submitted SPARQL queries, can be easily reformatted through an XSL stylesheet into a table where the column names are the variables

of the SELECT section of the query. Fig. 10 shows such a table for the query Q_1 . The four first rows concern the subject Patient4. First, an instance *PET-Dataset_1* is proposed which has for part

| Patient | Dataset | DatasetType | ROIType | AnatomicalEntityType | QualityType | Value |
|-----------|---------------|-------------|------------------------------|--------------------------------|--------------------------------|-------|
| Patient4 | MRI_Dataset_1 | MRI-dataset | free-form-region-of-interest | left-anterior-cingulate-gyrus | volume | 11770 |
| Patient4 | MRI_Dataset_1 | MRI-dataset | free-form-region-of-interest | right-anterior-cingulate-gyrus | volume | 8553 |
| Patient4 | PET_Dataset_1 | PET-dataset | free-form-region-of-interest | left-anterior-cingulate-gyrus | nuclear-medicine-tomo-activity | 21659 |
| Patient4 | PET_Dataset_1 | PET-dataset | free-form-region-of-interest | right-anterior-cingulate-gyrus | nuclear-medicine-tomo-activity | 21621 |
| Patient9 | MRI_Dataset_2 | MRI-dataset | free-form-region-of-interest | left-anterior-cingulate-gyrus | volume | 9808 |
| Patient9 | MRI_Dataset_2 | MRI-dataset | free-form-region-of-interest | right-anterior-cingulate-gyrus | volume | 8788 |
| Patient9 | PET_Dataset_2 | PET-dataset | free-form-region-of-interest | left-anterior-cingulate-gyrus | nuclear-medicine-tomo-activity | 19643 |
| Patient9 | PET_Dataset_2 | PET-dataset | free-form-region-of-interest | right-anterior-cingulate-gyrus | nuclear-medicine-tomo-activity | 21108 |
| Patient11 | MRI_Dataset_3 | MRI-dataset | free-form-region-of-interest | left-anterior-cingulate-gyrus | volume | 8191 |
| Patient11 | MRI_Dataset_3 | MRI-dataset | free-form-region-of-interest | right-anterior-cingulate-gyrus | volume | 6466 |
| Patient11 | PET_Dataset_3 | PET-dataset | free-form-region-of-interest | left-anterior-cingulate-gyrus | nuclear-medicine-tomo-activity | 21224 |
| Patient11 | PET_Dataset_3 | PET-dataset | free-form-region-of-interest | right-anterior-cingulate-gyrus | nuclear-medicine-tomo-activity | 21562 |

Fig. 10. Table gathering the results of query Q_1 .

| Patient | Dataset | Function | FunctionType | Var1Type | Var2Type | Var3Type |
|-----------|---------------|---|--|------------|------------|------------|
| Patient4 | MRI_Dataset_1 | T1-weighted-MR-signal-intensity-function_0 | T1-weighted-MR-signal-intensity-function | X-location | Y-location | Z-location |
| Patient4 | PET_Dataset_1 | nuclear-medicine-tomo-activity-function_0 | nuclear-medicine-tomo-activity-function | X-location | Y-location | Z-location |
| Patient9 | MRI_Dataset_2 | T1-weighted-MR-signal-intensity-function_10 | T1-weighted-MR-signal-intensity-function | X-location | Y-location | Z-location |
| Patient9 | PET_Dataset_2 | nuclear-medicine-tomo-activity-function_10 | nuclear-medicine-tomo-activity-function | X-location | Y-location | Z-location |
| Patient11 | MRI_Dataset_3 | T1-weighted-MR-signal-intensity-function_20 | T1-weighted-MR-signal-intensity-function | X-location | Y-location | Z-location |
| Patient11 | PET_Dataset_3 | nuclear-medicine-tomo-activity-function_20 | nuclear-medicine-tomo-activity-function | X-location | Y-location | Z-location |

Fig. 11. Table gathering the results of query Q₂.

an instance of *free-form-region-of-interest* on which is anchored an instance of *region-of-interest-annotation*. It states that the instance of *region-of-interest* exactly delimits the *left-anterior-cingulate-gyrus* whose *volume* is equal to *PET-Dataset_1*. Second, for the same instance *PET-Dataset_1*, the instance of *region-of-interest* that exactly delimits the *right-anterior-cingulate-gyrus*, whose *volume* is equal to *PET-Dataset_1*, is proposed. The same subject is concerned by *PET-Dataset_1* which has for part two instances of *free-form-region-of-interest*. On the first ROI is anchored an instance of *region-of-interest-annotation*, that states that the instance of *region-of-interest* exactly delimits the *left-anterior-cingulate-gyrus* whose *activity* is equal to *PET-Dataset_1*. Similarly, on the second ROI is anchored an instance of *region-of-interest-annotation* that states that the instance of *region-of-interest* exactly delimits the *right-anterior-cingulate-gyrus* whose *activity* is equal to *PET-Dataset_1*.

Note that, in Q₁ only the anatomical entity *anterior-cingulate-gyrus* was initially specified (line 16). However, because in the ontology left and right anterior cingulate gyrus classes are subsumed by *anterior-cingulate-gyrus*, the system can infer that the left and the right specializations are also anterior cingulate gyrus and

so retrieves them. The instruction at line 17 of the same query retrieves the direct type of this anatomical entity and informs the user that the anatomical entities are the left and the right specializations of the *anterior-cingulate-gyrus*.

Similarly, the general concepts of *dataset* (1) and *region-of-interest* (7) were initially indicated and the system retrieves for instance, *PET-Dataset_1*, an instance of a *PET-dataset*, and a subclass of *region-of-interest*, namely *free-form-region-of-interest*.

Use case 2: This second example illustrates the kind of reasoning that may be performed based on the explicit representation of the mathematical functions associated with *Datasets*. A registration between a PET and an MRI dataset can only be achieved if (1) the mathematical function attached to each dataset is a scalar function, and (2) their domain is defined according to the same three space variables. Such a verification may be automatically achieved by a file selector in order to filter only the datasets that meet these two conditions. The corresponding query Q₂ retrieves information about space variables (11–16) through the *Scalar Functions* (3–5), which represent the *Datasets* concerning a subject (1–2), and the *Intervals* (8–10), which compose the domain (6–7) of the *Scalar Function*.

```

Q2: SELECT ?Patient ?Dataset ?Function ?FunctionType ?Var1Type ?Var2Type ?Var3Type DISPLAY xml WHERE{
(1) ?Dataset rdf:type ds:dataset
(2) ?Dataset ds:concerns ?Patient
(3) ?Dataset ds:is-represented-by ?Function
(4) ?Function rdf:type fct:scalar-function
(5) ?Function rdf:type ?FunctionType
(6) ?Function fct:has-for-domain ?CartesianProduct
(7) ?CartesianProduct rdf:type cp:triple-cartesian-product
(8) ?CartesianProduct cp:has-for-first-set ?xInterval
(9) ?CartesianProduct cp:has-for-second-set ?yInterval
(10) ?CartesianProduct cp:has-for-third-set ?zInterval
(11) ?xInterval dol:q-location-of ?Var1
(12) ?Var1 rdf:type ?Var1Type
(13) ?yInterval dol:q-location-of ?Var2
(14) ?Var2 rdf:type ?Var2Type
(15) ?zInterval dol:q-location-of ?Var3
(16) ?Var3 rdf:type ?Var3Type
}

```

Q₂ retrieves *Datasets* represented by *Scalar Functions* to satisfy the first condition and as well as the semantics associated with the *Intervals*, which compose the domain of the associated *Mathematical Function*, to satisfy the second condition.

The result of query Q₂ is illustrated in Fig. 11.

Fig. 11 shows for all *Datasets* which are represented by *Scalar Functions*: the patient concerned, the direct type of the function which specializes the *scalar-function* (e.g. *nuclear-medicine-tomo-activity-function*), and the semantics associated with the *Intervals* which compose the *Cartesian Product* defining the domain of these functions.

These two use cases illustrate that, while a concise query is requested, the system, thanks to our conceptual model, automatically broadens the search to retrieve relevant data.

5. Discussion

Most existing databases for complex data types allow the user to retrieve data sets based on metadata. However, because of a lack of a rigorous modelling of complex imaging data and relations, of explicit and semantically precise metadata, queries are intrinsically limited to those anticipated by the database designers. Especially, it severely hampers the capabilities of data mining that could – if such metadata were present and accurate – give access to potentially relevant discoveries [4,24]. As mentioned in [25], several bases of neuroscience data are existing and available on the internet, but failed to be truly usable because of a lack of organization, annotation and association with appropriate search tools. They remain a convenient method for storing at individual re-

searcher level or sharing data for well defined multi-centre studies [26], but their integration into federated systems remains very challenging because of their intrinsic heterogeneity. In contrast, ontology allows making inferences based on the semantics of concepts and relations. This capability, mastered by the ontology designers, enlarges the set of possible queries compared to a standard data base. A first basic example consists in searching for *MRI Datasets* expressed in DICOM format. Most classical database implementation would simply use the DICOM format and image modality as search criteria. Our model makes it possible to query such *Datasets* at multiple levels of the *Dataset* hierarchy, e.g. either using a general class *MRI Dataset*, or using a more specific one like *T1-weighted MRI Reconstructed Dataset* or *FLAIR-weighted MRI Reconstructed Dataset*. Similarly, our model makes it possible to specify, either that any kind of *DICOM expression* is searched for, i.e. any kind of *DICOM Service Object Pair* (SOP) Class is allowed, or to specify a particular kind of *DICOM expression*, i.e. using the regular *DICOM MR Image Storage SOP Class* or using the *DICOM Enhanced MR Image Storage SOP Class*. A second example concerns a situation in which a user is searching for cases of patients whose MR Image Storage images show a *Brain Tumor locatedIn the Frontal Lobe*. Based on the ontology's knowledge, the system would be able to retrieve the case of a patient with a *Glioblastoma locatedIn the Inferior Frontal Gyus*, since a *Glioblastoma* is a *Brain Tumor* and the *Inferior Frontal Gyus isPartOf the Frontal Lobe*, and the relationship *locatedIn* satisfies the property that any entity₁, which is *locatedIn* an entity₂, is also *locatedIn* any entity that *hasForPart* entity₂. Clearly, this kind of result could not be obtained with a standard data base. Another way of exploiting the semantics embedded in the ontology amounts to calculate "semantic distances" between concepts (these distances rely on topological distances calculated in the graph corresponding to the structure of the ontology). Such semantic distances enable query engines to give approximate answers. Such a facility is for instance provided by the semantic research engine CORESE that we use in our project.

Thus, the use of ontologies helps overcoming the difficulties encountered within conventional databases, by providing precise definition of each concept and relation, and defining the common unified schema for the mapping of the local database schemas. Our work with OntoNeurobase brings two major contributions. The first is related to the general methodology we propose to build a multi-layered and multi-components application ontology; the second concerns the domain ontology we have designed for neuroimaging. These contributions are then put in perspective with similar works being carried out in the context of projects such as caBIG and BIRN.

5.1. Multi-layered and multi-components application ontology

Our approach aims at mastering two complexities. The first is a *conceptual complexity*, arising from the intrinsic nature of the entities belonging to our universe of discourse (i.e. medical images and their annotations). The second is a *design complexity* related to the need to articulate our model with other models, either existing or in development, addressing the needs of connected domains such as biology, clinical medicine, anatomy, physiology, etc. so that our contribution can be managed independently, while fitting into a consistent whole.

Our basic methodology is to model entities at different levels of abstraction, by re-using a set of core ontologies, based on a common foundational ontology.

Foundational ontologies provide rigorous logical axiomatisation explicating ontological commitments, making possible to reason about entities and to map ontologies in the future. DOLCE and the Basic Formal Ontology (BFO) [27], an ontology developed by IFOMIS and widely used in the biological sciences, are the favorite

candidates which propose rigorous foundational principles to model our domain. These ontologies have been elaborated in the context of the WonderWeb project [14], whose ultimate aim was to build a library of foundational ontologies, precisely to establish the foundations enabling the "negotiation of meaning" between agents. This work leads to the possibility of mapping ontologies conceived according to different philosophical approaches, as well as to a better understanding of the difficulties related to such mappings.

Our choice of DOLCE, considered as a reference by many authors, e.g. [28–30], was motivated by three major factors. The first is related to its rich and well-documented axiomatization as to location in space and time, dependence and parthood, and to the fact that it relies on explicit structuration principles [14]. Moreover, it is based on the OntoClean [9] methodology, thus providing a precious guide to structuring application ontologies, especially regarding taxonomic relationships. The second argument in favor of DOLCE is the availability of numerous extensions, such as DOLCE-Lite-Plus and many core ontologies, related to participation roles, semiotics, collections, artifacts, and manufactured objects, addressing difficult-to-model domains in which it would have been unrealistic to attempt significant work by ourselves. Finally, the third factor lies in the basic principles retained in DOLCE which we considered particularly relevant in our context. The deliberate choice of a "cognitive bias" (i.e. depending strongly on human perception and social conventions) proved relevant for modeling human artifacts such as dataset expressions, or mathematical concepts. Similarly, DOLCE's multiplicative approach (i.e. authorizing several entities to be co-localized in the same space-time) seemed appropriate for modeling spatio-temporally co-localized entities, such as an anatomical structure and the generator of a functional activity, although it is quite clear that both are inherent to the same biological reality.

BFO adopts a realist approach and then is reluctant to speak about categories which are language dependent. Our feeling is that annotating images and referring to brain pathology would probably require modelling entities such as language acts and cognitive states, which may be more difficult to introduce based on the BFO philosophical approach.

However our choice of DOLCE should not hide what is – in our vision – the most important aspect: the design of an ontology, should be based on a foundational ontology whatever it is. We claim that this is more important than the choice of a particular foundational ontology, since foundational ontologies provide the rigorous logical axiomatisation, making possible to reason about entities and to map ontologies in the future. In this regard, what is of paramount importance is to capture sufficient semantics in order to enable subsequent mappings between partly overlapping ontologies (since such overlapping seems inevitable). The use of a methodology such as OntoSpec, based on OntoClean, authorizing a semi-informal representation of semantics appears important in this respect.

It is certainly too early to judge of the added value of DOLCE and the reused core ontologies in facilitating the integration of multi-domain information (e.g. anatomy, physiology, pathology, image processing, biology), which is expected from any upper layer or foundational ontology. Clearly, only confrontation with experiments would support our claim, and the NeuroLog project (see Section 6) plans to carry out such work.

5.2. Domain ontology for neuroimaging

5.2.1. Image data

Our modeling helps clarify the various connotations attached to images. We distinguish what relates to physical entities, such as file materialization, or rendering on computer screens, from non

physical entities such as expressions according to various formats. Much remains to be done concerning this last point. In the present work, we have focused on what images refer to rather than rendering issues, such as windowing, 3D rendering or “blending” of multimodal values. Our current objectives lie in the sharing and reuse of data and image processing tools rather than in display applications. Data formats are insufficiently explicit, especially regarding the mathematical aspect, to allow reasoning about data and image processing and the composition of innovative image processing pipelines. Our categorization of datasets is a step towards achieving this goal.

Our accomplishments remain modest, especially when compared to a standard like DICOM in which the descriptions of image structure, semantics and metadata represent approximately 1000 pages of specifications, addressing the details and specificity of each imaging modality. However, the orientations we have developed have sufficient generality to enable revisiting the standard based on ontological principles. This is certainly a huge job, which should be conducted progressively to provide significant added value even at the early stages of its completion. Needs in this area have already emerged, for instance in the context of DICOM Working Group 23, “Application Hosting”, which addresses the issue of defining a standard API for image processing tools (such as plugins or Web services). This obviously requires that the semantics of the image data being processed be properly modeled and shared.

5.2.2. ROI annotations

For ROIs and annotations, the proposed models constitute a first step. The objective was to meet the most common requirements, such as referring to real-world entities, with relatively precise semantics. This allows distinguishing the case where an ROI exactly delimits an entity, e.g. an anatomical structure, from cases where it contains only a part of it, or conversely, where it belongs to a region of space that contains more than this structure. These relationships are intended to be used for spatial reasoning [31] in conjunction with formal ontologies of anatomy that support mereological properties, such as the Foundational Model of Anatomy (FMA) [32]. This point raises the issue of aligning FMA, or a brain-related subset of FMA, with foundational ontologies such as DOLCE or BFO, since they include their own Theory of Parts, whose compatibility with FMA should be assessed with caution.

One limitation of our ROI annotation model is that it requires representation of a separate instance of *ROI Annotation* for every *Quality* concerning an ROI. Thus, if we wish to represent both the mean and standard deviation of a signal intensity over an ROI, we must define two separate *ROI Annotations* associated with the same ROI. An alternative would have been to adopt a complex structured model of *ROI Annotations*, such as the one used in DICOM SR. This direction is currently being explored in the context of the “Annotation and Image Markup” project, a sub-project of the CaBIG initiative. Our feeling is that it may lead to over-complex implementation, compromising efficient querying in the most frequent cases.

5.2.3. ROI annotations and subjectivity

We consider important to establish relationships between the results of image processing (imaging biomarkers) applied to specific image regions, and real-world entities, while at the same time underlining the subjectivity of such relationships. This subjectivity concerns the whole observation context including the observer. Hence, although it may result from an automatic tool, the result of any processing is dependent on the specific tools used. For example, the numeric value obtained by the hippocampus volume computation from a structural image depends on the segmentation algorithm and if any, the pre-processing steps used, such as bias correction. We address this issue by using the “Participant Roles”

core ontology, very helpful in modeling the genealogy of the data. Subjectivity also concerns the categories of real-world entities that are referred to, which may depend on the observer – i.e. two observers may choose to refer to two different entities – but this may also evolve over time, since relevant new categories may appear based on the progress of knowledge [33]. This may lead to the creation of new annotations for the data, referring to these new categories. This possibility should not be underestimated for the future, since new categories, especially in the domain of pathology, are likely to emerge, e.g. based on genomic and proteomic data.

5.3. Relations with on going projects

Regarding other projects such as caBIG, interesting work is being performed in the context of the “In vivo Imaging workspace”, especially concerning annotation and image markup. Our work is clearly in line with this effort as well as complementary to the development of RADLex, a terminological resource for radiology developed by the Radiological Society of North America, although the ontological choices made in either projects are not explicit.

In the fields of neuroscience and neuroimaging, the Biomedical Informatics Research Network (BIRN) appears to date as the most advanced large-scale data integration effort. We share most of the general objectives described in the BIRN seminal paper [4], especially regarding the need to adopt a federated approach, and the need to found mediation on domain ontologies. A lot of efforts were deployed by the BIRN Ontology Task Force to reuse as much as possible existing terminologies such as UMLS (Unified Medical Language System), NeuroNames, SNOMED (Systematized Nomenclature of Medicine), GO (Gene Ontology), LOINC (Logical Observations Identifiers Names and Codes) etc., with a mapping between the different resources made via UMLS, or using the BONFIRE tool. With BIRN Lex, their most recent work is much more in line with our own approach, i.e. suggesting that such domain terminologies should be based on foundational ontologies [34]. However, we consider that UMLS has not a sufficient clear ontological foundation to support reasoning techniques as required for integrating heterogeneous data. Several difficulties have been reported using UMLS. For instance, Kumar and Smith in [35], based on a concrete example concerning the regulation of blood pressure, illustrate how well-formalized ontologies, contrary to a lightweight ontology such as UMLS, can detect and avoid conflicts. It appears that in UMLS the notion of cardiac output embraces both continuant and occurrent entities, due to a basic confusion between biomedical phenomena and their measurement in the context of a procedure.

Actually, their most recent achievements rely on work made under the auspices of the National Centre of Biomedical Ontology/Open Biomedical Ontologies foundry (Ontology for Biomedical Investigations, Common Anatomy Reference Ontology, BFO, Relation Ontology, GO, etc) and use BFO and RO as a foundational ontology. This alignment will certainly facilitate the inter-operability, by providing the semantic content which is needed to map the ontologies, eventually at different levels of abstraction. However, re-engineering existing terminologies to make them compliant to foundational ontologies such as BFO or DOLCE will take time. As far as we are concerned, our efforts go in the same direction, and we try to focus our contributions to those fields in which we are the most competent, i.e. imaging and image processing.

5.4. Extension and Interoperability

Extension and interoperability are key issues in knowledge engineering and ontology development in particular. As already mentioned, our ultimate aim is to define an ontology which is easily extensible, which allows integration of conceptualizations coming from different fields, and which ensures interoperability with

other ontologies. Currently, integration of conceptualization of domains such as anatomy or physiology is not realized yet. However, we plan in the next step to integrate FMA or some ontologies from OBO following a vertical strategy as proposed in [36]. According to this strategy, the most abstract concepts and relations defined in these ontologies (such as Anatomical structure, Pathological structure, Function, or Physiological state) are mapped to abstract concepts present in DOLCE (e.g. *Physical object*, *Feature*, *State*, and *Process*, respectively). Furthermore, the interoperability with ontologies conceived in the BFO framework can be facilitated by making a horizontal mapping between abstract concepts of DOLCE and BFO. On that purpose, we can indicate some correspondences between concepts which are either extensionally equivalent (e.g. *bfo:Object* = *dolce:Physical Object*, *bfo:Quality* = *dolce:PhysicalQuality*) or which hold subsumption relation (e.g. *bfo:Boundary of Object* < *dolce:Feature*).

6. Conclusion and perspectives

The next steps of this project will be carried out in the context of the NeuroLOG Project¹³, which received a grant from the French National Research Agency for 2007–2009. A specific workpackage concerns the consolidation and extension of the OntoNeuroBase ontology, especially with the aim of modeling image processing as well as processing tools, with the general perspective of sharing and reusing existing processing tools and creating new image processing pipelines. This extension will take into account three clinical application contexts, namely multiple sclerosis, stroke and brain tumors. A federated system will be deployed, associating research centers in Grenoble, Rennes, Sophia-Antipolis and Paris to share the images and processing tools “published” by these research centers. OntoNeuroBase will be used as a common semantic reference to align and query the potentially heterogeneous data available in the various repositories using OWL-Lite instances. This project gives us the opportunity to assess the validity of our choices and the relevance of OntoNeuroBase for large scale neuroimaging applications.

Acknowledgements

L. Temal is supported by a doctoral grant from the region of Brittany.

This work received financial support from the French Ministry of Research and Technology (“Action Concertée Incitative”).

References

- [1] Saltz J, Oster S, Hastings S, Langella S, Kurc T, Sanchez W, et al. caGRID: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics* 2006;22(15):1910–6.
- [2] Koslow SH, Hirsch MD. Celebrating a decade of neuroscience databases: looking to the future of high-throughput data analysis, data integration, and discovery neuroscience. *Neuroinformatics* 2004;2:267–70.
- [3] Gupta A, Ludäscher B, Grethe JS, Martone ME. Towards a formalization of disease-specific ontologies in neuroinformatics. *Neural Netw* 2003;16(9):1277–92.
- [4] Martone ME, Gupta A, Ellisman MH. e-Neuroscience: challenges and triumphs in integrating distributed data from molecule to brains. *Nat Neurosci* 2004;7(5):467–72.
- [5] Seeding the Europhysiome: a roadmap to the virtual physiological human. Co-ordination Action #027642, 2007. Available from: <http://www.europhysiome.org/>.
- [6] Barillot C, Benali H, Dojat M, Gaignard A, Gibaud B, Kinkingnehun S, et al. Federating distributed and heterogeneous information sources in neuroimaging: the NeuroBase project. *Stud Health Technol Inform* 2006;120:3–13.
- [7] Gangemi A, Borgo S, editors. Proceedings of the EKAW*04 workshop on core ontologies in ontology engineering. Northamptonshire (UK), 2004. Available from: <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-118/>.
- [8] Kassel G. Integration of the DOLCE top-level ontology into the OntoSpec methodology. LaRIA Research Report 2005–2008. Available from: <http://hal.ccsd.cnrs.fr/ccsd-00012203>. 2005.
- [9] Guarino N, Welty C. An overview of OntoClean. In: Staab S, Studer R, editors. Handbook on ontologies. Springer Verlag; 2004. p. 151–71.
- [10] Bell DS, Pattison-Gordon E, Greenes RA. Experiments in concept modeling for radiographic image reports. *JAMIA* 1994;1(3):249–62.
- [11] Fielding JM, Marwede D. The image as spatial region: location and adjacency within the radiological image. *Frontiers in artificial intelligence and applications. Formal ontology in information systems*. In: Bennett B, Fellbaum C, editors. Proceedings of the fourth international conference (FOIS 2006), vol. 150, 2006. p. 89–100.
- [12] Gemo M, Gouze A, Debande B, Grivegnée A, Macq B. A versatile knowledge-based clinical imaging annotation system for breast cancer screening. SPIE medical imaging: computer-aided diagnosis. San Diego, California, USA: The International Society for Optical Engineering; 2007.
- [13] Rector A. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In: Genari J, editor. Proceedings of the second international conference on knowledge capture (K-CAP 2003), ACM; 2003. p. 121–8.
- [14] Masolo C, Borgo S, Gangemi A, Guarino N, Oltramari A, Schneider L. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. WonderWeb Deliverable D18, Final Report (v. 1.0, 31-12-2003); 2003.
- [15] Fortier JY, Kassel G. Managing knowledge at the information level: an ontological approach. In: Proceedings of the ECAI'2004 workshop on knowledge management and organizational memories, Valencia, Spain; 2004. p. 39–45.
- [16] Sowa JF. Knowledge representation: logical, philosophical, and computational foundations. Pacific Grove: Brooks Cole; 2000.
- [17] Bottazi E, Catenacci C, Gangemi A, Lehman J. From collective intentionality to intentional collectives: an ontological perspective. *Cognitive Sys Res* 2006;7(2–3):192–208 [Special Issue on Cognition, Joint Action and Collective Intentionality].
- [18] Guarino N, Welty C. A formal ontology of properties. In: Dieng R, Corby O, editors. Proceedings of the 12th international conference on knowledge engineering and knowledge management: EKAW2000. Lecture notes on computer science. Springer Verlag; 2000. p. 97–112.
- [19] Bruaux S, Kassel G, Morel G. A Clarification of the ontological status of Knowledge Roles. In: Proceedings of the workshop on “Advances in Conceptual Knowledge Engineering”; 18th international conference on database and expert systems applications, DEXA 07, Regensburg, Germany; 2007. p. 529–33.
- [20] Temal L, Lando P, Gibaud B, Dojat M, Kassel G, Lapujade A. OntoNeuroBase: multi-layered application ontology in neuroimaging. In: Proceedings of the FOMI2006 workshop on formal ontology meet industry, Trento, Italy; 2006. p. 3–15.
- [21] Gruber T, Olsen G. An ontology for engineering mathematics. In: Doyle J, Torasso P, Sandewall E, editors. Fourth international conference on principles of knowledge representation and reasoning, Bonn, Germany: Gustav Stresemann Institut; 1994.
- [22] Corby O, Dieng-Kuntz R, Faron-Zucker C. Querying the semantic web with Core search engine. In: Proceedings of the 15th ECAI/PAIS, Valencia, Spain; 2004.
- [23] Haegelen C, Verin M, Aubert Broche B, Prigent F, Jannin P, Gibaud B, et al. Does subthalamic nucleus stimulation affect the frontal limbic areas? A single photon emission computed tomography study using a manual anatomical segmentation method. *Surg Radiol Anat* 2005;27:389–94.
- [24] Kotter R. Neuroscience databases: tools for exploring brain structure–function relationships. *Philos Trans R Soc Lond B Biol Sci* 2001;356:1111–20.
- [25] Ascoli GA. Mobilizing the base of neuroscience data: the case of neuronal morphologies. *Nat Rev Neurosci* 2006;7:318–24.
- [26] Hasson U, Skipper JJ, Wilde MJ, Nusbaum HC, Small SL. Improving the analysis, storage and sharing of neuroimaging data using relational databases and distributed computing. *NeuroImage* 2008;39:693–706.
- [27] Grenon P. BFO in a nutshell: a bi-categorical axiomatization of BFO and comparison with DOLCE, IFOMIS Report, Universität Leipzig, ISSN 1611-4019; 2003.
- [28] Yu AC. Methods in biomedical ontology. *J Biomed Inform* 2006;39:252–66.
- [29] Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. *Methods Inf Med* 2006;45(Suppl 1):124–35.
- [30] Héja G, Varga P, Pallinger P, Surjan G. Restructuring the foundational model of anatomy. *Stud Health Technol Inform* 2006;124:755–60.
- [31] Cohn AG, Hazarika SM. Qualitative spatial representation and reasoning: an overview. *Fundamenta Informaticae* 2001;46:1–29.
- [32] Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform* 2003;36:478–500.
- [33] Smith B. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. *J Biomed Inform* 2006;39(3):288–98.
- [34] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBI foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25(11):1251–5.
- [35] Kumar A, Smith B. The unified medical language system and the gene ontology: some critical reflections. In: Kruse R, Günter A, Neumann B, editors. KI 2003: advances in artificial intelligence. Springer Verlag; 2003. p. 135–48.
- [36] Rosse C, Kumar A, Mejino JLV, Cook DL, Detwiler LT, Smith B. A Strategy for improving and integrating biomedical ontologies. In: Proceedings of AMIA symposium 2005, Washington, DC, 2005. p. 639–43.

¹³ <http://neurolog.polytech.unice.fr/doku.php>.