Towards semantic-driven high-content image analysis: An operational instantiation for mitosis detection in digital histopathology

Daniel Racoceanu, F Capron

HAL Id: hal-01139965

https://hal.sorbonne-universite.fr/hal-01139965

Submitted on 8 Apr 2015

Towards Semantic-Driven High-Content Image Analysis.
An Operational Instantiation for Mitosis Detection
in Digital Histopathology

D. Racoceanu[a,b,c,*], F. Capron[d]

[a]*Sorbonne Universités, UPMC Univ Paris 06, LIB - UMR 7371 – UMR_S 1146, F-75013, Paris, France*
[b]*CNRS, UMR 7371, Laboratoire d'Imagerie Biomédicale, F-75013, Paris, France*
[c]*INSERM, UMR_S 1146, Laboratoire d'Imagerie Biomédicale, F-75013, Paris, France*
[d]*Sorbonne Universités, UPMC Pars 06, UH Pitié-Salpêtrière-CFx, Department of Pathology APHP, UIMAP, F-75013, Paris, France*

Abstract

This study concerns a novel symbolic cognitive vision framework emerged from the Cognitive Microscopy (MICO[1]) initiative. MICO aims at supporting the evolution towards digital pathology, by studying cognitive clinical-compliant protocols involving routine virtual microscopy. We instantiate this paradigm in the case of mitotic count as a component of breast cancer grading in histopathology. The key concept of our approach is the role of the semantics as driver of the whole slide image analysis protocol. All the decisions being taken into a semantic and formal world, MICO represents a knowledge-driven platform for digital histopathology. Therefore, the core of this initiative is the knowledge representation and the reasoning. Pathologists' knowledge and strategies are used to efficiently guide image analysis algorithms. In this sense, hard-coded knowledge, semantic and usability gaps are to be reduced by a leading, active role of reasoning and of semantic approaches. Integrating ontologies and reasoning in confluence with modular imaging algorithms, allows the emergence of new clinical-compliant protocols for digital pathology. This represents a promising way to solve decision reproducibility and traceability issues in digital histopathology, while increasing the flexibility of the platform and pathologists' acceptance, the one always having the legal responsibility in the diagnosis process. The proposed protocols open the way to increasingly reliable cancer assessment (i.e. multiple slides per sample analysis), quantifiable and traceable second opinion for cancer grading, and modern capabilities for cancer research support in histopathology (i.e. content and context-based indexing and retrieval). Last, but not least, the generic approach introduced here is applicable for number of additional challenges, related to molecular imaging and, in general, to high-content image exploration.

Keywords: cognitive microscopy, virtual microscopy, medical ontologies, cancer grading, breast cancer, semantic reasoning, high-content image exploration

---

*Corresponding author, E-mail: daniel.racoceanu@upmc.fr
[1]MICO - a French National Research Agency, Technologies for health and autonomy (ANR TecSan) project: `https://www.comp.nus.edu.sg/~danielr/projects.htm`

1. Introduction

Cancer is a deadly disease and major cause of death worldwide. Every year, 7.4 million people (about 13% of the people dying worldwide [1]) die because of cancer.

Breast cancer is a very common type of cancer, especially for woman, for whom it comprises 22.9% of all cancers [2]. Breast cancer is also a very deadly disease. In 2013, 234 580 new cases and 40 030 deaths from breast cancer are estimated in the United States [1]. The estimated new cases of breast cancer in the United States in 2013 were about 232 340 cases in female and 2 240 in male. The mortality was about 39 620 cases for female and 410 for male, with 1 woman out of 5,7 sick dying from breast cancer [1]. About 1 in 8 U.S. women (almost 12%) developed invasive breast cancer over the course of her lifetime[2] [1]. In Europe, breast cancer is the most current form of cancer for women, with 42 000 new cases a year [3].

The last few decades, digital technologies have been emerging on the market, like high resolution histopathological slide scanners or efficient software viewers for large-scale histopathological images. Similarly, pattern recognition image processing algorithms have been shown to be quite effective and mature for medical applications [4]. However, these technologies are still not being intensively used in pathology departments, and will not be adopted without being properly adapted to the medical domain and practices.

Cautious in creating and respecting new protocols in digital pathology, this study proposes a novel approach in cognitive vision, combining imaging, reasoning and ontologies. The work has been done in the framework of the Cognitive Microscopy project (MICO[1]), dealing with a cognition-driven visual explorer for breast cancer grading. The traceability of a second opinion being a key issue, the proposed system is so able to explicitly consider the visual (and contextual) medical knowledge, by indicating at each step, the rules and the ontologies used to generate the quantitative diagnosis support.

After presenting the challenge of the cognitive virtual microscopy, we analyse the problem of knowledge representation in digital histopathology, in order to further describe the effective use of reasoning, ontologies and imaging tools, in the context of breast cancer grading.

2. Cognitive Virtual Microscopy

Histopathology became widely accepted as a powerful exam for diagnosis and prognosis in mainstream diseases such as breast cancer. Currently, analysis of medical images in histopathology largely remains the work of human experts. For pathologists, this consists of hundreds of slides examined daily[3], a tedious repetitive work which will greatly benefit from reliable and traceable digital second opinion. Beside, the work of a pathologist is mostly based on an intensive visual micro-semiology training (at least a decade to reach a real expertise in most of specialties) and know-how, embedded in a very complex medical contextual knowledge.

---

[2]www.breastcancer.org

[3]For exemple, at the Pitié-Salpêtrière Hospital, Paris, about 1500 slides need to be examined daily by the Pathology Department.

In this sense, considering the cognitive part of the pathologist's work is essential when we envisage a reliable second opinion support. The Cognitive Microscope (MICO[1]) initiative aims at developing a cognition-driven visual explorer for histopathology, focusing on the use case of mitotic count in breast cancer grading. MICO proposes a radical change in medical practice by proposing a cognition-driven medical imaging environment, enabling safer decision support in histopathology.

The system integrates image visualisation and exploration support, as the use of specific visual medical knowledge, in order to set up the bases of a cognition-driven medical imaging, able to address pervasive adaptation and physical-virtual confluence challenges for histopathology. The analysis capabilities and results are made available to the pathologist through a platform combining virtual microscopy and cognitive reasoning. This allows medical staff to interact with the platform at the appropriate level of abstraction. The platform combines multi-scale whole slide image (WSI) exploration and analysis, as medical knowledge representation and inference using ontologies.

Among the existing cognitive vision approaches [5], we chose to go for the symbolic one, enabling a better interaction with the users (the pathologists - in our case), a traceability for verification and quality control, as an effective maintenance capability, by using referenced on-line medical thesauri (Snomed-CT[4], ADICAP[5] etc.).

3. Knowledge Representation in Histopathology

Histopathology is the study of signs of disease using the microscopic examination of a biopsy or surgical specimen processed and fixed onto glass slide [6]. Histopathologists often are experts able to analyse tissue samples on glass slide using light microscope to grade cancers [3]. Sample observation is important because it allows the expert to make a judgement on the state of the disease. Therefore, it helps to target the appropriate cure. Histology involves invasive procedures and is the most accurate way of grading cancers so far. The histopathological diagnosis plays also a legal role in most of advanced countries.

Accurate grading of invasive breast cancer requires good fixation, processing, section cutting, staining and careful application of grading criteria[6].

A pathologist may have to analyse different kind of biological, cytological or histological extractions. Since the cytology is the analysis of cells and no information is kept about their spatial organisation, the histology represents the study of a tissue. There are different techniques for tissue extraction. In the case of breast cancer, biopsy, tumorectomy, mastectomy or axillary curage may be operated, depending mostly on the nodule size. After its extraction, the tissue is analysed using a microscope. Figure 1 shows an example of lesion induced by breast cancer [7].

3.1. Breast cancer grading

In case of solid tumour, cancer diagnosis is morphological and made using human tissue samples [3]. Histological slides are extracted from these samples. Histopathologists exper-
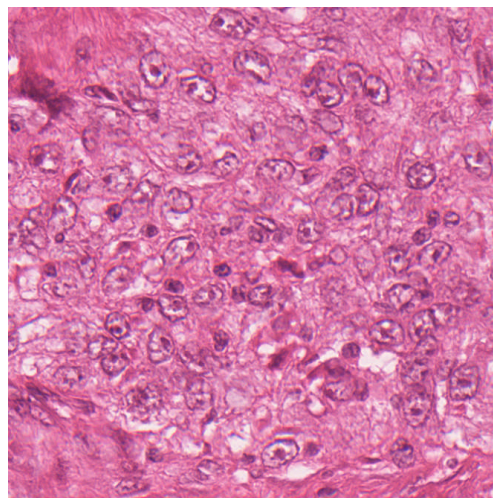
---

[4]SNOMED Clinical Terms (Snomed-CT): ihtsdo.org/snomed-ct

[5]French Association for the Development of the Informatics in Cytology and Anatomo-Patholog (ADICAP): adicap.asso.fr

[6]cancerscreening.nhs.uk.
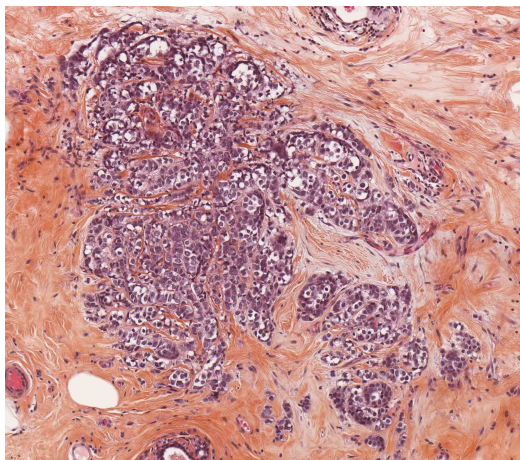
HES staining (Safran is used in some hospitals for the collagen fibers enhancement)

HE staining - classic international standard staining

Figure 1: Examples of histological images of breast lesion.

tise enables medical diagnosis by the observation of these histopathological slides. Figure 1 shows an example of what histopathologists may see using light microscope.

Histological grading helps to categorise the breast cancer, and therefore to treat it better. Breast cancer grading is now based on international standards. For a long period of time, the most widespread grading system has been the Scarff-Bloom-Richardson (SBR) system. Recently, the Nottingham Grading System (NGS) [8], a more reproducible variation of it [7], has become increasingly popular, being recommended -as mentioned- by the World Health Organisation.

The NGS consists in the study on three criteria, by establishing a cancer state evaluation protocol using of the respective scores. The three analysed criteria are tubules formation, nuclear pleomorphism and mitotic count [9]. A grade of 1 depicts a low level of cancer for the patient, while a grade of 3 means an advanced state of cancer. In the UK, in 2005, for example, about 20% of symptomatic breast cancers were grade 1, 30% grade 2, and 50% grade 3 [6].

Histopathological slide analysis is a complex task requiring advanced expertise. There are some drawbacks to the current histological grading process. Some grading tasks like counting mitosis can be very tedious and time consuming, inducing delay before diagnosis. Moreover, in most of the hospitals there are fewer and fewer experts, since increasingly important quantities of histopathological slides are produced every day[3]. Furthermore, one must admit that the criteria ruling histopathological analysis or grading are not fully formally described. Actually, there is currently no way to formally double-check judgements from histopathology experts about biological objects (for instance, it is very difficult to formalise how looks a mitosis and why a particular pattern corresponds to a mitosis). This lack of formalisation might be one reason of the long training and expertise time needed for the histopathologist to become a recognised pathologist expert in his medical area. Beside, even after reaching this expertise, one pathologist need to remain polyvalent, in order to insure the reactivity of the whole Pathology service/department if needed (extemporal
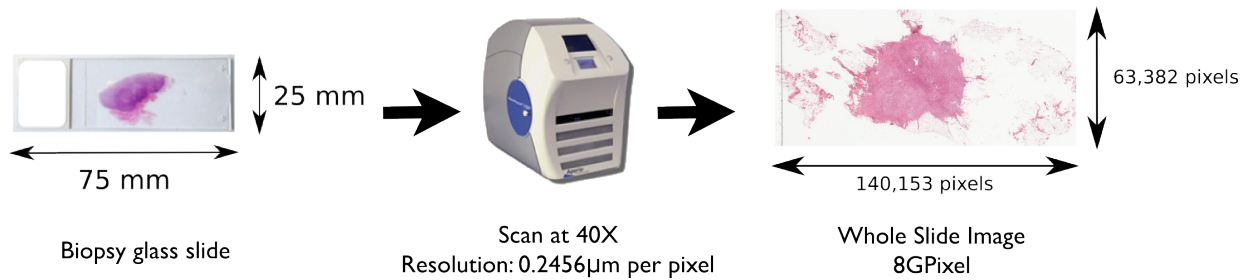
4

Figure 2: Slide digitisation process

histologic examination, holidays, expert absences, etc.).

Reproducibility of diagnosis is also an issue either between different pathologists (cross expertise in tele-pathology), post-operation exam, or longitudinal follow-up. In all these situations, a minimum of formalism supporting the pathological demarcate will be of great use, together with a flexible use of pattern recognition tools available on the market.

3.2. Medical technologies progress

Lastly, new medical technologies have been released on the market, such as high resolution slide scanners, or large medical image software viewers. These technologies are not currently being used – as a part of the routine protocol – in hospitals. However, they might help to improve the cancer grading process.

High resolution slide scanners. may be used to scan the histopathological slides (Figure 2). Some scanners are able to digitize several slides at a time. The images produced by a slide scanner are known as Whole Slide Images (WSI). A single image is generally about 140 000 by 60 000 pixels RGB (3 bytes per pixel), that is about 8 gigapixel. Each image is usually about 1 to 2GB when compressed, and 15 to 25GB when uncompressed. Such huge images require special softwares to be manipulated. However, WSI represents an efficient way to store the slides, protecting their critical information from degradation.

Slide visualisation softwares. Whereas histopathologists are using light microscope to observe the slides, slide visualisation softwares offer interesting capabilities to virtual microscopy. First, computerized view allows displaying different parts of the WSI next to each other, in order to compare them, which can be very useful (for example) during nuclear pleomorphism analysis. This live reference-structure is impossible with a microscope, were the histopathologist has to look at one part of the WSI, memorise it and then move to another part of the image, basing his/her judgement on his/her (somehow volatile) visual memory. Second, visualisation tools allow to precisely measure the size of the observed biological objects. Again, this might be useful in the case of nuclear pleomorphism analysis or mitosis candidates pre-selection. Third, the viewers are very promising for collaborative work. For instance, they allow the pathologist to add cross-annotations on the whole slide image (Figure 3), for a better collaboration and discussion between experts. Also, most of the WSI viewers platforms[7] are able to share WSI and the information associated to collab-

---

[7]CaloPix is an example of visualisation and data structure platform and Teleslide, a server application allowing sharing expertise and collaborating annotations on WSI : http://tribvn.fr/
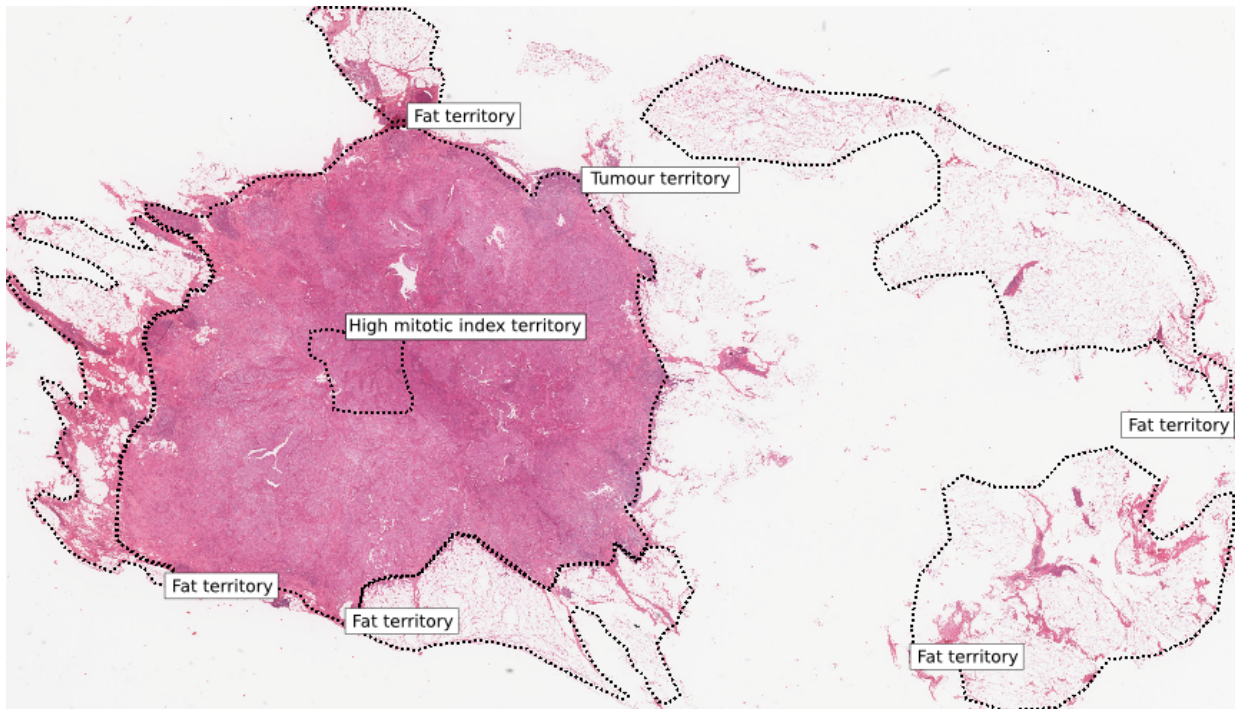
Figure 3: Example of WSI annotations using Calopix

orating pathologist, via internet. This is an important breakthrough in this field since, still in the current clinical practice, the histopathologists are sending slides to each other (for diagnosis confirmation) by postal mail. Postal mail is slow, and worse, the slide is broken in transit from time to time. Fourth, viewers are also great tools to display information from different origins. Unified information storage infrastructure, or PACS (Picture Archiving and Communication Systems)[8] store general information (age, sex, medical antecedents) and other information like screening (scanner, MRI, etc.) coming from different parts of the hospital. Allowing the histopathologists to access all the information about a patient (including the WSI) from the same interface, would help them to take into account relevant contextual information while grading the WSI. Finally, software viewers may also be considered as user interfaces for image processing macros or algorithms.

Due to these new technologies, histopathology is about to significantly evolve. Some procedures will probably need to be modified or completely revised to comply with the new tools, and it is hoped that new methods will also emerge (i.e. routine telepathology [3]), in the interest of both histopathologists and patients as, after all, in the interest of the quality of healthcare. However, the right tools still need to be built, and the pathologists need to learn how to use them in clinical routine.

3.3. Proposed Digital Pathology Protocol

At the initiative of the pathology department of La Pitié-Salpêtrière Hospital, we introduced a new digital pathology protocol for pattern collection and annotation. This Whole

---

[8]i.e. AGFA Healthcare' PACS http://www.agfahealthcare.com/global/en/main/
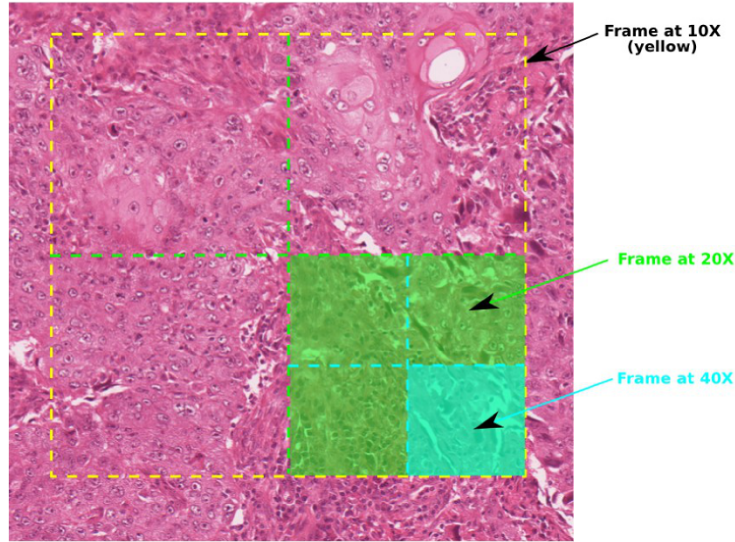
Figure 4: Collection and WSI exploration model for digital histopathology

Slide Image (WSI) computer-assisted analysis protocol uses a simple and efficient structured magnification scheme: the explored territory is sampled in 10X frames, embedding 4 frames 20X each, each of them embedding again 4 frames at 40X (See Figure 4). This will constitute a standard digital pathology protocol we are extending at the moment to a larger French consortium in telepathology (FlexMIm[9] FUI project) and we are planning to extend as a standard protocol, at a larger scale, in the future, for the benefit of a faster and wide clinical adoption of digital pathology and telepathology.

3.4. Knowledge representation and reasoning

3.4.1. Semantic Web technologies

As discussed, pathology is a highly (visual) cognitive medical speciality. However, most of information is currently available in a weakly structured form. Therefore, from the knowledge management perspective, the current information systems are not fully efficient for searching, extracting, maintaining and mining information. Indeed, the keyword-based search reaches its limits, and tedious human intervention is often required to extract relevant information from documents [10].

Semantic Web technologies enable designing operational knowledge management systems, presenting the advantage of a large community adoption and a dynamic evolution. By organising knowledge in conceptual spaces according to its meaning, keyword based search is replaced by query answering, automated tools being so able to check inconsistencies, to extract new knowledge, and to present it in human-friendly way. Therefore, it becomes possible to extract existing knowledge or to make inferences over several documents [11].

---

[9]FlexMIm project: `https://www.comp.nus.edu.sg/~danielr/projects.htm`

3.4.2. A matter of meta-data

Currently, most of the information targets human readers, rather than targeting programs. For instance, most of the services, technical documentation or web pages are written to be understandable by human only. In addition to containing formatted information aimed at human, they could also contain information about their content, or meta-data. Such representation is far more easily processable by machines. Indeed, the functionalities outlined in the previous paragraph will not happen using text-based manipulation of information, but by taking advantage of machine-processable meta-data [10]. The term meta-data refers to such information: data about data. Meta-data capture part of the meaning of data, thus the term semantic in Semantic Web.

3.4.3. Ontologies

There are various definitions of an ontology. Merriam Webster's dictionary [12] defines Ontology as *a branch of philosophy which studies what exists in all areas of reality.* In Artificial Intelligence (AI), the meaning of ontology (with a lower-case "o") can be slightly different. A suitable definition is given by Studer [13] : an ontology is a *formal, explicit specification of a shared conceptualization.* A 'conceptualization' refers to an abstract model of some phenomenon in the world by identifying the relevant concepts of that phenomenon. 'Explicit' means that the type of concepts used and the constraints on their use are explicitly defined. 'Formal' refers to the fact that the ontology should be machine understandable, which excludes natural language. 'Shared' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.

Concretely speaking, an ontology consists in a *finite list of concepts and relationships linking those concepts, formally describing a domain.* The concepts may be classes or instances, and the ontology may be viewed as a graph. Typically, the relationships include hierarchies of classes.

Ontological analysis. clarifies the structure of knowledge [14]. A domain ontology is the ineluctable core of any system of knowledge representation of the same domain. Without it, or similar conceptualization, there cannot be a vocabulary for representing knowledge. As ontologies help to increase the efficiency and consistency of describing resources, they allow for more sophisticated functionalities in knowledge management and information retrieval for application development [14]. Basically, the role of ontologies in the knowledge engineering process is to facilitate the construction of a domain model, by providing a vocabulary of terms and relations with which a domain can be modelled [13].

Ontology as a theory. Ontologies are content *theories* about the type of concepts and relations between concepts that are possible in a specified domain of knowledge [14]. In its strongest form, an ontology tries to capture universally valid knowledge, independent of its use, a view closely related to its philosophical origin. However, AI researchers quickly gave up this view, because it turned out that specific use of knowledge influenced its modelling and representation [15].

Within a given domain, an ontology is not just a representation – in a computer – of that domain but also claims to reflect a certain rate of consensus about the knowledge in that particular domain.

As all representations are surrogates for reality, and therefore distinct from it, they must be imperfect approximations of reality. Moreover, creating an ontology means selecting representations, in other words making a set of ontological commitments. These commitments of blurring or emphasising some aspects of the described objects are unavoidable (they are also useful because they allow to focus on the aspects relevant or useful to the domain application). As a consequence, a knowledge representation *is* a set of ontological commitments [16]. Thus, it may exist several concurrent ontologies describing the same objects. It doesn't mean one is better than the others, but maybe it is indeed better for a specific application domain.

Ontology levels. Depending on their generality degree, different kind of ontologies may be identified. Among others, we may distinguish the following types [14]:

- Upper level ontologies provide notions to which all concepts in existing ontologies are necessarily related. As a consequence, they are valid across several domains [13].

- General ontologies represent knowledge at an intermediate level of detail, independently of a specific task.

- Domain ontologies capture the knowledge valid for a particular domain. Within domain ontologies, the ontologies designed for specific tasks are called application ontologies.

Ontology design. While designing ontologies, it is a good practice to separately define assertions and terminological knowledge. Indeed, description logics traditionally split concepts and their relationships from the instances and their attributes and roles, expressed as fact assertions [17]. The concept split is known as the T-Box (terminological knowledge) and represents the schema or taxonomy of the domain at hand. The T-Box is the structural and intentional component of conceptual relationships. The second split of instances is known as the A-Box (assertions) and describes the attributes of instances, the roles between instances, and other assertions about instances regarding their class membership with the T-Box concepts. The T-Box is where the reasoning work occurs and the A-Box is where assertions occurs: While T-Box operations are based more on inferencing or verifying class memberships in the hierarchy, A-Box operations are more rule-based and govern fact checking.

The Resource Description Framework (RDF) is a language for representing information in the Web. Its purpose is to represent meta-data about Web resources, being designed to allow information to be processed by applications, instead of viewed / analysed by people [18]. RDF provides a common framework able to express this information in order to be shared between applications by keeping its original meaning. Being a common framework, application designers can leverage the availability of common RDF parsers and processing tools. The ability to exchange information between different applications means that the information may be made available to applications other than those for which it was originally created [18]. This characteristic represents a fondamental property of our system, initially designed for digital pathology, but usable for a large spectrum of high-content image/spaces analysis contextes.

RDF is based on the idea of identifying things using Web identifiers (called Uniform Resource Identifiers, or URIs), and describing resources in terms of simple properties and property values. This enables RDF to represent simple statements about resources as a graph of nodes and arcs representing the resources, and their properties and values.

RDF model encodes data in the form of triples [19]. Actually, the majority of data processed by machine may be described using triples [11]. It also enable us to translate human language into a model that machines can interpret. All sentences are sliced in triples, made of the following three entities [20]: *Subject*: what or who we are talking about ; *Predicate*: what property of the subject we are discussing ; *Object*: what we are saying about the property of the subject. The subject and object of a triple are both URIs that each identify a resource, or a URI and a literal string respectively. The predicate specifies how the subject and object are related, and is also represented by a URI [19] [11].

3.4.4. Resource Description Framework Schema

RDF properties may be thought of as attributes of resources and in this sense correspond to traditional attribute-value pairs. RDF properties also represent relationships between resources. RDF however, provides no mechanism for describing these properties, nor the relationships between these properties and other resources. That is the role of the RDF vocabulary description language, RDF Schema (RDF/S). RDF Schema is extending RDF vocabulary to define classes and properties that may be used to describe classes, properties and other resources. Such extension is required to allow the description of taxonomies of classes and properties, or to set the domain and range of properties. RDF Schema may be defined as a semantic extension of RDF [21].

RDF/S is often expressed using RDF/XML syntax. This syntax is often used to serialise a RDF graph as a XML document.

3.4.5. OWL

Web Ontology Language OWL is a semantic markup language designed to publishing and share ontologies on the World Wide Web [22]. OWL is developed as a vocabulary extension of RDF/S. OWL language provides two specific subsets. OWL Lite is designed for easy implementation and to provide a functional basic subset. OWL DL (where DL stands for "Description Logic") is designed to support the existing Description Logic business segment and to provide a language subset that has desirable computational properties for reasoning systems. The complete OWL language (called OWL Full to distinguish it from the subsets) relaxes some of the constraints on OWL DL so as to make available features which may be of use to many database and knowledge representation systems, but which violate the constraints of Description Logic reasoners [22].

3.4.6. Notation 3 & N3 Logic

RDF/ XML syntax is not very easily readable by humans. It is much more convenient to use the equivalent language Notation 3 (N3). N3 is a compact and easily readable alternative to RDF as XML syntax [23].

It has been mentioned previously that knowledge models such as Resource Description Framework (RDF/S) may provide a unifying representation of richly structured data. Adding logic to these data implies the use of rules to make inferences, to choose courses of action and answer questions. This logic must be powerful enough to describe complex

properties of objects. N3Logic is a first order based logic that allows rules to be expressed in an open environment. It extends RDF with syntax for nested graphs and quantified variables and with predicates for implication and accessing external resources. The main goal of N3Logic is to be a minimal extension to the RDF data model such that the same language can be used for logic and data [24]. N3Logic uses N3 syntax and extends RDF with a vocabulary of predicates. N3 aims to do for logical information what RDF does for data: to provide a common data model and a common syntax, so that extensions of the language are made simply by defining new terms in an ontology [24]. Since N3Logic uses first order logic, it is possible to do semantic inferences using this formalisation and a semantic inference engine. It exists several N3 compatible semantic reasoners, such as CWM[10] and EULER [11].

Nevertheless, there are are some particularities induced by the openness of the used knowledge bases: first, whereas a closed system may be built based on a single knowledge base of believed facts, an open Web-based system exists in an unbounded set of interconnected information resources. This requires that an entity be aware of the information source and responsible for its disposition. A language for use in this environment typically requires the ability to express which document or message asserts what. That is why N3Logic offers the possibility to quote facts [24]. Last but not least, to avoid paradoxes, N3Logic has been kept to limited expressive power: by excluding general first order negation [24].

Rules. So far, knowledge base facts have been described. Another kind of knowledge is described in this section: the rules. Rules allow one to infer new knowledge from existing knowledge [20]. Expressed using first order logic, a rule looks like:

$$\forall x, P(x) \Rightarrow Q(x) \tag{1}$$

The above reads "P implies Q", or in an instructive manner: "if the premise P is true, then the consequence Q is also true". It is possible to write such rules using N3.

Queries. Queries allow to extract targeted knowledge from the knowledge base. To be extracted, the knowledge must respect the constraints described within the query. Also, it is possible to choose the output format of the extracted individuals. Expressed using N3, queries take the same shape as rules, although their effect is quite different. The N3 queries premises describe the constraints that should be respected in order for individuals to be extracted. The N3 queries consequences describe the format of the results returned by the query.

3.5. Goal of the study

The goal of this study is to contribute to the cognition-driven aspect of the MICO platform. Indeed, until now ontologies and semantic tools have been used to exploit image processing results in order to infer medical knowledge, crucial to the pathologist. The vision of our approach also integrates the driving of image exploration & analysis using

---

[10]CWM (a forward chaining reasoner): http://www.w3.org/2000/10/swap/doc/cwm
[11]EULER (a semibackward reasoner): http://eulersharp.sourceforge.net/

semantic tools. Providing a semantic profile for each algorithm, gathering knowledge about the WSI to be analysed and matched with medical knowledge, it would be convenient to infer an adapted analysis process in order to get better and faster results. Besides the scientific deliverables, the MICO initiative leads to a Proof of Concept of a semantically driven medical image analysis platform based on the basic principles described here, and which will be deployed and validated by the pathologists. This platform will provide the opportunity to validate research concepts within an integrated system in order to allow moving forward at the clinical application level.

4. Operational Ontology-Driven High-Content Image Analysis

4.1. Conceptual graph

The first step in devising a knowledge representation based system is to perform an ontological analysis of the field or domain [14]. The following conceptual graph (Figure 5) is the result of a series of cross-knowledge meetings with the pathologists, involving researchers and computer scientists. It represents the premises of the structure used for the MICO semantic platform.

On the Figure 5, green and yellow colours represent concepts from histopathology, while grey and blue concepts may belong to imaging domain. Because there are overlaps between domains, the colour legend is approximative. The design of such conceptual graph shows the difficulty of merging views of experts coming from different domains within the same theory. This finding is responsible for some of the motivations going towards a split of MICO ontology into several ontological domains (Cf. Section 4.4.2).

4.2. Processes definition

Nottingham criteria evaluation on an histopathological slide image is a tedious and time consuming task for histopathologists. We address these issues and proposes a new approach for Nottingham grading, aiming at making it faster and more reliable, thanks to expert knowledge augmented computations. This section deals with proposals for Nottingham criteria automation. Nottingham criteria automated evaluation requires heavy image processing computations. However, it is not possible to compute the needed algorithms on a whole slide image within reasonable time. Strategies are proposed to overcome the processing power limitations. As the histopathology expert doesn't need to watch every part of the whole slide image in order to grade it, MICO aims to understand and reproduce this expertise during automated grading.

4.2.1. Mitosis analysis strategy

Mitotic score is one of the three Nottingham criteria. This score is quite easy to compute once the highest mitosis density within the histopathological slide is known. This section describes the digital strategy proposed for highest mitosis density evaluation without exhaustive analysis of the whole slide image (Fig. 6).

The digital protocol starts by a territory corresponding to the invasive area to be graded. This relevant territory is split into 40X frames (see Section 3.3). In the tumour, for the mitotic score, we are looking for the area having the highest concentration of mitotic nuclei. To perform this search faster, we operate a sampling on the set of frames covering
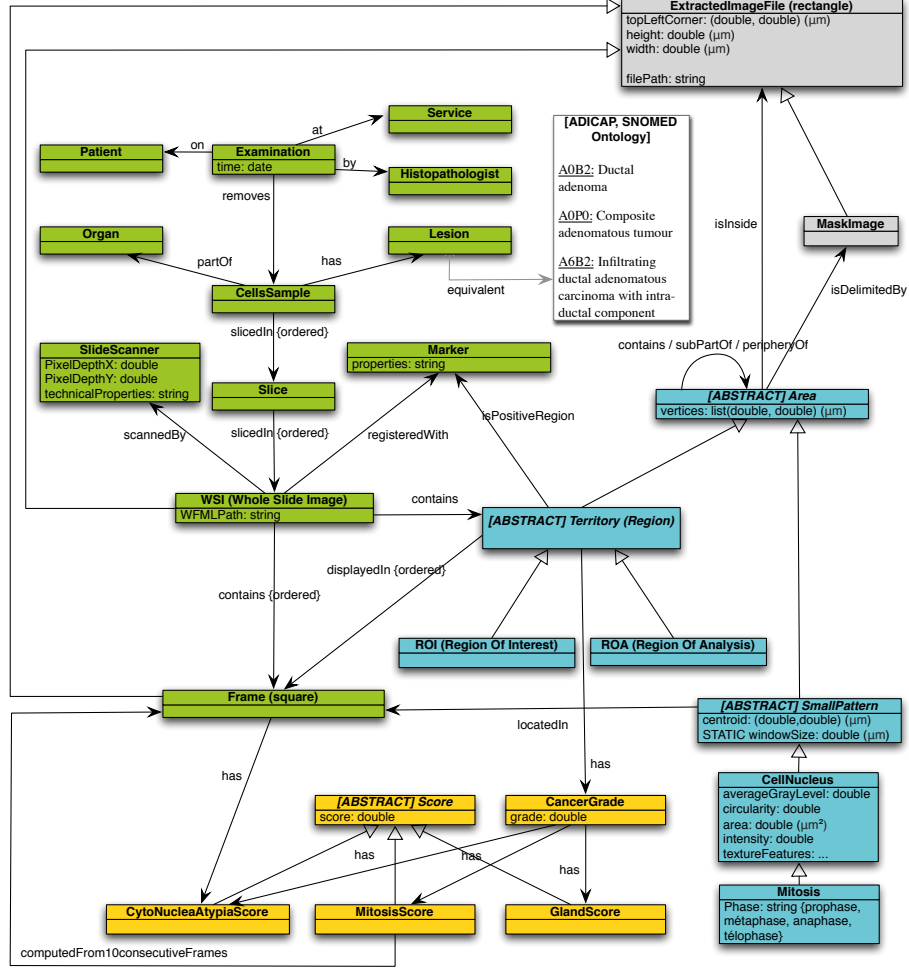
Figure 5: MICO conceptual graph {illustrative simplified example}

the tumour (see Fig. 6): the frames are grouped into blocks of 3×3 frames. In each 3×3 block, only the top left frame is analysed for mitosis detection. This sampling of one frame analysed out of nine gives us a broad picture of the concentration of mitosis in each area of the tumour. To select the area having highest number of mitosis, we now consider blocks of 4×4 frames such that the frames at each corner of a block have already been analysed for mitosis detection. We compute the sum of mitosis detected so far in each 4×4 block, and we select the block (or the blocks) having highest number of mitosis for further analysis. In this block of 16 frames, the four frames at the corners of the block have already been analysed. We now proceed to detection of mitosis on the 12 remaining frames of the block. So all the 16 frames of the block are now analysed. According to Nottingham Grading System, mitotic count is given for 10 adjacent frames. To comply with this definition, we select from our block of 16 frames the 10 frames with highest number of mitosis. The resulting mitotic count for the tumour territory will be equal to the sum of mitosis detected in these 10 frames. Of course, in case we had analysed more than one 4×4 block, the final mitotic count will be the highest one among all the 4×4 blocks so analysed.
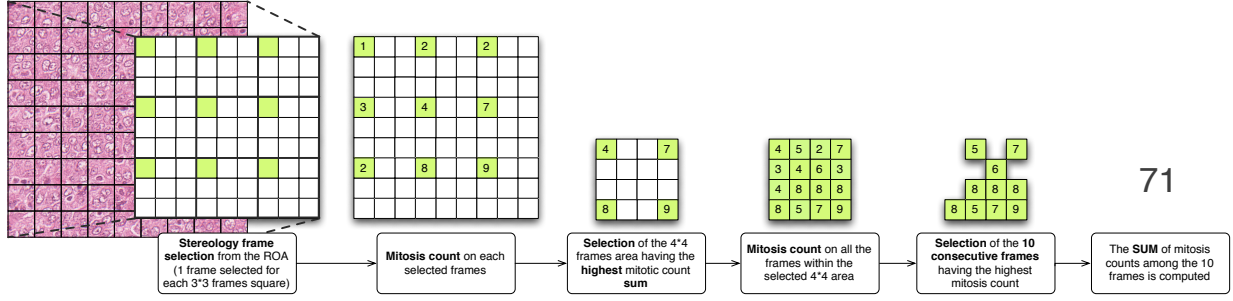
Figure 6: Mitosis count flow

This strategy gives an overview of what should be done to achieve a clinically acceptable mitosis score approximation. However, more details are needed for its realisation, like the different steps of the strategy, and the kind of data exchanged to go from one step to another. Therefore, mitotic scoring flow has to be split in modular algorithms, allowing a precise description in terms of inputs and outputs, in a more accurate manner.

4.3. Algorithms definition

Several modular algorithms are designed from the mitotic count strategy as a first step towards a cognitive platform realisation. In order to insure a broad usability of these modules, a particular attention is paid to the definition of the algorithms interfaces. Because its convenience for a semantic platform (Cf. section 4.4.3) and because its self-described caracteristics, easy to read and to maintain [25], Extensible Markup Language (XML) is chosen as the modality for information exchange with the modular algorithms. In addition of input output oriented design and well formed XML data samples, Document Type Declarations (DTD) are provided as definition of the grammar driving exchanges with the algorithms.

Seven algorithms are needed in the use case of the mitotic score analysis of an histopathological whole slide image:

TerritoryExtractor Extracts a labelled polygon from Calopix WFML file.

FrameGenerator Cuts a polygon into frames.

FrameSampler3x3 Outputs a sample of a given frame matrix (stereology) (Fig. 7).

FrameSelector4x4 Selects a 4x4 frames area from a given frame matrix according to the frames characteristics.

FrameSelector10 Selects ten frames from the 4x4 frame matrix according to the frames characteristics.

MitosisExtractor Extracts the mitosis of a given frame. This module is the most complex imaging module in the given example. The algorithm used here is issued from a recent quasi-exhaustive state of the art of our team [26] as the PhD work of Humayun Irshad [27] (Fig. 8).

ScoreCalculator Provides a mitotic score by aggregation of the mitotic count from the selected ten adjacent frames.
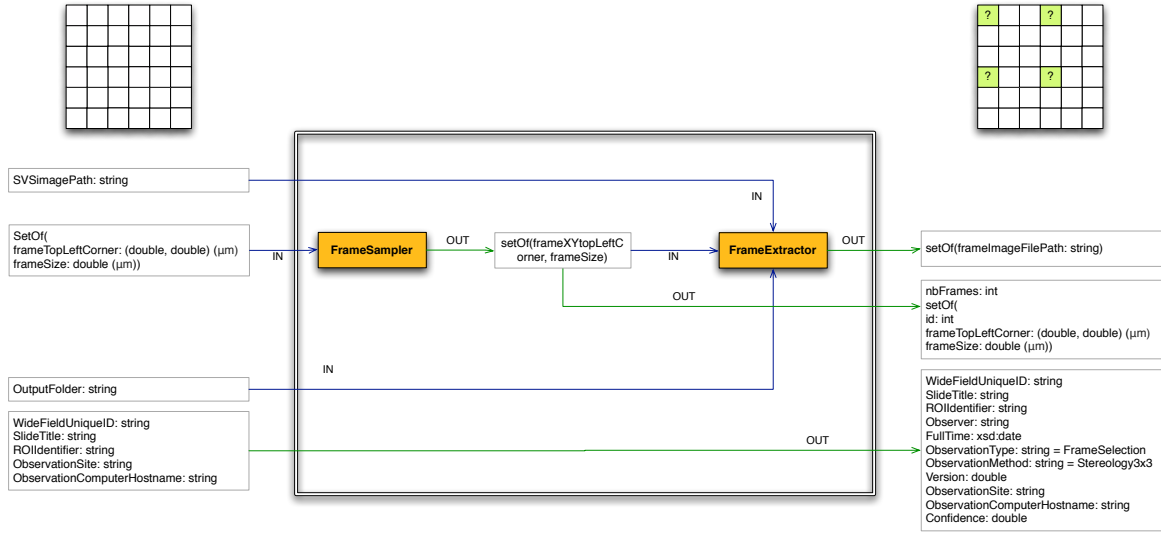
14

Figure 7: FrameSampler3x3 - module-actor of the setreology process
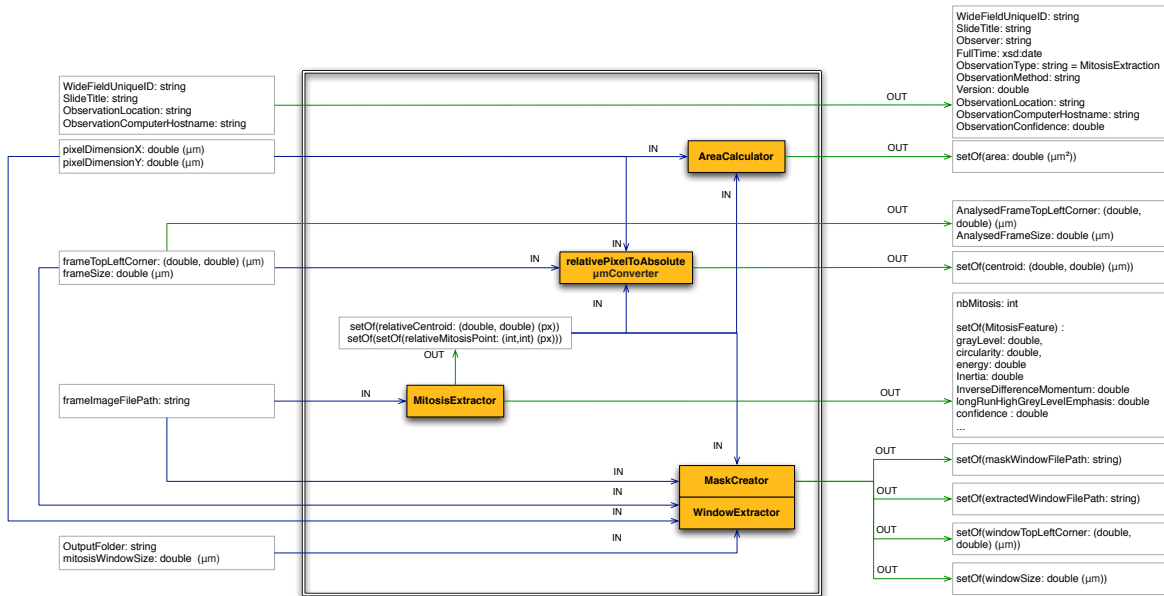

Figure 8: MitosisExtractor

4.4. MICO Platform: A semantic approach

MICO aims at supporting histopathologists during the cancer grading process, by taking benefit of the last medical and image processing technologies, while relying on histopathology experts' knowledge. Image processing algorithms allow fast image analysis and are expected to provide reliable and reproducible results. Hence, the use of such algorithms should help to make the grading process faster. MICO platform is also expected to reuse the best expert strategies in order to improve the grading process. NGS is to be implemented into the platform, as the non formally defined expert knowledge required to efficiently operate the same criteria analysis. Histopathological knowledge obviously plays a predominant role in our platform. Therefore, particular attention is given to the use of proper knowledge modelling and inferences technologies. Moreover, the issue of the impact of the knowledge on the platform behaviour is also addressed. How to ensure the knowledge has the position it deserve within the platform? This is the idea behind the proposed semantic approach: "Ontology at the helm". Indeed, MICO platform is built to be driven by the knowledge: each of MICO platform action should be induced by a decision, and each decision should be made within a formal knowledge based world. Such platform architecture offers several benefits like the improved traceability of decisions, the ability to be understood by the user, the ability to support him, a good flexibility, an improved technology acceptance and smarter use of image processing algorithms.

Traceability Medical tools must obey to strict regulations: because MICO platform is aimed to help histopathologists to take decisions by providing statements about medical cases, its decisions should obviously have some kind of traceability. Since semantic reasoning takes place in a formal world, each inference is proven: as a consequence, each decision is proven, and the proof trace may be viewed if necessary.

System understanding & Decision support MICO is definitely not aimed to replace the histopathologists, but to support them while taking important decisions, as to help them doing tedious and time consuming tasks. Thus, after the execution of low level tasks and inferences, the system should communicate relevant results to the expert, using suitable modality and formalization. These result should be human understandable and should be expressed with concepts from histopathology experts theory of histopathology, or histopathology domain ontology. Indeed, putting the user in the loop requires him to have a good understanding of what the system is doing. This may be archived by making sure expert and system use the same language, i.e. the expert activity language. MICO should support histopathologists by providing relevant information at the right time and using the right form, in order to minimize the effort needed to interpret the message and make higher level decisions.

Flexibility With a full semantic approach, all the facts and processes are expressed in an open manner. They are also fully described and therefore easily understood. A good understanding of the system knowledge is a first step toward its modification. Since the histopathologist user is not a computer scientist, semantic technologies only may allow him to modify the system knowledge. Compared to "hard coded" imaging systems, semantic systems are more flexible and easier to maintain.

Technology acceptance As mentioned previously, the use of semantic web technologies may help the user to understand what the system truly does, and therefore increase its perceived ease of use. In order for the system to take benefit of expert knowledge, it must be build in close collaboration with them. Communication with the user helps to understands his real needs, and to involve him in the project. By increasing the system perceived ease of use and its perceived usefulness, this approach will probably help the user to accept and use the technology.

Operational image processing Image processing algorithms might be effective, but they may also need quite a lot of processing power. Therefore, it may not be possible to compute a whole histopathological slide image (up to 25GB uncompressed) within reasonable time. A smarter approach is obviously needed, in order to grade the image without having to analyse every part of it. After all, the expert doesn't need to have a close look at every part of the slide in order to grade it. As a matter of fact, the histopathologists only use the 40X detailed view on very small parts of the slides. These analysis spots are selected using slide overview magnifications. Similarly, expert knowledge could be used to guide image processing algorithms, to target interesting spots in order to spare as much processing power as possible, and to make the overall grading faster.

4.4.1. Overview

MICO platform is composed of two main components: a semantic core, and a set of image processing algorithms. The core is made of an ontology interacting with a semantic inference engine: it may be considered as a glass box. On the contrary, image processing algorithms are binaries, and may be considered as black boxes.

4.4.2. Semantic core

MICO semantic core is the place where all the decisions are taken, and also the place where all the platform knowledge is held. It is made of two main components: a semantic reasoning engine (EULER) and an ontology. While EULER is in charge of the decisions, the ontology contains several layers of structured information about both image processing and histopathology domains. MICO semantic core is maybe the most important part of the platform, at least in regards to the semantic aspects of the project. It always works the same way, following a simple three steps life cycle, as illustrated in Figure 10.

First step During the knowledge acquisition step, external information is aggregated into the core ontology. It usually comes from MICO modules, and may be the result of image processing programs execution. If MICO semantic core is being launched for the first time, then a whole knowledge base is generated from seed files.

Second step Thanks to the inference engine, new knowledge may be discovered from existing facts, and queries are launched to select the knowledge relevant to further processing or user feedback.

Third step The last step of the MICO semantic core life cycle consists in the execution of the orders revealed in the previous step using queries, and the orders execution acknowledgement. Typically, an order is the parametrized launch of an image processing algorithm.
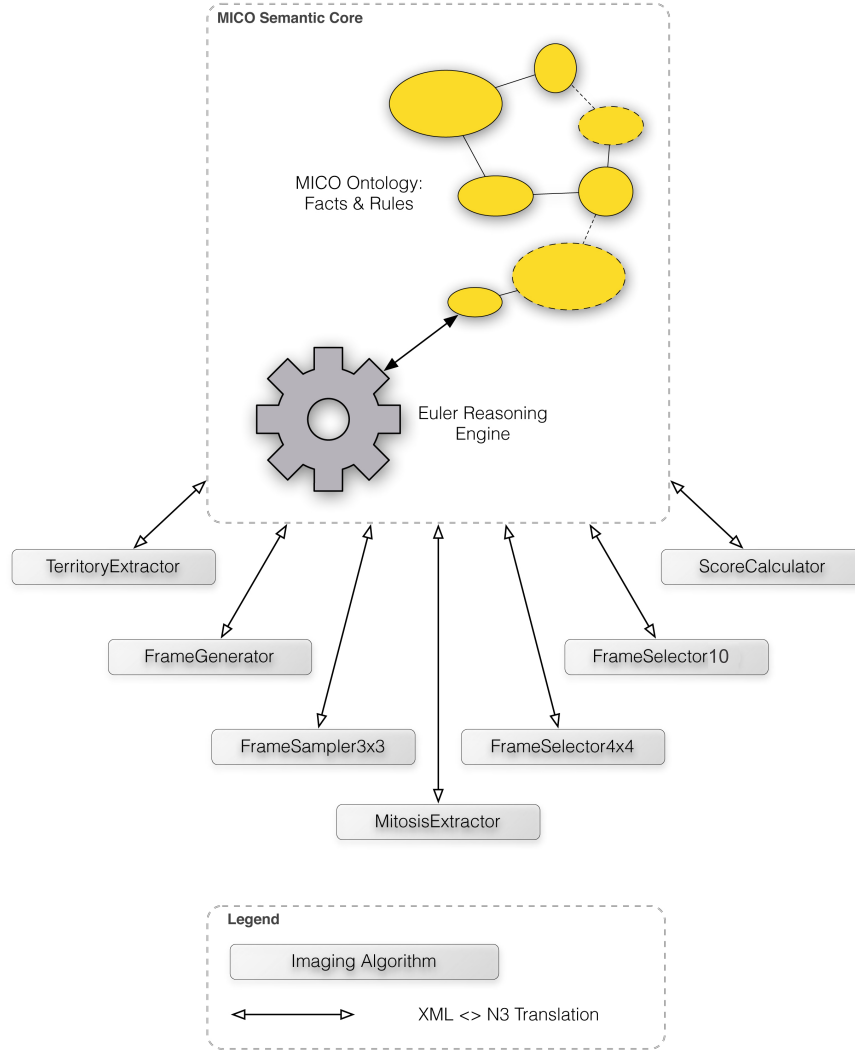
Figure 9: MICO semantic platform overview

A notable particularity of MICO semantic core is the way it manages knowledge. Indeed, some knowledge management systems or expert systems work with a frozen state knowledge base: once they are designed and developed, their knowledge base stays more or less the same. MICO semantic core approach differs in the sense its knowledge base is made to evolve with time: each time new knowledge is discovered thanks to inferences or programs outputs, it is incorporated into the knowledge base. As a result, the same inference never has to be made more than one time. Also, the concepts defining a knowledge state may be easier to identify, access, share, and link with external ontologies. Knowledge relevant to the user may be easier to monitor. Finally, it is a first step towards rules generation. However, there are drawbacks to being so exhaustive. Since the knowledge base grows with its use, it might become to large for the system to process it in reasonable time. In such a case, either the "useless" items should be removed, or an adapted database system should
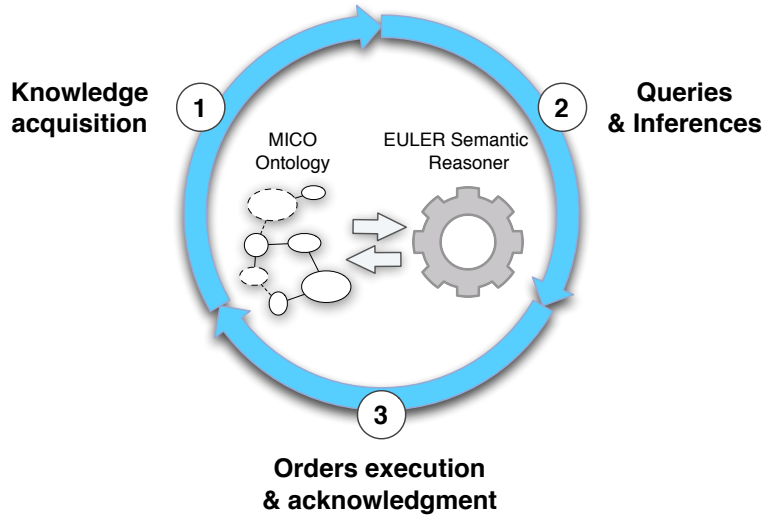
Figure 10: MICO semantic core life cycle

be used to store the ontology.

Ontologies. The ontology is the place where all the platform knowledge is stored. MICO ontology is actually made of several ontologies describing different topics. Each ontology is made of three kind of components: the structure, the individuals, and the rules. The individuals may be initialization knowledge, acquired or inferred knowledge: they compose the ontology A-Box. On the contrary, structure definition or rules belong to the ontology T-Box. MICO ontology is written using Notation3 and is paired with EULER semantic inferences engine. The current section describes the ontology structure, its logical rules and queries.

Overall structure Several kind of information are described into MICO ontology. For instance, a description for the initialization context is needed to start the processing of a new whole slide image. Similarly, a description of the information outputted by the image processing algorithms is needed to be able to interpret it when it comes up. Also, histopathologists vocabulary should be described to have the ability to provide results of the same language. Thus, the ontology should contains concepts from image processing domain to communicate with the image processing algorithms, and concepts from medical domain to exchange with histopathology experts during the grading process.

During the early steps of the project, it has been empirically found out it can be tough to merge or to unify concepts coming from different domains within the same ontology. Because such concepts may not perfectly overlap, and because people representation of their domain is tightly linked with the same domain traditional denominations, such compromise is complex (and maybe not reasonable ?) to achieve. On this matter, MICO approach is to consider every expert conceptualization as legitimate: experts using the platform should have the right to keep their habits wile defining and using their expertise. That is why it is
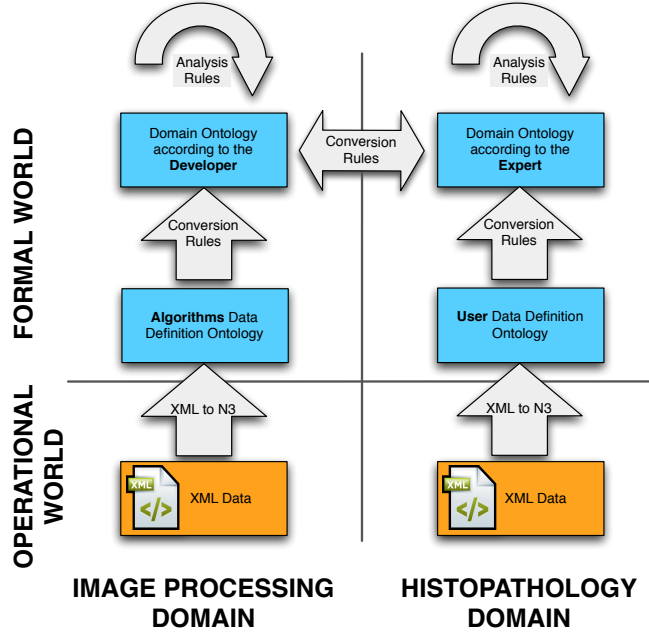
Figure 11: MICO ontologies

preferred to allow theories coming from different domains to coexist, even if they describe the same object. In such a case, translation rules should be written to give the system the ability to convert a knowledge from one ontology to another (Figure 11).

There is another reason to argue in favour of "splitting" the knowledge into several ontologies: it is about the difficulties induced by the evolution of image processing algorithms interfaces. Indeed, the input and output of these algorithms have been evolving and will probably continue to evolve. It would be interesting to minimize as much as possible the impact of this evolution on the platform, in particular on the analysis rules validity. To protect the reasoning rules from the algorithms format evolution, a distinction is made between the algorithms returning raw data and the higher knowledge level induced by it. This distinction implies conversion rules, to convert raw data to higher level knowledge of the same domain. Therefore, changes of an algorithm output format would only impact the description of its output and the conversion rules to convert the same output to higher level knowledge. The reasoning or conversion rules operating on higher level knowledge would not be impacted, because the used high level ontology structure would not be altered after modifications of algorithms output format. Thus, the platform would not be strongly affected by image algorithms input and output format evolutions.

To sum up about the two kind of separations between ontologies: as illustrated on the Figure 11, ontologies coming from different domains are all legitimate, and their existence is allowed in MICO platform. This is the case for the medical domain and the image processing domain. Within the same domain, the use of ontologies having different levels of abstraction is recommended, in order to make a proper distinction between changing data and the meaning of these data. This separation is indeed one of the good practices

of AGFA Healthcare R&D. According to AGFA: high level ontologies are called Domain Ontologies (DO) and low level ontologies are called Data Definition Ontologies (DDO).

The rules may have different purposes in the platform. Some rules aim to infer knowledge from measurements while others aim to convert knowledge from one theory to another, or to infer new knowledge from existing knowledge. According to the AGFA terminology, the rules allowing to translate knowledge from one ontology to another are called Conversion Rules, while the rules allowing to populate an ontology using knowledge from the same ontology are called Analysis Rules. The Listing 1 is a rule example from the developed platform prototype. This particular rule aims to detect when it is possible to launch a territory extraction analysis on an histopathological whole slide image WFML file. If the platform context fit the rule requirements, then an order is published in the ontology. Each order contains information about both the image processing algorithm to be launched (in this specific case, a territory extractor algorithm), and the parameters to be provided as algorithm input. These parameters are generated from ontology content.

Listing 1: TerritoryExtraction Launch Rule

```
{
?wfuid a ido:wideFieldUniqueID;
        ido:hasSlideTitle ?slideTitle;
        ido:hasWFMLAbsolutePath ?wfmlPath.
ng:mitosisCount ng:shouldBeAnalysedOn ?territory.
env:currentComputation a env:Computation;
        env:isComputedOn [env:isLocatedIn ?obsSite; env:hostname ?computerName].
?wfuid rdf:label ?wfuidLabel.
?territory rdf:label ?territoryLabel.
?slideTitle rdf:label ?slideTitleLabel.
?obsSite rdf:label ?obsSiteLabel.
} =>
{
algo:algorithm_territoryExtractor_v1-0 algo:mustBeLaunchedWithInput
        {
        xml:wideFieldUniqueID xml:hasValue ?wfuidLabel.
        xml:slideTitle xml:hasValue ?slideTitleLabel.
        xml:observationSite xml:hasValue ?obsSiteLabel.
        xml:computerHostname xml:hasValue ?computerName.
        xml:WFMLPath xml:hasValue ?wfmlPath.
        xml:territoryLabel xml:hasValue ?territoryLabel
        }.
}.
```

The queries take the same shape as the rules, even if their purpose is quite different. They are used to extract relevant information from the knowledge base. The extracted relevant individuals may be results to be displayed to the user or orders aiming to launch image processing algorithms. In the developed platform, the queries are typically used to isolate algorithm launch orders. The Listing 2 is a query example from the developed platform prototype. This particular query aims to detect when it is opportune to launch a territory extraction analysis on an histopathological whole slide image WFML file. In this

example, it is chosen to actually launch the algorithm when the known information allows its launch and it has not already been launched with the same context.

Listing 2: TerritoryExtraction Launch Query

```
{
algo:TerritoryExtractor algo:observationType ?observationType.
?algorithm a algo:TerritoryExtractor;
        algo:hasAbsoluteFilePath ?absolutePath;
        algo:mustBeLaunchedWithInput ?input.
?input log:includes {xml:wideFieldUniqueID xml:hasValue ?wfid}.
[]e:findall(?anotherInput{?algorithm algo:wasLaunchedWithInput...
 ...[log:includes {xml:wideFieldUniqueID xml:hasValue ?wfid}]}()).
} =>
{
?algorithm algo:mustBeLaunchedWithInput ?input;
        algo:hasAbsoluteFilePath ?absolutePath;
        algo:observationType ?observationType.
}.
```

4.4.3. Modules

MICO platform is made of both a semantic core and image processing algorithms. Since the image processing algorithms are not aware of any conceptualization domain, they are not able to directly communicate with the semantic core. And because they are generic tools and could be applied on different contexts, they should not be tight to a specific domain. Therefore, they should not interfere with non generic high level concepts. That is why a conversion layer is needed to allow communication between the generic image processing algorithms and the formal world describing the application domain, or MICO ontology. The combination of an image processing algorithm and its conversion layers is called a MICO module (Figure 12). Each MICO module conversion layer is actually made of two layers. The first layer (next to the image processing algorithm) is an XML interface. Such XML layer allows a first description of the algorithms input and outputs. The second layer is made of both a XML ↔ Notation3 converter and a semantic profile. For instance, when an algorithm produces data, it is first expressed using XML. Then, an XSL semantic profile is used by an XSLT engine to convert the XML to Notation3. The N3 is finally incorporated into MICO knowledge base. Such conversion from raw data to low level knowledge can be achieved because the algorithms outputs are highly described, highly contextualized. Because a data without context has no meaning, each algorithm measurement (output) contains various meta information like the measurement location, date, time, the kind of measurement, the method used, etc. Indeed, these meta information are crucial to make a proper N3 modelling. Thanks to the modules conversion layers, communication is possible between black boxes (the image processing algorithms) and white boxes (the semantic core).

5. Conclusion

The researches presented in this paper represent a considerable step towards a clinically viable digital pathology. The objective of our study is to make histopathological WSI
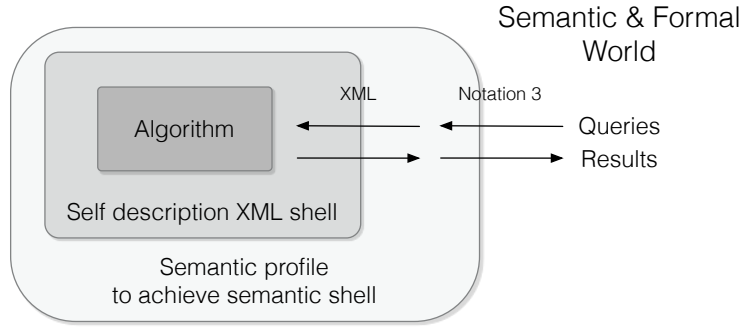
Figure 12: A MICO Module

analysis reliable and traceable, for a better second opinion, based on quantitative support.

A working prototype has been proposed for the use case of a WSI, focusing on cancer grading, and in particular, to the mitotic count. The implemented analysis is compliant with the Nottingham Grading System, supported by the World Health Organisation.

The prototype realised is composed of two main components: a semantic core, and a set of image processing algorithms. Generic image processing algorithms have been designed for the our MICO platform. They have been developed by different members of our team. A semantic description of each algorithm has been written, and they have been incorporated into the platform. MICO semantic core has also been designed and developed. Two main components have been implemented inside the core: a semantic reasoning engine and an ontology. While the semantic reasoning engine is in charge of the platform decisions, the ontology contains several layers of knowledge about both image processing and histopathology domains.

Being a very ambitious challenge, leading to a potential breakthrough in imaging, this work has also the meaning of its genericity. Indeed, the approach is applicable for number of additional challenges, related, for example, to cytology, biology, molecular imaging and, in general, to a large set of high-content high-throughput automatic screening approaches.

Acknowledgment

References

[1] "National cancer institute." http://www.cancer.gov/.

[2] A. Gonzalez-Angulo, J. Litton, K. Broglio, F. Meric-Bernstam, R. Rakkhit, F. Cardoso, F. Peintinger, E. Hanrahan, A. Sahin, M. Guray, *et al.*, "High risk of recurrence for patients with breast cancer who have human epidermal growth factor receptor 2–positive, node-negative tumors 1 cm or smaller," *Journal of Clinical Oncology*, vol. 27, no. 34, pp. 5700–5706, 2009.

[3] G. Le Naour, C. Genestie, L. Roux, A. Veillard, D. Racoceanu, F. Capron, *et al.*, "Un explorateur visuel cognitif (MIcroscope COgnitif-MICO) pour l'histopathologie. Application au diagnostic et à la graduation du cancer du sein.," 2011.

[4] P. Bamford and B. Lovell, "Unsupervised cell nucleus segmentation with active contours," *Signal Processing*, vol. 71, no. 2, pp. 203–213, 1998.

[5] D. Vernon, "Cognitive vision: The case for embodied perception," *Image and Vision Computing*, vol. 26, no. 1, pp. 127–140, 2008.

[6] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *Biomedical Engineering, IEEE Reviews in*, vol. 2, pp. 147–171, 2009.

[7] C. Genestie, "Mammary pathology." Hopital de la Pitie-Salpetriere, Paris, June 2011.

[8] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, 1991.

[9] J. J. Going, "Breast cancer grading - Nottingham criteria." 2005.

[10] G. Antoniou and F. Van Harmelen, *A semantic web primer*. the MIT Press, 2004.

[11] T. Berners-Lee and J. Hendler, "Scientific publishing on the semantic web," *Nature*, vol. 410, pp. 1023–1024, 2001.

[12] "Dictionary and thesaurus - merriam-webster online." `http://www.merriam-webster.com/`.

[13] R. Studer, S. Decker, D. Fensel, and S. Staab, "Situation and perspective of knowledge engineering," *Knowledge Engineering and Agent Technologies*, 2000.

[14] A. Tutac (ép. Branici), *Formal Representation and Reasoning for Microscopic Medical Image-Based Prognosis. Application to Breast Cancer Grading*. PhD thesis, International co-tutelle - Politehnica University of Timisoara, Romania and Université de Franche Comté, Besançon, France, 2010.

[15] R. Studer, V. Benjamins, and D. Fensel, "Knowledge engineering: principles and methods," *Data & knowledge engineering*, vol. 25, no. 1, pp. 161–197, 1998.

[16] R. Davis, H. Shrobe, and P. Szolovits, "What is a knowledge representation?," *AI magazine*, vol. 14, no. 1, p. 17, 1993.

[17] M. Bergman, "The fundamental importance of keeping an ABox and TBox split." `http://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2/`, May 2009.

[18] "Resource description framework (RDF): Concepts and abstract syntax." `http://www.w3.org/TR/rdf-concepts/`.

[19] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 1–22, 2009.

[20] R. Verborgh, "Semantic web reasoning with eye." `http://n3.restdesc.org/n3/`, 2011.

[21] "W3C RDF vocabulary description language 1.0: RDF schema." `http://www.w3.org/TR/rdf-schema/`.

[22] "W3C OWL web ontology language reference." `http://www.w3.org/TR/owl-ref/`.

[23] T. Berners-Lee and D. Connolly, "Notation3 (N3): A readable RDF syntax," *W3C Team Submission : http://www.w3.org/TeamSubmission/n3/*, no. 3, 1998.

[24] T. Berners-Lee, D. Connolly, L. Kagal, Y. Scharf, and J. Hendler, "N3logic: A logical framework for the world wide web," *Theory and Practice of Logic Programming*, vol. 8, no. 3, pp. 249–269, 2008.

[25] T. Bray, J. Paoli, C. Sperberg-McQueen, E. Maler, and F. Yergeau, "Extensible markup language (XML)," *World Wide Web Journal*, vol. 2, no. 4, pp. 27–66, 1997.

[26] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review ? current status and future potential," *IEEE Reviews on Biomedical Engineering*, no. 7, pp. 97–114, 2014.

[27] H. Irshad, *Automated Mitosis Detection in Color and Multi-spectral High-Content Images in Histopathology: Application to Breast Cancer Grading in Digital Pathology.* PhD thesis, Universités Grenoble-Alpes, Université Joseph Fourier, Grenoble 1, France, 2014.