Bridging the semantic gap between diagnostic histopathology and image analysis

Lamine TRAORE [a,b,1] , Yannick KERGOSIEN[a,c] and Daniel RACOCEANU[b,d]

[a] *Sorbonne Universités, UPMC Univ Paris 06, INSERM, Université Paris 13, Sorbonne Paris Cité, Laboratoire d'Informatique Médicale et Ingénierie des Connaissances en eSanté (LIMICS - UMR_S 1142), 15 rue de l'école de médecine, Paris, France;* [b] *Sorbonne Universités, UPMC Univ Paris 06, CNRS, INSERM, Laboratoire d'Imagerie Biomédicale (LIB), 75013, Paris, France;* [c] *Département d'Informatique Université de Cergy-Pontoise, Cergy-Pontoise, France;* [d] *Pontifical Catholic University of Peru, San Miguel, Lima 32, Peru*

Abstract. With the wider acceptance of Whole Slide Images (WSI) in histopathology domain, automatic image analysis algorithms represent a very promising solution to support pathologist's laborious tasks during the diagnosis process, to create a quantification-based second opinion and to enhance inter-observer agreement. In this context, reference vocabularies and formalization of the associated knowledge are especially needed to annotate histopathology images with labels complying with semantic standards. In this work, we elaborate a sustainable triptych able to bridge the gap between pathologists and image analysis engineers/scientists. The proposed paradigm is structured along three components: i) extracting a relevant semantic repository from the College of American Pathologists (CAP) organ-specific Cancer Checklists and associated Protocols (CC&P); ii) identifying - through the NCBO Bioportal – imaging formalized knowledge issued from effective histopathology imaging methods highlighted by recent Digital Pathology (DP) contests and iii) proposing a formal representation of the imaging concepts and functionalities issued from major biomedical imaging software (MATLAB, ITK, ImageJ). Since the first step i) has been the object of a recent publication of our team, this study focuses on the steps ii) and iii). Our hypothesis is that the management of available semantic resources concerning the histopathology imaging methods - issued from CAP documents - associated with effective methods highlighted by the recent DP challenges will facilitate the integration of WSI in clinical routine and support new generation of DP protocols.

Keywords. Histopathology image analysis, semantic annotation, formal representation.

1. Introduction

In this study, we continue our semantic cognitive virtual microscopy initiative[2,3] by proposing a sustainable way to bridge the content, features, performance and usability gaps [1] [2] between histopathology and WSI analysis. The MICO[2] project achieved a prototype system to perform some histopathology diagnosis related tasks on tissue slides where elementary imaging processes were combined by a logic engine [3], which could use formalized knowledge available as a set of rules. These rules, however, had been elaborated through local collaboration between pathologists and image scientists whereas sustainability calls for the use of publicly available knowledge gathered in standard formats from collaborative multi-centric efforts and periodic updates. A preliminarily work in this direction has been recently published by our team [3] proposing the use of the College of American Pathologists (CAP) organ-specific Cancer Checklists and associated Protocols (CC&P). Based on NCBO Bioportal and UMLS semantic types, the semantics generated represents a sustainable vocabulary, dedicated to histopathology, being able to effectively support daily work on whole slide images, in digital pathology. Semantic models and reference terminologies are essential in digital pathology, being able to support the reproducibility and quality of the diagnostic, to assist and standardize anatomopathological reporting, and to enable multi-center clinical collaboration or research, especially in the context of cancer grading [4]. Reference vocabularies and ontologies are especially needed for the annotation of histopathology images with labels complying with semantic standards. The availability of digital tools in pathology, especially WSI and the possibility to perform on them some image analysis tasks, call for an extension of semantic modeling to the realm of image processing and its integration with clinical semantics.

1 Corresponding authors; E-mails: laminet@gmail.com , daniel.racoceanu@upmc.fr

2 MICO project (COgnitive MIcroscopy) - French National Research Agency - Technologies for Health and Autonomy (ANR TecSan): http://daniraco.free.fr/projects.htm

3 FlexMIm project (Collaborative Pathology) - Consolidated Interministerial Fund (FUI - Fonds Unique Interministériel): http://www.systematic-paris-region.org/en/projets/flexmim

2. Challenge and objective

In this study, we complete the elaboration of a sustainable triptych, able to bridge the gap between pathologists and image analysis commmunity. The proposed path is structured along three major components: i) extracting a relevant semantic repository from the CAP's organ-specific CC&P; ii) identifying formalized imaging knowledge, issued from the effective histopathology imaging methods highlighted by recent Digital Pathology (DP) contests and iii) proposing a formal representation of the imaging concepts and functionalities extracted from major biomedical imaging software as MATLAB, ITK and ImageJ. Our present study focuses on the steps ii) and iii). Our hypothesis is that the management of available semantic resources concerning the histopathology imaging methods associated with effective methods highlighted by the recent challenges in DP will facilitate the integration of WSI in clinical routine and effectively support a new generation of DP protocols.

3. Materials and methods

The overall approach is presented in Figure 1. In this paper we treat the image analysis domain. Series of international benchmarking initiatives [5] have been launched for mitosis detection at MITOS 2012 (continued by AMIDA 2013, MITOS 2014 and TUPAC 2016), nuclear atypia grading at ATYPIA 2014 and glandular structures detection GlaS 2015. These initiatives allow envisaging a consolidated validation referential-database for DP.

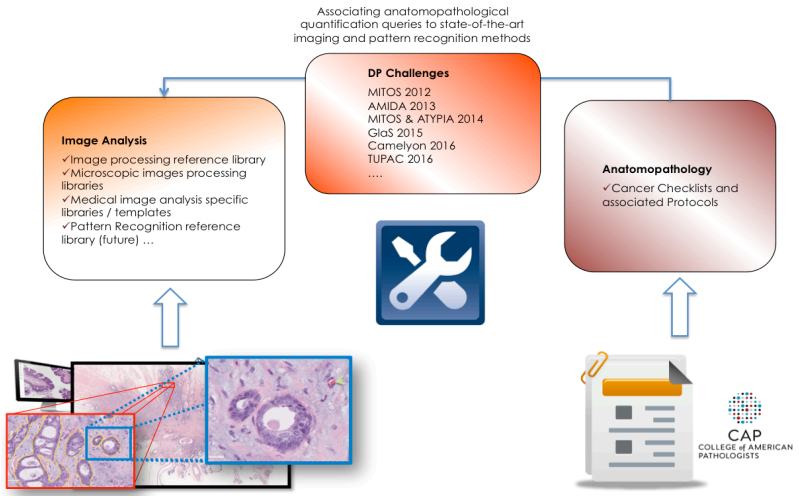

Figure 1. The overall proposed approach: use of the recent DP challenges to make an operational, instantiated link between anatomopathology and imaging.

3.1. Automatic annotation of corpus issued from contests with available semantic resources in the NCBO Bioportal.

We considered the 2012-2016 period and identified 5 international benchmarking contests related to 29 top performing histopathology-imaging methods. In accordance with our recent published work [3], 3 of the challenges are related to the breast cancer diagnosis and prognosis criteria. Corpuses were extracted from authors descriptions in articles [6], [7], and "Grand Challenge" platform [5]. Table 1 summarizes description of the corpus with associated contests and methods.

Table 1. Description of the corpus with associated contests, identified methods and word count.

Corpus Index	Associated Conference	Identified Challenges	Number of Methods	Word counts
C#1	ICPR 2012	MITOSIS (Mitosis detection in breast cancer histological images)	4	181
C#2	MICCAI 2013	AMIDA (Assessment of algorithms for mitosis detection in breast cancer histopathology images)	11	405
C#3	ICPR 2014	MITOS-ATYPIA (Detection of mitosis and high-grade atypia nuclei in breast cancer histology images)	4	627
C#4	MICCAI 2015	GlaS (Gland Segmentation in Colon Histology Images)	6	501
C#5	ISBI 2016	Camelyon 16 (cancer metastasis detection in lymph node)	4	896

For each corpus, by using Recommender [6] of NCBO Bioportal we obtained the ranking of the most pertinent ontologies individually or by sets of 4. The ontology-ranking algorithm used by Recommender evaluates the adequacy of each ontology to the input corpus using a combination of four evaluation criteria: Coverage, Acceptance, Detail of knowledge and Specialization. For each case, we adjusted these parameters by considering default weights (Coverage=0.55, Acceptance=0.15, Knowledge Detail=0.15, Specialization=0.15) and a focus on the coverage criterion (Coverage=1, others put to zero). We first annotated each corpus with the "imaging category" ontologies (n = 15) specified in NCBO Browse Tab. Then we redid the annotation by referring to "All ontologies" available (n = 668). In each case, the first 5 single ranked ontologies and the highest ranked ontology set (4 per set) were identified. Table 2, Table 3 and Table 4 report the results.

3.2. Visual representation of the imaging knowledge issued from MATLAB, ITK and ImageJ

Our visualization targeted the concepts issued from the three image analysis communities related to the use of MATLAB (image scientists and engineers), ITK (developers) and ImageJ (imaging biologists). We used corpuses extracted from the user manuals. Conserving the hierarchy levels from sources, we organized all identified concepts. Then, with Protégé® and its OWLviz plugin [8] we generated a visualization of concepts related to each source.

4. Results

4.1. Automatic annotation of corpus issued from contests with NCBO Bioportal resources

4.1.1. Automatic annotation with the 15 NCBO "imaging category" ontologies

The list of most pertinent "imaging category" ontologies found in Bioportal is reported in Table 2. Overall 10 ontologies were found ranked with respect to their popularity (number of visits). From NCBO "imaging category" ontologies, the maximum final annotation scores obtained with the coverage criterion (Coverage=1, others put to zero) were with Corpus#1: 9.0% for single ranked ontology (EDAM-BIOIMAGING) and 21.8% for ontology sets (EDAM-BIOIMAGING, NIDM-RESULTS, NEMO and IDQA). With the default configuration, single ranked ontology scores range from 11.4 (BIRNLEX) to 21.7% (NEMO).

Table 2. List of the most pertinent "imaging category" ontologies found in Bioportal with associated definitions and metrics

INDEX	NAME	CATEGORY	CLASSES
1	Radiation Oncology Ontology (ROO)	Development, Health, Human, Imaging, Vocabularies	1183
2	DICOM Controlled Terminology (DCM)	Imaging	3476
3	Information Artifact Ontology (IAO)	Biomedical Resources, Imaging, Other	180
4	Biomedical Informatics Research Network Project Lexicon (BIRNLEX)	Anatomy, Imaging	3580
5	Neural ElectroMagnetic Ontology (NEMO)	Anatomy, Biological Process, Experimental Conditions, Human, Imaging	1851
6	Biomedical Image Ontology (BIM)	Imaging	125
7	Cognitive Paradigm Ontology (COGPO)	Experimental Conditions, Human, Imaging	358
8	NIDM-Results (NIDM-RESULTS)	Imaging, Other	161
9	Image and Data Quality Assessment Ontology (IDQA)	Imaging	260
10	Bioimaging Ontology (EDAM-BIOIMAGING)	Imaging	130

3.1.2 Automatic annotation with all 668 ontologies available on the NCBO platform

From the results of the annotation with all ontologies available in NCBO Bioportal, we get the list of the ten (10) most relevant ontologies (with respect to their final scores) to be used for the annotation of the corpus describing imaging methods in histopathology domain. Table 3 reports the list with related definitions and metrics.
By considering the same example mentioned previously with Corpus#1 and the coverage criterion (Coverage=1, others put to zero) final results are 57.7% for single ranked ontology (NCIT) and 75.2% for ontology sets (NCIT, SNOMEDCT, SWEET and LOINC).

Table 3. List of the most relevant biomedical ontologies in NCBO Bioportal for the annotation of corpus describing imaging methods in histopathology domain

#	NAME	CATEGORY	CLASSES
1	Logical Observation Identifier Names and Codes (LOINC)	Health	187123
2	Material Rock Igneous (MATRROCKIGNEOUS)	Upper Level Ontology	3535
3	Medical Subject Headings (MESH)	Health	261990
4	Material Natural Resource (MNR)	Upper Level Ontology	3554
5	National Cancer Institute Thesaurus (NCIT)	Vocabularies	118941
6	Neuroscience Information Framework (NIF) Standard Ontology (NIFSTD)	All Organisms, Anatomy, Biological Process, Cell, Cellular anatomy , Dysfunction, Molecule, Neurologic Disease, Neurological Disorder, Other, Subcellular, Subcellular anatomy	124337
7	Orthology Ontology (ORTH)	All Organism, Genomic and Proteomic	4663
8	Read Codes, Clinical Terms Version 3 (CTV3) (RCD)	Not mentioned	140065
9	Systematized Nomenclature of Medicine - Clinical Terms (SNOMEDCT)	Health	324129
10	Semantic Web for Earth and Environment Technology Ontology (SWEET)	Not mentioned	4550

Table 4. Final score of the most relevant ontology set annotation by referring to “Imaging category” ontologies and “All ontologies" in NCBO Bioportal.

Corpus Index	Imaging (n=15), Coverage Configuration	All ontologies (n=668), Coverage Configuration	Imaging (n=15), Default Configuration	All ontologies (n=668), Default Configuration
C#1	EDAM-BIOIMAGING, NIDM-RESULTS, NEMO, IDQA => 21.8	NCIT, SNOMEDCT, SWEET, LOINC => 75.2	NEMO, DCM, EDAM-BIOIMAGING, NIDM-RESULTS => 21.7	NCIT, SNOMEDCT, MESH, SWEET => 74.2
C#2	EDAM-BIOIMAGING, NDM-RESULTS, NEMO => 15.7	NCIT, SNOMEDCT, MESH => 69.8 (NA Set 4)	NEMO, IAO, BIRNLEX, DCM => 16.7	NCIT, SNOMEDCT, MESH => 74.4
C#3	EDAM- BIOIMAGING, NEMO, DCM, IDQA => 21.3	NCIT, SNOMEDCT, RCD, ORTH => 80.9	NEMO, COGPO, DCM, NIDM-RESULTS => 22.4	NCIT, SNOMEDCT, RCD, ORTH => 78.7 (NA Set 4)
C#4	BIRNLEX, DCM, EDAM-BIOIMAGING, NEMO => 17.7	NCIT, SNOMEDCT, SWEET => 75.1 (NA Set 4)	NEMO, DCM, EDAM-BIOIMAGING, BIRNLEX => 17.9	NCIT, SNOMEDCT, RCD => 63.1 (NA Set 4)
C#5	BIOIMAGING, ROO, NEMO, BIRNLEX => 24.3	SNOMEDCT, LOINC, NIFSTD => 75.1	NEMO, ROO, DCM, EDAM-BIOIMAGING => 23.2	NCIT, SNOMEDCT, NIFSTD, SWEET => 77.4

3.2 Visual representation of concepts from MATLAB, ImageJ and ITK

Three (3) graphical tree representations reflecting the hierarchy and granularity of each source were obtained with respectively 565 concepts from MATLAB, 348 from ITK and 259 from ImageJ.

4 Discussion and conclusion

The proposed approach based on the annotation of benchmarking contests corpus with NCBO Bioportal aims to evaluate available semantic resources associated to the histopathology imaging domain. From the above results, we report that there is no ontology related to the imaging domain in NCBO Bioportal to annotate efficiently the identified histopathology imaging methods. With respect to the ontology lists in Table 2, Table 3 and annotation results in Table 4, we see that the most relevant ontologies annotating imaging concepts in Bioportal are SNOMEDCT, NCIT and other ontologies related to health, anatomy, biological process and similar categories. One should note that these huge resources are not specialized to the imaging domain even if they give the highest annotation scores. This also shows the need of an imaging domain ontology that will be built upon available image analysis concepts and functionalities.

Beyond NCBO Bioportal, we searched other ontology repositories such as OBO Foundry [9]. Out of the 181 ontologies in Ontobee, we could manually identify 17 ontologies related to the imaging domain. The selection criteria were based on the “Ontology Full name” and given definitions. Since, there is no annotating tool associated to Ontobee, we could not annotate our corpus with these semantic resources. In future work, we plan

to use these semantic resources "locally" with BioYodie[4] to annotate and evaluate the relevance of their concepts with respect to imaging methods from contests.
On another hand, we faced difficulties in getting "bigger" corpus. We could find few published papers in open access, describing contests' newly proposed methods. To complete this list, we sent requests to authors to obtain more descriptions. Publishing a description of competing methods is a requirement in most contests. However, in some cases patent restrictions limit the depth of the description related to a method. For example, one of the 7 highest-ranking methods in GlaS contest was not available.
By using Protégé® and OWLviz, we obtained the visual representation of concepts issued from Matlab, ImageJ and ITK image analysis communities. This helped us to better understand the hierarchy and granularity of the information contained in each source. At this stage, we considered the hierarchical organization of concepts and their respective definitions from different sources. Due to the limited number of pages, we could not include all annotation results and visual representation of issued concepts. Details related to all these materials are available upon request. This work is a step forward to answering the need to build a visualization and formal representation that integrates image analysis tasks with concepts related to the domain. It opens the perspectives of the Practical Image Processing Task Ontology (PIPTO) construction. PIPTO aims at capturing image domain knowledge in a generic way and providing a consensual understanding of concepts and functionalities identified in the standard tools from these communities.
To overcome the limits previously mentioned in the annotation process, we plan to consider concepts associated to the DICOM Controlled Terminology Ontology (CTO) and similar resources in the perspective of PIPTO construction. Since DICOM is the main standard in medical imaging, it would be interesting to consider existing descriptions in DICOM sources to enrich the definition of concepts in PIPTO. Additional efforts are needed to achieve a workable standard-based formal representation that will be clearly understandable by humans, machine processable, and sustainable.
Overall, we could identify and evaluate relevant ontologies associated to histopathology image analysis. Then by considering concepts from main biomedical imaging tools, we could propose a formal representation of the imaging knowledge from MATLAB, ImageJ, and ITK. Each of these software applications or libraries includes a set of concepts, definitions, functions and relations that are expected to cover most of the imaging methods.
Future anatomopathological services need to use digital technologies in valid routine pathological diagnosis and healthcare protocols, by integrating the WSI observation for diagnosis purposes in a whole large specific DP case record. This will generate an operational DP process in which the innovation relies in linking the microscopic exam of WSI to specific or generic annotations defined as micro-semiology semantic references. Such approach enables the generation of a structured and standardized image-related report. Through DP, the future of anatomopathology is on the way to reinforce its ethical and dynamical strengths.

References

[1] T. M. Deserno, S. Antani, and R. Long, "Ontology of gaps in content-based image retrieval," *J. Digit. Imaging*, vol. 22, no. 2, pp. 202–215, Apr. 2009.
[2] A. E. Tutac, *[Formal representation and reasoning for microscopic medical image-based prognosis] : [application to breast cancer grading]*. Besançon, 2010.
[3] L. Traore, C. Daniel, M.-C. Jaulent, T. Schrader, D. Racoceanu, and Y. Kergosien, "A sustainable visual representation of available histopathological digital knowledge for breast cancer grading," *Diagn. Pathol.*, vol. 2, no. 1, Jun. 2016.
[4] C. Daniel *et al.*, "Standards and specifications in pathology: image management, report management and terminology," *Stud Health Technol Inf.*, vol. 179, pp. 105–122, 2012.
[5] "grand-challenges - Home." [Online]. Available: https://grand-challenge.org/. [Accessed: 25-Oct-2016].
[6] "Mitosis Detection in Breast Cancer Histological Images | IPAL UMI CNRS - TRIBVN - Pitié-Salpêtrière Hospital - The Ohio State University." [Online]. Available: http://ipal.cnrs.fr/ICPR2012/. [Accessed: 17-Feb-2015].
[7] "Assessment of algorithms for mitosis detection in breast cancer histopathology images - 1-s2.0-S1361841514001807-main.pdf." [Online]. Available: http://ac.els-cdn.com. [Accessed: 22-Jul-2016].
[8] "Ontology Recommender Web service - NCBO Wiki." [Online]. Available: http://www.bioontology.org [Accessed: 10-Dec-2015].
[9] "protegeproject/owlviz," *GitHub*. [Online]. Available: https://github.com/protegeproject/owlviz. [Accessed: 07-Nov-2016].
[10] B. Smith *et al.*, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat. Biotechnol.*, vol. 25, no. 11, p. 1251, Nov. 2007.

4 Beyond semantic annotation, Bio-Yodie manages hierarchical disambiguation between the concepts and refers to UMLS to build updated reference resources. It is based on the GATE platform (General Architecture for Text Engineering) and offers a wide range of output format for annotated concepts.