the diagnostic pathology journal DIAGNOSTIC PATHOLOGY

L. Traore et al., diagnostic pathology 2016, 2:109
ISSN 2364-4893
DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109

Research

A sustainable visual representation of available histopathological digital knowledge for breast cancer grading

L.Traore[1, 2], C. Daniel[1, 3], M-C. Jaulent[1], T.Schrader[4], D. Racoceanu[2], Y. Kergosien[1, 5]

Affiliation:
1 – 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;
2 – Sorbonne Universités, UPMC Univ Paris 06, CNRS, INSERM, Laboratoire d'Imagerie Biomédicale (LIB), 75013, Paris, France;
3 – Assistance Publique-Hôpitaux de Paris (AP-HP), CCS SI Patient, Paris, France;
4 – University of Applied Sciences Brandenburg Magdeburger, Department Informatics and Media, Brandenburg, Germany;
5 – Département d'Informatique Université de Cergy-Pontoise, Cergy-Pontoise, France.

Abstract

Background

Recently, anatomic pathology (AP) has seen the introduction of such tools as slide scanners and virtual slide technologies, creating the conditions for broader adoption of computer aided diagnosis based on whole slide images (WSI). This change brings up a number of scientific challenges such as the sustainable management of the semantic resources associated to the diagnostic interpretation of AP images by both humans (pathologists) and computers (image analysis algorithms). In order to reduce inter-observer variability between AP reports of malignant tumours, the College of American Pathologists (CAP) edited more than 60 organ-specific Cancer Checklists and associated Protocols (CC&P). Each checklist includes a set of AP observations that are expected to be reported by pathologists in organ-specific AP cancer reports.

Our objectives were i) to identify the available histopathological formalized knowledge from the NCBO Bioportal in the scope of the CAP-CC&P for breast cancer grading and ii) to build a sustainable visual representation of this knowledge using UMLS semantic types.

Methods

Our methodology was applied to the two breast cancer CAP-CC&Ps dedicated to invasive carcinoma (IC) and ductal carcinoma in situ (DCIS). We focused on a subset of quantifiable AP observations of the CAP-CC&Ps, i.e. observable entities that could be computed by image analysis tools, and on the corresponding notes in the protocols that unambiguously describe how pathologists should derive a high-level observation (e.g. Nottingham score) from low-level morphological characteristics observed in images (e.g., mitotic count or glandular/tubular differentiation). The notes were annotated both manually by two AP experts (gold standard) and automatically by NCBO's Annotator using the 508 ontologies available on the NCBO platform. A subset of reference ontologies was algorithmically selected based on their capacities to identify concepts in the notes and compared to the subset of ontologies selected based on their capacities to identify the concepts identified by experts (gold standard). Once automatically extracted from the notes, the concepts belonging to different ontologies were integrated into a unique graph and organized according to UMLS semantic types.

Results

The most relevant biomedical ontologies to be used for the annotation of the notes describing quantifiable observable entities of breast cancer CAP-CC&Ps are SNOMED-CT, LOINC, NCIT, NCI CaDSR Value Sets and PathLex. A visual representation integrating 23 concepts from the 5 different ontologies organized according to 11 UMLS semantic types was built to support AP experts for building a formal representation of the low-level quantifiable entities automatically extracted from the CAP-CC&Ps notes. The whole process of producing the visual report about available semantic resources from selected texts was shown to be fully automatizable.

Conclusion

The proposed approach and tool, based on the CAP-CC&Ps, aim at supporting AP experts in building a standard-based representation of low-level morphological abnormalities observed in cancer that can be quantified using image analysis tools. This effort is complementary to the Integrating the Healthcare Enterprise (IHE) initiative building a standard-based representation of high-level AP observations required in cancer AP reports. Additional efforts are needed to achieve a workable standard-based formal representation of histopathological knowledge integrating both observable entities reported by humans (pathologists) and quantifiable entities automatically computed by machines. Providing such unique formal representation paves the way for more efficient use of computer aided diagnosis in AP as well as for the development of new biomarkers based on automatic analysis of whole slide images (WSI).

Keywords: Breast cancer grading, Semantic annotation, Knowledge formalization and modeling, Standardization, Computer aided diagnosis, High-content image exploration, Digital pathology.

Background

Recently, the introduction of several tools such as slide scanners and virtual slide technologies created the conditions for a broader adoption of computer aided diagnosis based on whole slide images (WSI) with the hope of a possible contribution to decreasing inter-observer variability in Anatomic Pathology (AP). These changes bring up a number of scientific challenges such as the sustainable management of the available semantic resources associated to the diagnostic interpretation of AP images by both humans (pathologists) and computers (image analysis algorithms).

In order to reduce inter-observer variability between AP reports of malignant tumours, the College of American Pathologists edited more than 60 organ-specific Cancer Checklists and associated Protocols (CAP-CC&P) [1]. Each checklist includes a set of AP observations that are relevant in the context of a given organ-specific cancer and have to be reported by the pathologist. The associated protocol includes interpretation guidelines for most of the required observations.

Based on the CAP-CC&Ps, a joint IHE and Health Level 7 (HL7) AP initiative defined a formal information model for Anatomic Pathology Structured Report (APSR) based on HL7 Clinical Document Architecture (CDA) that was published in March 2011[1]. The objective of the IHE/HL7 APSR template was to make it interoperable so that different healthcare facilities could collect, exchange and mine AP information at an international level. The current scope of the IHE APSR content profile addresses all fields of AP (inflammatory diseases as well as cancer). In the cancer domain, the IHE APSR value set appendix provides a list of organ-specific AP observations derived from the CAP-CC&Ps of the 20 most frequent cancers. The clinical content of APSR was encoded using reference terminologies [LOINC, SNOMED CT (including items from TNM UICC, 7th edition), ICD-O and PathLex].

Current terminology systems for AP structured reporting gather terms of very different granularity [2, 3] and have not yet been compiled in a systematic approach. Moreover, the IHE APSR template provides a formal representation of only high-level AP observations resulting from human interpretation of low-level morphological abnormalities. There is still a need to extend the scope of IHE APSR and to integrate in a unique formal representation both high-level AP entities observable by humans and the corresponding low-level morphological abnormalities, especially those that can be quantified using image analysis tools.

Semantic models are formal representations of knowledge in a given domain that allow both human users and software applications to consistently and accurately interpret domain terminology [4, 5]. The formalisms used to represent meaning and the protocols to interact with semantic stores have permitted humans to create and accumulate semantic data in a form that both machines and humans could use and reuse. Coming after technologies like semantic networks – UMLS still uses them – ontologies are nowadays the preferred way to formalize semantic knowledge and to convert it into a standard storable form (e.g. using

[1] http://www.ihe.net/Technical_Frameworks/#anatomic AND Daniel C, Macary F, Rojo MG, Klossa J, Laurinavičius A, Beckwith BA, et al. Recent advances in standards for Collaborative Digital Anatomic Pathology, Diagn Pathol. 2011, 6:(1):S17.

L. Traore et al., diagnostic pathology 2016, 2:109
ISSN 2364-4893
DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109

triples at a lower level). According to Gruber [6], an ontology is « an explicit specification of a conceptualization », where « conceptualization » means an « abstract, simplified view of the world that we wish to represent for some purpose ». Another requirement is that such a specification should be shared, e.g., published. Most available ontologies use Description Logics (DL), often hidden under specialized languages like OWL, a standard of the World Wide Web Consortium (W3C), to describe pieces of reality – domains – and to control the complexity of query processing, e.g., to forbid asking for undecidable or questions. Tools like Protégé [7] enable humans to create, check, and query ontologies. Portals like BioPortal are servers, which make ontologies available for queries by humans and machines alike, either through human-oriented Graphics User Interfaces (GUIs) that execute in browsers, or Application Programming Interfaces (APIs) that programmers can use to set up client machines. Portals also play the role of publishers as they accept ontologies to be uploaded by authors, which entails an important service related to concept identification: each author is only responsible for uniquely identifying each concept within her proposed ontology, the portal providing a unique identifier for each ontology it publishes and also its own unique portal identifier (the concatenation of the three identifiers results in a universal resource identifier (URI) for each concept in each ontology). A major aspect of ontology design is the effort to rely as much as possible on existing semantics by referring to available ontologies for concepts already modelled. That emphasis on collaboration backed by web standards is probably the main reason for the breakthrough of ontologies compared to former technologies. This entails investing time to explore how the domain of interest relates to existing semantic knowledge.

Modelling and standardizing the semantics of AP diagnostic interpretation requires a major input from AP experts and tools are welcome to partly relieve them from the burdens of identifying and integrating concepts from a complex and rapidly evolving domain. Our hypothesis is that it is possible to provide AP experts with a visual representation summarizing at any time the current state of the concepts available in existing biomedical ontologies in the scope of the AP of tumours. In particular, such tool is intended to support the development of a future AP Observation Ontology (APOO) including both observable entities reported by humans (pathologists) and quantifiable entities automatically computed by machines.

Our objectives were:

ISSN 2364-4893

DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109

i) to identify within the reference biomedical ontologies made accessible by the NCBO Bioportal [8, 9] and within the UMLS metathesaurus [10] the available histopathological formalized knowledge covering the scope of CAP-CC&Ps ii) to build a sustainable visual representation of this knowledge using the semantic types of the UMLS metathesaurus [11, 12].

Methods

We propose a methodology and some tools to build a sustainable visual representation of standard-based AP knowledge about AP observations. Our approach consists in two steps: i) identifying the set of reference biomedical ontologies that are most relevant for semantic annotation of low-level morphological abnormalities; ii) annotating CAP-CC&Ps notes using these reference ontologies and building for each high level observable entity an integrative visual representation of the concepts corresponding to low-level morphological abnormalities.

We first evaluated the methodology in the limited scope of the two CAP-CC&Ps dedicated to invasive carcinoma (IC) and ductal carcinoma in situ (DCIS) of the breast.

Step 1: defining the set of reference biomedical ontologies that are the most relevant for semantic annotation of low-level morphological abnormalities. We selected from the two CAP-CC&Ps a subset of five quantifiable AP observations - i.e. observable entities that could be computed by image analysis tools - and the corresponding notes in the protocols (4 notes from IC, 1 note from DCIS). Two senior pathologists independently identified in each note a list of key concepts that unambiguously describe how pathologists should derive a high-level observation from low-level morphological characteristics observed in images. The union of the lists provided by the pathologists was considered as a “gold standard”. The NCBO platform provides Recommender [13, 14], a service that proposes a selection of ontologies found to be relevant to a text. The ontology ranking algorithm used by Recommender evaluates the relevance of each ontology to the input using a combination of the following four evaluation criteria: coverage, acceptance, detail of knowledge, and specialization. For each of these four criteria, a score is computed, then the scores obtained are weighted and aggregated into a final score for each ontology [14]. The weights are modifiable by users, with default values: Coverage =0.55, Acceptance=0.15, Knowledge detail = 0.15 and Specialization = 0,15.

We tested Recommender with the full notes <Table 3> and the gold standard <Table 4>, using in each case either the default set of weights for the four criteria used by Recommender or giving full weight to the coverage coefficient. NCBO however does not make public the explicit definition of the computed criteria (their authors were contacted), so we decided to implement our own method for ranking ontologies.

We used NCBO Annotator [15, 16], a tool supporting the biomedical community in tagging raw texts automatically with concepts from the biomedical ontologies and terminologies hosted by Bioportal (an option is provided to annotate from a user defined subset of ontologies).

We automatically annotated the notes by NCBO Annotator using the 508 ontologies available on the NCBO platform. For each note i and each ontology j, we kept only one occurrence of each of the terms annotated by ontology j in note i (some terms appear several times in the same note and Annotator returns one hit per occurrence to permit contextual studies), getting a set $S_{i,j}$ with $n_{i,j}$ elements. For each note i we computed the set of terms annotated in note i by any ontology (merging hits from all ontologies and removing multiple occurrences) getting the set $S_{i,tot}$ with $n_{i,tot}$ elements. We then computed ratios $n_{i,j}$ / $n_{i,tot}$ which we call «coverage ratios» since an ontology that would hit every term in a note would get a ratio of 100% for that note. Ordering the ontologies in decreasing order of their coverage ratios led to the selection of the 5 best ontologies for each note. We also computed coverage rates averaged over the set of 5 notes to summarize our results and ease discussion. The results are presented in Table 1. Then the same procedure was followed for annotating the "gold standard". The results are presented in Table 2.

Step 2: building for each high level observable entity an integrative representation of the concepts representing the corresponding low-level morphological abnormalities. By using the Terminology Services REST APIs of the Unified Medical Language System (UMLS) [17] we queried the UMLS metathesaurus to recognize in the text of our 5 notes concepts belonging to the UMLS. Then, we identified their Concept Unique Identifiers (CUIs), and their semantic types as modelled in the UMLS semantic network, which is a semantic formalism different from ontologies.

To explore possible presentations, a first graphical visualization of the semantics associated to the notes was manually built from the lists of extracted concepts using the free version of the

L. Traore et al., diagnostic pathology 2016, 2:109
ISSN 2364-4893
DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109

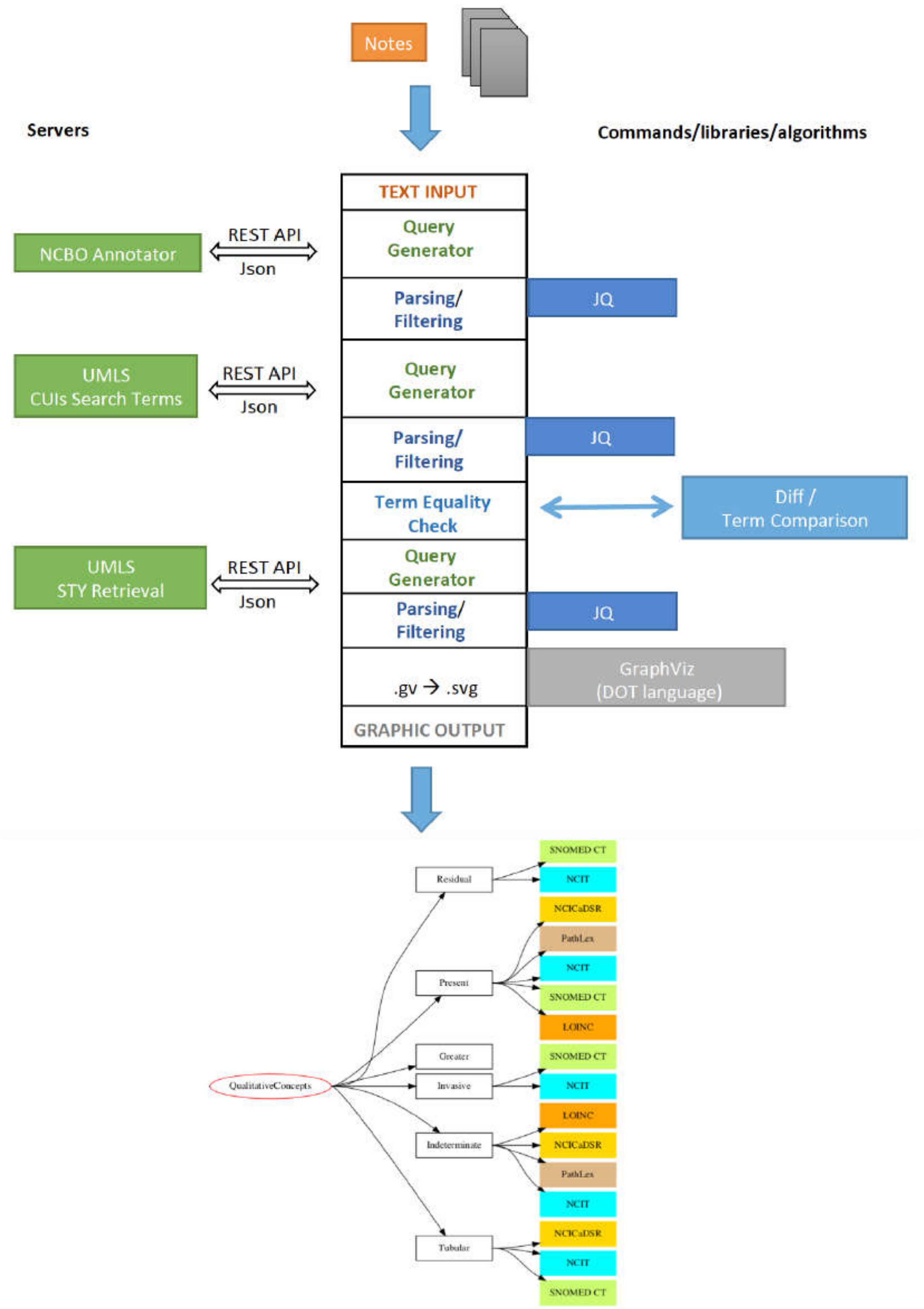


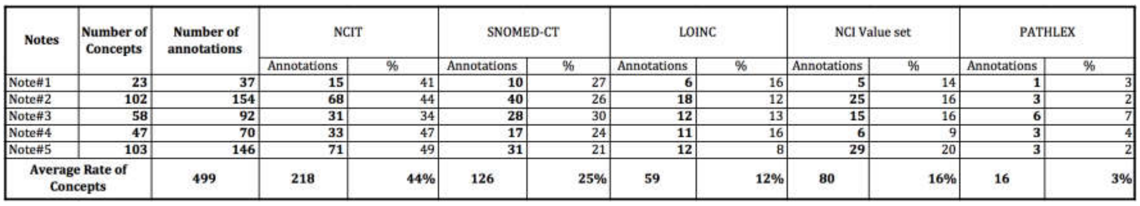
Figure 1: Automated workflow for the identification of available histopathological formalized knowledge from NCBO Bioportal and UMLS metathesaurus and building of the sustainable visual representation in the scope of the CAP-CC&P.

L. Traore et al., diagnostic pathology 2016, 2:109
ISSN 2364-4893
DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109

commercial visualization application MindMaple [18]. We then used the Python programming language [19], the jq tool [20], and GraphViz [21] as shown above in Figure 1 to automate each step of the workflow from text input to visual display, replacing MindMaple by GraphViz and producing a slightly different visualization.

Results

Our first result is the selection of the subset, at the time of writing, of the most relevant biomedical ontologies to be used for annotation of CAP-CC&P: SNOMED-CT, LOINC, NCIT, NCI CaDSR Value Sets and PathLex were found as the most appropriate reference ontologies in the context of the two notes related to breast cancer grading methods. For individual note annotation, the set of ontologies changes: NCIT and SNOMEDCT remains for all 5 Notes, LOINC for 4 notes, NCI caDSR for 3 notes, PATHLEX and CTV3 for 2 notes. However if we take the union of the first 5 ontologies in the annotation of each individual and Rank them, the order is as follows: SNOMEDCT, LOINC, NCIT, NCI CaDSR with PathLex and CTV3 ex æquo at the 5th position. Table 1 shows as percentages the coverage of the concepts of each note by the annotations of the reference ontologies. That these percentages can add to more than 100 for a single note reflects the possible overlap in ontologies coverage.



Notes	Number of Concepts	Number of annotations	NCIT		SNOMED-CT		LOINC		NCI Value set		PATHLEX	
			Annotations	%	Annotations	%	Annotations	%	Annotations	%	Annotations	%
Note#1	23	37	15	41	10	27	6	16	5	14	1	3
Note#2	102	154	68	44	40	26	18	12	25	16	3	2
Note#3	58	92	31	34	28	30	12	13	15	16	6	7
Note#4	47	70	33	47	17	24	11	16	6	9	3	4
Note#5	103	146	71	49	31	21	12	8	29	20	3	2
Average Rate of Concepts		499	218	44%	126	25%	59	12%	80	16%	16	3%

Table 1: Number of concepts and coverages of the reference ontologies in the annotation of observation notes of CAP-CC&P.

Table 2 uses the same format when only concepts from the gold standards are counted to quantify annotations. Minor changes in the average coverages of the 5 ontologies can be observed resulting – besides an unsurprising tie regarding low counts of the gold standards – in one swap between LOINC and SNOMED-CT in the ordered sequence. Overall the automated process reported in Table 1 captured well the quality scale deduced from the manually extracted gold standard in Table 2.

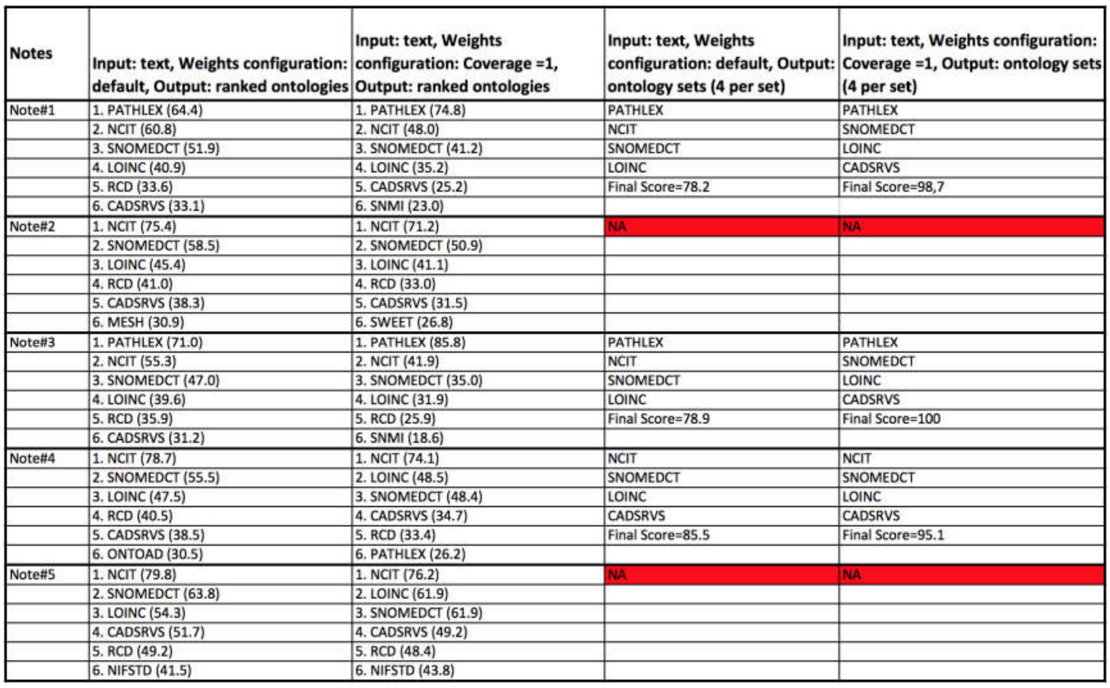
Gold standard		Ontologies									
		NCIT		LOINC		NCI Value set		SNOMED-CT		PATHLEX	
Notes	**Number of Concepts**	Concepts	Coverage (%)	Concepts	Coverage (%)	Concepts	Coverage (%)	Concepts	Coverage (%)	Concepts	Coverage (%)
Note#1	**2**	**2**	100	**2**	100	**0**	0	**0**	0	**0**	0
Note#2	**11**	**4**	36	**0**	0	**0**	0	**0**	0	**0**	0
Note#3	**16**	**7**	44	**1**	6	**0**	0	**1**	6	**1**	6
Note#4	**8**	**4**	50	**3**	38	**1**	13	**1**	13	**1**	13
Note#5	**29**	**10**	34	**1**	3	**6**	21	**4**	14	**1**	3
	Total	**Total**	Average coverage (%)	**Total**	Average coverage (%)	**Total**	Average coverage (%)	**Total**	Average coverage (%)	**Total**	Average coverage (%)
	66	**27**	41	**7**	11	**7**	11	**6**	9	**3**	5

Table 2: Number of concepts and coverages of the reference ontologies using the gold standard.

For each note, NCBO Recommender gave either a score for each ontology or set of 4 preferred ontologies, in both cases with adjustable weights for sub criteria. In Table 3, column 2 shows the 6 best-scored ontologies with the default weights, while column 3 shows the best 6 ontologies for all the weight put on the coverage criterion. Surprisingly, that second result is not exactly equal to our former procedure where we computed coverage after a query to Annotator. Indeed we could not find in the documentation or literature a precise formula for the coverage computed by Recommender. Columns 4 and 5 report the results for best sets of 4 of ontologies (4 being the maximal size of sets to be recommended). For the two longest notes, no answer (NA) came from the server after 30 minutes of multiple tries, after which we decided that no answer was available for such input size.

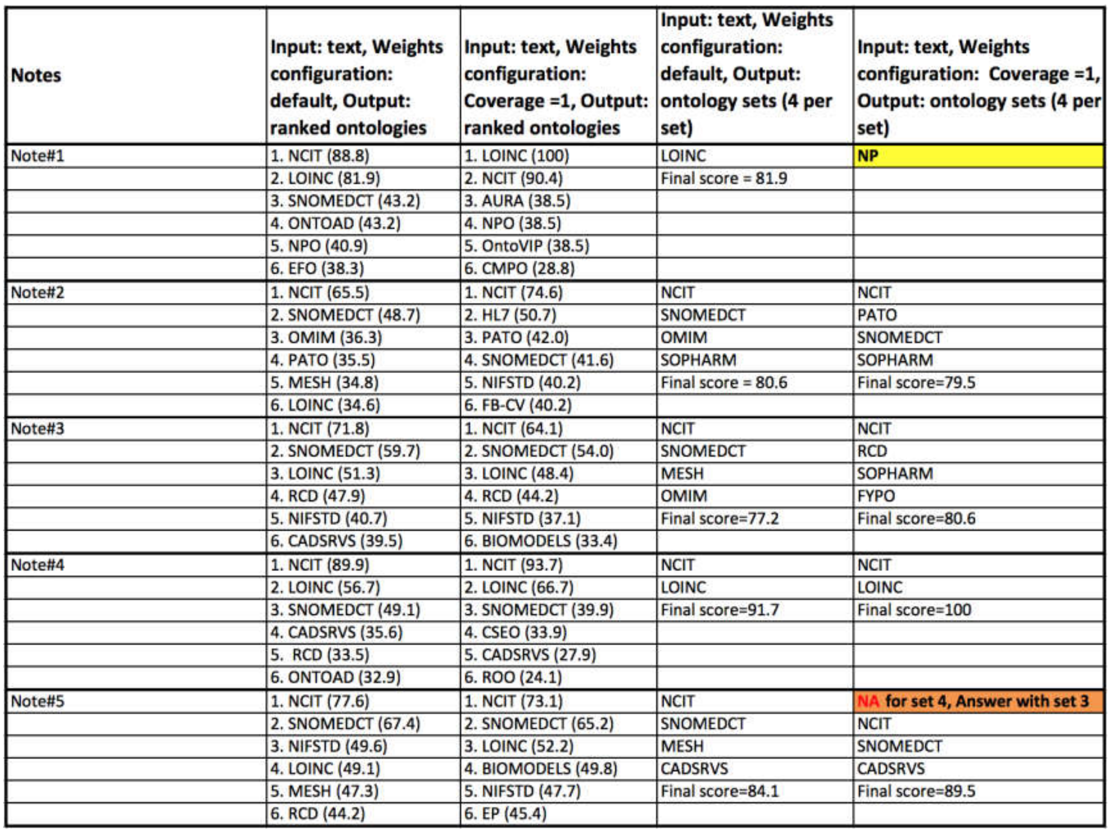
Similar results from Recommender for the gold standards as inputs are presented in table 4. Not Present (NP) means that there are no ontology sets recommended for the input provided. Please try the "Ontologies" output.Even if one recognizes ontologies selected with annotator at the first positions of Recommender rankings, differences appear soon, even more so for the gold standard results. Emphasizing coverage in the coefficients leads to unexpected ontologies.

For each note, the corresponding high-level observable entity term in the title was associated to a central node and peripheral nodes represented all the terms annotated in the note. Peripheral nodes were first linked to nodes representing their UMLS semantic types [12], these semantic nodes in turn linked to the central node. Clickable icons on links near term nodes permitted to pop up windows either to display the text of the notes where they appeared highlighted with the context of analysis and the concerned CAP-CC&P document, or to signal



Notes	Input: text, Weights configuration: default, Output: ranked ontologies	Input: text, Weights configuration: Coverage =1, Output: ranked ontologies	Input: text, Weights configuration: default, Output: ontology sets (4 per set)	Input: text, Weights configuration: Coverage =1, Output: ontology sets (4 per set)
Note#1	1. PATHLEX (64.4)	1. PATHLEX (74.8)	PATHLEX	PATHLEX
	2. NCIT (60.8)	2. NCIT (48.0)	NCIT	SNOMEDCT
	3. SNOMEDCT (51.9)	3. SNOMEDCT (41.2)	SNOMEDCT	LOINC
	4. LOINC (40.9)	4. LOINC (35.2)	LOINC	CADSRVS
	5. RCD (33.6)	5. CADSRVS (25.2)	Final Score=78.2	Final Score=98,7
	6. CADSRVS (33.1)	6. SNMI (23.0)		
Note#2	1. NCIT (75.4)	1. NCIT (71.2)	NA	NA
	2. SNOMEDCT (58.5)	2. SNOMEDCT (50.9)		
	3. LOINC (45.4)	3. LOINC (41.1)		
	4. RCD (41.0)	4. RCD (33.0)		
	5. CADSRVS (38.3)	5. CADSRVS (31.5)		
	6. MESH (30.9)	6. SWEET (26.8)		
Note#3	1. PATHLEX (71.0)	1. PATHLEX (85.8)	PATHLEX	PATHLEX
	2. NCIT (55.3)	2. NCIT (41.9)	NCIT	SNOMEDCT
	3. SNOMEDCT (47.0)	3. SNOMEDCT (35.0)	SNOMEDCT	LOINC
	4. LOINC (39.6)	4. LOINC (31.9)	LOINC	CADSRVS
	5. RCD (35.9)	5. RCD (25.9)	Final Score=78.9	Final Score=100
	6. CADSRVS (31.2)	6. SNMI (18.6)		
Note#4	1. NCIT (78.7)	1. NCIT (74.1)	NCIT	NCIT
	2. SNOMEDCT (55.5)	2. LOINC (48.5)	SNOMEDCT	SNOMEDCT
	3. LOINC (47.5)	3. SNOMEDCT (48.4)	LOINC	LOINC
	4. RCD (40.5)	4. CADSRVS (34.7)	CADSRVS	CADSRVS
	5. CADSRVS (38.5)	5. RCD (33.4)	Final Score=85.5	Final Score=95.1
	6. ONTOAD (30.5)	6. PATHLEX (26.2)		
Note#5	1. NCIT (79.8)	1. NCIT (76.2)	NA	NA
	2. SNOMEDCT (63.8)	2. LOINC (61.9)		
	3. LOINC (54.3)	3. SNOMEDCT (61.9)		
	4. CADSRVS (51.7)	4. CADSRVS (49.2)		
	5. RCD (49.2)	5. RCD (48.4)		
	6. NIFSTD (41.5)	6. NIFSTD (43.8)		

Table 3: NCBO Recommender results for Note#1 to Note#5 processed as text, with ontology ranking or set of ontologies as output.



Notes	Input: text, Weights configuration: default, Output: ranked ontologies	Input: text, Weights configuration: Coverage =1, Output: ranked ontologies	Input: text, Weights configuration: default, Output: ontology sets (4 per set)	Input: text, Weights configuration: Coverage =1, Output: ontology sets (4 per set)
Note#1	1. NCIT (88.8)	1. LOINC (100)	LOINC	NP
	2. LOINC (81.9)	2. NCIT (90.4)	Final score = 81.9	
	3. SNOMEDCT (43.2)	3. AURA (38.5)		
	4. ONTOAD (43.2)	4. NPO (38.5)		
	5. NPO (40.9)	5. OntoVIP (38.5)		
	6. EFO (38.3)	6. CMPO (28.8)		
Note#2	1. NCIT (65.5)	1. NCIT (74.6)	NCIT	NCIT
	2. SNOMEDCT (48.7)	2. HL7 (50.7)	SNOMEDCT	PATO
	3. OMIM (36.3)	3. PATO (42.0)	OMIM	SNOMEDCT
	4. PATO (35.5)	4. SNOMEDCT (41.6)	SOPHARM	SOPHARM
	5. MESH (34.8)	5. NIFSTD (40.2)	Final score = 80.6	Final score=79.5
	6. LOINC (34.6)	6. FB-CV (40.2)		
Note#3	1. NCIT (71.8)	1. NCIT (64.1)	NCIT	NCIT
	2. SNOMEDCT (59.7)	2. SNOMEDCT (54.0)	SNOMEDCT	RCD
	3. LOINC (51.3)	3. LOINC (48.4)	MESH	SOPHARM
	4. RCD (47.9)	4. RCD (44.2)	OMIM	FYPO
	5. NIFSTD (40.7)	5. NIFSTD (37.1)	Final score=77.2	Final score=80.6
	6. CADSRVS (39.5)	6. BIOMODELS (33.4)		
Note#4	1. NCIT (89.9)	1. NCIT (93.7)	NCIT	NCIT
	2. LOINC (56.7)	2. LOINC (66.7)	LOINC	LOINC
	3. SNOMEDCT (49.1)	3. SNOMEDCT (39.9)	Final score=91.7	Final score=100
	4. CADSRVS (35.6)	4. CSEO (33.9)		
	5. RCD (33.5)	5. CADSRVS (27.9)		
	6. ONTOAD (32.9)	6. ROO (24.1)		
Note#5	1. NCIT (77.6)	1. NCIT (73.1)	NCIT	NA for set 4, Answer with set 3
	2. SNOMEDCT (67.4)	2. SNOMEDCT (65.2)	SNOMEDCT	NCIT
	3. NIFSTD (49.6)	3. LOINC (52.2)	MESH	SNOMEDCT
	4. LOINC (49.1)	4. BIOMODELS (49.8)	CADSRVS	CADSRVS
	5. MESH (47.3)	5. NIFSTD (47.7)	Final score=84.1	Final score=89.5
	6. RCD (44.2)	6. EP (45.4)		

Table 4: NCBO Recommender results for Gold Standard terms from Note#1 to Note#5 processed as text, with ontology ranking or set of ontologies as output.

ISSN 2364-4893
DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109

and open each ontology annotating them with the corresponding NCBO or UMLS resources.

The visual report built using Mindmaple is shown on Figure 2 for the note on Glandular/tubular differentiation. It federates semantic knowledge from different sources, either from Bioportal's ontologies or from the UMLS metathesaurus and semantic network. The layout proposed by MindMaple after some manual interaction is satisfactory. Notice how the semantic types provide some hierarchical organizations of the peripheral concepts (e.g., the nodes linked to QualitativeConcepts). Figures 3 and 4 show how popup windows provide complementary information from source ontologies or the context where annotated terms appear in the notes with the title, ID, version and exact pages of the concerned CAP&CCP.

Automating the whole workflow from text input to visual display of the graphical representation was shown to be possible. We addressed the very similar APIs provided by BioPortal and ULMS (both use a REST architecture) and used the Python scripts provided in their documentations to automate all the necessary queries from the note input, obtaining answers in the common JSON file format. The jq tool was used to parse the results and extract the data we needed to build the graphical representation. GraphViz, and especially it's «dot» program were used to produce the graphical representation <Figure 5> from a text file which can be automatically written from the two outputs of jq using by a python program.

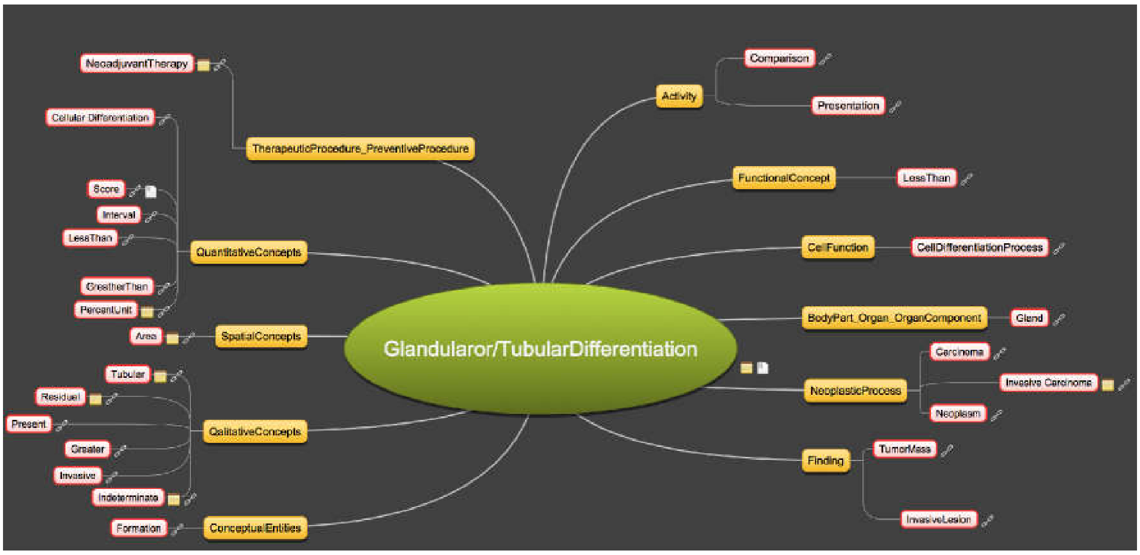
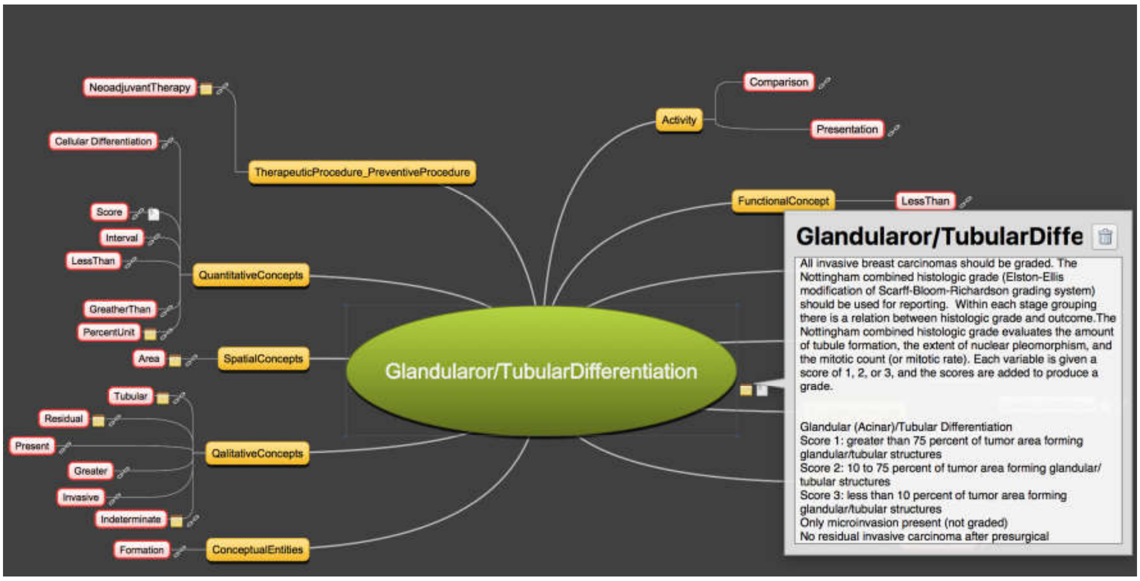


Figure 2: Graphical view of the sustainable semantic modelling approach in the context of Glandular/Tubular differentiation.

ISSN 2364-4893

DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109



Glandularor/TubularDiffe

All invasive breast carcinomas should be graded. The Nottingham combined histologic grade (Elston-Ellis modification of Scarff-Bloom-Richardson grading system) should be used for reporting. Within each stage grouping there is a relation between histologic grade and outcome.The Nottingham combined histologic grade evaluates the amount of tubule formation, the extent of nuclear pleomorphism, and the mitotic count (or mitotic rate). Each variable is given a score of 1, 2, or 3, and the scores are added to produce a grade.

Glandular (Acinar)/Tubular Differentiation
Score 1: greater than 75 percent of tumor area forming glandular/tubular structures
Score 2: 10 to 75 percent of tumor area forming glandular/ tubular structures
Score 3: less than 10 percent of tumor area forming glandular/tubular structures
Only microinvasion present (not graded)
No residual invasive carcinoma after presurgical

Figure 3: the popup window permits reading the term in text context within the note modelled here.

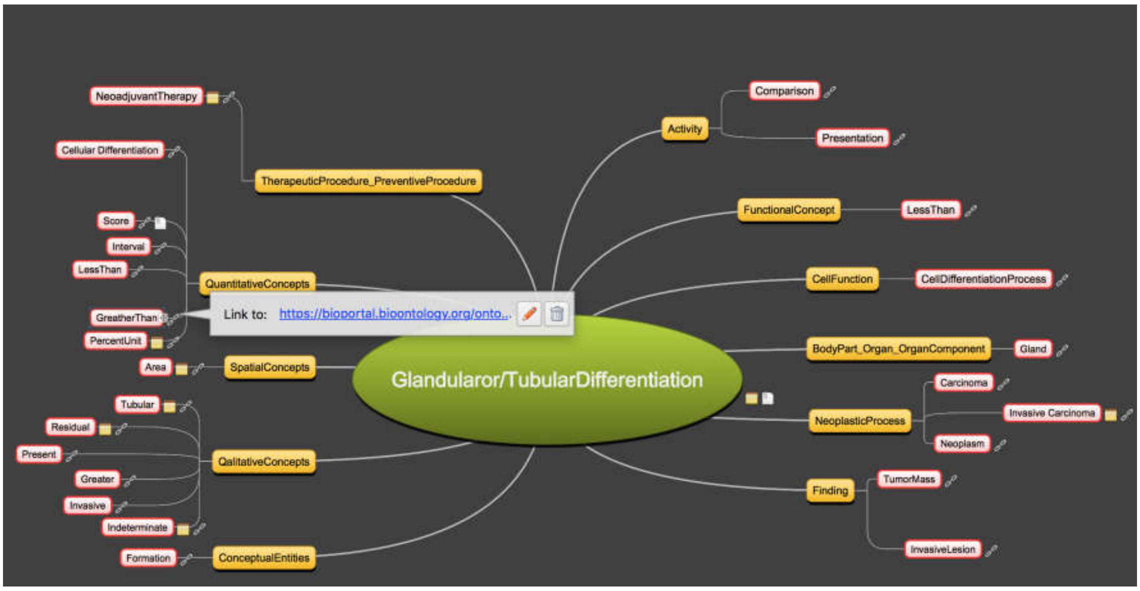


Figure 4: access to source ontologies is readily available for further exploration of the semantic modelling of concepts annotated in this note.

Discussion

Our objectives of sustainability address robustness to resource updates and domain extensibility. As for extensibility to other CAP-CC&Ps, the issue of manually extracting gold standards for the remaining 65 protocols seems less serious considering the good agreement between Table 1 and Table 2 : using direct input of the whole text of notes led to a quite satisfactory selection of ontologies for the two protocols studied. If gold standards are thus not necessary for ontology selection, the only manual task needed before an extension of the present work to the totality of CAP-CC&Ps would consist in selecting notes or some other paragraphs within the protocols. We would use our method of ontology selection based on Annotator as long as the issues we found with Recommender are not cleared. Updating the visual report to follow the evolution of the source ontologies or the UMLS metathesaurus and semantic network is addressed by simply rebuilding the visual report often enough. The workflow proposed here is compatible with complete automation. Each query we first performed manually has an API counterpart, using standard formats such as JavaScript Object Notation (JSON) or Extensible Mark-up Language (XML) for data exchange. We chose JSON [22] for simplicity. To build a graphical representation, data from lists of annotated terms (Annotator output) and semantic types (UMLS output) had first to be correlated. In particular one had to check that the preferred term of the UMLS file returned for a term already known as a hit for Annotator was equal to that term. A visual alert could be triggered only if the equality test fails, but one could also exploit the other terms, such as synonyms, that UMLS returns, and build complementary visualizations in further work. Once the results of Annotator and UMLS were integrated in a common data structure, writing the GraphViz source file was straightforward and would only require a simple algorithm. That file was converted by the «dot» utility of Graphviz into a svg file displayable in standard browsers. The svg format was chosen because of its simplicity for inserting hyperlinks from graphviz. The visualization presented can be extended in many ways, for instance to replace the role of UMLS semantic types by an ontology specific semantic object.

Even at this basic stage we found the presentation quite informative in our quest for links to image processing tasks.

The novelty of this approach is the federation of the knowledge issued from different ontologies and the sustainable management that automation eases. This formal representation is based on the UMLS semantic types of the concepts and will refer to source ontologies for future maintenance. Figure 2 shows the proposed semantic modelling in the context of glandular/tubular differentiation. For each concept we have information related to its Concept Unique Identifier (CUI), semantic type, source ontology, semantic relation and links

ISSN 2364-4893

DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109

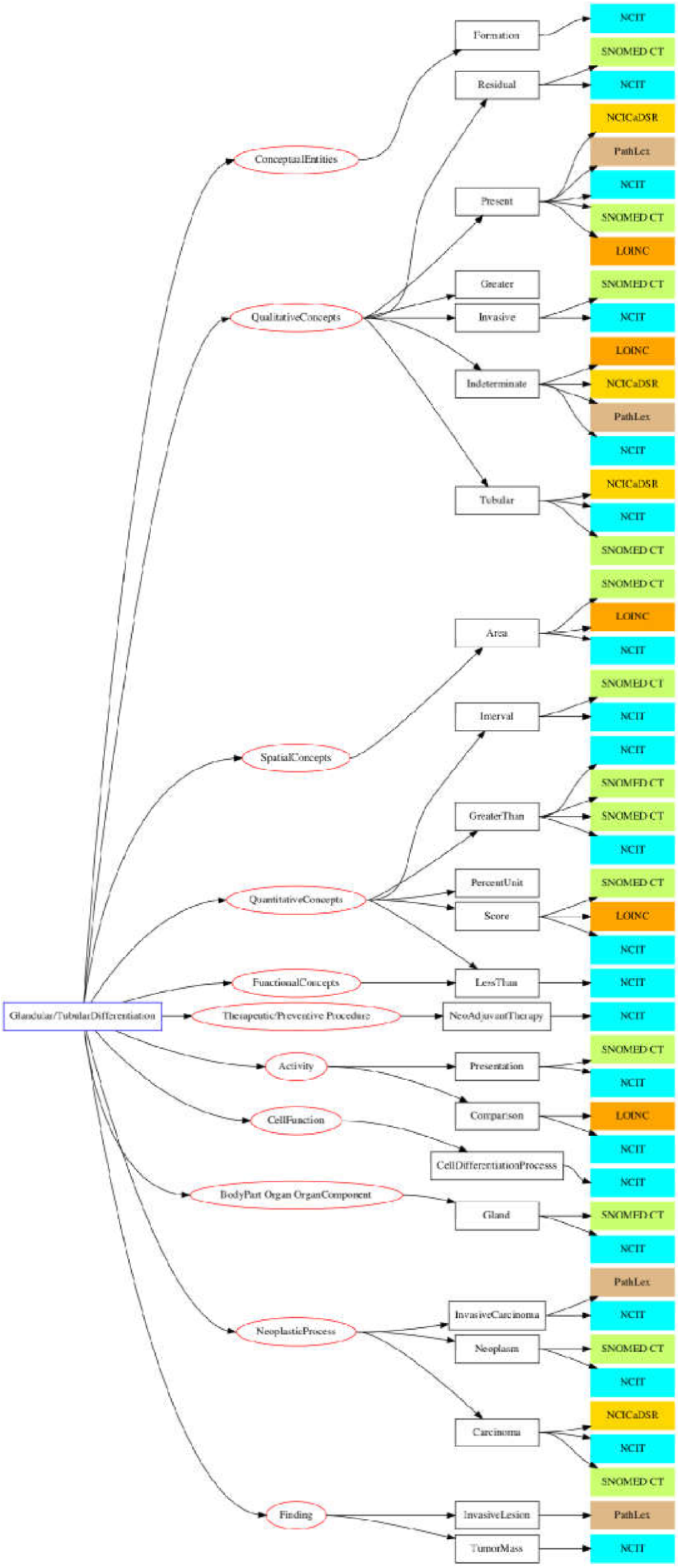


Figure 5: Graphical view of the sustainable semantic modelling approach in the context of Glandular/Tubular differentiation obtained with GraphViz.

to related metadata. These preliminary results open the prospect of building an Anatomic Pathology Observation ontology that will allow an accurate representation of AP reports understandable by both human and software applications.

The current proposed model includes relevant terms corresponding to the various features defining the grades and scores of breast tumours. It provides a sustainable formal representation of the knowledge involved during the AP diagnostic process. Extending the scope of such resource would benefit from the involvement of an international consortium of pathologists provided with supportive tools enabling community members to contribute terminological content and provide feedback on existing classes and properties.

Conclusions

This study proposed a formal representation of histopathological knowledge related to breast cancer grading, underpinning AP-focused informatics tools for patient care and clinical research.

We described the role of this semantic approach in bridging the gap between the CAP-CC&Ps data elements, NCBO ontologies, the UMLS Metathesaurus and the UMLS Semantic Network. Greater participation of the AP community is needed in the development, adoption, and maintenance of such a source in a sustainable manner. The proposed approach and tools, based on the CAP-CC&Ps, aim at supporting AP experts in building a standard-based representation of low-level morphological abnormalities observed in cancer that can be quantified using image analysis tools. This effort is complementary to the Integrating the Healthcare Enterprise (IHE) initiative building a standard-based representation of high-level AP observations required in cancer AP reports. Additional efforts are needed to achieve a workable standard-based formal representation of histopathological knowledge integrating both observable entities reported by humans (pathologists) and quantifiable entities automatically computed by machines. Providing such unique formal representation paves the way for a more efficient use of computer aided diagnosis in AP. Sustainable management of the explicit and unambiguous semantics associated to the diagnostic interpretation of AP image by both humans (pathologists) and computers grading process, (image analysis algorithms) will support a better use of existing image analysis algorithms such as the ones

ISSN 2364-4893
DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109

elaborated in the MICO[2] [23] and their adaptation to other contexts (same type of cancer but different organs, e.g., from breast to prostate, or same organ but different types of cancer).

References

1. “CAP - Cancer Protocol Templates.” [Internet], College of American Pathologists, 2016. Available from: http://www.cap.org
2. Daniel C., Booker D., Beckwith B., Della Mea V., García-Rojo M., Havener L., Kennedy M., Klossa J., Laurinavicius A., Macary F., Punys V., Scharber W., Schrader T., Standards and specifications in pathology: image management, report management and terminology, *Stud Health Technol Inf. 2012*, 179: 105–122.
3. Haroske G., Schrader T., A reference model based interface terminology for generic observations in Anatomic Pathology Structured Reports, *Diagnostic Pathology 2014*, 9(1): 4.
4. Bodenreider O., “Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support,” *Yearb. Med. Inform. 2008*, 67–79.
5. Rubin D. L., Shah N. H., Noy N. F., “Biomedical ontologies: a functional perspective,” *Brief. Bioinform. 2007*, 9(1):75–90.
6. Gruber T. R., A Translation Approach to Portable Ontology Specifications, *Knowl Acquis 1993*, 5(2):199–220.
7. “Protege-OWL 3.x Support - Ontologies on Image Processing... any idea!” [Internet]. [Accessed: 16-Oct-2015]. Available: http://protege-project.136.n4.nabble.com/Ontologies-on-Image-Processing-any-ideatd417.html
8. Musen M. A., Noy N. F., Shah N. H., Whetzel P. L., Chute C. G., Story M.-A., Smith B., and the NCBO team, The National Center for Biomedical Ontology, *J. Am. Med. Inform. Assoc. 2012*, 19:2, 190–195.
9. Whetzel P. L., Noy N. F., Shah N. H., Alexander P. R., Nyulas C., Tudorache T., Musen M. A., “BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications,” *Nucleic Acids Res. 2011*, 39:W541–W545.

[2] ANR TecSan project MICO (Cognitive MIcroscopy) : http://www.agence-nationalerecherche. fr/?Project=ANR-10-TECS-0015

10. Bodenreider O., "The Unified Medical Language System (UMLS): integrating biomedical terminology." [Internet]. [Accessed: 17-Dec-2015]. Available: http://nar.oxfordjournals.org
11. "Semantic Types and Groups." [Internet]. Bethesda (MD): National Library of Medicine (US) [Updated: 27-Apr-2015; Accessed: 17-Apr-2016]. Available: https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml
12. "Current Semantic Types." [Internet]. Bethesda (MD): National Library of Medicine (US). [Accessed: 17- Apr-2016]. Available: https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html
13. Jonquet C., Musen M. A., Shah N. H., Building a biomedical ontology recommender web service, *J. Biomed. Semant. 2010*, 1:1, S1.
14. "Ontology Recommender | bioontology.org." [Internet]. Stanford (CA): The National Center for Biomedical Ontology (US). [Accessed: 10-Dec-2015]. Available: http://www.bioontology.org/ontology-recommender
15. "Annotator | NCBO BioPortal." [Internet]. Stanford (CA): The National Center for Biomedical Ontology (US). [Accessed: 11-Dec-2015]. Available: http://bioportal.bioontology.org/annotator
16. Shah N. H., Bhatia N., Jonquet C., Rubin D., Chiang A. P., Musen M. A., Comparison of concept recognizers for building the Open Biomedical Annotator, *BMC Bioinformatics 2009*, 10:9, S14.
17. "UMLS REST API Home Page." [Internet]. Bethesda (MD): National Library of Medicine (US) [Accessed: 26-May-2016]. Available: https://documentation.uts.nlm.nih.gov/rest/home.html
18. "MindMaple - Mind Mapping Software - Improve Brainstorming Techniques." [Internet]. Santa Clara (CA): MindMaple Inc. [Accessed: 17-Apr-2016]. Available: http://www.mindmaple.com/Products/Features/
19. "Welcome to Python.org," *Python.org*. [Internet]. Beaverton (OR): Python Software Foundation. [Accessed: 26-May-2016]. Available: https://www.python.org/
20. "jq." [Internet]. GITHub. [Accessed: 26-May-2016]. Available: https://stedolan.github.io/jq/
21. Emden R., Gansner, "Graphviz | Graphviz - Graph Visualization Software." [Internet]. [Accessed: 20-May-2016]. Available: http://www.graphviz.org/

ISSN 2364-4893

DOI: http://dx.doi.org/10.17629/www.diagnosticpathology.eu-2016-2:109

22. “JSON.” [Internet]. ECMA International, 2013. [Accessed: 26-May-2016]. Available: http://json.org/
23. Racoceanu D., Capron F., Towards Semantic-Driven High-Content Image Analysis. An Operational Instantiation for Mitosis Detection in Digital Histopathology, *Comput. Med. Imaging Graph.*, 2014.
24. Traore L., et al. Sustainable formal representation of breast cancer grading histopathological knowledge, Diagnostic Pathology 2016, 1:154.