Sémantique des données d'observation en neuro-imagerie selon un point de vue réaliste

Emna Amdouni *, Bernard Gibaud **
,

* Institut de Recherche Technologique B<>com, Rennes, France
emna.amdouni@uni-lyon2.fr,
** LTSI Inserm 1099, Université de Rennes 1, Rennes, France
bernard.gibaud@univ-rennes1.fr

Résumé. L'objectif de ce travail est de décrire avec une approche réaliste la signification des données d'observation en neuro-imagerie sous un format formel pour faciliter leur interprétation par les cliniciens et leur réutilisation dans d'autres contextes.

1 Introduction

Dans le cadre de ce travail, nous nous focalisons sur l'étude de la sémantique des données d'observation associées aux tumeurs cérébrales. Associer une sémantique à une donnée d'observation consiste à mettre la donnée en relation avec d'autres entités participant, soit au phénomène observé (par exemple, la température corporelle d'un sujet), soit à un processus d'observation (par exemple, cette même valeur est reliée à une action d'observation, à son observateur, à l'instrument de mesure utilisé et l'instant de la mesure).

En termes de modélisation ontologique, il existe deux approches de modélisation qui sont adoptées pour décrire sémantiquement le contenu sémantique d'une image : l'approche cognitive (Cimino, 2006) et l'approche réaliste (Smith, 2006). L'approche cognitive oriente sa modélisation autour des "concepts" décrits par des "termes" faisant partie d'un lexique spécifique que nous manipulons pour l'attribution des propriétés (des qualités ou des dimensions) des données d'observation ; ces termes sont construits selon notre perception et connaissance des entités du monde réel. Contrairement à l'approche cognitive, l'approche réaliste aligne les termes des terminologies aux entités qui existent dans le monde indépendamment d'agents cognitifs reconnaissant leur existence. De plus, l'approche réaliste considère qu'il n'y a "qu'une seule réalité objective universelle" ; "chaque attribut du patient est lui-même une entité unique en réalité et on lui attribue son propre identifiant. Ainsi, les entités réelles référencées peuvent être de différents types : entités réelles, mesures, etc.

Jusqu'à aujourd'hui, les principaux formats existants pour la description des données d'observation sont : les comptes rendus radiologiques DICOM (Digital Imaging and Communications in Medicine) SR (Structured Report) (Clunie, 2000) et le modèle AIM (Annotation and Imaging Markup) (Channin et al., 2010) :

— Le compte rendu DICOM SR est une structure de données qui est définie dans le standard DICOM. DICOM SR permet de formaliser la représentation des observations

d'imagerie dans les rapports cliniques en introduisant un ensemble de règles qui limitent l'organisation des concepts et un vocabulaire (c'est-à-dire des codes et des significations de code associées) couvrant le domaine des observations d'imagerie. Les observations d'imagerie incluses en DICOM SR concernent principalement les modalités d'images DICOM, les images dérivées, les résultats de segmentation, les mesures (taille, surface, volume, etc.), les évaluations qualitatives, etc.

— Le modèle AIM est un modèle d'information et un format de fichier basé sur XML qui décrit les informations minimales nécessaires pour enregistrer des annotations d'images. Ce modèle d'information a introduit les entités les plus pertinentes dans l'annotation des images médicales (des annotations radiologiques qui se réfèrent à des mesures, des textes, des observations, des formes graphiques délimitant des régions d'intérêt, etc).

Ces formats informatiques permettent de décrire le contenu des images médicales, mais ils ne sont pas adaptés pour supporter un raisonnement logique. En effet, ils utilisent des termes issues de terminologies normalisées (par exemple SNOMED CT), mais n'exploitent pas les définitions formelles de ces termes. Par conséquent, seules les recherches basées sur des mots-clés peuvent être traitées sur des outils de prise de décision exploitant les données de ces formats. Rubin et al., Levy MA et de nombreux autres chercheurs ont encouragé l'utilisation d'ontologies formelles (Rubin et al., 2009) (Van Soest et al., 2014) pour assurer un raisonnement automatique sur ces modèles de données.

Notre travail est basé sur l'hypothèse que l'utilisation des technologies du Web sémantique, en particulier les ontologies et leurs capacités de raisonnement, peut rendre plus explicite la sémantique des données d'observation en neuro-imagerie et faciliter leur exploitation et leur interprétation «avancées». La couverture de toutes les informations impliquées dans l'évaluation des tumeurs cérébrales est impossible car aucune source consensuelle n'existe pour spécifier les exigences précises de ce domaine. Pour surmonter cette difficulté, nous avons limité notre étude au domaine couvert par la terminologie VASARI [1] (Visually Accessible Rembrandt Images). VASARI constitue un cas d'utilisation représentatif de l'ensemble minimaliste des connaissances basiques qui nécessitent une modélisation formelle.

La terminologie VASARI est un vocabulaire d'annotation des gliomes cérébraux de haut grade en particulier le glioblastome multiforme (GBM) dans les images IRM (Imagerie par Résonance Magnétique). Son objectif principal consiste à normaliser la description des tumeurs cérébrales et à faciliter leur interprétation par les neuro-radiologues. La terminologie VASARI contient trente critères d'imagerie et elle a été développée par des experts du domaine qui ont considéré la majorité des évaluations possibles des tumeurs cérébrales en IRM. Chaque critère d'imagerie de la terminologie VASARI est référencé par un numéro (F_1, F_2, etc.) et un ensemble de valeurs de scores possibles. Par exemple, le critère "F_1:tumor location" de VASARI évalue la localisation de l'épicentre géographique et il définit sept valeurs d'étiquette possibles = frontal, temporal, insulaire, pariétal, occipital, tronc cérébral, cervelet.

L'objectif principal de ce travail est de faciliter l'identification, le partage et le raisonnement sur les résultats d'observation des tumeurs cérébrales via la formalisation de leurs significations sémantiques.

1. https://wiki.cancerimagingarchive.net/display/Public/VASARI+Research+Project

2 Matériel et méthode

L'annotation des caractéristiques d'imagerie des tumeurs cérébrales implique différents types d'entités : les entités physiques, les qualités liées aux objets physiques et les mesures de volume et de taille. Selon la terminologie VASARI, les entités physiques qui caractérisent certaines anomalies des tissus cérébraux sont : la tumeur cérébrale, l'épicentre de la tumeur cérébrale, les composantes de la tumeur cérébrale (à savoir : région de prise de contraste, région de non prise de contraste, partie nécrotique, composante oedémateuse et la bordure de la tumeur cérébrale) et la partie périphérique d'une tumeur cérébrale ou d'une partie de la tumeur cérébrale.

L'ontologie VASARI a été conçue selon l'approche réaliste. Notre méthodologie de modélisation se compose de cinq étapes principales qui peuvent être décrites comme suit : tout d'abord, nous avons analysé la signification de l'aspect étudié par chaque critère VASARI F_i et nous avons trié ses configurations possibles pour établir la liste des valeurs possibles autorisées pour chaque critère. Deuxièmement, nous avons identifié les principales entités observées qui sont impliquées dans chaque critère. Troisièmement, nous avons défini les entités observées, soit avec des classes d'ontologies existantes soit avec de nouvelles classes ontologiques. Quatrièmement, nous avons spécifié les axiomes caractérisant ces entités. Enfin, nous nous sommes assurés que toutes les configurations possibles pour chaque critère F_i sont bien modélisées de manière formelle.

Après une analyse de la signification des caractéristiques étudiées par les critères VASARI et de l'identification des différentes entités qu'elles impliquent, nous avons procédé à leur description formelle. Cette étape n'a pas été une tâche très simple pour nous étant donné que nous avons rencontré certains problèmes de modélisation que nous avons soulevés et discutés dans ce travail (Amdouni et Gibaud, 2016). Ces problèmes de modélisation concernent : les données d'observations négatives, les données de représentation de la connaissance spatiale et la représentation d'entités complexes touchant notamment à des mesures dérivées de mesures élémentaires.

Toute la construction ontologique (taxonomies des classes et des propriétés) s'appuie sur la version 2 de l'ontologie BFO (Basic Formal Ontology) (Grenon et al., 2004), ce qui facilite l'intégration d'ontologies spécialisées issues de la fonderie OBO (Smith et al., 2007). En particulier, nous avons réutilisé les ontologies suivantes : FMA, IAO, PATO, OBI, OGMS, UO et l'ontologie RO (toutes ces ontologies sont disponibles sur Bioportal du National Center of Biomedical Ontology (NCBO) [2]). Nous avons développé l'ontologie VASARI en format OWL2 en utilisant la version 5 de l'outil Protégé. Nous avons utilisé Ontofox [3] pour l'extraction des ontologies OBO.

Dans notre travail expérimental, nous avons développé un outil d'annotation sémantique des données fondé sur les classes et relations de l'ontologie VASARI. Cet outil permet à l'utilisateur de transformer la description informelle des 30 critères VASARI en une description formelle. Pour faire le test, nous l'avons appliqué au corpus de données REMBRANDT [4]. REMBRANDT contient les observations relatives à 30 critères VASARI formulées par 3 radiologues et relatives à 34 patients atteints de GBM. Les résultats d'observation sont représentés

2. https://bioportal.bioontology.org/
3. ontofox.hegroup.org/
4. https://wiki.cancerimagingarchive.net/display/Public/VASARI+Research+Project

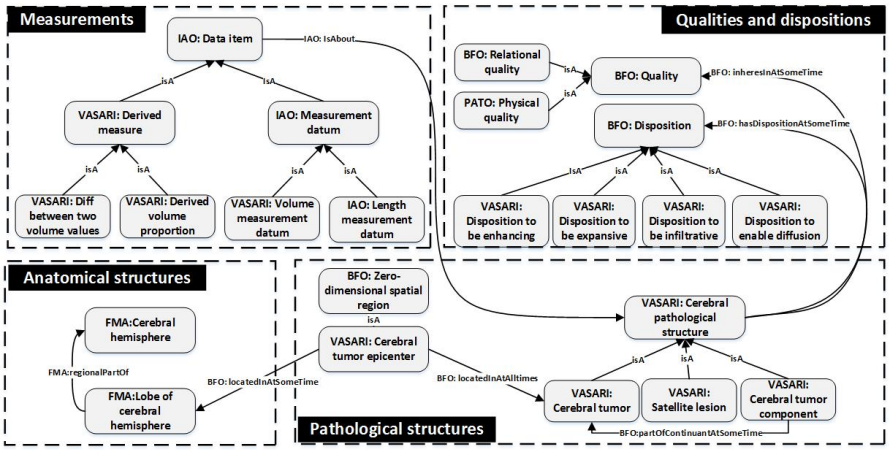

FIG. 1 – *Le modèle de base des classes principales dans l'ontologie VASARI*

dans un fichier Excel où chaque feuille de calcul contient des évaluations soumises par un radiologue. L'ensemble de données sémantiques résultant a été utilisé pour monter l'importance de la représentation réaliste des données d'observations en neuro-imagerie.

3 Résultats

L'ontologie VASARI est composée de huit modules d'ontologies et contient environ 570 classes OWL et 120 propriétés. La figure 1 met en évidence les quatre aspects sémantiques majeurs qui décrivent le domaine VASARI, à savoir : les structures pathologiques, la localisation anatomique, les qualités et dispositions, et les mesures.

Le logiciel d'annotation prend comme donnée d'entrée l'ensemble des valeurs des critères d'imagerie de la base REMBRANDT ainsi que le schéma de l'ontologie VASARI. Ensuite, pour annoter sémantiquement les données, le logiciel réalise quatre tâches principales : tout d'abord, il crée des instances des classes de l'ontologie VASARI en se basant sur les valeurs des critères. Deuxièmement, il décrit les critères d'imagerie en générant des triplets RDF qui établissent des liens sémantiques entre les instances. Troisièmement, il crée dans un graphe RDF des affirmations à partir des triplets. Quatrièmement, il sérialise les données en RDF/XML et enregistre le graphe RDF en mémoire. On note que le logiciel stocke séparément le schéma de l'ontologie et les données d'instances (dans les prochains paragraphes, on emploie le terme Tbox pour faire référence au schéma et Abox pour désigner la base d'instances). En terme de performance, le logiciel génère le graphe RDF de l'ensemble de données contenues dans la base REMBRANDT en 1.06 s (soit $\approx$ 0.47s par feuille).

La figure 2 montre un exemple d'annotation de 10 observations du patient 9000_00_5316 de la base REMBRANDT avec notre ontologie : F1.tumor location= parietal, F2.side of tumor= left, F3.eloquent brain= none, F5.proportion of enhancing= 6-33%, F8.cyst= no ; F12.definition of the enhancing margin= well defined, F14.proportion of edema= 6-33%, F16.hemorrhage= no, F29.lesion size= 6,5cm. Considérons cet exemple pour illustrer quelques capacités d'in-

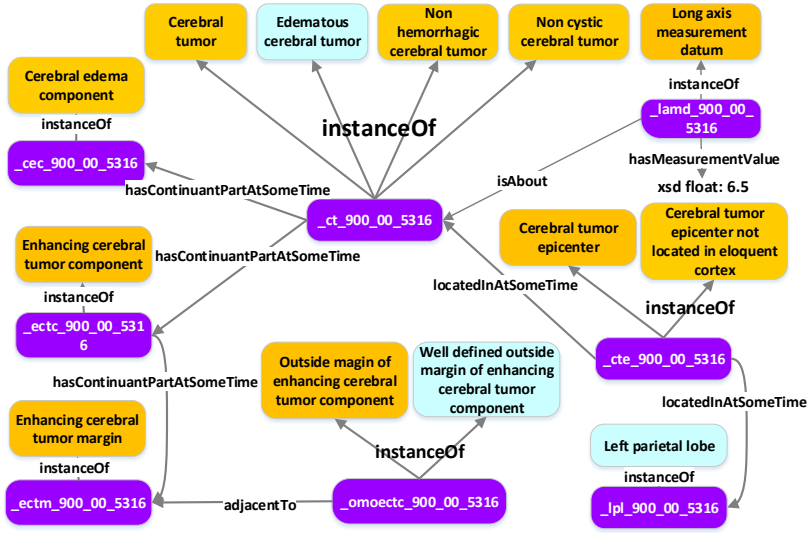

FIG. 2 – *Représentation sémantique des observations du patient 9000_00_5316 de la base REMBRANDT. Les classes déclarées sont représentées en jaune, celles inférées sont représentées en bleu clair et les instances en violet.*

férence et démontrer comment la représentation sémantique permet d'exploiter des connaissances sur différents aspects de la tumeur. Comme le montre la Figure le raisonneur a déduit les affirmations suivantes :

— L'épicentre de la tumeur cérébrale de ce patient est localisée dans le lobe pariétal gauche vu que elle est déclarée située dans son lobe pariétal et dans son hémisphère cérébral gauche. L'axiome mis en jeu est : "fma:left temporal lobe" ≡ def. "fma:parietal lobe" and "fma:regionalPartOf" some "fma:left cerebral hemisphere".
— La tumeur de ce patient est de type oedémateuse vu qu'elle contient une partie de type oedémateux. L'axiome mis en jeu est : "vasari:edematous cerebral tumor" ≡ def. "bfo:has continuant part at some time" some "vasari:cerebral edema component".
— La périphérie de la tumeur qui prend la contraste est bien définie. L'axiome mis en jeu est : "vasari:well defined outisde margin of enhanncing cerebral tumor component" ≡ def. "vasari:outside margin of enhancing cerebral tumor component" "obi:has quality at some time" some "vasari:well defined".

Pour détecter les incohérences contenues dans la base de connaissances nous pouvons utilisé l'objet "ValidityReport" de l'API JENA. Cette structure encapsule tous les axiomes et assertions inconsistants qui sont détectés. Pour générer des explications sur les causes d'incohérences, nous avons utilisé la méthode "explainconsistency()". Cette méthode énumère tous les axiomes impliqués.

4 Conclusion

En résumé, nous pensons que les données de la neuro-imagerie devraient être codées dans un format structuré qui rend leurs significations explicites et facilitent ainsi leur compréhension ainsi que leur gestion. La prise en considération de la sémantique des critères d'imagerie (valeurs de mesure, qualités, composantes de la tumeur, localisation des lésions, etc.) est nécessaire pour deux raisons principales : (1) améliorer la qualité des soins cliniques qui ont tendance à fournir des traitements personnalisés aux patients par l'utilisation de recommandations cliniques qui sont basées sur les critères d'évaluation (2) soutenir la recherche clinique sur le développement de nouveaux biomarqueurs d'imagerie en combinant les données cliniques avec des informations provenant de différents domaines médicaux.

Références

Amdouni, E. et B. Gibaud (2016). Concept-based versus realism-based approach to represent neuroimaging observations. In *Keod : Proceedings of The 8th International Joint Conference On Knowledge Discovery, Knowledge Engineering and Knowledge Management-Vol. 2*, pp. 179–185.

Channin, D., P. Mongkolwat, V. Kleper, K. Sepukar, et D. Rubin (2010). The caBIG annotation and image markup project. *Journal of digital imaging 23*(2), 217–225.

Cimino, J. (2006). In defense of the desiderata. *Journal of biomedical informatics 39*(3), 299–306.

Clunie, D. (2000). *DICOM structured reporting*. PixelMed Publishing.

Grenon, P., B. Smith, et L. Goldberg (2004). Biodynamic ontology : applying bfo in the biomedical domain. *Studies in health technology and informatics*, 20–38.

Rubin, D. L., P. Mongkolwat, et D. S. Channin (2009). A semantic image annotation model to enable integrative translational research. *Summit on translational bioinformatics 2009*, 106.

Smith, B. (2006). From concepts to clinical reality : an essay on the benchmarking of biomedical terminologies. *Journal of biomedical informatics 39*(3), 288–298.

Smith, B., M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. Goldberg, K. Eilbeck, A. Ireland, C. Mungall, et al. (2007). The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology 25*(11), 1251–1255.

Van Soest, J., T. Lustberg, D. Grittner, M. S. Marshall, L. Persoon, B. Nijsten, P. Feltens, et A. Dekker (2014). Towards a semantic pacs : Using semantic web technology to represent imaging data. *Studies in health technology and informatics 205*, 166.

Summary

The aim of this work is to describe with a realistic approach the meaning of neuroimaging data in a formal format to facilitate their interpretaion by the clinicians and their reuse in other contexts.