Atelier DAHLIA

DigitAl Humanities and cuLtural herItAge: data and knowledge management and analysis

Comité d'organisation et du programme :

Claudia Marinica (MIDI - ETIS UMR8051, CY Université, ENSEA, CNRS, Institut IDHN)
Fabrice Guillet (DUKe – LS2N, Polytech'Nantes)
Florent Laroche (IS3P – LS2N, Ecole Centrale de Nantes)
Julien Velcin (DMD - ERIC, Université de Lyon 2)

organisé par le **groupe de travail DAHLIA** soutenu par l'Association EGC

conjointement avec la conférence
Extraction et Gestion des Connaissances (EGC2020)

le 28 janvier 2020 à Bruxelles, Belgique

Editeurs :

Claudia Marinica
Laboratoire ETIS UMR8051, CY Université, ENSEA, CNRS, Institut IDHN
page web : https://perso-etis.ensea.fr//marinica/
email : claudia.marinica@u-cergy.fr

Fabrice Guillet
Laboratoire LS2N, équipe DUKe - Polytech'Nantes
page web : http://www.univ-nantes.fr/site-de-l-universite-de-nantes/fabrice-guillet--2320.kjsp
email : fabrice.guillet@univ-nantes.fr

Florent Laroche
Laboratoire LS2N, équipe IS3P - Ecole Centrale de Nantes
page web : http://www.florentlaroche.net/
email : florent.laroche@ec-nantes.fr

Julien Velcin
Laboratoire ERIC - Université Lyon 2
page web : http://mediamining.univ-lyon2.fr/velcin/
email : julien.velcin@univ-lyon2.fr

Accès en ligne :

Atelier DAHLIA : http://dahlia.egc.asso.fr/atelierDAHLIA-EGC2020.html
Groupe de travail DAHLIA : http://dahlia.egc.asso.fr
Mailing liste : gt-dahlia@egc.asso.fr

Table de matières

Ce que le numérique fait à l'archéologie et aux archéologues

Christophe Tufféry*

*Inrap, Direction Scientifique et Technique 121, rue d'Alésia, CS 20007
75685 Paris Cedex 14
christophe.tuffery@inrap.fr
http://www.inrap.fr

Résumé. La communication présente les premières réflexions d'une thèse d'histoire par le projet engagée cette année à l'Université de Cergy-Pontoise en partenariat avec l'Institut National du Patrimoine. Ce travail de recherche emprunte l'essentiel de sa méthodologie et de ses corpus bibliographiques de référence à l'histoire des sciences et techniques et à l'épistémologie de l'archéologie. L'objectif est de tenter de décrire les effets de l'usage du numérique, considéré comme un ensemble de dispositifs techniques, sur l'archéologie comme discipline et sur les archéologues dans leurs pratiques et leurs identités professionnelles.

Résumé

L'archéologie s'est assez tôt inspirée d'une division scientifique du travail. Depuis plus d'un siècle, les archéologues n'ont eu de cesse d'optimiser, de rationaliser leurs activités. Avec le développement de l'archéologie préventive, le secteur professionnel a adopté des techniques, des méthodes et des organisations de travail permettant d'accroître cette rationalisation de l'activité des archéologues, pour répondre autant à des objectifs économiques que scientifiques. L'utilisation de l'informatique (mécanographie) en archéologie est presque aussi ancienne que l'informatique elle-même (Gardin 1991, Collectif 2017). Depuis une trentaine d'années, avec l'apparition de la micro-informatique, l'usage d'outils informatiques en archéologie n'a cessé de se développer, sur le terrain ou dans les laboratoires et les centres de recherche (Desachy 2008). Depuis une dizaine d'années, la multiplication des outils numériques sur le terrain s'accélère (tablettes, smartphones, drones). Un nouveau terme et un nouveau domaine de compétences sont apparus, l'archéomatique. Des formations spécifiques se sont multipliées dans ce domaine. Des pratiques systématisées autour de tel ou tel type de progiciel se sont manifestées en archéologie comme dans d'autres disciplines : SIG, 3D, photogrammétrie, réalité virtuelle, etc. Dans ce contexte, il est légitime de s'interroger sur ce que l'informatique fait à l'archéologie et aux archéologues. Plus précisément, qu'est-ce que les outils informatiques font à la discipline archéologique en termes épistémologiques ? Quels sont leurs effets sur les pratiques scientifiques, les identités professionnelles (Dubar 1991) ou encore les organisations professionnelles des archéologues ? Tel est le sujet de la thèse qui, par des questionnements de nature historique, épistémologique, sociologique doit aider à porter un regard réflexif d'ensemble, encore peu développé sur cette évolution. Le thème de la thèse nécessite de commencer par s'interroger sur le numérique, considéré ici comme un ensemble de moyens techniques et de compétences pour les mettre en œuvre, en réponse à des objectifs

scientifiques et des obligations réglementaires. Mais s'interroger sur le numérique doit aussi conduire à s'interroger sur ce qu'est la technique, un terme polysémique qui recouvre des notions multiples. Le recours au numérique en archéologie ne doit pas être considéré uniquement sous l'angle des moyens. Il doit être appréhendé dans le cadre d'une approche multiple, historique, scientifique, technologique, épistémologique, etc. Dans la thèse, nous ne pourrons pas aborder tous ces points de vue même si le sujet mérite une approche pluridisciplinaire. Sur le plan méthodologique, nous avons prévu :

- Des lectures dans divers domaines susceptibles de contribuer à l'approche pluridisciplinaire
- Des entretiens avec des archéologues de l'Inrap et d'autres organisations professionnelles
- Des observations d'archéologues en situation dans leurs « lieux de savoir » : terrain, laboratoires, colloques, etc.)
- Des études d'outils informatiques utilisés en archéologie (matériels, logiciels, applications, etc.)
- Des études de corpus de données archéologiques à divers moments du cycle de vie des données
- Etc.

Aujourd'hui, les pratiques discursives liées au numérique en archéologie cherchent à faire admettre ces usages comme synonyme de progrès, de révolution incontournable et bienfaisante et forcément vertueuse pour les organisations professionnelles de l'archéologie et pour les archéologues eux-mêmes. Partant de ce constat, la thèse cherchera à mobiliser la méthode archéologique foucaldienne pour tenter de révéler les différents niveaux de ces pratiques discursives, en tenant de dépasser l'apparente évidence de la raison qu'elles défendent. Le numérique en archéologie introduit-il une « rupture épistémologique », un « seuil de scientificité » ou n'est-il qu'une étape dans le changement des pratiques par l'usage de nouveaux moyens techniques (Tufféry et Augry 2019)? Peut-il créer de nouvelles conditions de véridiction en faisant évoluer les critères de vérité de notre époque (Foucault 1969) ? Peut-il faire advenir un nouveau régime de vérité et ouvrir vers une nouvelle dimension anthropologique par la mise en œuvre de nouvelles formes de relations entre des chercheurs et leurs objets de savoirs ou relèverait-il d'une forme de croyance (Gollac et Kramarz 2000) ?

Références

Jean-Claude Gardin (1991), Le calcul et la raison. Essais sur la formalisation du discours savant. Editions de l'EHESS, coll. Recherches d'histoire sociale, 296 pages

Collectif (2017), Jean-Claude Gardin (1925-2015), Les nouvelles de l'archéologie, 144 | 2016 [En ligne] URL: https://journals.openedition.org/nda/3448; DOI : 10.4000/nda.3448 (consulté le 15 janvier 2020)

Desachy B. (2008), De la formalisation du traitement des données stratigraphiques en archéologie de terrain. Sciences de l'Homme et Société. Université Panthéon-Sorbonne - Paris I, 2008. Français. https://tel.archives-ouvertes.fr/tel-00406241

Dubar C. (1991), La socialisation, construction des identités sociales et professionnelles, Paris, Armand Collin, Collection U, 256 pages

Foucault M, (1969), L'archéologie du savoir. Paris, Éditions Gallimard, Bibliothèque des Sciences humaines. 294 pages

Gollac M. et Kramarz F. (2000) L'informatique comme pratique et comme croyance. In: Actes de la recherche en sciences sociales. Vol. 134, septembre 2000. L'informatique au travail. pp. 4-21

Tufféry C. et Augry S. (2019), « Harmonisation de l'acquisition des données d'opération d'archéologie préventive. Retours d'expérience et perspectives à partir de l'application EDArc », Actes de l'atelier DAHLIA (DigitAl Humanities and cuLtural herItAge : data and knowledge management and analysis) dans le cadre de la conférence EGC (Extraction et Gestion des Connaissances) 22 janvier 2019 à Metz, pp. 21-27 http://dahlia.egc.asso.fr/atelier_DAHLIA2019_actes.pdf (consulté le 15 janvier 2020)

Summary

The paper presents the first reflections of a history thesis by the project launched this year at the University of Cergy-Pontoise in partnership with the Institut National du Patrimoine. This research work borrows most of its methodology and bibliographical corpus of reference from the history of science and technology and the epistemology of archaeology. The objective is to attempt to describe the effects of the use of digital technology, considered as a set of technical devices, on archaeology as discipline and on archaeologists in their professional practices and identities.

Valorisation du patrimoine culturel matériel et immatériel des Pyrénées : retour d'expérience

Landy Rajaonarivo*, Marie-Noelle. Bessagnet*, Christian Sallaberry*, Philippe Roose**, Annig Lacayrelle*, Patrick Etcheverry**, Christophe Marquesuzaà**

*Université de Pau et des Pays de l'Adour, E2S UPPA, Laboratoire LIUPPA, 64000 Pau, France
prenomcomplet.nom@univ-pau.fr
http://liuppa.univ-pau.fr
** Université de Pau et des Pays de l'Adour, E2S UPPA, Laboratoire LIUPPA, IUT Informatique, 64100 Bayonne, France
prenomcomplet.nom@univ-pau.fr
http://liuppa.univ-pau.fr

Résumé étendu

L'accès à des ressources numériques du patrimoine culturel à travers des applications mobiles ou des sites Web est aujourd'hui dans l'ADN de tout touriste. De plus, un système de recommandation intégré dans de telles applications est un autre atout auquel les personnes sont actuellement rompues et attachées.

Nous présentons le projet européen FEDER TCVPYR (http://tcvpyr.iutbayonne.univ-pau.fr/) dont le but est de promouvoir le tourisme dans les régions pyrénéennes françaises en mettant en valeur des éléments méconnus de son patrimoine culturel matériel et immatériel, en particulier celui en relation avec la villégiature et le thermalisme de la chaine pyrénéenne. TCVPYR est un projet pluri-disciplinaire impliquant des scientifiques de domaines variés des SHS (historiens, géographes, anthropologistes, chercheurs des Inventaires régionaux, par exemple) et de l'informatique.

Pour atteindre ces objectifs, les chercheurs en SHS collectent sur le terrain des données du patrimoine culturel matériel et immatériel dans des zones ciblées des Pyrénées. Ces données sont ensuite stockées dans les bases de données (distinctes) des Inventaires régionaux. Après traitement, ces données sont agrégées pour être valorisées notamment dans une application mobile afin de promouvoir le patrimoine et le tourisme dans cette région. Cette application mobile permet aux touristes mais également aux scientifiques impliqués dans le projet d'accéder aux données du patrimoine culturel matériel et immatériel sous la forme de points d'intérêts (POI). En outre, ces POI sont présentés selon le profil de l'utilisateur et de son contexte environnemental, incluant les aspects spatio-temporels. Par exemple, l'application peut suggérer à un touriste un itinéraire avec plusieurs POI prenant en compte ses intérêts, son moyen de transport, le temps dont il dispose et sa position actuelle. D'autres paramètres sont automatiquement collectés comme son niveau de batterie, la qualité de la connexion internet sur la zone de visite proposée, afin d'ajuster le contenu et d'éventuellement anticiper des chargements de données. L'originalité de notre approche réside dans l'accès à des données expertes du patrimoine pyrénéen par les touristes à travers une application mobile se basant sur un algorithme de recommandation qui consiste non seulement à suggérer les POI pertinents selon le profil et le contexte du touriste mais

également à encourager la découverte de POIs qui doivent surprendre l'utilisateur dans le cadre d'heureuses découvertes.

Le projet TCVPYR nous a permis de travailler sur plusieurs verrous scientifiques que nous avons pu valoriser dans des articles de recherche aussi bien du côté SHS qu'informatique.

Côté informatique, nous avons élaboré :

- Un ensemble de modèles :
  - un modèle de données pivot et fédérateur permettant de gérer l'hétérogénéité des sources de données ;
  - un modèle utilisateur définissant le profil de l'utilisateur caractérisé par le genre, la catégorie d'âge et les préférences. Un utilisateur peut avoir plusieurs préférences thématiques (parc, musée, par exemple) et historiques (XVème siècle, par exemple) ;
  - un modèle de contexte composé des trois facettes utilisateur, physique et ressource. Le contexte utilisateur vient compléter le profil utilisateur avec des informations propres à une visite donnée ;
  - un modèle d'itinéraire, un itinéraire étant une succession d'étapes dans lesquelles figurent les POI à visiter. Il est défini par sa date de génération, le moyen de déplacement utilisé et l'ensemble des étapes à enchainer pendant la visite. Chaque étape désigne un ou plusieurs POI correspondant à l'étape.
- Des algorithmes originaux implémentés dans un prototype d'un système client/serveur pour une application touristique ;
- Des thésaurus utilisateurs qui sont des simplifications des thésaurus définis par les experts. L'idée de simplification réside dans la diminution des concepts à utiliser en les regroupant, ainsi que sur le nommage des concepts afin de les rendre compréhensible au grand public ;
- Un démonstrateur Web permettant de visualiser les POI patrimoniaux au fur et à mesure de la mise à jour de la base de données ;
- Un système de recommandation hybride : l'originalité de notre approche consiste à rapprocher les POI déjà parcourus par le touriste de ceux également parcourus par d'autres touristes de profil similaires afin de lui proposer de nouveaux POI selon le principe « un autre touriste qui a visité les mêmes POI que vous a aussi visité les suivants » ;
- Un système de génération d'itinéraires prenant en compte une fonction de scoring originale basée sur 3 aspects : la pertinence de chaque POI de l'itinéraire proposé par rapport (i) aux préférences de l'utilisateur, (ii) au trajet et à la durée de visite et (ii) au comportement des autres utilisateurs ;
- Un premier prototype d'une application mobile Pyr-AT a été testé avec succès sur le terrain.

Le lecteur intéressé pourra se reporter aux références bibliographiques suivantes : (Bessagnet et al, 2018), (Fonteles et al, 2018), (Rajaonarivo et al, 2019a), (Rajaonarivo et al, 2019b) ainsi qu'au site web du projet http://tcvpyr.iutbayonne.univ-pau.fr/.

Côté SHS, les chercheurs ont produit des articles de recherche, organisé des colloques, conférences et journées d'étude liés au patrimoine pyrénéen. Les chercheurs du projet,

chargés d'inventaire, ont également réalisé des restitutions publiques auprès des instances locales.

Enfin, toutes les données de ce projet ainsi que l'application finale seront publiées en open data et open source.

Références

(Bessagnet et al, 2018) Bessagnet M.-N., Etcheverry P., Lacayrelle A., Marquesuzaà C., Rajaonarivo L., Roose P. et al.(2018, octobre). Leveraging heterogeneous Cultural Heritage data to promote tourism. In Open Source Geospatial Research and Education Symposiums (OGRS). Lugano, Switzerland. Consulté sur https://hal.archives-ouvertes.fr/hal-01976405

(Fonteles et al, 2018) Fonteles A. S., Bessagnet M.-N., Lacayrelle A., Sallaberry C. (2018, November). Un environnement pour la valorisation de données patrimoniales hétérogènes. In Spatial Analysis and GEOmatics (SAGEO). Montpellier, France. (To be published)

(Rajaonarivo et al, 2019a) L. Rajaonarivo, A. Fonteles, C. Sallaberry, P. Roose, M-N. Bessagnet, P. Etcheverry, A. Lacayrelle, C. Marquesuzaa, C. Cayere & Q. Coudert. *Recommandation et valorisation d'objets patrimoniaux hétérogènes*, Inforsid 2019

(Rajaonarivo et al, 2019b) Landy Rajaonarivo, André Fonteles, Christian Sallaberry, Marie-Noëlle Bessagnet, Philippe Roose, Patrick Etcheverry, Christophe Marquesuzaà, Annig Le Parc-Lacayrelle, Cécile Cayèré, Quentin Coudert, Recommendation of Heterogeneous Cultural Heritage Objects for the Promotion of Tourism. ISPRS Int. J. Geo-Information 8(5): 230 (2019)

Remerciements

Ce projet est réalisé dans le cadre du programme de recherche Européen TCVPYR (2017-2020), financé par l'Union Européenne (FEDER) en partenariat avec les régions Occitanie et Nouvelle-Aquitaine.

Towards a terminological knowledge base on Islamic archaeology: linguistic and conceptual aspects

Bruno Almeida*, Rute Costa*

*NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal
{brunoalmeida,rute.costa}@fcsh.unl.pt
http://clunl.fcsh.unl.pt

Abstract. This paper describes work carried out for the OntoAndalus project at NOVA CLUNL. This project aims at establishing the foundations for a terminological knowledge base (TKB) on Islamic archaeology. The main part of the work carried out so far is the development of an ontology of artefact types in al-Andalusian pottery studies, which was done by reusing the DOLCE+DnS Ultralite foundational ontology. Subsequently, Portuguese and Spanish terms for artefact types were extracted from a corpus of specialised texts and represented by means of lexical networks. The Lexicon Model for Ontologies (Lemon), recently developed by a W3C Community Group, was put forward as a promising framework for integrating language-specific and language-independent information in a future TKB on the domain. The possibility of aligning OntoAndalus with Lemon is also discussed in this paper.

1 Introduction

Developing multilingual terminological resources in the Semantic Web and integrating them in the Linked Open Data Cloud have motivated closer ties between terminology work and ontology development. The recognition of applied ontology as a domain in its own right and the acknowledgement of computational ontologies as an integral part of the Semantic Web have brought new light to the foundations of terminology as a discipline (Durán-Muñoz and Bautista-Zambrana, 2013; Roche, 2012; Santos and Costa, 2015; Temmerman and Kerremans, 2003). Furthermore, the development of ontologies representing shared domain knowledge has been featured in several recent terminological projects, from healthcare to cultural heritage (Brin-Henry et al., 2019; Roche et al., 2019).

The notion of a knowledge-based and computational terminology is, however, far from new, having an important precedent in the notion of 'terminological knowledge base' (TKB) put forward in the 1990's (Meyer et al., 1992). The development of such resources is often complex and relies on interdisciplinary work involving, for example, terminologists/linguists, domain experts and computer scientists. TKBs, however, have a number of advantages, such as the possibility of drafting natural language definitions based on formal descriptions of classes and other ontological predicates, which should lead to more consistent terminological resources. Furthermore, TKBs may provide richer conceptual and linguistic information about specialised domains and their terms in multiple languages.

This paper describes work carried out in the context of the OntoAndalus project, with the purpose of establishing the foundations of a TKB in a subdomain of Islamic archaeology, namely the pottery artefacts of al-Andalus. This work was originally motivated by terminological issues in the community of Portuguese and Spanish archaeologists working on this domain, particularly with respect to the harmonisation of terms denoting artefact types in pottery studies (Bugalhão et al., 2010; Gómez Martínez, 2004; Torres et al., 2003). In order to have a TKB interoperable in the Semantic Web, it was decided to develop a domain ontology focussing on the relevant artefact types, which was eventually named OntoAndalus. [1] Further work was carried out with regard to the extraction and representation of artefact designations in Portuguese and Spanish, as well as on their relationship with the relevant concepts of the domain, already present in OntoAndalus. Lemon, the Lexicon Model for Ontologies developed by the W3C Ontology-Lexica Community Group (2016), was adopted for this purpose, since – as will be shown – it allows to represent rich semantic and grammatical information about terms and their relationship to ontology predicates.

2 Overview of OntoAndalus

OntoAndalus is an ontology of relevant artefact types in pottery studies of al-Andalus. The ontology was developed based on the interpretation of a bilingual (Portuguese and Spanish) corpus of specialised texts in the domain, in which the artefact typologies put forward by the CIGA group (Bugalhão et al., 2010) in Portugal and by Rosselló-Bordoy (1978, 1991) in Spain were paramount. Other sources of information included English textbooks and other reference works on archaeology and pottery analysis, such as Rice (1987).

DOLCE+DnS Ultralite (DUL) was selected as a foundational ontology for the development of OntoAndalus following a review of literature on philosophy and applied ontology (Almeida, 2019). DUL is a version of DOLCE for the Semantic Web with the addition of the Descriptions and Situations ontology (DnS) for modelling non-physical objects (e.g. organisations, tasks, mental concepts) (Gangemi, 2016). DUL as a number of practical advantages, such as its stability, completeness and being made available in OWL format. More importantly, DUL allowed for a rich conceptualisation of the domain, in which important concepts in archaeology were sufficiently explicated (e.g. artefact, part, quality, function) (Almeida and Costa, 2019).

OntoAndalus was developed in Protégé (Musen, 2015) by importing the DUL ontology through the `owl:imports` property. At this time, OntoAndalus consists of 161 classes, 30 object properties and 135 individuals (excluding elements in the DUL namespace). OntoAndalus includes 72 artefact types, which are organised in the functional collections shown in Fig. 1.

In the following section, lighting artefacts will be outlined as a case study of how artefact types can be modelled in OntoAndalus.

2.1 Modelling artefact types in OntoAndalus

Lighting artefacts include some of the more representative objects in the archaeology of al-Andalus. According to Gómez Martínez (2004), lighting artefacts may fall within four pottery

1. OntoAndalus is made available through `https://github.com/brunoalmeida81/OntoAndalus`.

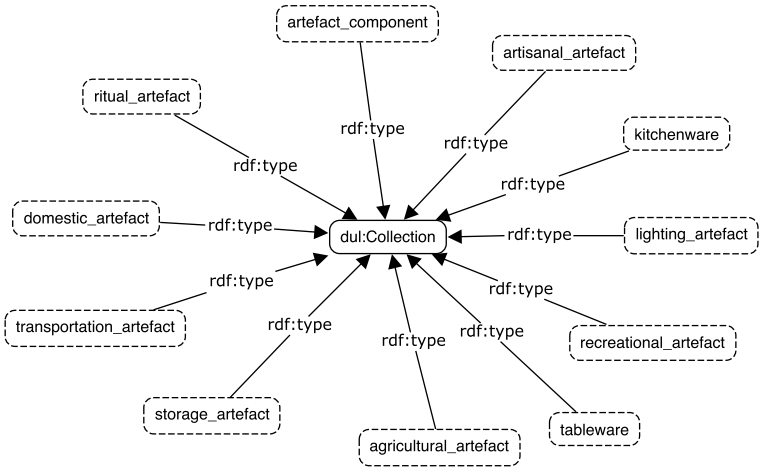

FIG. 1 – *Functional collections declared in OntoAndalus. Dashed boxes denote individuals and solid boxes denote classes.*

types (or *series*), which are denoted by the following Spanish terms: (i) *candil*; (ii) *almenara* or *policandela*; (iii) *lamparilla*; and (iv) *fanal* or *linterna* (Fig. 2).

In OntoAndalus, the above-mentioned pottery types were represented as classes subsumed by `dul:DesignedArtifact`. The following English identifiers were attributed to each class: (i) `Lamp`; (ii) `MultipleLamp`; (iii) `StationaryLamp`; and (iv) `Lantern`. Each class was then described according to the following pattern: **superordinate class + collection + function + part(s) or component(s)**. This pattern was also the basis for drafting natural language definitions, which were associated to each class via the `skos:definition` property. The formal and natural language definitions or descriptions of each class are the following:

Lamp (*candil*). *Def.* Artefact for lighting in closed spaces composed of at least one spout and a single chamber for liquid fuel.

```
Lamp rdfs:subClassOf dul:DesignedArtifact

Lamp rdfs:subClassOf dul:isMemberOf some lighting_artefact

Lamp rdfs:subClassOf hasFunction some containing_fire_for
_lighting_in_closed_spaces_using_liquid_fuel

Lamp rdfs:subClassOf hasComponent exactly 1 LampFuelChamber

Lamp rdfs:subClassOf hasComponent min 1 Spout
```

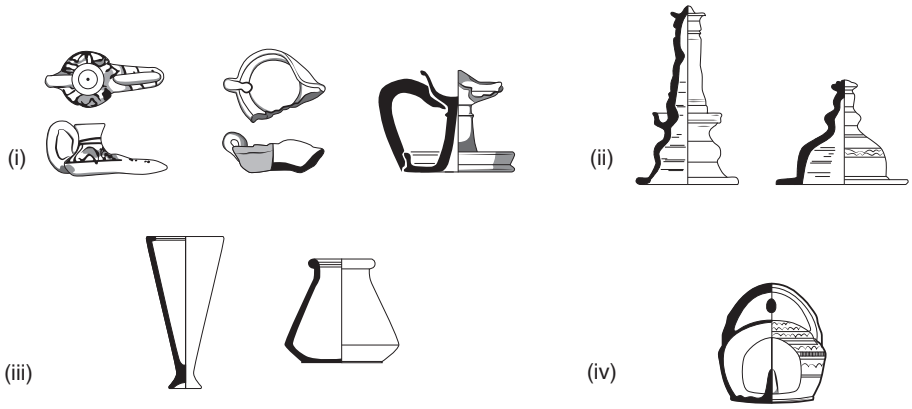

FIG. 2 – *Lighting artefact types. Vector illustrations by Mariana Tavares.*

Explanation: The function of the *candil* was that of domestic lighting. It was typically composed of a fuel chamber, a single spout for holding a wick and a handle. There are, however, objects in this series which were designed to be suspended and, as a consequence, do not have any handles. Furthermore, less typical instances of this type have more than one spout (Gómez Martínez, 2004; Rosselló-Bordoy, 1991).

MultipleLamp (*almenara* or *policandela*). *Def.* Artefact for stationary lighting in closed spaces composed of more than one chamber for liquid fuel unified by a structure.

```
MultipleLamp rdfs:subClassOf dul:DesignedArtifact

MultipleLamp rdfs:subClassOf dul:isMemberOf some lighting
_artefact

MultipleLamp rdfs:subClassOf hasFunction some containing
_fire_for_stationary_lighting_in_closed_spaces_using
_liquid_fuel

MultipleLamp rdfs:subClassOf hasComponent min 2
LampFuelChamber

MultipleLamp rdfs:subClassOf hasComponent some
MultipleLampStructure
```

Explanation: Not much is known on this artefact type, whose instances would have consisted of several fuel chambers unified by a base for stationary or, possibly, suspended lighting (Gómez Martínez, 2004; Rosselló-Bordoy, 1991).

StationaryLamp (*lamparilla*). *Def.* Artefact for stationary lighting in closed spaces composed of a single chamber for liquid fuel.

```
StationaryLamp rdfs:subClassOf dul:DesignedArtifact

StationaryLamp rdfs:subClassOf dul:isMemberOf some lighting
_artefact

StationaryLamp rdfs:subClassOf hasFunction some containing
_fire_for_stationary_lighting_in_closed_spaces_using
_liquid_fuel

StationaryLamp rdfs:subClassOf hasComponent exactly 1
LampFuelChamber
```

Explanation: This possible lighting artefact type would have been left on a table stand or on a discoidal plate, which could have held several instances (Gómez Martínez, 2004; Vallejo Triano and Escudero Aranda, 1999). In OntoAndalus, the `StationaryLamp` class is further divided according to shape: either inverted frustum or bifrustum-shaped.

Lantern (*fanal* or *linterna*). *Def.* Artefact for lighting in open spaces composed of a single chamber for solid fuel.

```
Lantern rdfs:subClassOf dul:DesignedArtifact

Lantern rdfs:subClassOf dul:isMemberOf some lighting
_artefact

Lantern rdfs:subClassOf hasFunction some containing
_fire_for_lighting_in_open_spaces_using
_solid_fuel

Lantern rdfs:subClassOf hasComponent exactly 1
LanternFuelChamber
```

Explanation: The function of this artefact was to provide a light source outdoors by using a solid fuel, such as wax. Its typical features include having a closed shape, globular body and a single handle, which are based on isolated findings (Bugalhão et al., 2010; Gómez Martínez, 2004).

Since the `Lamp` class corresponds to the most widely studied type of lighting artefacts in the archaeology of al-Andalus, the ontology puts forward a classification of possible subtypes according to multiple criteria of subdivision, namely: (i) vessel form (e.g. open, closed); (ii) type of spout (e.g. channelled, pinched); (iii) inclusion of a discus or neck; and (iv) inclusion of a tall foot. Relevant classes were defined based on these criteria and were made pairwise

disjoint. Since the more salient criterion is that of vessel form (i.e. open or closed), this was chosen as the primary criterion for subdividing the `Lamp` class (Bugalhão et al., 2010; Gómez Martínez, 2004). Only the disjoint classes `ClosedLamp` and `OpenLamp` are further subdivided in the asserted hierarchy of OntoAndalus. The full classification can be inferred by a reasoner (Figs. 3 and 4)

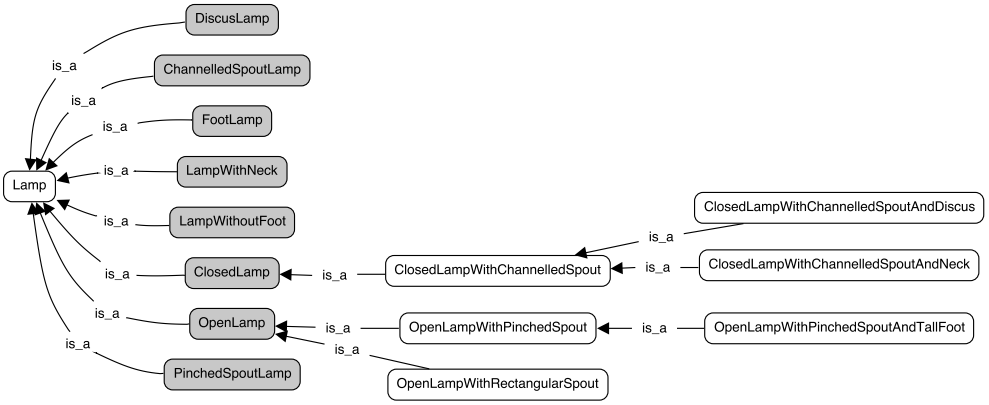

FIG. 3 – *Asserted hierarchy of* `Lamp`*. Gray boxes indicate defined classes.*

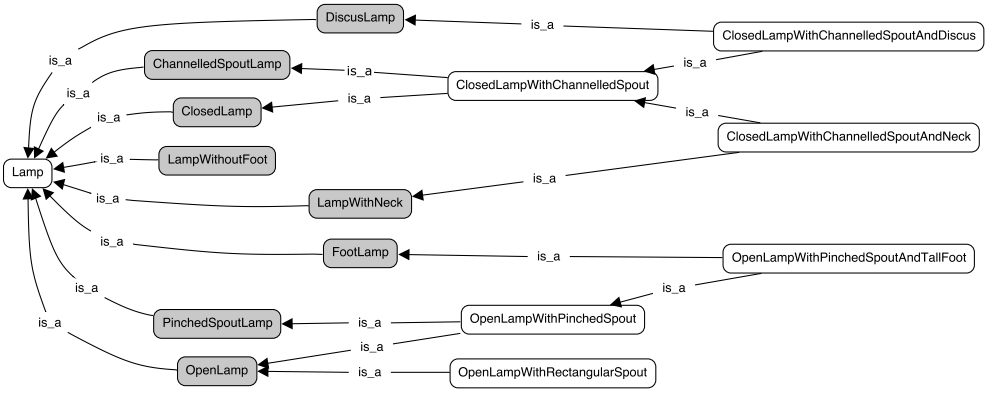

FIG. 4 – *Inferred hierarchy of* `Lamp`*.*

3 Representation of artefact terms

One of the more important steps in our work consisted of representing the extracted Portuguese and Spanish terms for lighting artefacts by means of lexical networks. [2] In terminology work, these networks are considered to be useful for organising and representing language-specific information about terms and other linguistic units (Santos and Costa, 2015).

Figs. 5 and 6 show the more relevant terms for lighting artefact types in Portuguese and Spanish. The lexical relations used in these networks are those of taxonomy (i.e. a specialisation of hyponymy between terms denoting types) and synonymy (Cruse, 1986). Terms arising from the same criterion of subdivision are represented through divided taxonomic arcs, for example *candil* and *candeia* in Fig. 5, which are both motivated by vessel form.

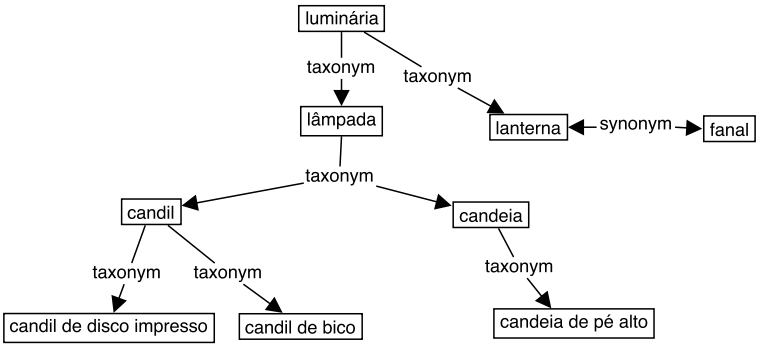

FIG. 5 – *Lighting artefacts terms in Portuguese.*

The question remains of how to relate information about terms with language-independent information about the concepts of the domain in a multilingual resource. A promising model for achieving this is Lemon, which will be briefly described in the following section.

3.1 Integrating linguistic and conceptual information with Lemon

Lemon is an acronym for 'Lexicon Model for Ontologies'. The model has been in development over the past few years (W3C Ontology-Lexica Community Group, 2016), and has been made available through several OWL files. [3]

The purpose of Lemon is to provide linguistic grounding for ontologies, which is often nonexistent or limited to labels represented by means of the `rdfs:label` annotation property. In order to achieve this purpose, Lemon makes use of the following core modules: [4]

— *Ontolex*. For relating a lexicon with an ontology;

— *Synsem*. For representing information at the syntactic and semantic levels;

2. A thorough description of the corpus as well as of the tools and methods used for terminology extraction is provided in a forthcoming article (Almeida et al., 2019).

3. `https://github.com/ontolex/ontolex/`.

4. Recently, Lemon has been extended by means of lexicog, the lexicographic module of Lemon, which works in conjunction with the modules presented here (particularly ontolex) for describing dictionaries and similar resources (Bosque-Gil et al., 2019).

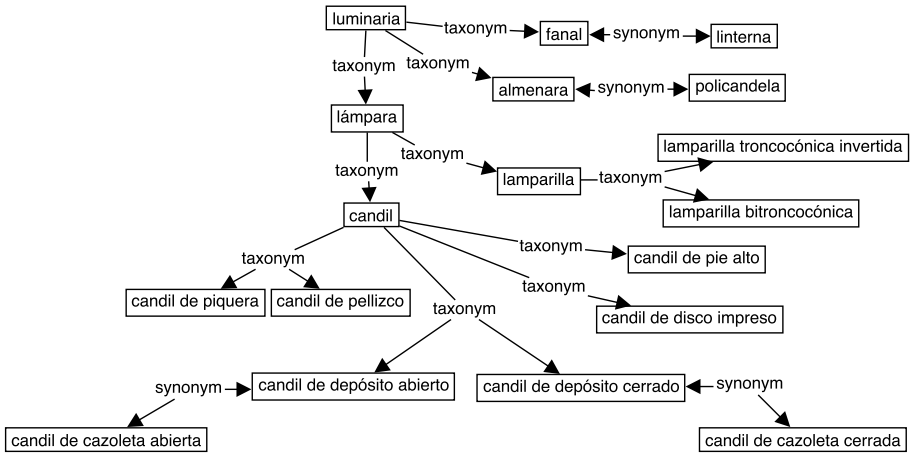

FIG. 6 – *Lighting artefacts terms in Spanish.*

— *Decomp*. For representing information on the decomposition of multiword expressions;
— *Vartrans*. For representing information regarding variation and translation;
— *Lime*. For representing linguistic metadata.

Ontolex is the primary module of Lemon, since it allows to draw relations between lexical entries (i.e. single words, multiword expressions and affixes) and ontology elements (e.g. classes, object properties, individuals). A lexical entry can be realised as a series of forms in a language, including its canonical form – typically the lemma – and other forms (e.g. the plural of nominals). Each form may have several written and/or phonetic representations.

For our present purposes, the relationship between lexical entries and ontology elements can be drawn directly, via the `denotes`/`isDenotedBy` object properties or indirectly through instances of the `LexicalSense` class.[5] The latter can be linked to instances of *LexicalEntry* by means of the `sense`/`isSenseOf` object properties and, subsequently, to ontology elements through the `reference`/`isReferenceOf` object properties. While the direct method is considerably simpler, using the `LexicalSense` class and the previously mentioned object properties allows to model lexicosemantic relations (e.g. taxonomy, synonymy), which may hold between instances of `LexicalSense`.

3.2 Aligning OntoAndalus with Lemon

The possibility of including predicates from Lemon directly in OntoAndalus would allow for a detailed representation of linguistic information, namely with regard to terms in different languages and their lexicosemantic relations. This would involve importing predicates from

5. There is also the possibility of using the `LexicalConcept` class and its associated object properties (i.e. `evokes`/`isEvokedBy` and `concept`/`isConceptOf`), but this is not relevant to our purposes, since OntoAndalus already includes the relevant concepts of our domain of interest.

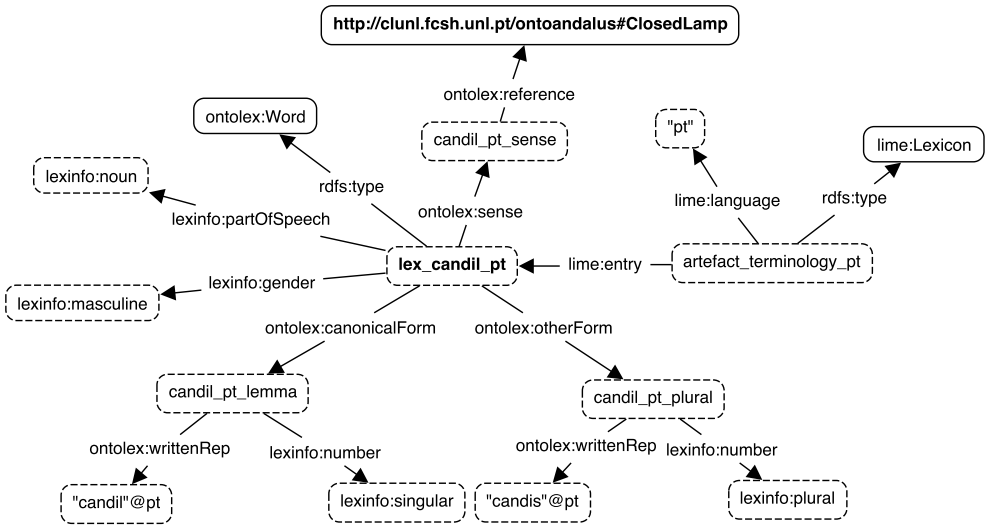

FIG. 7 – *Entry for the Portuguese term* candil.

(at least) the ontolex, lime and vartrans modules, but also from the lexinfo ontology, which provides several categories for Lemon (e.g. the `hyponym` relation between lexical senses). [6]

Fig. 7 shows a possible entry for the Portuguese term *candil* using predicates from the above-mentioned namespaces. In this example, the terminology of artefacts in Portuguese is declared as an instance of `lime:Lexicon`. Grammatical information about the entry includes its part of speech, gender, canonical and plural forms. Finally, semantic information includes reference to the `ClosedLamp` class within OntoAndalus.

With regard to lexicosemantic relations, Fig. 8 shows an instance of the equivalence relation between terms in different languages, as well as an instance of the hyponymy relation. The former is drawn between the Portuguese term *candil* and the Spanish term *candil de depósito cerrado*. This is done by simply pointing the reference of the senses of both terms to the same ontology element, namely the `ClosedLamp` class. Using the relevant object properties in the vartrans module, the hyponymy relation is established between the Portuguese terms *candil* and *candil de disco impresso*. [7] In this example, the sense relation is reified, i.e. represented as an individual, which allows to identify its source (the sense of *candil*), target (the sense of *candil de disco impresso*) and category (hyponymy).

Lemon, therefore, allows to represent diverse information about the terms, including grammatical information (e.g. part of speech, gender) and semantic information (e.g. lexicosemantic relations, reference to ontology elements).

6. `https://lexinfo.net`.

7. Since the more specialised taxonomy relation is not present in lexinfo, hyponymy was used instead in this example.

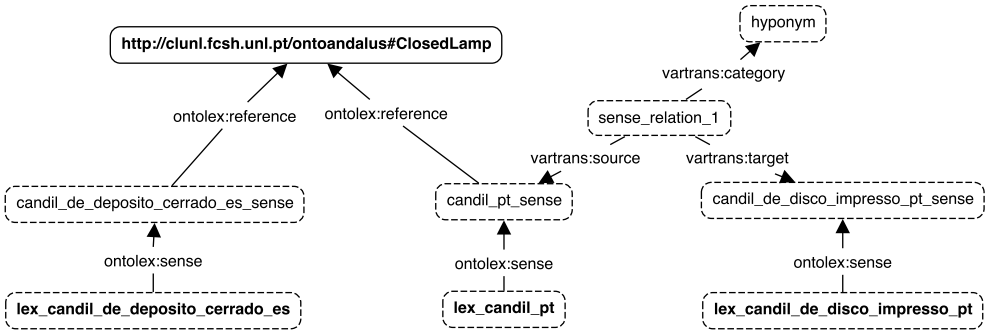

FIG. 8 – *Representing lexicosemantic relations with Lemon.*

4 Conclusion

This paper provided an overview of work carried out for establishing the foundations of a terminological knowledge base in Islamic archaeology. Domain knowledge was represented by means of OntoAndalus, which focusses on relevant artefact types in al-Andalusian pottery studies. OntoAndalus reuses DOLCE+DnS Ultralite, which allows for a detailed representation of domain knowledge, which was exemplified with the case study of lighting artefacts. Information regarding the artefact designations in Portuguese and Spanish was then extracted from a specialised corpus and the Lexicon Model for Ontologies (Lemon) was used in order to represent information about the terms and their relation to the relevant concepts within the domain ontology. Lemon was shown to be capable of representing rich grammatical and semantic information about the terms, and remains a promising model for terminographic applications.

Acknowledgement. Research financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020.

References

Almeida, B. (2019). *Terminology and knowledge representation: ceramic artefacts of al-Andalus*. Doctoral thesis, Universidade NOVA de Lisboa.

Almeida, B. and R. Costa (2019). OntoAndalus: an ontology of Islamic artefacts for terminological purposes. Manuscript submitted for publication.

Almeida, B., R. Costa, and C. Roche (2019). The names of lighting artefacts: extraction and representation of Portuguese and Spanish terms in the archaeology of al-Andalus. *Revue TAL 60*(3), 113–137.

Bosque-Gil, J., J. Gracia, J. P. McCrae, P. Cimiano, S. Stolk, F. Khan, K. Depuydt, J. Does, F. Frontini, and I. Kernerman (2019). The OntoLex Lemon Lexicography Module: final community report. Technical report, Ontology-Lexicon Community Group.

Brin-Henry, F., R. Costa, and S. Desprès (2019). TemPO: towards a conceptualisation of pathology in speech and language therapy. In *Proceedings of the TALN-RECITAL 2019 conference*, Toulouse, pp. 69–80.

Bugalhão, J., H. Catarino, S. Cavaco, J. Covaneiro, I. C. Fernandes, A. Gomes, S. Gómez Martínez, M. J. Gonçalves, M. Grangé, I. Inácio, G. Lopes, and C. Santos (2010). CIGA: projecto de sistematização para a cerâmica islâmica do Gharb al-ândalus. *Xelb 10*, 455–476.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Durán-Muñoz, I. and M. Bautista-Zambrana (2013). Applying ontologies to terminology: advantages and disadvantages. *Hermes 51*, 65–77.

Gangemi, A. (2016). Dolce+D&S Ultralite and its main ontology design patterns. In P. Hitzler, A. Gangemi, A. Janowicz, A. A. Krisnadhi, and V. Presutti (Eds.), *Ontology engineering with ontology design patterns: foundations and applications*, pp. 81–103. Amsterdam: IOS Press.

Gómez Martínez, S. (2004). *La cerámica islámica de Mértola: producción y comercio*. Doctoral thesis, Universidad Complutense de Madrid.

Meyer, I., D. Skuce, L. Bowker, and K. Eck (1992). Towards a new generation of terminological resources: an experiment in building a terminological knowledge base. In *COLING '92 Proceedings of the 14th conference on Computational linguistics*, Volume 3, Stroudsburg, PA, pp. 956–960. Association for Computational Linguistics.

Musen, M. A. (2015). The Protégé project: a look back and a look forward. *AI Matters 1*(4), 4–12.

Rice, P. (1987). *Pottery analysis: a sourcebook*. Chicago, Il: The University of Chicago Press.

Roche, C. (2012). Ontoterminology: how to unify terminology and ontology into a single paradigm. In *LREC 2012*, Madrid, pp. 2626–2630.

Roche, C., R. Costa, S. Carvalho, and B. Almeida (2019). Knowledge-based terminological e-dictionaries: the EndoTerm and al-Andalus pottery projects. *Terminology 25*(2), 262–294.

Rosselló-Bordoy, G. (1978). *Ensayo de sistematización de la cerámica árabe en Mallorca*. Palma de Mallorca: Institut d'Estudis Baleàrics.

Rosselló-Bordoy, G. (1991). *El nombre de las cosas en al-Andalus: una propuesta de terminología cerámica*. Palma de Mallorca: Museo de Mallorca.

Santos, C. and R. Costa (2015). Domain specificity: semasiological and onomasiological knowledge representation. In H. J. Kockaert and F. Steurs (Eds.), *Handbook of terminology*, Volume 1, pp. 153–179. Amsterdam: John Benjamins.

Temmerman, R. and K. Kerremans (2003). Termontography: ontology building and the sociocognitive approach to terminology description. In *Proceedings of CIL17*, Prague. Matfyzpress.

Torres, C., S. Gómez Martínez, and M. B. Ferreira (2003). Os nomes da cerâmica medieval: inventário de termos. In *Actas das 3as Jornadas de Cerâmica Medieval e Pós-Medieval*, Tondela, pp. 125–134. Câmara Municipal de Tondela.

Vallejo Triano, A. and J. Escudero Aranda (1999). Aportaciones para una tipología de la

cerámica común califal de Madinat al-Zahra. *Arqueología y territorio Medieval 6*, 133–176.

W3C Ontology-Lexica Community Group (2016). Lexicon model for ontologies: community report. Technical report, W3C Ontology-Lexica Community Group.

Résumé

Cet article décrit le travail effectué dans le cadre du projet OntaAndalus développé au NOVA CLUNL. Ce projet a pour objectif de jeter les bases d'une base de connaissances terminologiques (BCT) concernant l'archéologie islamique. La partie centrale de cette recherche a été le développement d'une ontologie des types d'artefacts de la poterie de l'al-Andalus qui a été réalisée en réutilisant l'ontologie DOLCE + DnS Ultralite. Par la suite, les termes en portugais et en espagnol désignant les types d'artefacts ont été extraits d'un corpus de textes spécialisés et représentés par des réseaux lexicaux. Le Lexicon Model for Ontologies (Lemon), récemment développé par un groupe communautaire du W3C, est présenté comme un cadre prometteur pour l'intégration d'informations linguistiques et extralinguistiques dans une future BCT du domaine. La possibilité d'aligner OntoAndalus et Lemon est abordée dans cet article.

Thésaurus et terminologies à la source de l'interopérabilité des données archéologiques

Emmanuelle Perrin*, Jean Pierre Girard*
Sébastien Durost**, Marie-Odile Rousset*

*Archéorient
5/7 rue Raulin 69365 Lyon cedex 07
emmanuelle.perrin.touche@gmail.com
http://www.archeorient.mom.fr
**Bibracte EPCC
58370 Gllux-en-Glenne
s.durost@bibracte.fr
http://www.bibracte.fr

Résumé. HyperThésau aborde l'interopérabilité des données archéologiques sous l'angle du vocabulaire, dans une démarche *bottom-up* : conserver toute la richesse et accepter les imprécisions sémantiques des données saisies sur le terrain, mais leur offrir une passerelle : une terminologie alignée sur les référentiels du web sémantique *via* le gestionnaire de thesaurus Opentheso. Les contraintes formelles de construction du thesaurus ouvrent alors la possibilité d'un dialogue effectif machine-machine.

1. L'enjeu : partager des données massivement hétérogènes

Le projet HyperThesau, financé par le Labex IMU de l'Université de Lyon, propose une approche originale du problème de l'hétérogénéité structurelle et sémantique des données archéologiques. Cette approche repose d'une part sur la création d'une architecture informatique – le « lac de données » –, qui préserve la diversité des formats des bases de données et des vocabulaires utilisés par les équipes de recherche (Sawadogo et al. 2019) ; le projet entend en effet, d'autre part, fonder l'interopérabilité des données sur le vocabulaire et prévoit la constitution d'un thésaurus-pivot de l'archéologie.

À la suite des travaux du Centre for Archaeological Information Domain pour la déclinaison de l'ontologie CIDOC-CRM au patrimoine anglais (CRM-EH), plusieurs initiatives et programmes (ARIADNE-ARIADNE Plus en Europe, MASA en France) ont fait porter leurs efforts sur l'accès aux jeux de données archéologiques et la modélisation du processus d'acquisition des données sur le terrain avec CIDOC-CRM (Doerr et al. 2016). Néanmoins, d'autres travaux ont montré la difficulté intrinsèque à construire une méthode de réutilisation des données archéologiques appuyée sur un modèle général, ainsi que l'intérêt d'approches ciblées ou spécifiquement adaptées à un terrain particulier (Lukas et al. 2018). Quelles qu'elles soient, les interfaces de consultation doivent en effet s'adapter au « flou sémantique » des données elles-mêmes.

Un thésaurus est une liste organisée de termes contrôlés. Cet outil documentaire cherche à résoudre le problème de l'équivocité du langage naturel (polysémie et homonymie). Un terme descripteur ou préférentiel doit notamment y décrire de manière univoque un concept. Ce type de vocabulaire sert à l'indexation des ressources et vise à améliorer la recherche

documentaire en augmentant le taux de rappel de documents pertinents au regard d'une requête. Strictement définie par une norme (ISO 254964-1 et 254964-2), la structure d'un thésaurus permet d'exprimer l'ensemble des relations sémantiques d'un concept (relation d'équivalence, relation hiérarchique, relation associative).

Le thésaurus élaboré dans le cadre du projet HyperThesau vise à harmoniser le vocabulaire scientifique et technique de l'archéologie. Il est construit avec le gestionnaire de thesaurus multilingue Opentheso, développé par la plateforme technologique Têtes des Réseaux Documentaires, située à la Maison de l'Orient et de la Méditerranée, qui permet notamment un export en Skos, format standard pour la publication des thésaurus dans le web sémantique. Il se fonde principalement sur le recueil et le traitement de la terminologie employée dans les bases de données, issue des pratiques des archéologues et il doit prendre en compte l'ensemble de la chaîne opératoire des données, du terrain jusqu'à leur publication. Pour désambiguïser les termes, il suppose un important travail de retour sur leurs définitions, attestées par des sources de qualité (Dictionnaire de l'Académie française, Littré, Trésor de la langue française, manuels d'archéologie, publications scientifiques, index et glossaires) qui sont systématiquement citées. En ce qui concerne le mobilier archéologique par exemple, bien souvent la description morphologique prime sur la définition fonctionnelle de l'objet, qui paraît implicite. On constate également un emploi assez vague et parfois désuet de termes généraux (télédétection, parure) parallèlement à un usage de termes techniques très spécialisés issus de publications anciennes (bélière[1], furgeoire[2], barbacane[3]). L'anglais paraît aussi avoir une certaine influence dans la formation du vocabulaire de l'archéologie (forceps/pince).

2. S'aligner sur les référentiels-matières du web sémantique

La voie choisie par HyperThésau consiste à s'attacher à construire un thesaurus-pivot, rigoureusement structuré mais buissonnant, afin de permettre des alignements thématiques, à un niveau très fin, par « rebond » entre les vocabulaires des archéologues et des référentiels pérennes, extérieurs à la communauté disciplinaire mais ouverts sur le web sémantique et dotés des moyens et d'une autorité compatibles avec leur vocation universelle.

Ce thésaurus doit jouer un rôle de pivot pour l'interopérabilité des données archéologiques en ménageant des liens avec les systèmes d'information internationaux et les grands référentiels publiés dans le web de données par la communauté des bibliothèques (IdRef, data.bnf.fr, *Library of Congress Subject Heading*). Il est conçu comme un outil de médiation entre des vocabulaires « locaux » ou idiolectes et des vocabulaires documentaires plus généraux. Il s'agit notamment de faire coexister et communiquer les concepts d'usage

[1] Anneau qui maintient le battant d'une cloche ; par analogie, anneau de suspension d'une lampe d'église, d'une montre, d'une boucle d'oreille, d'un fourreau de sabre (CNRTL).
[2] Nom donné à divers outils de toilette (cure-dents, cure-oreille et autres) réunis en manière
[2] Nom donné à divers outils de toilette (cure-dents, cure-oreille et autres) réunis en manière de trousse, ils se suspendaient parfois à la ceinture (V. Gay, Glossaire archéologique du Moyen-Age et de la Renaissance, Paris, Librairie de la société bibliographique, 1887).
[3] Barbacane ou porte d'agrafe : anneau destiné à recevoir une agrafe (C. Enlart, Manuel d'archéologie française depuis les temps mérovingiens jusqu'à la Renaissance. III. Le costume, Paris, Picard, 1916).

courant en archéologie avec le langage d'indexation matière de la BnF (Rameau : Répertoire d'autorité-matière encyclopédique et alphabétique unifié).

Destiné à l'interaction machine-machine, ce thésaurus obéit à des contraintes formelles et logiques qui sont très différentes de l'interprétation scientifique. L'application stricte de la relation hiérarchique comme relation genre-espèce conduit notamment à « découper » un objet scientifique en plusieurs concepts qui vont entretenir entre eux des relations d'associations. Selon le principe du langage postcoordonné, l'indexation d'un bracelet fera appel à plusieurs descripteurs : matériau, morphologie, fonction, décor, technique de fabrication, période, lieux. De même, la description de l'acquisition des données associe plusieurs branches ou sous-domaines du thésaurus : méthode, champ d'observation, instrumentation et documents produits. À terme, l'alignement sur les vedettes-matière de Rameau et de la *Library of Congress* permet l'interconnexion avec d'autres jeux de données par les vocabulaires et ainsi d'enrichir les données et de produire de nouvelles connaissances.

Références

Doerr M., Theodoridou M., Aspöck E., Masur A., « Mapping archaeological databases to CIDOC-CRM » in CAA2015 Keep the revolution going, Proceedings of 43rd Annual Conference of Computer Applications and Quantitative Methods in Archaeology, Archaeopress Archaeology, 2016, 443-451.

Lukas D., Engel C., Mazzucato C., « Towards a Living Archive: Making Multi Layered Research Data and Knowledge Generation Transparent », Journal of Field Archaeology, vol. 43, 2018 (doi.org/10.1080/00934690.2018.1516110).

Rabinowitz A., « It's about time: historical periodization and Linked Ancient World Data » in ISAW Papers : Current Practice in Linked Open Data for the Ancient World, 7/22, 2014 (dlib.nyu.edu/awdl/isaw/isaw-papers/7/rabinowitz/).

Sawadogo P.N., Kibata T., Darmont J., « Metadata Management for Textual Documents in Data Lakes », 21st International Conference on Enterprise Information Systems (ICEIS 2019), Heraklion, 2019, 72-83; INSTICC, Setúbal, Portugal (Vol. 1).

Summary

HyperThésau addresses the interoperability of archaeological data from a vocabulary perspective, in a bottom-up way; it aims to preserve all the richness of the raw ground data and to accept their semantic inaccuracies, but offers a gateway: a terminology aligned with semantic web repositories via Opentheso, a thesaurus manager. The formal constraints of thesaurus construction then open the possibility of an effective machine-machine dialogue.

Evaluer la crédibilité des sources historiques

Jacky Akoka*, Isabelle Comyn-Wattiau**
Cédric du Mouza*

*CNAM, lab. CEDRIC, Paris, France
{akoka,dumouza}@cnam.fr
**ESSEC Business School
wattiau@essec.edu

Résumé. La recherche en histoire s'appuie principalement sur l'étude des sources d'information historique. Les résultats de cette recherche dépendent largement de la qualité des sources d'information. L'objectif de cet article est de décrire les premiers éléments d'une approche d'évaluation automatique de la crédibilité des sources d'information historique numérisées. Fondée sur une approche des sciences de conception (design science), notre contribution comporte un modèle conceptuel décrivant les caractéristiques principales des sources d'information historique et une démarche algorithmique d'estimation de la crédibilité fondée sur ce modèle. La suite de cette recherche consistera en l'application de cette approche à la recherche prosopographique médiévale.

1. Introduction

Les historiens fondent leurs travaux sur la base de sources d'information. Ces dernières sont de deux types : les sources primaires sont les reliques ou les témoignages recueillis dans des manuscrits alors que les sources secondaires sont des écrits ou des études analysant ces informations primaires. La qualité de leurs analyses dépend directement de la qualité de ces sources d'information historique. Cette qualité recouvre de nombreux aspects, notamment la crédibilité. La crédibilité d'une source d'information historique se définit comme un degré de confiance accordée par les historiens en la capacité de cette source à fournir une information fiable. A l'heure où les historiens disposent de plus en plus de sources numérisées, la mesure automatique de cette crédibilité est un enjeu majeur.

La crédibilité d'une information est un concept dont la définition n'est pas standardisée. Le terme lui-même possède plusieurs synonymes et concepts voisins, notamment la réputation, l'objectivité et la vérifiabilité. Dans cet article, nous contribuons à affiner ces définitions dans le contexte des informations historiques. Nous proposons un modèle conceptuel regroupant les éléments principaux décrivant les sources d'information et les aspects pouvant contribuer à l'évaluation de leur crédibilité. Ce modèle est utilisé comme fondement pour définir une approche automatique d'évaluation de la crédibilité.

Le reste de l'article est organisé comme suit. La section 2 présente un état de l'art qui synthétise les travaux antérieurs relatifs à l'évaluation de la crédibilité des sources d'information. La section 3 décrit notre modèle conceptuel. Notre approche d'estimation de la crédibilité fait l'objet de la section 4. Enfin, la section 5 conclut l'article et propose des pistes de recherche future.

2. Etat de l'art

Nous distinguons dans cet état de l'art les méthodes utilisées par les chercheurs pour évaluer la crédibilité des sources d'information puis celles utilisées dans la recherche en histoire.

2.1 Evaluation de la crédibilité des sources d'information

Le processus d'évaluation de la crédibilité des sources d'information est un sujet qui a connu un développement important depuis l'avènement d'Internet et des réseaux sociaux (Metzger et al., 2010), ouvrant la voie à la conceptualisation de ce processus. Des outils pratiques sont proposés pour faciliter l'évaluation des sources. CARS est un exemple de *checklist* très référencé (Harris, 2018). CARS consiste en l'évaluation de la crédibilité (*Credibility*), l'exactitude (*Accuracy*), la raisonnabilité (*Reasonableness*) et le renfort (*Support*). D'autres *checklists* existent, par exemple la liste des cinq questions (Radom *et al.*, 2014).

La crédibilité est l'un des nombreux facteurs de qualité souvent mentionnés dans les recherches sur la qualité de l'information (Nurse *et a*l., 2011). Un cadre conceptuel a été proposé pour les données liées et ouvertes (Zaveri *et al*., 2016). Il recense toutes les terminologies utilisées par les différents cadres conceptuels et études précédentes. Plusieurs termes anglais très voisins sont définis et comparés. La crédibilité (en anglais *credibility*) d'une source y est définie comme le jugement d'un utilisateur sur l'intégrité de la source. Il est synonyme de réputation (en anglais *reputation*) Y sont aussi définis les termes de vérifiabilité, confiance, objectivité et crédibilité (en anglais *believability*).

2.2 Crédibilité des sources d'information historique

A notre connaissance, il n'existe pas de méthode systématique d'évaluation de la crédibilité des sources historiques. En revanche, on trouve beaucoup de guides méthodologiques à l'usage des chercheurs ou des étudiants en histoire. Ces guides abordent, entre autres, le sujet de la qualité des sources historiques et fournissent des directives utilisables dans le processus d'évaluation. Par exemple, l'université du Kentucky propose une démarche d'évaluation des sources historiques qui diffère selon que ces dernières sont primaires ou secondaires (UKY, 2019). L'évaluation d'une source primaire (manuscrit, lettre, journal intime, mémo, autobiographie, etc.) s'appuie sur l'auteur, l'audience, la logique, le cadre de référence, le lien avec d'autres sources, etc. L'évaluation d'une source secondaire (livre historique, article de recherche historique, etc.) est fondée sur l'analyse de la structure, de la thèse défendue, de l'argumentation, etc. Le terme le plus utilisé est celui de fiabilité (en anglais *reliability* ou *trustworthiness*). Là encore, on trouve des aides à l'évaluation de la crédibilité comme OPCAM pour Origin, Perspective, Context, Audience et Motive ou TOMACRU pour Type, Origin, Motive, Audience, Content, Reliability et Usefulness (Vest, 2007).

Comme on peut le constater, il n'y a pas de méthode structurée pour faciliter l'évaluation systématique de la crédibilité des sources d'information. A notre connaissance, il n'existe pas d'effort de conceptualisation des dimensions essentielles décrivant la crédibilité accordée aux

sources historiques. C'est précisément l'objet de notre article de structurer les éléments permettant de définir ce concept dans un modèle qui servira de base à l'évaluation automatique des sources historiques numériques et d'illustrer comment celui-ci peut servir de base à un algorithme calculant automatiquement la crédibilité des sources d'information numérisées.

3. Modéliser la crédibilité

A notre connaissance, il n'existe pas, dans la recherche en histoire, une typologie standardisée des concepts définissant la fiabilité et/ou la crédibilité des sources d'information disponibles. Nous proposons, dans cet article, de reprendre le cadre de (Zaveri *et al*., 2016) mentionné plus haut. Il permet de clarifier ces concepts tous regroupés dans la catégorie *Trust* (Figure 1). La crédibilité (en anglais *believability*) y est définie comme le degré auquel une information est considérée comme correcte, vraie, réelle et crédible. La réputation (en anglais *reputation* ou *credibility*) est définie comme un jugement porté par quelqu'un quant à l'intégrité d'une source d'information. Les deux concepts sont dans l'intersection entre la catégorie *Trust* et la catégorie *Contextual*. Ils sont reliés dans la mesure où la réputation d'une source impacte sa crédibilité.

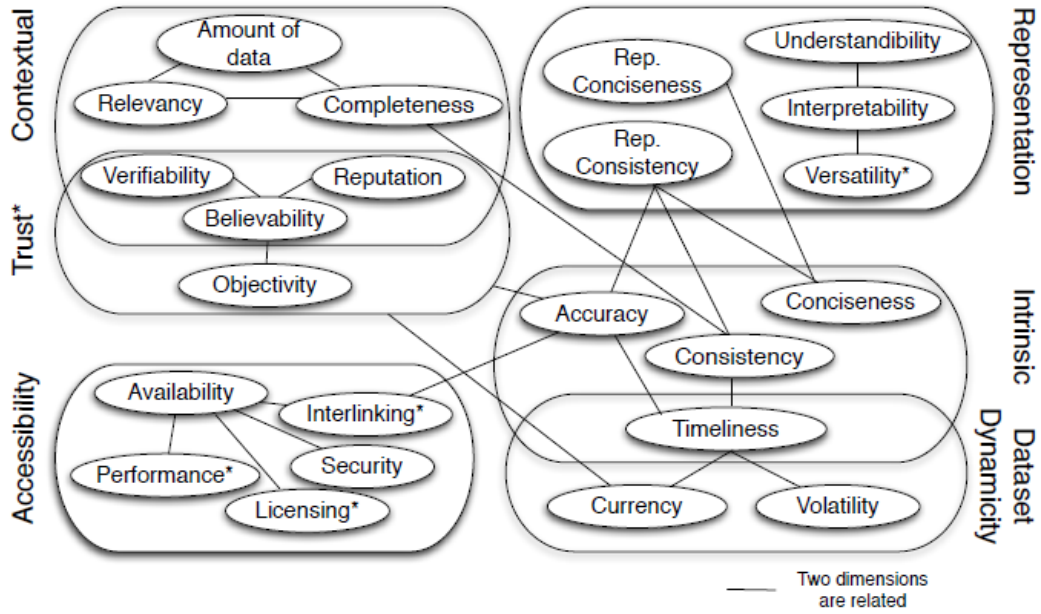

FIG 1. *Les dimensions de la qualité du Linked Open Data Quality et les relations entre elles (Zaveri et al., 2016)*

Ce sont précisément ces deux concepts que l'on va tenter d'évaluer automatiquement en identifiant les paramètres pertinents. Cette évaluation ne porte pas sur les informations elles-

mêmes mais sur les sources qui les contiennent. La typologie la plus fréquente des sources historiques comporte deux catégories : primaires et secondaires auxquelles s'ajoutent plus rarement les sources tertiaires.

3.1 Les sources historiques

Les sources historiques sont divisées généralement en trois catégories suivant le support sur lequel elles reposent : sources primaires, secondaires ou tertiaires. Suivant leur nature, les critères d'évaluation de la crédibilité d'une source peuvent différer.

3.1.1 Les sources primaires

Une **source primaire** est un document original donnant des informations directes sur le sujet de la recherche. Ce document date le plus souvent de l'époque où le ou les événements rapportés ont eu lieu (mais pas obligatoirement cependant). Les sources primaires sont des lettres, des textes, des images, des rapports, des registres, etc.

La crédibilité d'une source primaire dépend de plusieurs facteurs. Avant tout, elle dépend de l'auteur de cette source : est-il connu comme fiable et précis dans sa production ? Cela est souvent lié à sa place dans la société (rang, classe, fonction, …), la date de production de la source par rapport à la date de l'événement relaté, la motivation pour l'auteur à produire cette source, ses convictions personnelles, etc. Cela dépend aussi de la nature du fait décrit : s'agit-il de faits rapportés ou de faits observés/constatés par l'auteur ? La crédibilité de la source peut être attestée directement par des experts du domaine, ou parce qu'elle est citée dans d'autres sources considérées comme sérieuses.

3.1.2 Les sources secondaires

Les **sources secondaires** désignent des documents qui utilisent des sources primaires, et souvent la consultation d'autres sources secondaires, dont ils constituent une analyse, une synthèse, une explication ou une évaluation. Les biographies ou les ouvrages de recherche en Histoire sont des exemples de sources secondaires.

Ici aussi la crédibilité de la source dépend de son auteur : est-il reconnu comme fiable par les experts du domaine ? Quelle est la thèse qu'il cherche à défendre et reste-t-il objectif ? Quelle est la crédibilité des sources qu'il cite ? Cette source secondaire est-elle citée souvent, par des experts reconnus ?

3.1.3 Les sources tertiaires

Une **source tertiaire** est une sélection et une compilation de sources primaires et secondaires. Les bibliographies, les catalogues de bibliothèques, les répertoires, les listes de lectures conseillées et les articles proposant des tours d'horizon sont des exemples de sources tertiaires.

La crédibilité des sources tertiaires s'évalue de manière semblable aux sources secondaires.

D'autres typologies existent (Howell *et al.*, 2001) : parmi les sources primaires, on distingue les reliques des témoignages oraux ou écrits. Ces derniers peuvent être manuscrits ou publiés. Pour évaluer leur crédibilité, on tient compte de l'intention et du contexte, qui s'évalue par rapport au moment où la source est créée. (Kipping *et al.*,2014) opposent la validité et la crédibilité. La première s'évalue à l'aide d'informations externes alors que la seconde s'appuie sur une analyse interne. La triangulation est une technique permettant de réduire le biais afin d'accroître la robustesse des résultats de recherche. L'herméneutique situe la source dans son contexte historique et dans sa relation aux autres textes.

C'est sur ces typologies qu'est fondé le modèle conceptuel proposé ci-après.

3.2 Modèle conceptuel de la crédibilité des sources historiques

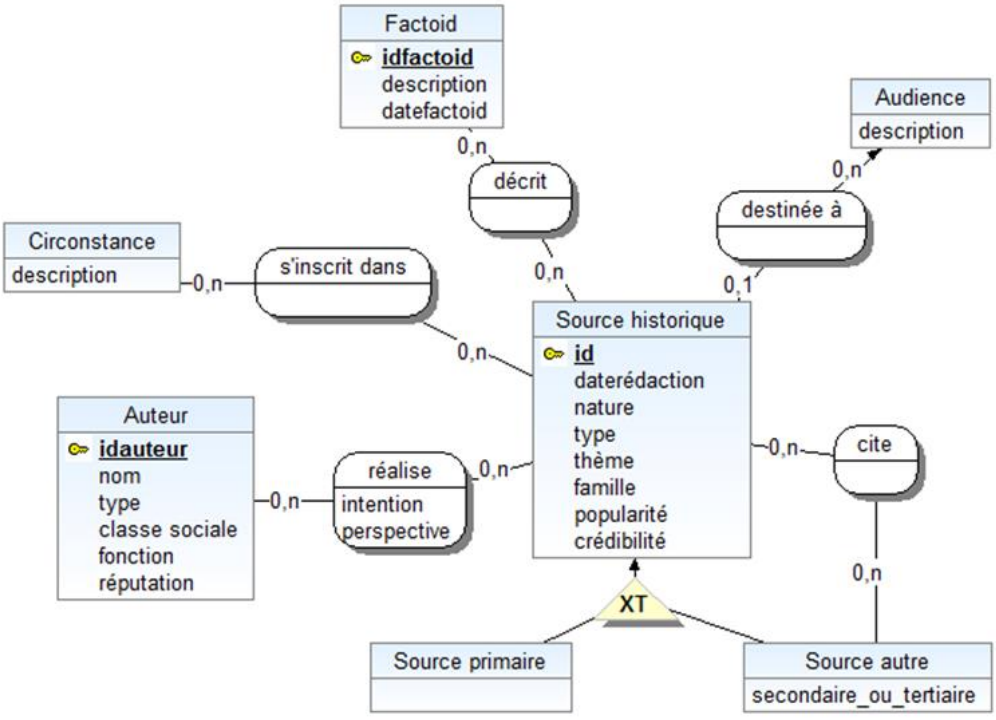

FIG 2. *Modèle conceptuel des sources historiques*

L'entité principale du modèle est la source dont on souhaite évaluer la crédibilité. Cette source est caractérisée par une référence, une date de rédaction, une nature (manuscrit, publication), un type (lettre, article de journal académique, article de presse, film, etc.), un thème, une famille. Une source peut être primaire ou autre. Une source autre peut être secondaire ou tertiaire. Elle cite des sources qui peuvent être primaires ou non. La popularité d'une source est liée au nombre de fois où elle est citée par d'autres sources. La crédibilité d'une source résulte de l'agrégation des différents jugements émis par des historiens. Une source est utilisée par un historien parce qu'elle mentionne des *factoïds* caractérisés par une véracité. Une source peut être réalisée par plusieurs auteurs ou créateurs. On distingue différents types d'auteurs notamment les témoins oculaires, les chercheurs, les écrivains, les journalistes, etc. Un auteur occupe une fonction, appartient à une classe sociale et a une réputation.

La source est liée à un contexte dans lequel elle a été réalisée. Au-delà de la date, le contexte est décrit par les circonstances dans lesquelles elle naît. Enfin, elle est réalisée dans une certaine intention (informer, convaincre, contredire, etc.) pour une audience (privée, familiale, publique, etc.) avec une perspective traduisant le point de vue de l'auteur lors de la réalisation de la source. Une source décrit des *factoïds* que les historiens étudient.

Le modèle présenté ci-dessus permet d'offrir un cadre à l'évaluation de la crédibilité. Il regroupe à la fois les concepts utilisés pour l'analyse des sources historiques (intention, perspective, audience, contexte, auteur ou origine) et les mesures de qualité de l'information (crédibilité, réputation) qui peuvent en être déduites.

4. Estimer la crédibilité

4.1 Notations

Nous introduisons tout d'abord les notations utilisées dans la suite de cette section. Nous considérons un ensemble de sources $\mathbf{S} = \{s_1, s_2, \ldots, s_N\}$ et un ensemble d'auteurs $\mathbf{A} = \{a_1, a_2, \ldots, a_M\}$. Nous distinguerons dans **S** l'ensemble des sources primaires ($\mathbf{S^{prim}}$), secondaires ($\mathbf{S^{sec}}$) et tertiaires ($\mathbf{S^{ter}}$). Nous adoptons dans la suite les notations suivantes :

Notation (paternité) : la paternité est le fait pour un auteur d'avoir réalisé une source (primaire, secondaire ou tertiaire). Nous supposons l'existence d'une fonction, pour la représenter, notée dans la suite $pater : A \times S \rightarrow \{0,1\}$, 1 signifiant que l'auteur a réalisé la source considérée.

Notation (citation) : une citation apparait dans une source secondaire ou tertiaire uniquement et peut faire référence à une source primaire, secondaire ou tertiaire. Nous supposons l'existence d'une fonction, pour la représenter, notée dans la suite $cite : (S^{prim} \cup S^{sec}) \times S \rightarrow \{0,1\}$.

L'objectif de notre approche est de proposer un modèle qui permet d'estimer :

- La crédibilité d'une source, $cred : S \rightarrow [0,1]$.

- La réputation d'un auteur, *repu : A → [0,1].*

4.2 Evaluation des sources : Approche PageRank

Pour estimer le score d'une source, nous proposons de nous appuyer sur le modèle de citations de PageRank. Ainsi plus une source est citée par d'autres sources très citées, plus cette source a un score de crédibilité élevé. Cette approche a déjà été proposée dans le contexte des documents à l'aide de graphes de citations comme dans (Yan et Ding, 2010) et (Dunaiski, 2014).

Nous proposons d'estimer le score de crédibilité comme suit :

Définition (score de crédibilité) : *considérons une source s ∈ **S**. Son score de crédibilité est :*

$$cred(s) = (1-d) \times \frac{\sum_{1 \le i \le N} cite\ (s_i, s)}{\sum_{1 \le i,j \le N} cite(s_i, s_j)} + d \times \sum_{1 \le i \le N} cite\ (s_i, s) \times \frac{cred(s_i) \times context(s_i)}{\sum_{1 \le j \le N} cite(s_i, s_j)}$$

Le facteur d'amortissement *d* dans notre contexte représente la probabilité qu'un auteur décide de passer à une autre « famille » de sources non liée avec la source précédente, parce qu'il considère un nouveau thème ou simplement souhaite changer de « famille de sources » (par exemple passer des manuscrits de l'université aux registres d'un évêché). Cela permet aussi de traduire le fait que lorsqu'on aboutit à une source primaire, qui n'a donc pas de citation (ou rarement) d'autres sources, on passe à une autre source non-liée à la précédente, sur la base de sa « popularité ». $\sum_{1 \le i \le N} cite\ (s_i, s)$ désigne le nombre de sources qui citent s, donc intuitivement une source déjà très citée par d'autres sources a une plus grande probabilité d'être citée par quelqu'un changeant de thème ou de famille de sources, d'où le facteur $(1-d) \times \frac{\sum_{1 \le i \le N} cite\ (s_i, s)}{\sum_{1 \le j,j \le N} cite(s_i, s_j)}$.

Cette formulation montre que le score de crédibilité d'une source se calcule en fonction du score de crédibilité des sources qui la citent ainsi que du « contexte » de la source, noté *context(s)*. Le score de contexte couvre de nombreuses dimensions : *temporelle* (la date de rédaction est un élément de contexte dans la mesure où les faits sont plus précis et vérifiés à l'époque contemporaine et où elle est plus ou moins éloignée de la date des factoïds rapportés, etc.), la *classe sociale* ou *fonction* de l'auteur, les *circonstances (*événements tels que guerre, épidémie et croyances*)* de l'époque, etc. Il est donc particulièrement complexe à évaluer. Son évaluation peut être automatisée si une base de connaissances et de règles est constituée avec les historiens.

Modélisons le problème sous une forme matricielle. Considérons un nombre N de sources,

- le vecteur unitaire $U = (U_i)_{1 \le i \le N}$ avec $u_i = 1, \forall i \in [1, N]$, et
- la matrice d'adjacence A = $(a_{ij})_{1 \le i,j \le N}$ avec $a_{ij} = cite(s_i, s_j)$, et
- le vecteur de popularité P = $(P_i)_{1 \le i \le N}$ avec $p_i = \frac{\sum_{1 \le k \le N} cite\ (s_k, s_i)}{\sum_{1 \le k,j \le N} cite(s_k, s_j)}, \forall i \in [1, N]$, et

Evaluer la crédibilité des sources historiques

- le vecteur de contexte C = $(\boldsymbol{C_i})_{\mathbf{1 \le i \le N}}$ avec $c_i = context(s_i), \forall i \in [1, N]$

Le calcul de la crédibilité des sources revient à trouver le vecteur X = $(x_i)_{1 \le i \le N}$ tel que

$$X = (1-d) \times P.U^T + d \times A.C \times X$$

Ce problème se résout de manière itérative en posant :

$$\begin{cases} X^0 = (p_i)_{1 \le i \le N} \quad \text{où } p_i \text{ est la popularité telle que vu ci-dessus} \\ X^{k+1} = (1-d) \times P.U^T + d \times A.C \times X^k \end{cases}$$

Le vecteur de crédibilité X s'obtient à la convergence qui est garantie par le théorème de Perron-Frobenius. Cette approche permet donc d'automatiser le calcul du score de crédibilité sous réserve que les règles de calcul du contexte soient établies par les experts. A noter qu'il existe des propositions pour appliquer PageRank en distribué comme (Das Sarma et al., 2015), ce qui permet un passage à l'échelle d'une part, mais aussi d'appliquer cette approche dans un domaine où les sources sont souvent réparties sur de nombreux sites.

4.3 Évaluation de la réputation d'un auteur

Pour estimer la réputation d'un auteur, nous proposons de nous appuyer sur la crédibilité des sources qu'il a réalisées et d'en agréger les scores. Plusieurs approches peuvent être envisagées :

1) La moyenne de la crédibilité des différentes sources produites par un auteur : $repu(A) = \frac{\sum_{s \in S} pater(A,s) \times cred(s)}{\sum_{s \in S} pater(A,s)}$ pour un auteur A. L'avantage de cette méthode est sa facilité à mettre en place une fois que le calcul des crédibilités est effectué. Cependant il atténue l'importance d'une source considérée comme très crédible si l'auteur a produit d'autres sources moins crédibles à côté. La variance et l'écart-type peuvent compléter l'estimation en relevant l'uniformité ou non dans la qualité de sa production.

2) Une combinaison des scores permettant de donner un poids plus important pour les sources importantes. Cela peut être une combinaison linéaire avec des poids qui sont fixés par des experts du domaine sur chaque source. Le problème de cette solution est qu'elle implique l'expert et donc n'est pas entièrement automatisable.

3) Une moyenne quadratique des crédibilités des sources produites : : $repu(A) = \sqrt{\frac{\sum_{s \in S} pater(A,s) \times (cred(s))^2}{\sum_{s \in S} pater(A,s)}}$ pour un auteur A. L'avantage de cette méthode est qu'elle n'implique pas l'expert. Elle est donc automatisable, et permet de renforcer l'importance d'une source avec un score élevé de crédibilité.

D'autres approches sont proposées par exemple dans (Moreira *et al*., 2015) et pourraient être appliquées. Des agrégations plus pertinentes peuvent être réalisées en analysant la nature des sources pour, par exemple, les regrouper par époque et/ou par sujet. En effet un historien peut ainsi être considéré comme un expert sur les universitaires du XIIIème siècle de l'université de Paris et moins sur les officiers poitevins au XVème siècle. Par conséquent la prise en compte, lors de l'agrégation des scores, des thèmes et époques sur lesquels portent les sources produites permettrait d'établir les domaines d'expertise des différents auteurs.

5. Conclusion

Dans cet article, nous avons proposé un modèle conceptuel regroupant les éléments principaux décrivant les sources d'information et les aspects pouvant contribuer à l'évaluation de leur crédibilité. Sur la base de ce modèle nous avons présenté un algorithme de PageRank étendu afin de pouvoir estimer de manière automatique la crédibilité d'une source et nous avons proposé plusieurs stratégies d'agrégation des scores des sources afin d'estimer la réputation d'un auteur.

Nous souhaitons désormais travailler sur une (semi-)automatisation du calcul de score de contexte historique en nous appuyant sur des experts historiens pour établir les différentes règles à considérer. Par ailleurs, nous validerons expérimentalement cette approche à l'aide de ces experts.

Références

Das Sarma, Atish et al. "Fast Distributed PageRank Computation." Theoretical Computer Science 561 (2015): 113–121.

Dunaiski, Marcel. (2014). Analysing ranking algorithms and publication trends on scholarly citation networks.

Harris, R. Evaluating Internet Research Sources Version Date: October 11, 2018 www.virtualsalt.com

Howell, M.C. Prevenier W. (2001) From reliable sources: An introduction to historical methods, Cornell University Press.

Kipping M., Wadhwani R.D., Bucheli M. (2014) Analyzing and interpreting historical sources: a basic methodology, in: Organization in time: History, theory, methods, pp. 305-329.

Moreira C., Calado P., and Martins B.. 2015. Learning to rank academic experts in the DBLP dataset. *Expert Sys: J. Knowl. Eng.* 32, 4 (August 2015), 477-493.

Miriam J. Metzger, Andrew J. Flanagin, Ryan B. Medders, Social and Heuristic Approaches to Credibility Evaluation Online, Journal of Communication, Volume 60, Issue 3, September 2010, pp 413–439

Nurse, J. R. C., Rahman, S. S., Creese, S., Goldsmith, M., & Lamberts, K. (2011). Information quality and trustworthiness: a topical state−of−the−art review (pp. 492–500). IEEE ICCANS2011.

Rachel Radom and Rachel W. Gammons, “Teaching Information Evaluation with the Five Ws: An Elementary Method, an Instructional Scaffold, and the Effect on Student Recall and Application,” Reference & User Services Quarterly 53, 4 (2014): 334-47.

Singh A. P., Shubhankar K. and Pudi V., "An efficient algorithm for ranking research papers based on citation network," *2011 3rd Conference on Data Mining and Optimization (DMO)*, Putrajaya, 2011, pp. 88-95.

UKY 2019, Evaluating Historical Sources, www.uky.edu/~dolph/HIS316/handouts/sources.html

Vest Kathleen. Using Primary Sources in the Classroom. Upper Saddle River: Shell Education 2007

Yan, Erjia & Ding, Ying. (2010). Discovering author impact: A PageRank perspective. Information Processing & Management. 47. 125-134.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. Semantic Web, 7(1), 63-93.

Summary

Research in History relies on the study of sources of historical information. The results of this research depend largely on the quality of the sources of information. The purpose of this article is to describe the first elements of an approach to automatically assess the credibility of digitized historical information sources. Based on a design science approach, our contribution includes a conceptual model describing the main characteristics of historical information sources and an algorithmic approach to estimating credibility based on this model. The next step of this research will be the application of this approach to medieval prosopographic research.

Cartographies des formations en Humanités numériques en France

Orélie Desfriches Doria*, Elie Allouche**

*Adresse postale complète
orelie.desfriches-doria@univ-paris8.fr
http://www.une-page.html
**Autre adresse
elie.allouche@education.gouv.fr

Résumé. Dans cet article, nous donnons à voir le développement des formations en humanités numériques, à travers la mise en cartographies des contenus et discours associés à ces formations. Dans une première partie nous esquissons brièvement une caractérisation des formations en HN en France. La deuxième partie présente des cartographies des formations en HN en France par type de formation, par discipline, par objets et par techniques, d'après des données issues de la Dariah. Enfin une dernière cartographie des Licences d'origine des étudiants, issue de « Trouver mon master » est proposée. Pour chaque cartographie, les données sont présentées, puis les cartographies analysées. Ce travail vise à donner une vue globale des formations à travers leur description et les métadonnées associées à ces entités, dans des bases de données qui recensent des formations.

1. Introduction

Les humanités numériques, aussi bien dans la recherche que dans l'enseignement, s'affirment à la fois comme un champ d'étude et comme une communauté de praticiens qui s'organise autour d'une approche interdisciplinaire des questionnements et méthodes scientifiques en lettres-arts-sciences humaines et sociales, en s'appuyant sur les bouleversements épistémologiques apportés par l'utilisation des technologies numériques.

Une partie du travail présenté ici vise à donner à voir la répartition non pas géographique des formations en Humanités Numériques (H.N.) en France, cartographie géographique déjà accessible à partir de la base de données Dariah, mais bien d'avoir une vue globale des formations à travers leur description et les métadonnées associées à ces entités, dans des bases de données qui recensent des formations.

Dans cet article, nous donnons à voir le développement des formations en humanités numériques, à travers la mise en cartographies des contenus et discours associés à ces formations. Dans une première partie nous esquissons brièvement une caractérisation des formations en HN en France. La deuxième partie présente des cartographies des formations en HN en France par type de formation, par discipline, par objets et par techniques, d'après des données issues de la Dariah. Enfin une dernière cartographie des Licences d'origine des étudiants, issue de « Trouver mon master » est proposée. Pour chaque cartographie, les données sont présentées, puis les cartographies analysées.

2. Présentation et caractéristiques globales des formations en H.N.

Au cœur de ces formations se posent notamment les questions de la connexion entre l'évolution des méthodes scientifiques, la construction des savoirs, l'évolution des pratiques didactiques et l'évolution de la formation professionnelle. Elles constituent une hybridation entre objets d'étude traditionnels, tradition de formation des institutions/unités de recherche et pratiques des technologies numériques.

Ces formations traduisent ainsi un effort d'articulation entre compétences techniques, compétences disciplinaires et démarche projet dans une optique le plus souvent interdisciplinaire. Cette articulation a été présentée par exemple en ces termes lors du THATCamp 2012 :

> *"La tentative d'identification du digital humanist passe par une identification des compétences de base, qui seraient notamment la maîtrise des langages HTML, TEI et Javascript."*
>
> *"Mais le répertoire de compétences du digital humanist dépasse le cadre de ces compétences techniques. Il est donc, de fait, difficile de trouver une base commune à l'ensemble des compétences des digital humanists."*[1]

Dans le même sens, nous pourrons constater dans la partie suivante la diversité des technologies mentionnées dans la description des formations en humanités numériques, en France.

En effet, l'offre de formation en humanités numériques est d'abord fondée sur le rapport ancien de la science et de la technique et des conditions techniques de la construction des savoirs (Olivier Le Deuff, 2012).

Dans l'environnement numérique contemporain, les compétences mobilisées relèvent à la fois d'un effort de distanciation et de compréhension des mécanismes à l'œuvre nécessitant des compétences techniques (recueil et traitement des données, maîtrise d'outils, de normes), des compétences en culture numérique (littératie/translittératie), en design et des compétences organisationnelles (veille, documentation, travail collaboratif, gestion de projet, etc.).

L'analyse de ces formations confirme que les humanités numériques reposent à la fois sur des méthodes et compétences disciplinaires et transdisciplinaires (techniques, informationnelles, médiatiques), explorant des territoires déjà balisés ou de nouveaux territoires mis au jour par la mise en données numériques des savoirs, mais également sur des valeurs de partage et de diffusion des compétences et des connaissances (rappelées notamment par le Manifeste de 2010), et enfin, sur l'organisation d'une communauté de praticiens à dimensions nationale et internationale.

---

[1] Source : BERRA, Aurélien et LE DEUFF, Olivier, 2012. THATCamp Paris 2012 - Quelles compétences et littératies pour les humanités numériques ? - Éditions de la Maison des sciences de l'homme. In : THATCamp Paris 2012 : Non-actes de la non-conférence des humanités numériques [en ligne]. Éditions de la Maison des sciences de l'homme. Paris : Éditions de la Maison des sciences de l'homme. [Consulté le 29 mai 2019]. ISBN 978-2-7351-1527-3. Disponible à l'adresse : https://books.openedition.org/editionsmsh/334

3. Cartographie des formations en humanités numériques

Au cours de la présentation de l'analyse des visualisations, qui sont présentées ci-après, il faut garder à l'esprit que les données sont partielles et incomplètes, elles correspondent à un état à un instant T des données présentes dans les bases prise comme sources. D'autre part, certaines formations changent d'intitulé et de programmes au fur et à mesure des évolutions des disciplines, mais également des technologies numériques, et ce, au fil du déploiement international des conceptions des H.N.

Les questions dans ce cadre concernent l'opportunité ou la possibilité de produire et décrire les formations du champ, à l'aide de référentiels de description des formations et en particulier dans le champ émergent des Humanités Numériques. Ainsi, nombreuses sont les définitions co-existantes des H.N., elles-mêmes proposées par des enseignants chercheurs qui sont aussi les porteurs des formations étudiées ici, qui embarquent donc des contenus et descriptions qui correspondent à leur vision des H.N. De plus la description de ces formations dans les bases étudiées relève de processus de communication à l'intention des futurs étudiants notamment. La question du niveau de complexité et de détail adopter pour quels publics et selon quelles stratégies nous semble cruciale. Par ailleurs, dans les descriptions des formations dans les bases utilisées comme source, les pratiques de description des formations sont hétérogènes. Ainsi, les formations repérées dans ces sources sont renseignées avec plus ou moins de détails (nombre de mots clés associés qui varient fortement par exemple), et avec des niveaux conceptuels eux aussi variables. Les stratégies d'indexation des formations (non explicites) nous semblent donc relever alternativement d'une volonté d'exhaustivité (avec un grand nombre de termes utilisés pour décrire une formation pour certains items) ou d'une stratégie d'indexation à visée prototypique (quelques mots-clés censés être représentatifs pour donner un aperçu).

3.1 Quel périmètre pour répertorier les formations en HN ?

Nous recensons ici les formations en humanités numériques qui s'affichent comme telles dans leur intitulé et qui sont répertoriées sur les portails présentés ci-dessous, ainsi que les sites des universités et unités de recherche. Pour établir une cartographie des formations en HN, au moins deux outils systématiques peuvent être utilisés :

- Dariah Course Registry[2] : recensement et cartographie géographique contributifs des formations proposée par l'infrastructure européenne de recherche numérique pour les arts et les sciences humaines (DARIAH). Dans cette base, la recherche peut s'opérer par les filtres suivants : par pays, par ville, par institution, par niveau. L'entrée dans les contenus de formation s'effectue quant à elle selon trois entrées : disciplines, techniques, objets.

Cette base *"vise à aider les étudiants, les chercheurs, les enseignants et les institutions (de DARIAH et au-delà) pour trouver, promouvoir et se connecter à des activités d'enseignement et de formation en humanités numériques. […]* [Elle] *"peut être utilisé librement et sans inscription par les étudiants, les enseignants et les personnes intéressées. Pour ajouter vos propres cours à la base de données, il vous suffit de vous inscrire sur le site*

[2] [2] https://www.dariah.eu/tools-services/tools-and-services/tools/digital-humanities-course-registry/

Web. Une fois approuvé par le modérateur national de votre pays respectif, le cours est visible sur notre carte interactive."

Au 20/05/19 la base recense 209 formations en H.N., dont 28 pour la France[3].

Pour la communauté francophone, le travail de recensement et de modération est assurée par le groupe de travail formations de l'association Humanistica[4] (association francophone des humanités numériques) qui *"propose de formuler des recommandations concernant les formations des étudiants et des jeunes chercheurs en humanités numériques."*

- Trouver mon master[5] : site officiel du Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation qui présente l'offre de formation nationale au niveau master.

L'accès aux formations s'effectue par recherche libre ou selon une série de critères (mention, parcours/spécialité, établissement, localisation, périmètre géographique).

Au 20/05/19 cette base recense 26 masters en humanités numériques, en résultat à la simple requête "humanités numériques".

	Dariah Course Registry	Trouver mon Master
Nombre de résultats bruts en France	28	26
Nombre d'établissements différents	18	23
Nombre de Masters	13	23
Nombre de Parcours	/	57

TAB 1 – *Comparatif quantitatif des données récoltées dans les deux sources*

La base de données Dariah recense des formations de Master, de Licence, et aussi des Cours, à l'international, ici les chiffres sont tirés d'une extraction de la BDD, basé sur l'unique filtre "France" appliqué au champ Pays.

Dans la base de données Dariah, l'entité de référence dans la structuration de la base est une "formation" (de quelque nature qu'elle soit dans la nomenclature). Certains établissements apparaissent plusieurs fois comme ils ont renseigné plusieurs formations dans cette base.

Le nombre de formations de Master dans cette extraction est de 13 Masters, auxquels il faut ajouter 2 Masters intitulés "recherche" qui n'entrent pas dans la même catégorie sur ce site, à cause de la dimension internationale de cette source. Les parcours différenciés et entrant sous le même intitulé de Master sont difficilement détectables et parfois non mentionnés. Ainsi, le Master "Humanités Numériques" de l'Université Paris 8 n'apparait qu'une seule fois dans la base, bien qu'il compte 5 parcours différents en réalité. Les 15

[3] https://registries.clarin-dariah.eu/courses/statistic
[4] http://www.humanisti.ca/groupes-de-travail/
[5] https://www.trouvermonmaster.gouv.fr/

autres formations se distribuent entre “Licence” et “Course” dans la nomenclature de la Dariah.

« Trouver mon Master » ne recense que les formations de Master en France. L’entité de référence dans cette base est l’établissement d’enseignement, et dans chaque occurrence d’une fiche on trouve les différents parcours proposés sous le même intitulé de Master avec des descriptions qui varient. Dans ce cadre, le passage de 26 résultats à 23 établissements s’explique par des doublons ou des fiches vides. Il y a donc en France 23 Établissements d’enseignement supérieur qui proposent des Masters en Humanités Numériques, repérables dans cette BDD avec la requête “humanités numériques”. Chacune de ces formations compte un ou plusieurs parcours et 57 parcours différents sont proposés en France actuellement.

Il est important de préciser les éléments présentés ci-dessus puisque ces données sont hétérogènes et structurées de manière différente, par conséquent, difficilement comparables. De plus, ces données restent incomplètes : nous avons connaissance d’un Master dispensé à l’INA sur le patrimoine audiovisuel et numérique[6] qui n’apparaît dans aucune de ces bases de données, et qui n’est certainement pas un cas isolé.

L’outil de visualisation utilisé pour mener ce travail est un outil de cartographie dite sémantique (Desfriches Doria et Lavenir, 2017), il s’agit de GraphCommons[7] qui est un outil en ligne, libre, qui permet de faire des cartographies sous forme de graphes, à petite échelle (1000 noeuds max). Ses principaux avantages sont d’être gratuit, facile d’utilisation (interactions avec les graphes et les données via la carte ou des tableaux; facilité d’import des données), et de pouvoir embarquer des modèles de connaissances au sens de l’ingénierie des modèles, à la manière des ontologies.

3.2 Statistiques et visualisations de la base de données Dariah

Pour la réalisation de ces visualisations les données ont été extraites via l’API proposée sur la base de données Dariah et téléchargées, puis nettoyées et reformatées afin de les rendre importables dans GraphCommons.

Etant donné la quantité et qualité des données qui sont présentées et analysées dans cet article, nous tenons à préciser que la nature des analyses qui ont été menées restent volontairement à un niveau descriptif par souci de rigueur scientifique, afin de ne pas extrapoler des interprétations sur ces données imprécises, parcellaires, et qui posent de nombreuses questions procédurales, quant à leur production (pour la base de données Dariah notamment).

3.2.1 Cartographie des formations par type de formation

Cette première cartographie[8], présentée en figure 1, établie d’après les données de la base de données Dariah donne une vision globale des formations recensées dans cette base, par type de formation (Master/ Master Recherche/ Bachelor Programme/ Course).

[6] https://www.ina-expert.com/masters-audiovisuels-et-diplomes-de-deuxieme-cycle/diplome-ina-gestion-de-patrimoines-audiovisuels.html

[7] https://graphcommons.com/

[8] Consultable en ligne : https://graphcommons.com/graphs/9372006f-1640-47d5-b23d-1456d39d330f

On peut observer une surreprésentation des formations en master (en bleu) dans cette cartographie globale. D'après les données de cette base, c'est donc le type de formation majoritairement proposé en H.N. en France. Cela révèle une vision des DH qui correspond à une formation en double cursus, un cursus initial axé sur une des disciplines des humanités classiques et une formation en HN qui complète la première.

Figure 1. Cartographie globale des formations en H.N en France, par type de formations

3.2.1 Cartographie des formations par discipline

Cette cartographie[9] (en figure 2) présente une visualisation dont les données sont extraites de la base de données Dariah, (les 28 formations recensées en France par cette base). Cette visualisation donne une vue globale axée sur la description des formations selon les disciplines (en rouge dans la figure 2) mentionnées pour chaque formation.

La cartographie est spatialisée avec l'algorithme "Force Atlas 2" qui vise à rendre le graphe le plus lisible possible en optimisant l'espace afin qu'il y ait le moins possible de croisements entre les liens. Les nœuds les plus visibles (les plus grands) sont mis en avant à travers un algorithme qui donne un poids relatif en fonction du nombre de liens entrants vers un nœud par rapport au plus petit nombre de liens entrants vers un nœud.

[9] Consultable en ligne : https://graphcommons.com/graphs/525db979-df48-4d66-b370-49f7ca28f6a9

Figure 2. Cartographie des formations en H.N en France, par disciplines

3.2.2 Analyse des données de la cartographie par discipline

Il existe de fortes disparités dans ces données et elles suscitent des questionnements :

- Les formations sont-elles décrites avec des mots clés qui décrivent les disciplines dont les formations en H.N. se réclament, ou ces termes décrivent-ils des disciplines connexes, ce qui n'implique pas la même interprétation.
- Quelle est cette nomenclature de disciplines ? Par qui est-elle proposée ?
- Ces descriptions à base de mots-clés, qui décrivent les disciplines dans lesquelles les formations s'inscriraient, peuvent-elles être équivalentes aux données issues de "Trouver Mon Master" concernant les licences d'origine des étudiants pour entrer dans les formations ?
- Lorsqu'on télécharge ces données, à l'état "brut", chaque ligne correspond à une formation, et chaque discipline est classée dans une colonne ordonnée. Cela signifie-t-il que les disciplines sont ordonnées par ordre d'importance lors de leur saisie ? Ceci pose la question de l'aspect contributif de la base de données Dariah, et de l'encadrement des pratiques d'indexation et de description des formations.
- Ces 28 formations sont décrites avec une fourchette en termes de nombre de disciplines mentionnées qui va de 1 mot-clé à 8 mots-clés pour décrire les disciplines. Or, dans le détail, 7 formations ne présentent qu'un mot clé, quand 3 formations en présentent 8. Les formations qui ne sont décrites qu'à l'aide d'un seul mot clé, sont exclusivement de type "course" ou "bachelor

programme". Cela pose la question de l'hétérogénéité des données en fonction du type de formation, et donc de la représentativité de la quantité des citations de mots clés.
- Les Masters font aussi l'objet de disparités dans la description, leur nombre de mots-clés varie de 2 à 8.
- Enfin, au niveau de la généricité des mots-clés employés, que signifie l'utilisation d'un mot-clé "Social Science" alors que d'autres mots-clés plus précis existent dans les données comme "Histoire", "Archéologie", "Philosophie" par exemple. Cette nomenclature est proposée pour décrire les disciplines au niveau international, mais peut-être existe-t-il des subtilités dans la considération de ces dernières, dans leur positionnement et même dans leur intitulé dans des contextes nationaux.

3.2.3 Analyse de la cartographie par disciplines

Dans la figure 2, les disciplines les plus citées pour décrire les formations du corpus ressortent visuellement.

On observe donc d'après cette cartographie, la hiérarchisation en 5 niveaux dans la fréquence de citation des disciplines pour décrire les formations en HN en France, ces rangs sont calculés en fonction du nombre de liens entrants vers les noeuds :

- en premier rang (les noeuds les plus visibles et les plus grands) les disciplines les plus citées pour décrire les formations sont: "Literary and Philological studies", "Linguistic and Language Studies", Theory and Methodology of DH" et "History" ;
- en deuxième rang, les disciplines suivantes apparaissent : "Arts and Cultural studies", et "Computer sciences ;
- en troisième rang, on observe les disciplines suivantes : "Media and Communication Studies", "Human Language" et "Archeology".
- en quatrième rang apparaissent : "Social sciences" et "Library and Information Science" ;
- et en cinquième rang, on trouve "Musicology" et "Philosophy".

On peut donc constater que les SIC ("Library and Information Science" et "Media and Communication Studies") apparaissent en retrait, puisqu'elles se situent en quatrième et cinquième rang, dans ces descriptions conceptuelles des formations à l'aide d'intitulés disciplinaires. Cette sous-représentation des SIC apparente peut être aussi résonner comme un appel à contributions pour compléter les données dans la base Dariah.

Au-delà de ces observations, on remarque que les disciplines mentionnées sont nombreuses. Cette nébuleuse qui se révèle étendue renvoie aussi aux questions des fondations méthodologiques communes et des pratiques convergentes à travers cette diversité disciplinaire au sein des H.N. (Desfriches Doria et al., 2018).

3.2.4 Cartographie par objet

La cartographie des objets des formations, présentée en figure 3[10], a été élaborée avec l'outil GraphCommons, et la spatialisation a été réalisée selon les mêmes paramètres que la carte précédente.

[10] Consultable en ligne : https://graphcommons.com/graphs/4bc0e2c1-58c0-470b-af78-1c45207abea1

3.2.5 Analyse des données de la cartographie par objets

La description des formations sur la base du vocabulaire des objets des formations est manifestement inégale sur la quantité de mots-clés renseignés, générant donc une disparité sans doute trompeuse de ces objets dans la visualisation.

La nomenclature même de ces objets pose question car certains objets désignent aussi des technologies. La généricité de certains termes pose question : en effet, l'utilisation des termes "data" ou "files" ou "tools" ne peut pas être spécifique aux H.N., nombreuses sont les disciplines ou les formations qui pourraient être décrites avec le même vocabulaire.

Par ailleurs, on retrouve ici des questionnements liés aux usages des langages documentaires : la description des formations à l'aide des termes utilisés est-elle vue comme un système de descripteurs où chaque terme pris séparément fait sens précisément pour chaque item décrit, ou, est-ce l'adjonction conjointement exhaustive des termes qui permet la description complète des items ? Ceci pose la question de la pertinence du traitement mené ici : le découpage en fonction de la fréquence des citations des mots-clés pour réaliser la cartographie. Le sens des termes utilisés, s'il se définit en référence à l'univers sémantique par proximité avec les autres termes utilisés pour un item, est perdu à l'occasion de ce découpage, par la suppression, via ce procédé, de l'environnement sémantique créé par les autres mots-clés utilisés pour décrire un item.

Figure 3. Cartographie des formations en H.N en France, par objets

3.2.6 Analyse de la cartographie par objets

On observe d'après cette cartographie, la hiérarchisation selon 3 niveaux principaux dans la fréquence de citation des objets pour décrire les formations en HN en France :

- le premier niveau porte sur les objets suivants : "Text", "DH", "Data",
- le deuxième niveau comprend : "Literature" et "Language" ;
- le troisième niveau plus fréquent de citation des objets inclue : "metadata", "multimedia", "image", "manuscript", "file" et "tools".

On peut remarquer la généricité des termes, et leur absence de précision. Une des dominantes observées porte ainsi sur le travail sur les données, métadonnées, les fichiers, le texte et plus globalement tout ce qui touche au traitement de corpus, sans que ce dernier terme (corpus) soit inclus dans les objets cités.

Les humanités numériques apparaissent elles-mêmes comme objet d'étude, confirmant ainsi le côté réflexif des pratiques développées et la dynamique interdisciplinaire. Ce lien entre interdisciplinarité et dynamique réflexive a été étudié par certains articles récents par exemple (Bénel, 2014). La notion de "Projet" fondamentale en H.N. (Desfriches Doria et al., 2018), est sous-représentée. Cet aspect renvoie à la question des pédagogies mobilisées dans les formations en H.N. En effet, dans les H.N., les projets en eux-mêmes déterminent en grande partie les compétences à mobiliser et à transmettre dans le cadre d'une formation, et non exclusivement l'origine disciplinaire des acteurs ou des institutions. Les enseignements par projet en H.N. sont, selon nous, profitables, notamment dans l'apprentissage du travail en collaboration interdisciplinaire, interdisciplinarité qui est une dimension forte dans l'identité des H.N., comme nous l'avons montré dans (Desfriches Doria et al., 2018).

3.2.7 Cartographie des techniques

La cartographie des techniques[11], présentée en figure 4, a été élaborée avec l'outil GraphCommons, et la spatialisation a été réalisée selon les mêmes paramètres que les cartes précédentes.

3.2.8 Analyse des données de la cartographie par techniques

On observe le même type de disparités que précédemment, le nombre de termes pour décrire une formation du point de vue des techniques allant de un seul terme à dix-huit termes.

3.2.9 Analyse de la cartographie par techniques

On observe une structuration en plusieurs niveaux, comme dans les cartographies précédentes, pour la dimension « techniques » de description des formations, dans la cartographie présentée en figure 4 :

- Le premier niveau est constitué d'un seul terme central qui ressort (le nœud le plus visible) : « Encoding » (cité 16 fois)

[11] Consultable en ligne : https://graphcommons.com/graphs/51ef1758-d8d2-4dbe-917b-885f59583b82

- Le deuxième niveau comprend les termes : « Searching » et « Text Mining » (cités 10 fois) d'une part et « Information retrieval » et « Linked open data » (cités 9 fois)
- Le troisième niveau inclut les termes : « Concordancing », « Preservation metadata » et « POS-Tagging » et « Named entity recognition » (cités 7 fois), et aussi « Machine learning » et « Topic modelling » (cités 6 fois)
- Le quatrième niveau concerne les termes : « Principal component analysis », « Cluster analysis », « commenting » et « Web crawling » (cités 5 fois)
- Le reste des nœuds se distribue comme suit : « Open archival information systems », « Collocation analysis » (4 citations) ;
- « Mapping », « Pattern recognition » et « Gamification » (3 citations) ;
- « Scanning », « Sentiment analysis », « Georeferencing », « Photography », « Brainstorming », « Distance measures », « Debugging » et « Sequence alignement » (2 citations) ;
- et enfin « Migration » et « Browsing » (1 citation).

Figure 4. Cartographie des formations en H.N en France, par techniques

On constate une forte hétérogénéité des sortes de techniques inclues dans les termes pour décrire les formations concernant cette dimension « Techniques ».

Ainsi, certains termes comme le « Linked open data » ou « Brainstorming » relèvent-t-ils bien d'une technique ? La variabilité dans la spécificité des termes pose elle aussi question :

est-il possible de mettre au même niveau des techniques de « machine learning » et d' « information retrieval » ? Que signifie l'emploi d'un terme « Photography » dans le champ « Techniques » ? Ce terme ne serait-il pas plus approprié pour décrire les « Objets » des formations ? Enfin, un terme « Named entity recognition » décrit une technique qui peut être plus généralement inclue dans les techniques de « Text-mining ».

Cela pose à nouveau la question de la qualité des données utilisées pour réaliser ces visualisations et de la variabilité dans la précision des termes due aux aspects contributifs de l'indexation dans la base de données Dariah. Ces problématiques de cohérence inter-indexeurs est une problématique classique en Sciences de l'Information-Documentation.

Plus génériquement, une autre limite du travail présenté tient à l'identification des compétences visées à l'issue de ces formations, qui ne se réduisent pas nécessairement à des objets ou des techniques.

3.3 Traitement et analyse des données de "Trouver mon master"

Les données qui servent de base à la réalisation de cette cartographie proviennent de "Trouver Mon Master". Dans cette base, chaque établissement d'enseignement peut renseigner des détails sur chaque parcours de formation et notamment les licences d'origine des étudiants, considérés comme valides pour entrer dans ces parcours.

Cela apporte un éclairage sur le positionnement du parcours en termes de programme de formation et de positionnement dans le champ des H.N.

3.3.1 Cartographie des Licences d'origine

Les données ont été récoltées, et reformatées pour leur insertion dans l'outil qui a servi à la réalisation de cette carte : GraphCommons. Le paramétrage a été réalisé selon les mêmes principes que les cartes précédentes.

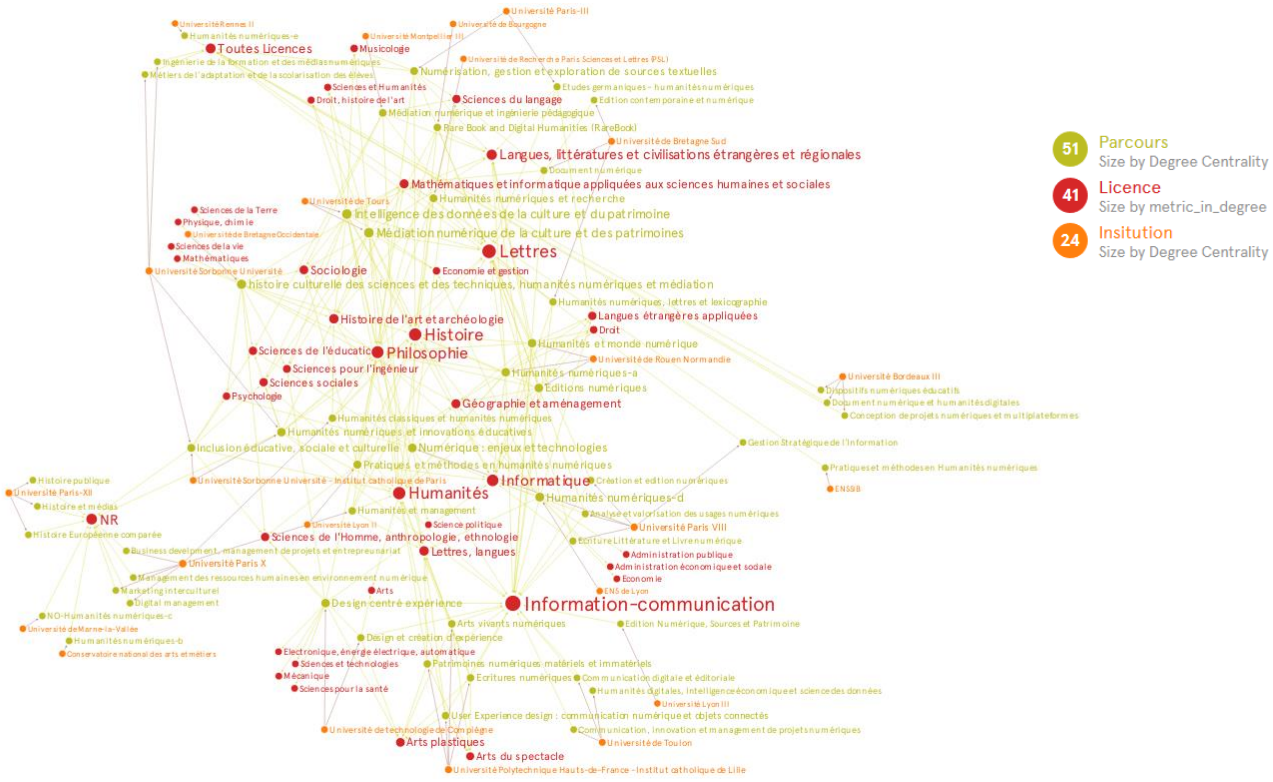

Figure 5. Cartographie des licences d'origine des étudiants admissibles en Master H.N. en France

3.3.2 Analyse de la cartographie des Licences d'origine

A travers cette cartographie présentée en figure 5, on remarque que l'Infocom est en premier rang dans les formations dont peuvent être issus les étudiants en HN en France. Le deuxième rang de cette cartographie est composé des Lettres, de l'Histoire, de la Philosophie, et des Humanités. En troisième rang arrive l'Informatique, et en quatrième, les Langues (sous divers intitulés) et des regroupements de filières de type "Lettres, Langues" ou "LEA", et en cinquième rang arrivent les Arts.

Il nous apparaît ici intéressant d'opérer des rapprochements entre la cartographie des disciplines (Figure 2) présentée plus haut, et celle-ci. En effet, si la cartographie des disciplines présente les disciplines dont se réclament les HN, ou ses disciplines connexes, celle des parcours d'origine des étudiants décrit les formations qui constituent le vivier des futurs praticiens des HN, qui feront évoluer les pratiques en HN au fil des promotions. On constate des disparités notoires. En effet, la place de l'Infocom est en tête concernant les filières d'origine pour l'entrée dans les masters HN, alors que dans la cartographie des disciplines l'Infocom apparaît nettement en retrait. Il pourrait être judicieux d'investiguer davantage cet aspect : est-ce l'appétence pour le numérique des étudiants de cette filière qui en fait un vivier favorable pour l'entrée dans des Master HN, ou bien la nature interdisciplinaire de cette section académique, ou encore l'intérêt pour l'écrit, le texte, la philologie, liée à la branche documentation de l'infocom ?

On observe également que les Arts se situent en 5ème rang dans la cartographie des filières tandis qu'elle est en deuxième rang dans la cartographie des disciplines. La Philosophie passe du dernier rang dans la cartographie des disciplines au deuxième rang dans la cartographie des filières d'origine des étudiants. Cela pose aussi question, ces positions reflètent-elles un effet d'aubaine pour l'affichage de débouchés porteurs pour les étudiants de filières dont on sait qu'elles sont difficiles en termes d'insertion professionnelle en France ?

Ces questionnements ne doivent cependant pas occulter les disparités dans les données rapprochées ici, et une requalification des disciplines mentionnées pour les filières d'origine des étudiants dans la nomenclature des disciplines de la base de données Dariah (ou l'inverse) permettrait d'affiner encore l'analyse.

4. Conclusion

Le travail présenté ici fait émerger des questions : Comment (si c'est pertinent) travailler à recenser des formations dans un domaine hétérogène tel que celui des H.N., et comment proposer des langages documentaires de description de ces formations, qui fassent consensus dans ce domaine émergent, et en recomposition constante dans les contextes locaux de mise en oeuvre et de projets ?

Quelles sont les entrées pertinentes pour décrire ces formations : concernant les H.N. spécifiquement, est-il souhaitable d'opter pour une approche par compétences ? En effet, si pour les outils numériques et les technologies, ces compétences sont repérables, bien qu'elles peuvent déjà prêter à débat, concernant les dimensions "humanités classiques", elles apparaissent plus complexes à typologiser.

Enfin pour parfaire ce travail de cartographie du champ des H.N. à travers les formations, il nous semble qu'un prolongement de ce travail est nécessaire, sur le recensement des contenus de formations, et leur décryptage à l'aide d'entretiens avec les responsables de ces

formations, afin de sortir des aspects liés aux discours sur les formations, et d'entrer dans une véritable possibilité de recenser la diversité des approches, des contenus de formations, et des visions des H.N. qui leur sont corrélées. Cette poursuite du travail sous forme d'enquête qualitative pourrait être complétée d'un travail maïeutique sur les trajectoires professionnelles des enseignants responsables de ces formations et sur leur profil académique.

Dans cet article nous avons proposé des visualisations de données à propos de ces formations. Ce travail de spatialisation relève lui-même des humanités numériques, confirmant une fois de plus la dimension réflexive de ce champs de recherches, hétérogène, mais riche de sa diversité. Il est important de souligner que les données utilisées correspondent à un état des descriptions des formations au moment de la collecte des données utilisées. Ces données sont, par ailleurs, issues de recherches dans des bases, qui ont été formulées avec des termes simples, elles sont donc potentiellement non-exhaustives. Ces données peuvent également évoluer dans le temps, puisqu'elles sont renseignées de manière contributive et déclarative. Il reste intéressant de constater les difficultés liées à l'hétérogénéité des descriptions et des pratiques d'indexation des objets étudiés, les formations en H.N., qui, par définition, sont vouées à évoluer.

Concernant la description des formations sur la base de données de la Dariah, un travail comparatif des données à l'international permettrait d'observer l'existence hypothétique de caractéristiques nationales dans l'appréhension de ce champ émergent des H.N.

Références

BENEL, A. (2014). Quelle interdisciplinarité pour les «humanités numériques»?. Les Cahiers du numérique, 10(4), 103-132.

BERRA, Aurélien et LE DEUFF, Olivier, 2012. THATCamp Paris 2012 - Quelles compétences et littératies pour les humanités numériques ? - Éditions de la Maison des sciences de l'homme. In : THATCamp Paris 2012 : Non-actes de la non-conférence des humanités numériques [en ligne]. Éditions de la Maison des sciences de l'homme. Paris : Éditions de la Maison des sciences de l'homme. [Consulté le 29 mai 2019]. ISBN 978-2-7351-1527-3. Disponible à l'adresse : https://books.openedition.org/editionsmsh/334

DESFRICHES-DORIA, O., SERGENT, H., TRAN, F., HAETTICH, Y., & BOREL, J. (2018, October). What is Digital Humanities' identity in interdisciplinary practices?: An experiment with digital tools for visualizing the francophone DH network. In Proceedings of the 2nd International Conference on Web Studies (pp. 39-47). ACM.

DESFRICHES DORIA, O. ; LAVENIR, C. (2017), « Document et Annotation : le cas des cartographies d'informations », dans M. Hassoun, K. Zreik, O. Larouk, G. Besacier (Dir), Actes du 20e Colloque International sur le Document Électronique - CiDE.20 : "Le document ?", 23-25 novembre 2017, ENSSIB, Lyon, France, p. 297-308.

Summary

Merci de ne pas changer le titre de cette section qui doit contenir la traduction en anglais du résumé présenté sur la première page.

Utilisation de la 3D pour des médiations scientifiques et culturelles multiplateformes

Éric Desjardin*, Hervé Deleau*, Stéphanie Prévost*

*Université de Reims Champagne-Ardenne, CReSTIC, Reims, France
{eric.desjardin,herve.deleau,stephanie.prevost}@univ-reims.fr

Résumé. Lors de l'analyse d'une œuvre d'art, l'interaction physique, ou tout au moins visuelle, peut être un apport important à sa compréhension. De nombreuses situations se présentent alors pour lesquelles cette approche est rendue difficile (exposition en salle d'accès restreint…) voire impossible (mobilité réduite, objet détruit…). Alors que d'une manière générale il est légitime de se questionner sur la nécessité de l'utilisation d'une acquisition 3D, elle est dans ces cas un apport intéressant qui ouvre de plus de nombreuses autres perspectives. Se pose alors la question d'une faisabilité à faible coût tant de l'acquisition que de la modélisation, de restitutions visuelles enrichies et du travail collaboratif autour de l'œuvre. En nous appuyant sur le travail d'un groupe de recherche autour du Tombeau de Jovin, exposé au musée Saint-Rémi de Reims, nous présentons un ensemble de problématiques qui apparaissent tout au long de ce processus et de l'ouverture vers des publics divers.

1. Introduction

Ce premier travail s'inscrit dans une recherche interdisciplinaire sur le Tombeau de Jovin exposé au Musée Saint-Rémi de Reims, où notre objectif final est de pouvoir construire d'une manière collaborative plusieurs calques 3D thématiques pour une visualisation/virtualisation enrichie. Les informations traitées pourront être les résultats d'analyse sur certaines parties de marbre, des informations sur les personnages de la scène issues de sources textuelles, des ajouts (sous forme d'hypothèses) de parties manquantes, les fractures, etc. L'objectif sera aussi de pouvoir fournir ces couches en réalité augmentée tant pour les chercheurs en histoire et archéologie que lors de médiations scientifiques vers le grand public.

2. Acquisition du modèle par photogrammétrie

L'acquisition de modèle d'objet peut être menée en utilisant de nombreuses approches techniques : laser, scanner à lumière structurée, photogrammétrie, etc. Le choix à opérer est dépendant de plusieurs facteurs liés à l'ouvrage : dimension, nature physique, mobilité, mais environnement, et des nécessités techniques pour les utilisations : résolutions, précision, coût d'acquisition, compétences, matériels, etc. Notre approche était d'évaluer la faisabilité d'acquisition et de traitement d'un objet de fort volume dont le contenu était suffisamment riche pour pouvoir proposer plusieurs types d'analyse. Avec l'évolution fortement favorables des coûts des matériels de prise de vue et la réapparition fulgurante d'outils de traitement multi-images pour reconstruire des modèles en 3 dimensions, le choix que vous avons réalisé à porter sur l'expérimentation d'un traitement par photogrammétrie. Pour cela, une session de photographie a été organisée en dehors des horaires d'accès au public. Une série a été

réalisée avec un smartphone ordinaire, une deuxième avec un reflex sur pied à plusieurs hauteurs et selon 3 orientations constituant au total un ensemble d'une centaine de clichés en lumière naturelle des 3 faces du tombeau (quelques exemples sont présentés en Figure 1).

3. Construction du modèle

Actuellement, la construction du modèle 3D a été réalisée en utilisant la suite logiciel Recap d'Autodesk. Elle permet de sélectionner les images source, de fabriquer le nuage de points 3D de l'objet avec son environnement et de proposer son texturage à partir des photos.

4. Supports de médiation multiplateformes

4.1 Boîte à effet hologramme

Une mise en œuvre simple, très peu onéreuse et efficace auprès du grand public est d'utiliser une simple boite à effet hologramme constituée d'un caisson noir, d'un écran, d'un verre plexiglas à 90° sur lequel se reflète l'affichage de écran (voir Figure 2). La visualisation est réalisée par une application permettant une mise en mouvement de l'objet afin que l'effet perceptuel de flottement prenne effet. Une souris sans fil apporte l'interaction avec les caractéristiques de l'objet (zoom, déplacement latéral, rotation) favorisant un échange riche et interactif avec le public que nous avons pu expérimenter lors d'événements tout public comme la Fête de la Science.

4.2 Salle d'immersion

La plateforme « Centre image » de l'Université de Reims Champagne-Ardenne dispose d'un mur d'immersion constitué d'un écran 4x2 mètres en stéréo-active multi-utilisateur et suivi de l'utilisateur principal (voir Figure 2). Ce cadre d'utilisation est très intéressant car il permet à plusieurs visiteurs de partager l'accès à l'œuvre dans une impression de « in situ » malgré son absence. De nombreux problèmes de visualisation se posent cependant car, de par l'effet d'immersion et la dimension de l'écran, la qualité du modèle est très importante pour une perception acceptée et agréable à différentes échelles.

4.3 Réalité virtuelle / augmentée

Pour valider notre approche d'enrichissement des outils par virtualisation de l'œuvre et immersion de l'utilisateur, nous avons aussi développé une application Unity couplée à un casque Hololens sans fil qui permet une restitution du modèle 3D en surimpression au monde réel (voir Figure 3). Cette première phase nous a permis d'identifier les challenges technologiques afférents à cet environnement. L'étape à venir est son intégration dans le musée en vraie réalité augmentée.

4.4 Accès web

Pour offrir un accès beaucoup plus large, nous avons développé une application web permettant la visualisation et les premiers outils d'interaction dans un navigateur internet.

5. Conclusion

L'évolution des matériels et de la puissance de calcul des ordinateurs permet maintenant d'accéder à des coûts raisonnables à une modélisation 3D ouvrant de nombreuses nouvelles approches des objets historiques et patrimoniaux. De nombreux challenges sont toutefois encore devant nous comme la mise en place de l'accès à l'information à partir de la visualisation.

Remerciements

Ce travail a été réalisé avec le concours du Centre Image, plateforme technologique de l'URCA, les « Musées de Reims », la société savante « Reims-Histoire-Archéologie ».

Figure 1 : Exemples d'images source

Figure 2 : Boite à effet hologramme

Figure 3 : Mur d'immersion du Centre Image

Figure 4 : Hololens + Unity

Summary

During the analysis of a work of art, physical, or at least visual, interaction can be an important contribution to its understanding. Then, many situations arise for which this approach is made difficult (exposure in restricted access rooms...) or even impossible (reduced mobility, destroyed object...). While in general it is legitimate to question the need to use a 3D acquisition, in these cases it is an interesting contribution that opens many other perspectives. This raises the question of a low-cost feasibility of both acquisition and modelling, enriched visual restitution and collaborative work. Based on the work of a research group around the Jovin Tomb, exhibited at the Saint-Rémi Museum in Reims, we present a set of issues that appear throughout this process and the opening towards various audiences.