Atelier DAHLIA

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

Comité d'organisation et du programme :

Claudia Marinica (DUKe – LS2N, Polytech'Nantes)
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 (EGC2021)

le 26 janvier 2021 à ~~Montpellier~~ – en ligne pour cause de la Covid-19

Editeurs :

Claudia Marinica
Laboratoire LS2N, équipe DUKe - Polytech'Nantes
page web : https://claudia-marinica.polytech.univ-nantes.fr/
email : claudia.marinica@univ-nantes.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-EGC2021.html
Groupe de travail DAHLIA : http://dahlia.egc.asso.fr
Mailing liste : gt-dahlia@egc.asso.fr

Méthodes de classification non supervisée pour l'analyse de données mobilières en archéologie

Arthur Coulon*, Philippe Husi*, Lise Bellanger**

*UMR 7324 CITERES – Laboratoire Archéologie et Territoires, 37204 Tours cedex 3
40, rue James Watt, ActiCampus 1, 1er étage 37200 TOURS
arthur.coulon@univ-tours.fr ; philippe.husi@univ-tours.fr
**Université de Nantes, Laboratoire de mathématiques Jean Leray UMR CNRS 6629
2, rue de la Houssinière BP 92208 – F-44322 Nantes cedex 03
lise.bellanger@univ-nantes.fr

1. Résumé étendu

Cette présentation se place dans le contexte d'un projet d'archéologie sur la céramique médiévale et moderne du Bassin de la Loire Moyenne. Ce projet s'inscrit dans la longue durée, puisqu'il a débuté en 1996 et a déjà fait l'objet de deux ouvrages sur le sujet (Husi P. dir. 2003 et 2013) et une troisième publication à venir (2021). C'est l'un des premiers projets à l'échelle nationale à avoir abordé la question de l'étude de la céramique médiévale et moderne de manière collective et systémique. Autrement dit, le choix fait dès l'origine a été de mobiliser un important corpus céramique issu de divers sites archéologiques et suivant une méthodologie commune, la vaisselle en terre cuite étant une des sources principales de connaissance de la vie quotidienne des populations préindustrielles. Ce mobilier archéologique a comme avantage, d'être indestructible, omniprésent dans les niveaux archéologiques, avec des changements typo-stylistiques évoluant rapidement au cours du temps. Ces caractéristiques en font un objet de recherche archéo-statistique pertinent. La masse de données produites a rendu indispensable la mise en place de la base de données ArSol (Archives du Sol), système d'enregistrement développé par le Laboratoire Archéologie et Territoires (UMR CITERES-LAT)). Parallèlement, ont également été mis en place des outils typologiques structurés, hiérarchisés, évolutifs et dynamiques à l'origine d'un réseau d'information sur la céramique médiévale et moderne (ICERAMM). Une collaboration de longue date entre archéologues du laboratoire CITERES-LAT et statisticiens du Laboratoire de Mathématiques Jean Leray (UMR 6629, CNRS/Université de Nantes) permet d'explorer les données recueillies et d'en extraire des connaissances archéologiques grâce à l'utilisation ou au développement de méthodes statistiques adaptées. Les résultats de ces recherches ont fait l'objet de nombreux articles dans des revues internationales (Bellanger, Husi 2012 ; Bellanger et al. 2006, Bellanger et al. 2008).

Le développement de méthodes d'analyses archéo-statistiques est indispensable à l'étude fine de corpus de données mobilières de grande dimension (céramique, objets, verre…). Cette recherche interdisciplinaire se poursuit avec la construction de méthodes originales de classification non supervisée. Celles-ci ont été utilisées récemment (i) pour la construction de la périodisation générale du Bassin de la Loire Moyenne, avec hclustcompro, une méthode de classification ascendante hiérarchique (CAH) compromis (Bellanger et al. 2020) ; (ii) pour la définition d'aires culturelles, avec : mapclust, une méthode de classification descendante hiérarchique avec contraintes géographiques (Bellanger et al., soumis). Ces méthodes de classification sont implémentées dans le package R **S**tatistical **PA**ttern **R**ecognition and da**T**ing using **A**rchaeological **A**rtefacts assemblage**S** (SPARTAAS), dédié au traitement statistique des données archéologiques mobilières. L'enjeu de cette intervention est d'exposer,

à partir d'un corpus de données céramique choisi l'ensemble des étapes formant la démarche interdisciplinaire adoptée allant de l'extraction des données de la base ArSol à la spatialisation avec mapclust en passant par la périodisation avec hclustcompro.

Dans notre démarche, hclustcompro est la première méthode appliquée. C'est une méthode de classification ascendante hiérarchique par compromis qui prend en compte deux sources d'information, potentiellement sujettes à erreur et associées aux mêmes observations (contextes archéologiques). L'une reflète les caractéristiques étudiées (corpus céramique) et l'autre la structure des contraintes temporelles (datations associées). Une approche basée sur la distance est adoptée pour modifier la mesure de dissimilarité dans l'algorithme CAH classique en utilisant une combinaison convexe pour prendre en compte les deux matrices de dissimilarité initiales. Le dendrogramme associé à cette CAH peut être interprété comme le résultat d'un compromis entre chaque source d'information analysée séparément. A la fin de cette étape, on aboutit à l'élaboration d'une périodisation archéologique, préalable indispensable à tout discours historique. Dans une deuxième étape, mapclust, une méthode de classification descendante hiérarchique avec contraintes géographiques, est appliquée à un sous corpus. Ce nouveau jeu de données est associé à une des périodes obtenues à l'étape précédente et à une question archéologique précise. La répartition spatiale des données peut être hétérogène et présenter des agrégations locales appelé patch. Woillez et al. (2007) ont proposé un algorithme pour les identifier. La classification mapclust fonctionne de manière divise en séparant les observations en deux à chaque étape utilisant ces patchs spatiaux. Le dendrogramme obtenu peut être coupé afin d'obtenir une partition optimale en utilisant des critères d'évaluations comme l'inertie intra-groupe ou l'indice de silhouette. Cette approche permet, à partir d'une classification, de répondre aux problèmes récurrents en archéologie de répartition spatiale des contextes archéologiques caractérisés par leurs données mobilières.

Références

Bellanger L., Coulon A., Husi P. (2020), PerioClust: a new Hierarchical agglomerative clustering method including temporal or spatial ordering constraints. Springer Series, Studies in Classification, Data Analysis, and Knowledge Organization.

Bellanger L., Coulon A., Husi P. (soumis), Determination of cultural areas based on medieval pottery using an original hierarchical divisive clustering method with geographical proximity constraint (MapClust), JAS

Bellanger L., Husi P. (2012), Statistical Tool for Dating and interpreting archaeological contexts using pottery. Journal of Archaeology Science, 39(4), pp. 777-790.

Bellanger L., Husi P., Tomassone R. (2006), Statistical aspects of pottery quantification for dating some archaeological contexts. Archaeometry, Volume 48, pp. 169-183.

Bellanger L., Tomassone R., Husi P. (2008), A statistical approach for dating archaeological contexts. Journal of Data Science (JDS), 6(2).

Husi P. dir. (2003), La céramique médiévale et moderne du Centre-Ouest de la France (11e – 17e s.). Chrono-typologie de la céramique et approvisionnement de la vallée de la Loire moyenne, 20e supplément à la RACF, Tours, FERAC, 1 cédérom, 110 p. [En ligne].

Husi P. dir. (2013), La céramique du haut Moyen Âge dans le Centre-Ouest de la France : de la chrono-typologie aux aires culturelles, 49e supplément à la RACF, Tours, ARCHEA/FERACF, 1 dvd, 268 p. [en ligne].

Woillez, M., Poulard, J.C., Rivoirard, J., Petitgas, P., Bez, N. (2007), Indices for capturing spatial patterns and their evolution in time, with application to European hake (Merluccius merluccius) in the Bay of Biscay. ICES J. Mar. Sci. 64, 537–550.

Annotations sémantiques de textes liés à l'héritage culturel français

Solibia Pazimna*, Jean-Claude Moissinac*

*19 place Marguerite Perey F-91120 Palaiseau
adresse@telecom-paris.fr,
https://www.telecom-paris.fr/

Résumé. Dans le cadre du projet Data&Musée, nous avons souhaité annoter sémantiquement des textes du domaine culturel français. Nous avons mis en oeuvre une méthodologie pour construire un système d'annotation de textes pour un vocabulaire spécifique d'un domaine. Les méthodes classiques d'annotation s'intéressant à l'annotation de mentions de personnes, de lieux, d'organisations... et ne couvrent donc pas une bonne partie des termes d'un vocabulaire spécifique. Nous nous sommes donc attaché à mettre au point une méthode qui s'appuie sur des résultats récents en matière de modélisation de textes.

1 Introduction

Ce travail a été réalisé dans le cadre du projet DataMusée, dont le but était la collecte et l'exploitation de données concernant des musées de France. Ce travail a porté sur l'annotation sémantique de textes de ce domaine, par exemple des descriptions d'expositions.

L'annotation sémantique ajoute des informations complémentaires à des textes non-structurés, elle peut permettre en particulier d'identifier et de relier les entités du texte avec des données du Web sémantique. L'annotation sémantique de textes liés à l'héritage culturel français a pour objectif de repérer des mots ou des syntagmes, ainsi que diverses expressions associées décrivant des références du vocabulaire culturel français. Pour ce travail, nous avons utilisé les vocabulaires définis par le service des Musées de France [1]. Ces annotations peuvent par exemple contribuer à une mise en valeur des musées et expositions françaises.

Les principales difficultés pour concevoir un système d'annotation automatique sont les ambiguïtés inhérentes au langage naturel, en particulier ici à celles liées au vocabulaire culturel français. Un nombre important de types du vocabulaire culturel existe, tel que ceux mis à disposition par le service des musées de France. Sans être la seule, cette diversité typologique est l'une des sources provoquant des situations d'ambiguïté. La désambiguïsation est considérée comme une sous-tâche de l'annotation sémantique, elle consiste à attribuer une identité unique aux termes mentionnés dans le texte.

La mise à disposition des vocabulaires scientifiques servant à l'interrogation et à l'alimentation des divers champs de la base Joconde (voir sectiondonnées) par le service des Musées de France, constitue une avancée importante pour la mise en valeur des musées et expositions

1. `http://data.culture.fr/thesaurus/page/vocabulaires`

français. Grâce à ces vocabulaires, nous avons accès à un certain nombre de termes et expressions qui font références à des types liés à l'héritage culturel français.

L'objectif de ce travail est de mettre en place un outil permettant de reconnaître et lier automatiquement dans un corpus textuel en français, les différents types du vocabulaire Joconde.

Nous présentons dans cet article un système d'annotation sémantique de texte liés à l'héritage culturel français, basé sur une méthodologie que nous proposons et applicable à d'autres vocabulaires. La section 2 présente le problème général que nous avons cherché à traiter, suivie de la section 3 qui présente des travaux intéressants pour traiter cette problématique. La méthodologie proposée est décrite en section 4. La section 7 présente les résultats de l'évaluation de notre système . Enfin la section 8 conclut cet article et présente des perspectives.

2 Problématique

Dans le projet Data&Musée, nous disposons de textes décrivant notamment des expositions, des oeuvres, des musées et des monuments. Nous avons souhaité relier les entités décrites par ces textes avec d'autres données collectées par ailleurs sous forme de graphes de connaissances. Pour cela, nous avons décidé d'annoter ces textes en créant des liens entre des parties de ces textes et des entités sélectionnées du web sémantique.

L'annotation sur des types usuels telles que personnes, lieux, ou organisations est une tâche correctement effectuée avec des outils comme DBpedia SpotLight (Mendes et al., 2011) ou AIDA (Hoffart et al., 2011). Pour ces types, de tels outils fonctionnent pour l'anglais, et, de plus en plus, pour d'autres langues comme le français, mais ont des résultats moins bons pour le français que pour l'anglais. Ils ne prennent pas en compte les vocabulaires de l'héritage culturel français. Ce travail vise à mettre à disposition de la communauté un outil qui annote sémantiquement un corpus textuel spécifiquement sur ces vocabulaires.

3 Etat de l'art

Dans cette section, nous abordons plusieurs travaux en rapport avec notre objectif.

3.1 Annotation

L'annotation de textes est un problème classique. L'annotation par un liage avec des entités de graphes de connaissances a connu des développements significatifs dans la dernière décennie. On trouve de nombreux outils pour cela : OpenCalais, Zemanta, ANNIE... Par exemple, AIDA assure des liens entre des parties d'un texte avec Yago Hoffart et al. (2011) tandis que DBpedia SpotLight Mendes et al. (2011) assure des liens avec DBpedia. L'un comme l'autre traitent un nombre limité de types de mentions : lieux, personnes célèbres, organisations... Ces types ne couvrent que très partiellement les vocabulaires que nous voulons couvrir. De plus, ils sont beaucoup moins efficaces en français qu'en anglais, notamment parce que DBpedia et Yago couvrent moins bien les entités de culture française que des entités d'autres cultures que ne le fait Wikidata. Nous avons identifié très peu de travaux qui font de l'annotation avec Wikidata, par exemple Delpeuch (2020) qui ne répond pas à notre besoins puisque d'abord

le système est conçu spécialement sur du texte anglophone, ce qui ne marche pas bien sur du texte français, mais aussi ne prend pas en compte notre vocabulaire.

3.2 Reconnaissance d'entités nommées

Le concept d'entités nommées est apparue dans les années 90 à l'occasion de conférence d'évaluation MUC (Message Understanding Conference). Ces conférences avaient pour but de promouvoir la recherche en extraction d'information. Les tâches proposées consistaient à remplir de façon automatique des formulaires concernant des événements. Certains objets textuels, ayant une importance applicative particulière dans plusieurs domaines du TALN, ont été regroupés sous le nom d'entités nommées. La reconnaissance de ces dernières est donc considérée comme une sous-tâche à part entière de l'extraction d'information (Eshkol-Taravella, 2015).

La reconnaissance d'entités nommées est donc une technique permettant de reconnaître une expression linguistique qui fait référence à une entité du monde réel ou à un concept. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mots) catégorisables dans des classes.

Deux approches sont utilisés pour implémenter un système de reconnaissance d'entité nommée : l'approche des règles prédéfinies ou celle des méthodes d'apprentissage machine.

De récentes études ont montré que les modèles pré-entraînés se sont avérés efficaces pour améliorer de nombreuses tâches de traitement du langage naturel (Suárez et al., 2020).

— BERT (Devlin et al., 2019) : acronyme anglais de Bidirectional Encoder Representations from Transformers, est un modèle de langage développé par Google en 2018. Cette méthode a permis d'améliorer significativement les performances en traitement automatique des langues. Basé sur les transformers Vaswani et al. (2017), BERT est plus performant en terme de résultats et Base en terme de rapidité d'apprentissage que ses prédécesseurs. Une fois pré-entraîné, de façon non supervisée, il possède une "représentation" linguistique qui lui est propre, constituant un modèle linguistique du langage traité. Il est ensuite possible, sur la base de ce modèle initial, de le personnaliser pour une tâche particulière. Il peut être entraîné en mode incrémental (de façon supervisée cette fois) pour le spécialiser rapidement et avec peu de données complémentaires.

— Camembert (Martin et al., 2020) : est un modèle de langage de pointe pour le français basé sur l'architecture RoBERTa Liu et al. (2019) pré-entraîné sur le sous-corps français du corpus multilingue OSCAR [2] nouvellement disponible. [3] La différence entre Camembert et BERT réside dans leur pré-entraînement. CamemBERT a été pré-entraîné sur un corpus francophone et avec des hyper-paramètres différents.

Un récent état de l'art sur la reconnaissance d'entités nommées sur du texte français mené par Suárez et al. (2020) compare les performances de Camembert modèle de langage de type BERT français à des modèles de référence, y compris des modèles multilingues basés sur BERT (mBERT et UDify) et un modèle qui n'utilise pas de plongement contextuel (UDPipe Future). Les auteurs démontrent la valeur ajoutée d'un modèle français en considérant diverses tâches de traitement de la langue naturelle. Leurs résultats montrent que, sur la tâche de recon-

2. https://oscar-corpus.com
3. https://camembert-model.fr

naissance d'entités nommées en français, la performance, mesurée par le F-score, s'améliore considérablement.

Nous verrons plus loin comment Camembert a contribué à notre projet.

3.3 La liaison d'entité

La tâche de liaison d'entité consiste à attribuer une identité unique aux entités mentionnées dans un texte. Il s'agit de relier le contenu d'un texte à des entités dans une ontologie.

Cette tâche qui semble simple d'un point de vue général est en pratique confrontée à plusieurs défis tels que la variation de nom, l'ambiguïté, l'absence, les langues multiples, etc.

Plusieurs recherches récentes, notamment (Delpeuch (2020), Ling et al. (2020)) abordent le problème comme une tâche de classification. Delpeuch (2020) utilise la comptabilité locale, la similarité sémantique, pour construire un classifieur binaire qui prédit pour chaque mention m $\in$ d, d un document, si m doit être lié ou pas à une entité **e** donnée. Celle proposée par Ling et al. (2020) nommée RELIC est une classification multiclasse qui utilise le contexte des entités.

RELIC, acronyme de Representations of Entities Learned in Context, est une méthode de liage d'entités. L'idée principale de la méthode est de représenter des entités par un contexte afin d'effectuer un système d'apprentissage qui va correspondre aux représentations. Nous verrons plus loin comment RELIC nous a permis d'entrainer un modèle qui associe un terme à une classe définie par l'entité Wikidata qui la définit.

— Données d'entrées de RELIC : Soient V = {[MASK] ; [Es] ; [Ee] ; w_1, …, w_m} un vocabulaire, et E = {e_0, …., e_n} un ensemble distinct d'entité. Un contexte X={x_0, x_1, …., x_p} est une séquence de mots. X contient x_i.., x_k mots correspondant à la mention d'entité avec k>i. Un mot spécial [Ee] insérer avant le mot x_i et un autre mot spécial [Es] après x_k.
RELIC prend en entré une liste de paire (x_i ; e_i) avec $y_i \in E$

— Encodeur de contexte et plongement des entités : chaque sortie [CLS] de BERT et chaque entité est linéairement plongé dans un vecteur de dimension R^d ; d est l'ensemble de définition des entités.

— Fonction de perte de RELIC : en apprentissage automatique, les fonctions de perte sont utilisées pour décider de la capacité de prédire le modèle. RELIC utilise cross entropy comme fonction de perte. En théorie de l'information, le cross entropy entre deux lois de probabilité mesure le nombre de bits moyen nécessaires pour identifier un événement issu de l'« ensemble des événements ».

4 Approche

Nous avons vu section 4 les avantages de modèles de type BERT et notamment Camembert pour le français. Notre approche va donc consister à nous appuyer sur Camembert pour l'entraîner de façon supervisée à l'aide de notre vocabulaire. Nous considérons la tâche d'annotation sémantique comme un processus ordonné des tâches d'extraction de terminologie et de liaison d'entité. La première tâche consiste à extraire automatiquement une liste de termes du texte à partir d'un corpus spécialisé, et la deuxième consiste à attribuer une identité unique aux termes mentionnés. Elle consiste donc dans les étapes suivantes :

— Le texte à annoter est passé à un modèle de reconnaissance des mentions d'entités sur notre vocabulaire culturel. Le modèle de reconnaissance d'entité nous fourni en sortie des mentions d'entités avec leur type, c'est-à-dire le sous-vocabulaire auxquelles elles se rattachent (étape 3 ci-dessous).
— Les mentions et leurs types obtenus à l'étape précédente sont cherchés dans Wikidata. Si le résultat de la recherche est nul, alors la mention n'est pas associée à une entité (étape 4 ci-dessous).
— Si le résultat de la recherche est supérieur à un, alors nous passons la mention trouvée et son contexte, c'est à dire la phrase dans laquelle se trouve la mention dans le texte, au modèle de désambiguïsation pour nous fournir l'entité wikidata exacte correspondante (étape 7 ci-dessous).

Pour traiter ce problème, nous avons adopté la méthodologie suivante :

1. constitution d'un corpus de phrases dans lesquelles des expressions de nos vocabulaires de référence sont identifiées (voir 6.4)
2. utilisation de ce corpus pour construire un système qui propose des expressions candidates à une association sémantique à partir d'un texte à annoter
3. utilisation du modèle pour trouver les mentions dans un texte
4. pour chaque expression, recherche d'entités candidates à l'association dans le graphe de connaissance retenu
5. construction d'un corpus pour la désambiguïsation, composé de textes associés à une entité de Wikidata (voir 5.5)
6. construction d'un modèle qui permet de choisir parmi les entités candidates celle qui convient dans le contexte de la phrase à annoter (désambiguïsation)
7. utilisation de ce modèle pour choisir parmi les résultats de l'étape 4

Ainsi, nous avons un processus qui peut s'appliquer à nos vocabulaires -mais pourrait s'appliquer à d'autres- et annoter sémantiquement un texte relativement à un vocabulaire de référence à partir des étapes 3, 4 et 7, une fois que les modèles sont construits à l'aide des autres étapes. Nous verrons à la section 6 l'implémentation retenue.

5 Données

5.1 La base de données Joconde

La base de données Joconde décrit 589278 œuvres d'art de collections françaises. Elle est établie par le ministère français de Culture. Une extraction a été mise à disposition en Open Data via la plateforme ouverte des données publiques françaises . Une extraction est disponible dans plusieurs formats dont JSON. C'est l'extraction dans ce format que nous avons utilisée. Une licence ouverte permettant une réutilisation gratuite est associée à ces données. Chaque oeuvre est décrite par 14 champs.

Dans cet article, nous nous intéressons particulièrement aux champs ci-après. DENO désigne un type de création, par exemple 'bas-relief' et 'amphore'. DOMN désigne un domaine, par exemple 'mosaïque' et 'musique'. EPOQ désigne une époque, par exemple 'néolithique

moyen' et 'XXIIe dynastie'. PERI désigne un période, par exemple '1er siècle av JC' et '2e millénaire'.

Le tableau 1 montre dans la colonne "Termes de Joconde" le nombre de valeurs différentes pour les champs : AUTR, DOMN, DENO, LOCA, EPOQ, PERI.

La base Joconde a été construite par agrégation des données d'un ensemble de musées. Les données provenaient des conservateurs de ces musées et nous considérons donc que les vocabulaires utilisés peuvent être considérés comme une utile référence.

Categorie (field)	*Validé*	*Termes de Joconde*	*%*
Créateurs (AUTR)	5220	37828	13.79
Domaines (DOMN)	168	168	100.
Types d'objets (DENO)	337	5766	5.84
Lieux (LOCA)	49	3593	1.36
Epoques (EPOQ)	510	831	61.37
Périodes (PERI)	60	346	17.34

TAB. 1. GROUND TRUTH (14/7/2020).

Table 1 shows the state of the ground truth at 14/7/2020. Corresponding files are available on github (and other files related to this article) githubsemjoconde (2020).

5.2 Wikidata

Wikidata est un grand ensemble de données décrivant des entités liées aux pages de Wikipédia ; au 12/12/2020, Wikidata comporte plus de 91M d'éléments (Wikidata, 2020). Wikidata est un projet de la fondation Wikimedia. Comme Wikidata offre la meilleure couverture des musées et monuments partenaires du projet Data & Musée, nous avons privilégié les liens avec Wikidata. Wikidata permet des requêtes sophistiquées pour obtenir des représentations RDF décrivant les entités ; elles peuvent notamment être obtenus à l'aide de requêtes SPARQL. C'est donc avec Wikidata que nous chercher à annoter des textes du domaine culturel.

5.3 Projet Wikidata Vocabulaires Joconde

Le projet Wikidata Vocabulaires Joconde JocondeProject (2020) a pour but d'associer chacun des termes des vocabulaires de la base Joconde avec une entité de Wikidata, avec une vérification humaine de chacune des associations proposées. Au 12/12/2020, le tableau 1 montre dans la colonne "Validé", le nombre de termes du vocabulaire qui ont été validés par des contrôleurs humains et la colonne % indique le pourcentage de termes validés. Ces données constituent pour nous un ensemble de référence avec des associations terme/entité contrôlées.

5.4 Construction d'un corpus pour la recherche de mentions

Nous avons identifié le besoin d'un corpus de textes dans lesquels sont distinguées les termes ou expressions -mentions- que nous voulons annoter. Nous avons établi une liste de textes dans lesquels sera identifié une mention à annoter et son type.

L'annotation d'un corpus se fait à la main ou automatiquement. Une annotation manuelle a pour risque les incohérences, beaucoup de travail,... et une annotation automatique a pour risque de produire trop d'annotations erronées.

En pratique, il est conseillé de faire une pré-annotation automatique puis une vérification par l'annotateur Benoît Sagot (2012). Nous utilisons la méthode de Benoît Sagot (2012) afin de pouvoir obtenir un corpus de référence. Pour pré-annoter notre corpus, nous proposons un algorithme de recherche des mots lématisés. Nous recherchons la forme lématisée, car nous avons certaines mentions qui peuvent être au pluriel, ou accordées, par exemple **sculptures et sculpture**. Après l'étape de pré-annotation, nous passons à une vérification humaine afin de corriger certaines imperfections.

Pour avoir un corpus riche sur le vocabulaire culturel, nous avons récupéré les textes issus des descriptions d'expositions françaises, obtenues dans le projet Data&Musée et des descriptions d'entités du vocabulaire Joconde identifiées comme telles dans Wikidata. Ces textes nous fournissent des mentions riches en vocabulaires culturels. Une fois ces textes annotés, nous avons obtenu un corpus de référence sur lequel nous avons appliqué les techniques de reconnaissance des mentions afin de créer le modèle numérique qui nous permettra de trouver les mentions intéressantes dans de nouveaux textes (voir section 7).

5.5 Construction d'un corpus pour la désambiguïsation

Nous avons identifié le besoin de disposer d'une liste de textes auquel sera associé une entité de Wikidata. Nous nommons contexte une association (texte, entité), où le texte est une phrase dans laquelle se trouve la mention de l'entité. On s'assure que la mention est bien mise en exergue via des marqueurs **[Es]** au début, et **[Ee]** à la fin de la mention.

Par exemple dans la phrase : ***L'impressionnisme est un point de départ pour Georges Seurat et Paul Signac.***, avec ***impressionnisme*** qui correspond à l'entité wikidata ***Q40415***.

Pour avoir un corpus riche dans sur les entités de notre vocabulaire déjà validé par wikidata, nous avons fait un dump wikidata de toutes les entités des vocabulaires Joconde. Le dump comprend le ***Label*** ainsi que le ***Also known as*** de l'entité wikidata. Par exemple l'entité wikidata **Q40415** a pour label français ***impressionnisme*** et "Also known as" français ***impressionisme***.

Nous avons ensuite sélectionné sur wikipedia la description française correspondante de chaque entité . Par exemple la description correspondante de l'entité wikidata **Q40415** est **https ://fr.wikipedia.org/wiki/Impressionnisme**. Dans le texte récupéré sur wikipedia de la description correspondante à l'entité wikidata, nous cherchons le nom du label français et son Also known as français dans le texte de sa description correspondante sur wikipedia pour former une liste de contexte.

6 Implémentation

Nous montrons ici comment nous avons mis en oeuvre l'approche proposée à la section 4.

6.1 Reconnaissance des mentions d'entités

L'objectif de cette partie est de trouver dans un texte des termes ou expressions candidates et d'associer à un terme ou expression un type dans le vocabulaire culturel. Pour pouvoir

construire un tel modèle de reconnaissance des mentions, nous devons disposer d'un corpus de référence annoté sur les mentions culturelles. Nous avons vu à la section 6.4 comment nous avons construit un tel corpus.

6.1.1 Modèle

Nous optons pour l'architecture du modèle proposé par Suárez et al. (2020) pour atteindre l'état de l'art dans la reconnaissance des entités nommées en français. Il s'agit d'une architecture $LSTM-CRF+CamemBERT_{Base}$ dont l'efficacité sur du texte français a été prouvée avec un un score moyen de ~93%.

Nous utilisons camembert-cased comme modèle de pré-entraînement. Le cased signifie que durant le pré-entraînement de ce modèle, il y avait une sensibilité à la casse.

Le corpus utilisé sur ce modèle est celui de la recherche de mentions. Le texte en entré est tokenizé avec le tokenizer de camembert. Nous utilisons camembert pour plongé le texte tokenizé. La sortie issue du plongement est utilisée pour alimenter une couche LSTM-CRF.

Le modèle ainsi construit permet de prendre un texte en entrée et d'annoter chaque token du texte par un type défini par notre typologie. Les types qui nous intéressent dans le cadre de cet article sont : DOMAINE","EPOQUE", "PERIODE", "TECHNIQUE".

6.2 Recherche d'entités candidates dans le graphe de connaissance

Nous présentons ici la méthode simple que nous avons adoptée pour trouver des entités candidates pour le liage sémantique de chaque expression trouvée à l'étape précédente.

Nous avons choisi Wikidata comme graphe de connaissances vers lequel créer des annotations sémantiques. Pour cela, nous faisons appel au Wikidata Query Service (WDQS), un service qui permet d'effectuer des requêtes SPARQL sur Wikidata. Pour le travail précédent , nous avons mis au point une requête SPARQL qui nous renvoie des entités ayant un label correspondant à une chaîne de caractère choisie (mention). Nous appliquons cette requête pour la mention, la mention en minuscules, la mention en majuscule, la mention en mode Titre (chaque mot avec le premier caractère en majuscule), et enfin si la mention comporte plusieurs mots, on tente de déplacer le premier mot en dernière position (utile pour certains éléments comme les personnes). Ensuite, nous filtrons par le type que nous a donné le modèle à l'étape précédente.

— Si aucune entité trouvée, il n'y aura pas de liage
— Si une seule entité est trouvée, nous la conservons ;
— si plusieurs entités sont trouvées, nous sollicitons le modèle de désambiguïsation décrit à l'étape suivante pour trouver la bonne entité.

6.3 Construction du modèle de désambiguïsation : RELIC

Pour représenter les entités dans un modèle numérique, RELIC prend en entrée des contextes textuels des entités et construit un modèle numérique à partir de l'ensemble des contextes de l'ensemble des vocabulaires qui nous intéressent.

L'entrée de RELIC sera une liste de **(X, y)**. Par exemple :

— X = ***L' [Es] impressionnisme [Ee] est un point de départ pour Georges Seurat et Paul Signac.***
— y = ***Q40415***.

Comme vu à la section 5.5, nous avons construit un ensemble de contextes des entités qui nous intéressent. Cet ensemble de contextes nous a permis d'entraîner un modèle RELIC.

Chaque identifiant unique (Q-id) d'entité de Wikidata est incorporé sur un vecteur dans R^d. Ensuite, chaque contexte -paire texte/entité- est plongé dans un vecteur de taille fixe à l'aide d'un encodeur de contexte. Les paramètres de l'encodeur sont initialisés par CamemBERT.

Nous prenons la sortie groupée de l'encodeur de contexte qui correspond au jeton initial [CLS] dans la représentation de séquence de camembert comme notre contexte encodée, et nous la projetons linéairement dans R^d en utilisant une matrice de poids apprise W $\in$R^{d*768} pour obtenir un contexte incorporant dans le même espace comme nos entités. La taille 768 dans la matrice s'explique par la taille de la sortie de $camembert_{Base}$. que nous avons utilisé.

7 Expérimentation

Dans cette section, nous expérimentons notre approche sur les vocabulaires **"DOMAINE", "EPOQUE", "PERIODE", "TECHNIQUE"** de Joconde.

7.1 Reconnaissance des mentions d'entités

7.1.1 Préparation des données

Nous avons récupéré la liste des mentions qui correspondante à chaque vocabulaire de la base Joconde, que nous avons lématisé.

Le texte utilisé dans notre travail provient prindipalement de wikipedia (voir section 6.4). Il s'agit des textes issu des descriptions des expositions françaises et des descriptions correspondantes aux entités wikidata du vocabulaire joconde sur lequel nous expérimentons.

Pour pré-annoter, nous avons :

— divisé les textes en phrases avec l'outil sentence_splitter. [4]

— vérifié avec langdetect [5] que toutes les phrases sont en français.

— tokenisé les phrases avec l'outil spacy. [6]

— Nous avons stocké les données au format IOB.

Le texte des descriptions des expositions était limité pour élaborer un corpus pour former un modèle. Précisément sur le texte des descriptions des expositions, nous avons après pré-annotation et correction 2384 phrases qui contiennent au moins une entité dont 1174 domaines, 660 techniques, 366 époques, 12 périodes.

Ce corpus des expositions est insuffisant pour avoir des résultats satisfaisants.

Pour améliorer le corpus, nous avons récupéré les textes des entités moins nombreux (époques, périodes) sur wikipedia. Par exemple l'entité wikidata `http://www.wikidata.org/entity/Q166713` correspond à la description française wikipedia `https://fr.wikipedia.org/wiki/Postimpressionnisme`

Pour cela, nous avons utilisé le dataset wiki40b/fr [7] de google. Du fait des résultats précédents sur les annotations des expositions, la récupération du texte des entités s'est plus concentré sur les types périodes et époques. Au total **4207** phrases utiles (contenant au moins une

4. `https://pypi.org/project/sentence-splitter/`
5. `https://pypi.org/project/langdetect/`
6. `https://spacy.io/`
7. `https://www.tensorflow.org/datasets/catalog/wiki40b#wiki40bfr`

entité) ont été récupérées de wikipedia et ajoutées au corpus des expositions ; nous avons ainsi **6591** phrases qui contiennent au moins une mention d'entité.

Après pré-annotation et correction humaine, nous avons les statistiques suivantes 5512 domaines, 4259 techniques, 2638 époques, 1395 périodes.

	%	*DOMN*	*TECH*	*EPOQ*	*PERI*
Entrainement	72	3298	1625	3649	778
Validation	08	597	265	181	161
Test	20	1617	748	429	456

TAB. 2. RÉPARTITION DES DONNÉES (AVEC % DES PHRASES)

Le tableau 2 montre la répartition des données.

7.1.2 Résultats

Nous avons entraîné le modèle (voir section 7.1.1) sur un batch_size de 8, une taille de la séquence au maximal 512, un learning rate de 2e-5 sur 50 epochs. Les résultats ci-dessous nous montrent la performance du modèle final sur les données de validation et de test.

	Evaluation données de test			Evaluation données devalidation		
	Precision	**Rappel**	**F-score**	**Precision**	**Rappel**	**F-score**
EPOQUE	0.87160	0.90979	0.89029	0.90244	0.90244	0.90244
DOMAINE	0.95025	0.96869	0.95938	0.94333	0.96587	0.95447
PERIODE	0.97360	0.97039	0.97199	0.94393	0.99020	0.96651
TECHNIQUE	0.93098	0.89296	0.91157	0.89520	0.82329	0.85774
micro avg	**0.93769**	**0.94331**	**0.94049**	**0.92727**	**0.92643**	**0.92685**
macro avg	**0.93788**	**0.94331**	**0.94040**	**0.92641**	**0.92643**	**0.92596**

Au vue de ces résultats, nous pouvons conclure que :
— Le modèle est plus précis sur les PERIODES
— Les EPOQUES sont les éléments qui offrent la moins bonne performance
— En moyenne nous avons un modèle de score 94% sur nos données de test

7.2 Désambiguïsation

7.2.1 Préparation des données

Nous avons construit le corpus des données de désambiguïsation suivant le processus décrit en section 6.5. Certaines entités par exemple Q40719766 (époque Joconde) dans wikidata n'ont pas de description correspondante sur wikipedia. Nous n'avons donc pas pris en compte ces entités. Au total, nous avons **552** entités reparti dans **12252** contextes dont **9687** contextes en training, **1175** contextes en validation, et **1390** contextes en test.

7.2.2 Résultats

Nous avons avons entraîné le modèle (voir section 7.3) sur un batch_size de 32, une taille de la séquence au maximal 128, un learning rate de 2e-5 sur 50 epochs. Les résultats ci-dessous nous montrent les résultats du modèle final sur les données de validation et de test.

	Evaluation données de test			Evaluation données de validation		
	Precision	Rappel	F-score	Precision	Rappel	F-score
micro avg	0.99928	0.99928	0.99928	0.99830	0.99830	0.99830
macro avg	0.99515	0.99676	0.99569	0.99582	0.99699	0.99615

7.3 Web service

Un web service est un ensemble de fonctionnalités exposées sur internet ou sur un intranet, par et pour des applications ou machines, sans intervention humaine, de manière synchrone ou asynchrone. Afin de pouvoir mettre en service le travail, nous avons mis au point un serveur qui renvoi le texte annoté à la suite d'une requête d'annotation.

```python
import requests

url = "http://127.0.0.1:8000/annotate/"
lng = "fr"
encoding_type = "utf-8"
params = {"language": lng, "encoding_typ": encoding_type}
data = {"content": doc, "params": params}

doc = "Le postimpressionnisme caracterise une periode de
l'histoire."

response = requests.post(url=url, data=data)
```

```
{
    "code": 200,
    "result": [
        {
            "sentence": "Le postimpressionnisme caracterise une
    periode de l'histoire.",
            "entities": [
                {
                    "mention": "postimpressionnisme",
                    "type": "EPOQUE",
                    "beginOffset": 3,
                    "wikidata":
    "http://www.wikidata.org/entity/Q166713"
                },
                {
                    "mention": "histoire",
                    "type": "DOMAINE",
                    "beginOffset": 52,
                    "wikidata": "http://www.wikidata.org/entity/Q309"
                }
            ]
        }
    ]
}
```

8 Conclusion et perspectives

Nous avons mis en oeuvre une méthodologie pour construire un système d'annotation de textes pour un vocabulaire spécifique d'un domaine. Nous avons validé cette méthode sur des vocabulaires du domaine culturel français. Notre méthode utilise des techniques récentes de modélisation de textes.

Dans notre article précédent (Moissinac et al., 2020), nous avons traité des vocabulaires utilisés par les champs "TECH" (techniques), "AUTR" (créateurs), "LOCA" et "STAT" (localisations). Dans le présent article, nous avons traité des champs "DENO" (type d'objet),"DOMN" (domaine), "EPOQ" (époque), "PERI" (période). De prochains travaux vont s'atteler à améliorer la couverture de ces vocabulaires, notamment par un renforcement croisé des vocabulaires rencontrés sur une même création, ainsi que la corrélation des vocabulaires de Joconde avec les vocabulaires de (getty2018, 2018), qui font référence dans le domaine.

Par ailleurs, nous prévoyons d'appliquer le méthode ci-dessus pour l'annotation de textes du domaine informatique.

Références

Benoît Sagot, Marion Richard, R. S. (2012). Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. In *La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral. Linguistique. Université d'Orléans, 2015. ffftel-01250650ff*, pp. 10–17.

Delpeuch, A. (2020). Opentapioca : Lightweight entity linking for wikidata.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). Bert : Pre-training of deep bidirectional transformers for language understanding.

Eshkol-Taravella, I. (2015). La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral. In *La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral. Linguistique. Université d'Orléans, 2015. ffftel-01250650ff*, pp. 10–17.
english

getty2018 (2018). Getty Research : Editorial Guidelines. URL : `http://www.getty.edu/research/tools/vocabularies/guidelines/index.html` [retrieved : September, 2020].
english

githubsemjoconde (2020). Repository for SemJoconde. URL : `https://github.com/datamusee/semjoconde` [retrieved : September, 2020].

Hoffart, J. et al. (2011). Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland*, pp. 782–792.
english

JocondeProject (2020). Wikidata :WikiProject Vocabulaires Joconde. URL : `http://www.wikidata.org/wiki/Wikidata:WikiProject\_Vocabulaires\_Joconde/en` [retrieved : September, 2020].

Ling, J., N. FitzGerald, Z. Shan, L. B. Soares, T. Févry, D. Weiss, et T. Kwiatkowski (2020). Learning cross-context entity representations from text.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, et V. Stoyanov (2019). Roberta : A robustly optimized bert pretraining approach.

Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, de la Clergerie, D. Seddah, et B. Sagot (2020). Camembert : a tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Mendes, P. N., M. Jakob, A. García-Silva, et C. Bizer (2011). Dbpedia spotlight : Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, New York, NY, USA, pp. 1–8. Association for Computing Machinery.

Moissinac, R. et al. (2020). Toward a Semantic Representation of the Joconde Database. In *SEMAPRO 2020, The Fourteenth International Conference on Advances in Semantic Processing, SEMAPRO 2020, Nice, France*, pp. 62–67.

Suárez, P. J. O., Y. Dupont, B. Muller, L. Romary, et B. Sagot (2020). Establishing a new state-of-the-art for french named entity recognition.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, et I. Polosukhin (2017). Attention is all you need.
english

Wikidata (2020). Wikidata. URL : `https://www.wikidata.org/wiki/Wikidata:Statistics/fr` [retrieved : Décembre, 2020].

Summary

As part of the Data&Musée project, we wanted to semantically annotate texts from the French cultural domain. We have implemented a methodology to build a text annotation system for a domain specific vocabulary. The classic annotation methods deals with the annotation of mentions of people, places, organizations ... and therefore do not cover a good part of the terms of our specific vocabulary. We have therefore set out to develop a method which is based on recent results in the field of text modellization.

Ariane: dispositif de fouille et de lecture synthétique de textes

Motasem Alrahabi

OBVIL, Maison de la recherche, Sorbonne Université, 75005 Paris
motasem.alrahabi@gmail.com

Résumé. La croissance exponentielle des documents textuels en format numérique et le développement des moyens informatiques de plus en plus sophistiqués offrent de nouveaux moyens pour l'analyse sémantique et discursive des textes. Dans cette communication, nous présentons Ariane, une interface web basée sur des annotations sémantiques et permettant de fouiller des corpus textuels selon des modalités linguistiques: opinions, sentiments, émotions, perceptions, etc. L'annotation automatique est assurée par un outil à base de règles, et les ressources linguistiques sont manuellement crées et vérifiées. Un scénario d'utilisation concret est présenté pour montrer l'intérêt de l'application dans le domaine des humanités numériques.

1. Présentation

L'architecture générale d'Ariane[1] permet d'interroger des corpus textuels en combinant un modèle d'annotation sémantique et un modèle de lecture synthétique de textes. Les annotations sémantiques sont obtenues à l'aide d'un système d'annotation de patrons de surface. Le modèle de lecture synthétique permet la mise en œuvre de parcours de lecture allant d'une vision globale des modalités (au niveau du corpus entier) à une vision locale exprimée au niveau du texte, voire de la phrase[2].

L'utilisateur a le moyen de croiser ces informations avec une technique classique de recherche d'information: requêtes par mots clés et filtrage par métadonnées (auteur, date, titre…). À n'importe quel moment d'un parcours de fouille, l'utilisateur a le moyen d'afficher les statistiques en cours, de les visualiser par des diagrammes ou de les exporter sous forme de tableurs csv.

L'annotation automatique des textes indexés dans Ariane est assurée par Textolab, un outil à base de règles qui permet de manipuler les patrons linguistiques de surface et de créer les règles heuristiques.

[1] https://obvil.huma-num.fr/ariane/

[2] Ariane s'inspire directement de l'application e-quotes (Alrahabi, 2015).

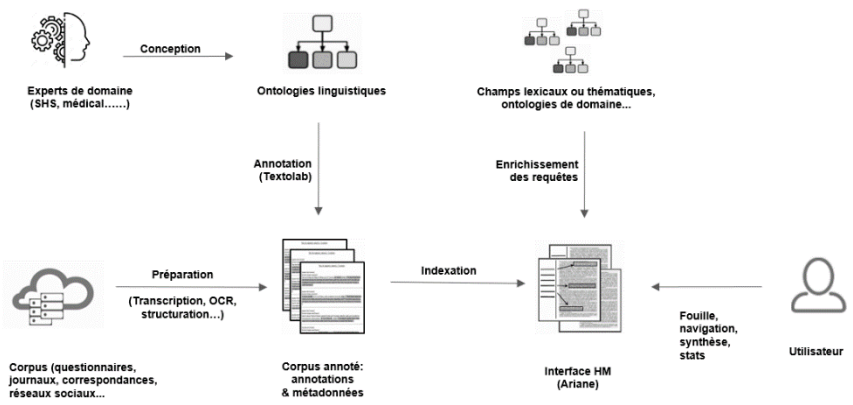

FIG. 1 – *Architecture générale du système*

Le processus d'étiquetage dans Textolab est incrémental et permet de localiser, grâce à des heuristiques, des patrons sous forme de chaînes de caractères, d'expressions régulières ou de métadonnées (annotations existantes). Les règles sont déclaratives et ordonnables selon la priorité. Chaque règle procède à la désambiguïsation des marqueurs identifiés, par des indices contextuels confirmatifs qui valident l'annotation, ou par des indices infirmatifs qui annulent l'annotation en cours. Lorsque toutes les conditions sont satisfaites, au moins une des étiquettes suivantes est attribuée au passage textuel concerné: modalité, polarité, intensité. Cette procédure prend en compte le traitement de la négation, de l'interrogation et du mode hypothétique.

2. Scénario d'utilisation

En traitement automatique du langage et fouille de textes, les travaux relatifs aux modalités recoupent souvent avec l'analyse des opinions, des sentiments et des émotions (Liu, 2015).

Différents parcours de fouille ont été expérimentés avec Ariane dans le domaine des humanités numériques et médicales (Riguet et Alrahabi, 2020 ; Alrahabi et al., 2020 ; Alrahabi et Bordry, 2020). Nous nous proposons ici d'explorer l'interface d'Ariane à travers une étude de cas qui porte sur le jugement critique des écrivaines au XIXe siècle (Riguet et Alrahabi, 2020). Notre hypothèse était de savoir si les jugements des critiques du XIX^{e} siècle étaient formulés de la même façon pour les écrivaines que pour les écrivains. Quels registres sont convoqués, et surtout quels types de valeurs littéraires (esthétique, éthique, intellectuelle, psychoaffective…) leur sont particulièrement attachés ? L'étude du discours critique de la seconde moitié du XIX^{e} siècle, formulé par des auteurs exclusivement masculins, parle en majorité d'hommes écrivains. Cependant, et dans une perspective d'étude du genre, il nous a semblé intéressant de nous pencher sur le traitement particulier de certaines femmes qui font exception dans cette disparité proportionnelle, comme George Sand, Germaine de Staël ou Madame de Lafayette.

À partir de la bibliothèque numérique du labex OBVIL[3], nous avons isolé 134 ouvrages de critique littéraire publiés entre 1850 et 1914. Le corpus, composé de 487 fichiers (environ 33000 phrases), a été annoté par Textolab à l'aide d'une ontologie des modalités linguistiques (Alrahabi, 2010 ; Alrahabi et Riguet, 2017) qui comprend des classes comme l'opinion, le jugement, le désaccord, l'appréciation, la louange, l'indignation, etc.

À partir des résultats obtenus[4], deux constats ont été observés. D'une part, les proportions entre jugements positifs et jugements négatifs semblent très proches, qu'ils portent sur les hommes ou les femmes: avec 71% (hommes) et 65% (femmes) de catégories globalement connotées positives ; et 29% (hommes) et 34% (femmes) de catégories globalement connotées négatives.

D'autre part, si les grands types de jugement se répartissent de façon similaire, c'est davantage la nature des *valeurs* associées qui se distinguent. En effet, la lecture à l'échelle de la phrase rendue possible par Ariane montre que, contrairement à la variété des valeurs d'appréciation (esthétique, intellectuelle, éthique, référentielle, etc.) repérées dans le corpus des hommes, les appréciations du corpus des femmes mettent nettement en avant la dominance de l'émotionnel, du psycho-affectif et de l'esthétique sur la dimension intellectuelle.

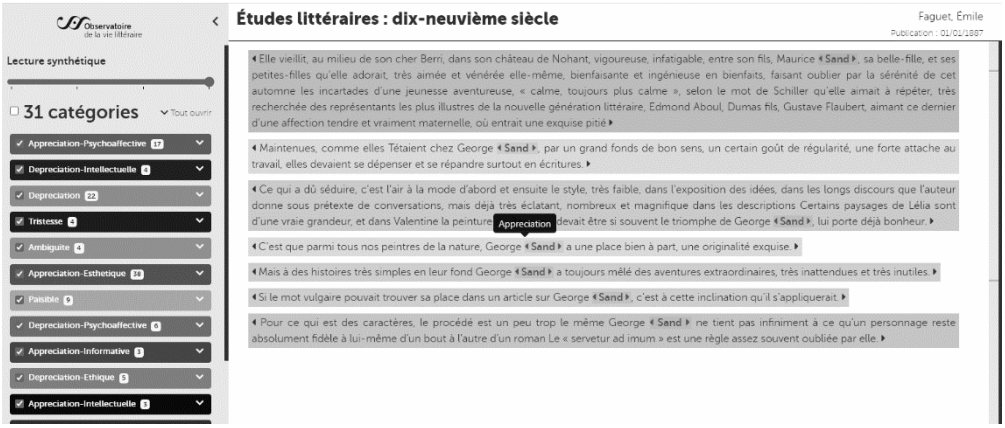

FIG. 2 – *Capture d'écran d'un texte annoté sur Ariane*

À l'aide d'Ariane, il est possible d'interroger les résultats par sous-corpus (métadonnées), par polarité ou par modalité. L'utilisateur a le moyen d'effectuer des requêtes par mots clés au sein des passages annotés, de visualiser les résultats document par document, ou par concordance, avec la possibilité d'accéder au contexte d'origine de chaque mots clé. Des fonctionnalités de lecture synthétique permettent de réduire le texte autour des passages annotés.

Cette approche sémantique nous permet de rendre apparents, dans le discours, la part que les écrivaines occupent et les types de traitements dont elles font l'objet ; elle nous permet de comparer les modalités et les valeurs potentiellement distinctes par le biais desquels les critiques évaluent les œuvres selon le genre de l'auteur.

[3] https://obvil.sorbonne-universite.fr/corpus/critique/
[4] https://obvil.huma-num.fr/ariane/etudeGenre/search

3. Conclusion

En combinant les annotations sémantiques et la lecture synthétique à des échelles différentes, du corpus au texte, l'objectif final d'Ariane est d'être à la fois un outil d'exploration et un support à l'interprétation des textes dans le domaine des humanités. Cette application permet de mettre en lumière des agencements linguistiques et participe à l'analyse du discours en aidant l'utilisateur à formuler des hypothèses, susceptibles par la suite d'être précisées ou infirmées au regard des résultats d'annotation.

Notre prochain objectif est de permettre aux utilisateurs de charger leurs propres textes dans le système, de les annoter et de les exploiter via l'interface.

Références

Alrahabi, Motasem. 2010. « Excom-2 : plateforme d'annotation automatique de catégories sémantiques : conception, modélisation et réalisation informatique : applications à la catégorisation des citations en arabe et en français ». These de doctorat, Paris 4. http://www.theses.fr/2010PA040005.

—. 2015. « E-Quotes: Enunciative Modalities Analysis Tool for Direct Reported Speech in Arabic ». Dans *Computational Linguistics and Intelligent Text Processing*, édité par Alexander Gelbukh, 479-490. Lecture Notes in Computer Science. Cham : Springer International Publishing. https://doi.org/10.1007/978-3-319-18111-0_36.

Alrahabi, Motasem et Marguerite Bordry. 2020. « L'ironie dans la critique littéraire : quelques pistes pour un traitement automatique ». Dans *Humanités numériques et Digital Studies*. Montpellier.

Alrahabi, Motasem, Pauline Flepp et Camille Koskas. 2020. « Polémiques dans le rituel épistolaire : les cas de la correspondance Ponge et Paulhan ». *Revue Épistolaire 46, Honoré Champion.*

Liu, Bing. 2015. *Sentiment Analysis: mining sentiments, opinions, and emotions.* Cambridge University Press, 2015. doi:10.1017/CBO9781139084789.

Riguet, Marine et Motasem Alrahabi. 2017. « Pour une analyse automatique du Jugement Critique: les citations modalisées dans le discours littéraire du XIXe siècle ». Dans *DHQ: Digital Humanities Quarterly*.

—. 2020. « Analyse automatique pour une étude du genre : quels jugements des écrivaines au XIXe siècle ? » Dans *Digital Humanities Conference*. Ottawa, Canada.

Summary

The exponential growth in the digitization of texts and the development of increasingly sophisticated computing tools offer new opportunities for the semantic and discursive analysis of texts. In this talk, we present Ariane, a web interface based on semantic annotations and allowing to mine textual corpora according to linguistic modalities (opinions, sentiments, emotions, perceptions…). Automatic annotation is provided by a rule-based tool, and linguistic resources are manually created and verified. A concrete use case is presented to show the potential of the application in the field of Digital Humanities.

Traitement avancé des narratives iconographiques, quelques suggestions

Gian Piero ZARRI*

*STIH, Sorbonne Université
1, rue Victor Cousin - 75005 Paris
zarri@noos.fr, gianpzarri@gmail.com

Résumé. Cet article concerne la nécessité d'utiliser une nouvelle génération de procédures de digitalisation dans le domaine de l'Héritage Culturel, moins préoccupées de reproduire l'aspect purement physique des entités de ce domaine et plus sensible à la possibilité d'une description détaillée de leur contenu sémantique. Cette possibilité est spécialement importante pour tous ces objets culturels qui véhiculent un contenu du type « iconographique » – tableaux, dessins, fresques, mosaïques, sculptures etc. Pour représenter ces « narratives visuelles » en allant au-delà des limitations habituelles, on suggère de se servir d'un système comme NKRL (*Narrative Knowledge Representation Language*) qui est, en même temps, un outil pour la représentation du contenu sémantique des narratives ainsi qu'un environnement informatique pour l'exploitation de ces représentations. La description NKRL de la « narrative » illustrée par la scène centrale de la « Reddition de Breda » de Velázquez est utilisée comme exemple dans l'article.

1. Introduction

Les idées exposées dans cet article à propos de la façon d'effectuer un traitement informatique avancé de « l'Héritage Culturel » ont trouvé un écho inattendu dans certaines préoccupations récemment exprimées par la Commission Européenne. La Commission a dernièrement indiqué, en effet, qu'elle aimerait voir l'émergence d'une *nouvelle génération de procédures de digitalisation*, moins préoccupées de reproduire l'aspect purement « *physique* » (dimensions, techniques de production, origine, lieu de conservation, etc.) des entités qui font partie de cet héritage, et plus sensible à la possibilité d'une description fidèle et complète du « *message* » qu'elles véhiculent ainsi que de leur contexte historique, social et culturel – voir les appels à propositions qui concernent la composante « Transformations » du volet « *Europe in a Changing World* » d'Horizon 2020 (European Commission, 2020).

Cette nouvelle attitude de la Commission confirme ainsi l'importance d'effectuer une représentation formelle adéquate de la « *signification profonde* » (*inner meaning*) propre à chaque objet culturel afin de pouvoir réaliser un traitement informatique exhaustif de ces objets. Cette possibilité est d'une importance particulière pour toutes les entités qui véhiculent un contenu de type « *iconographique* » – tableaux, dessins, fresques, mosaïques, sculptures, murales mais aussi, de façon plus générale, posters, bandes dessinées, films ou illustrations publicitaires – et qui constituent, entre autres, une composante tout à fait fondamentale (et économiquement importante, il suffit de penser au tourisme) de l'héritage culturel. Ces objets relatent en effet des *histories complexes sous une forme visuelle* (de la bataille de Hastings et les préparatifs pour l'invasion de l'Angleterre dans la Tapisserie de Bayeux aux événements tragiques relatés par le Radeau de la Méduse) qui, pour être modélisées de façon réellement complète, ne peuvent pas se limiter à une description sous forme de simples mots-clés/métadata. Cet article suggère ainsi d'utiliser pour la représentation informatique des « *narratives iconographiques* » un langage spécialisé comme NKRL, le *Narrative Knowledge Representation Language* (Zarri, 2009). Créé grâce à plusieurs projets européens, NKRL est en effet, dans le même temps, un outil conceptuel particulièrement riche pour la représentation de narratives complexes comme les narratives iconographiques, et un environnement informatique complet pour l'exploitation de ces représentations.

Cet article est structuré de la façon suivante. La Section 2 fournit un état de l'art sommaire des pratiques courantes à propos de représentation formelle d'objets culturels reconductibles à des narratives iconographiques. La Section 3 décrit brièvement les caractéristiques principales du langage NKRL, tandis que la Section 4 fournit un exemple concret de l'utilisation possible de ce langage dans un contexte d'héritage culturel. La Section 5 constitue la « Conclusion » de l'article.

2. Représentation de narratives iconographiques, état de l'art

NKRL utilise une sorte « *d'interprétation opérationnelle* » de la notion de narrative qui est tout à fait conforme aux définitions théoriques les plus récentes voir, par exemple, (Bal, 1997 ; Jahn, 2017). Dans un contexte NKRL, une narrative est ainsi conçue comme une *structure informationnelle complexe qui concerne une séquence ("stream") d'événements élémentaires logiquement et chronologiquement liés* – une narrative peut se réduire à un seul événement élémentaire. Ces événements impliquent à leur tour la mise en place d'une *description détaillée des activités, expériences et relations mutuelles propres des entités (pas nécessairement humaines) impliquées dans l'événement* et qui se déroulent à l'intérieur d'un *contexte spatio-temporel* bien défini. L'élément « *dynamique/structuré* » qui caractérise la description des narratives en termes de NKRL est, ainsi, tout à fait évident ; cela est valable aussi, bien évidemment, pour ces histoires complexes exposées sous forme visuelle qui correspondent aux narratives iconographiques.

Toute description de ces narratives fondée uniquement sur une superposition purement *statique* aux images de mots-clés ou d'autres entités conceptuellement équivalentes serait ainsi à éviter. Cela concerne, par exemple, l'utilisation de thesaurus de type ICONCLASS (http://www.iconclass.nl), *Art and Architecture Thesaurus*, AAT (http://www.getty.edu/research/tools/vocabularies/aat/index.html), *Union List of Artist Names*, ULAN (http://www.getty.edu/research/tools/vocabularies/ulan/index.html), etc., ainsi que celle de métadonnées fondées sur des modèles de données particulièrement simples comme le *Dublin Core* d'origine (http://dublincore.org/) ou même de modèles plus complexes du type VRA (*Visual Resources Association*) Core 4 XML Schema (https://www.loc.gov/standards/vracore/schemas.html). Même dans ce dernier cas, en effet, la description d'une entité du type agent se réduit en pratique à une succession *non structurée* d'entités binaires du type « propriété-valeur » comme name-Peter Paul Rubens ou culture-Flemish, accompagnées de simple attributs XML La conversion de plusieurs de ces modèles en recommandations RDF-compatibles, voir DCMI (*Dublin Core Metadata Initiative*) Abstract Model (Nilsson *et al.*, 2008) ou la version RDF(S) de l'outil CIDOC CRM (http://www.cidoc-crm.org/rdfs/5.0.4/cidoc-crm) – un outil tout à fait remarquable, par ailleurs, d'un point de vue essentiellement taxonomique, voir Le Boeuf et al. (2018) pour la dernière version – ne change rien, bien évidemment, au fond du problème.

Un exemple paradigmatique de ce qu'on pourrait appeler *l'approche « statique »* aux modalités d'annotation/représentation des narratives iconographique est fourni par l'utilisation comme "*case study*", dans le document *Image Annotation on the Semantic Web W3C Incubator Group Report*, voir Troncy et al. (2007), d'un ensemble d'informations en RDF/XML à propos du tableau de Claude Monet, « La terrasse à Sainte-Adresse » ainsi que de deux reproductions de ce tableau utilisées pour les opérations d'annotation. Ces informations, très complètes, nous renseignent, par exemple, à propos de la localisation actuelle du tableau ainsi que de tous les emplacements précédents. Elles fournissent aussi, en utilisant le lexique ATT, toute une série détaillée d'éléments techniques à propos de ce tableau, sa dimension, le type de support, la peinture, le style, la technique utilisée, le mouvement artistique impliqué, ainsi que des renseignements sur les deux reproductions

associées (la résolution par exemple), etc. La description de la narrative qui constitue la *raison d'être* du tableau se réduit toutefois aux quatre lignes suivantes :

```
<!-- Subject matter: (who/what is depicted by this work -->
<vra:subject>Jeanne-Marguerite Lecadre (artist's cousin)</vra:subject>
<vra:subject>Madame Lecadre (artist's aunt)</vra:subject>
<vra:subject>Adolphe Monet (artist's father)</vra:subject> ;
```

se limitant ainsi à fournir *l'identification* des trois personnages qui apparaissent dans le tableau sans livrer le moindre renseignement sur le « *contenu sémantique* » du tableau en question, quelques indications par exemple à propos de la disposition des trois personnages et de leur attitude en face de la mer et du paisible et lumineux paysage, la profusion de fleurs, etc. Nous sommes loin, ici, de ces exigences exprimées dans la Section précédente à propos de la reproduction exhaustive des « *messages* » véhiculés par les narratives iconographiques ainsi que de leur « *contenu profond* ».

Le rapport du groupe « Incubateur » du W3C mentionné ci-dessus date de 2007, mais la situation ne semble pas avoir évolué depuis lors. Dans un article récent (Lodi *et al*., 2017) se référant au projet *Cultural-ON*(tologies), l'on trouve la description formelle d'une narrative iconographique, une sculpture du Giambologna conservée au musée Capodimonte de Naples qui représente l'enlèvement d'une femme Sabine par un Romain. La représentation utilisée se limite à donner le nom de l'œuvre et celui de son réalisateur, et à signaler que l'œuvre en question fait partie de la Collection Farnese du musée. Aucune indication n'est fournie par contre à propos des trois personnages représentés – la femme effrayée, l'enleveur, l'homme qui cherche à empêcher l'acte, leurs attitudes et relations réciproques etc. D'autres exemples de ce type sont très faciles à repérer, voir la description formelle du tableau de Mona Lisa (la Gioconda) effectuée en 2013, dans un contexte Europeana, en utilisant le *European Data Model (EDM)*, où la description du contenu sémantique de cette œuvre semble se réduire à des affirmations du type « le sujet du tableau est une femme », voir Isaac *et al.*, 2011), etc.

Le problème signalé par l'insuffisance des exemples ci-dessus se ramène au problème général qui concerne *le manque « d'expressivité »* (*expressiveness*) des outils formels – OWL, RDF(S), SPARQL, SWRL, SKOS, FOAF etc. – utilisés dans un contexte Web Sémantique (WS). Ce problème est bien connu à l'intérieur même de la communauté WS, voir par exemple la contestation de certains principes fondamentaux du Web Sémantique opérée par quelques pères (et mères) nobles du mouvement (Bernstein et al., 2016) et, sur un plan plus technique, l'analyse effectuée par Trame et collègues (Trame et al., 2013) des échecs répétés de toute tentatives de rendre « *n*-aires » des structures de données banalement « binaires ». Dans la mesure, en effet, où les outils WS se fondent sur un *simple modèle binaire*, ils ne sont pas très utiles quand il s'agit de modéliser de façon simple, compréhensible et reproductible des situations, scripts, scénarios, storyboards, narratives etc. impliquant différentes entités qui entretiennent des relations multiples conditionnées par des contraintes spatio-temporelles. Dans un modèle binaire, les propriétés/les attributs associés à une notion ou à un concept donné *peuvent uniquement associer un individu à un autre individu ou à une valeur*. Les résultats obtenus sont ainsi susceptibles de résulter *indûment simplifiés*, comme dans les exemples précédents, par rapport aux situations réelles. Et, par ailleurs, il est bien connu aussi que l'obligation de *tout modéliser avec des outils de base trop simplistes* produit à l'arrivée des systèmes *structurés de façon extrêmement complexe, enchevêtrés et difficilement utilisables* – voir encore, à ce propos, la modélisation du tableau de Mona Lisa dans Europeana. L'utilisation pour les narratives iconographiques – et l'Heritage Culturel en général – d'un outil *n*-aire crée exprès pour le traitement avancé des structures narratives complexes comme NKRL peut représenter ainsi une alternative intéressante aux procédures standard utilisées jusqu'à présent.

3. Les caractéristiques essentielles de NKRL

NKRL dans sa forme actuelle est le résultat d'activités de recherche appliquée développées, entre 1990 et 2005, dans le cadre de plusieurs projets Européens. NKRL a été

utilisé depuis avec succès dans plusieurs applications dans les domaines le plus différents, de l'analyse de notices d'agence dans un contexte terrorisme au raisonnement causal sur les accidents industriels etc., voir par exemple (Zarri, 2010 ; 2011 ; 2013 ; 2014a).

3.1. Les deux ontologies

L'innovation essentielle introduite par NKRL par rapport aux procédures ontologiques habituelles concerne l'introduction, à côté de la traditionnelle « ontologie des concepts », d'une nouvelle « *ontologie des événements* ».

L'ontologie des concepts, dans sa version NKRL, est appelée HClass (hiérarchie des classes) et inclut à présent environ 5.000 concepts « standard » – standard signifie ici que la représentation formelle utilisée pour ces concepts suit le modèle binaire habituel. D'un point de vue strictement formel HClass n'est pas très différente, ainsi, des ontologies que l'on peut créer en se servant de la version « *frame* » originale de Protégé (Noy *et al.*, 2000) – voir, toutefois, Zarri (2009 : 103-137) à propos des originalités qui caractérisent ses définitions. L'ontologie des événements élémentaires représente, par contre, un nouveau type de structure hiérarchique où les nœuds correspondent à des structures *n*-aires complexes appelées « *templates* (patrons, modèles) » ; l'ontologie est appelée ainsi HTemp, la hiérarchie des *templates*. A l'opposée des fonctions ontologiques éminemment « statiques » assignées aux « concepts » de HClass, celles déléguées aux *templates* sont *typiquement dynamiques*. Plus précisément, les *templates correspondent à la représentation formelle de catégories homogènes d'événements élémentaires* du type « déplacer un objet physique », « avoir une attitude positive/négative envers quelqu'un/quelque chose », « demander un avis », « rendre un service », « se situer à un certain endroit à un moment donné », « transmettre un message », etc. HTemp inclut environ 150 *templates* ; la hiérarchie est aisément extensible et personnalisable.

Dans la mesure où les *templates* sont caractérisés par une structure formelle complexe, ils ne peuvent pas être représentés en utilisant une *simple approche binaire* comme dans le cas des concepts de HClass. La formule *n*-aire dénotée par Eq. 1 fournit une représentation très schématisée de l'expression de base d'un *template* :

$$(L_i\ (P_j\ (R_1\ a_1)\ (R_2\ a_2)\ \ldots\ (R_n\ a_n)))\ . \qquad (1)$$

Dans Eq. 1, L_i dénote « *l'étiquette symbolique* » qui identifie le schéma *n*-aire correspondant à un *template* donné. P_j est le « *prédicat sémantique* » qui caractérise le *template* et qui définit en la *catégorie conceptuelle globale* (mouvement, attitude, propriété, production etc.) propre à l'ensemble des événements élémentaires correspondant à ce *template*. R_k est un "rôle fonctionnel", voir Zarri (2019), utilisé pour indiquer la *fonction* (sujet, objet, destinataire etc.) de l'entité/des entités a_k (l'argument ou les arguments du prédicat) par rapport à l'action/état/situation… défini par le prédicat en question. a_k indique ainsi un élément ou des éléments de HClass (concepts ou individus, c.-à-d., *instances de concepts*) qui, dans les événements élémentaires associés au template, *définissent le type/les types d'entité/s réelles impliquée/es dans ces événements*. Ces entités peuvent être plusieurs (argument complexe) dans le cas, par exemple, d'une action qui est menée en même temps (coordination) par différents acteurs.

Pour éviter tout problème « *d'explosion combinatoire* » (Zarri, 2009 : 57-61), *le nombre des prédicats sémantiques ainsi que celui des rôles fonctionnels utilisables dans un template est limité à sept*. Les prédicats sont les suivants : BEHAVE, EXIST, EXPERIENCE, MOVE, OWN, PRODUCE, RECEIVE, tandis que les rôles fonctionnels sont : SUBJ(ect), OBJ(ect), SOURCE, BEN(e)F(iciary), MODAL(ity), TOPIC, CONTEXT. Les sept prédicats définissent les sept « branches » de l'arborescence HTemp, dans le sens que les *templates* appartenant à une branche donnée se caractérisent par le fait d'être tous construits à partir d'un même

prédicat sélectionné parmi les sept introduits ci-dessus. TAB 1 reproduit, dans sa première partie, le *template* Move:TransferMaterialThingsToSomeone, et dans la deuxième, une instance de ce *template* qui formalise un événement élémentaire très simple, « Aujourd'hui, Pierre a fait cadeau d'un livre d'art à Marie ».

Dans le jargon NKRL, ce type *de description formalisée d'un événement* prend le nom « *d'occurrence prédicative* ». A partir de TAB 1, l'on peut déduire que les termes a_k de l'Eq. 1 sont implémentés en pratique en utilisant des variables var_i et des contraintes sur les variables. Les constraints sont exprimés en utilisant des termes de HClass ; cela confirme que les due ontologies de NKRL sont *strictement intégrées*. PIERRE_, MARIE_ et ART_BOOK_1 sont des « *individus* », instances de concepts de HClass et indiqués par des *lettres majuscules* dans la couche « externe » du métalangage de NKRL. as_a_gift est par contre *un concept (en lettre minuscules)* spécifique du terme activity_related_property qui fait partie de la sous-branche qualifier_ de HClass.

---

name: Move:TransferMaterialThingsToSomeone
father: Move:TransferToSomeone
position: 4.21
natural language description: "Transfer a Material Thing (e.g., a product, an object, a letter…) to Someone"

MOVE	SUBJ	*var1*: [*var2*]
	OBJ	*var3*
	[SOURCE	*var4*: [*var5*]]
	BENF	*var6*: [*var7*]
	[MODAL	*var8*]
	[TOPIC	*var9*]
	[CONTEXT	*var10*]
	{ [modulators], ≠abs }	

var1	=	human_being_or_social_body
var3	=	artefact_
var4	=	human_being_or_social_body
var6	=	human_being_or_social_body
var8	=	process_, sector_specific_activity, service_
var9	=	sortal_concept
var10	=	situation_, symbolic_label
var2, *var5*, *var7*	=	location_

ex.c1)	MOVE	SUBJ	PIERRE_
		OBJ	ART_BOOK_1
		BENF	MARIE_
		MODAL	as_a_gift
		date-1:	2018-09-20-08:26
		date-2:	

TAB 1 – *Dérivation d'une occurrence prédicative à partir d'un template*.

Quand plusieurs événements élémentaires sont reliés par des "*opérateurs de connexion*" – du type causalité, finalité, discours indirect, coordination/subordination etc. – pour donner lieu à des « *événements complexes* » (scripts, scénarios etc.), NKRL se sert pour les activités de formalisation correspondantes de structures logiques du deuxième ordre crées par des opérations de « *réification* ». Ces dernières sont implémentées en utilisant les "*étiquettes symboliques*" des occurrences prédicatives impliquées – c.-à-d., les termes L_i de l'Eq. 1 – selon deux différents mécanismes conceptuels. Le premier concerne la possibilité de se référer à un événement, élémentaire ou complexe, en tant *qu'argument* d'un événement élémentaire. Un exemple pourrait ainsi concerner un événement élémentaire *X* dans lequel un individu donné s'exprime à propos de *Y*, où *Y* est à son tour un événement élémentaire ou complexe. A la différence de ce premier cas où les deux événements s'intègrent de façon très stricte – *Y* est comme « absorbé » par *X* – le deuxième mécanisme associe ensemble, en utilisant différents types de relations logico-sémantiques, des événements, élémentaires ou

complexes, *X*, *Y*, *Z* … *qui se situent au même niveau logique et qui conservent totalement leur indépendance*. Par exemple, on peut considérer un événement *X* que l'on découvre en relation avec un deuxième événement par l'existence, au niveau linguistique, « *d'indices* » (*cues*) évidents de connexion logico-sémantique du type causalité, finalité, coordination, alternative, etc. Dans NKRL, le premier mécanisme d'association est appelé « construction complétive », et le deuxième « occurrences de type liaison (*binding occurrences*) », voir Zarri (2009 : 86-98).

En simplifiant, la construction complétive est implémentée en associant à l'un des quatre rôles fonctionnels OBJ, MODAL, TOPIC et CONTEXT d'une occurrence prédicative un « *filler* » (un argument du prédicat) du type L_i, où cette étiquette peut dénoter *tant l'expression formel d'un événement élémentaire que d'un événement complexe (binding occurrence)*. L'étiquette est associée à l'opérateur « # », $\#L_i$, ainsi à signaler son statut de *filler* d'un type particulier. Un exemple concerne le Tableau 2 de la Section 4 : les deux occurrences prédicatives breda.c5 et breda.c6 sont associées dans le cadre d'une construction complétive qui indique que l'action effectuée par le SUBJ(ect) de breda.c5 se déroule dans le CONTEXT de l'événement décrit par breda.c6.

Les occurrences de type liaison sont représentées par des listes étiquetées formées par un opérateur de liaison Bn_i et les arguments L_i. Une occurrence de liaison bc_i peut être ainsi formulée comme :

$$(Lb_k \ (Bn_i \ L_1 \ L_2 \ \dots \ L_n)) \ , \qquad (2)$$

où Lb_k est l'étiquette symbolique qui désigne la structure de liaison dans son ensemble. Les opérateurs Bn_i sont : ALTERN(ative), COORD(ination), ENUM(eration), CAUSE (cet opérateur dénote une « causalité stricte »), REFER(ence, l'opérateur de « causalité faible »), GOAL (l'opérateur qui dénote une « intentionnalité stricte »), MOTIV(ation, l'opérateur de « intentionnalité faible »), COND(ition). Deux exemples d'occurrences de liaison de type COORD sont fournis par les occurrences breda.c1 et breda.c2 du Tableau 2 de la Section 4. Les modalités d'utilisation des opérateurs Bn_i sont décrites dans Zarri (2009 : 91-98). Par exemple, dans le cas d'occurrences de liaison du type ALTERN, COORD et ENUM, aucune restriction n'est imposée à propos de la cardinalité de la liste, c.-à-d., à propos du nombre des possibles termes L_j. Par contre, pour les occurrences de liaison étiquetées avec CAUSE, REFER, GOAL, MOTIV et COND, deux seuls arguments, L_m et L_n, sont admis. Ces occurrences ont ainsi la forme $(Lb_k \ (Bn_i \ L_m \ L_n))$ au lieu de celle indiquée en général par l'Eq. 2.

3.2. Les techniques NKRL d'interrogation et d'inférence

« Raisonner », dans NKRL, s'étend du *questionnement direct* d'une base de connaissances en format NKRL en se servant de "modèles de recherche" (*search patterns*) p_i qui unifient l'information contenue dans la base grâce à l'utilisation d'un « Module pour l'unification et le filtrage », *FUM* (*Filtering Unification Module*), voir Zarri (2009 :183-201), à des *opérations de raisonnement complexes qui utilisent un moteur d'inférence en chainage arrière*. Les modèles de recherche p_i sont particulièrement importants dans un contexte NKRL dans la mesure où, en plus de la possibilité d'offrir à l'utilisateur la possibilité de poser directement des questions (selon des modalités du type « *information retrieval* intelligent ») à une base de connaissances NKRL, ils peuvent aussi *être générés automatiquement* par les moteurs d'inférence en tant qu'étape finale des différents pas de raisonnement qui matérialisent les procédures d'inférence de haut niveau. D'un point de vue formel, ces modèles correspondent à des *templates* de la hiérarchie HTemp, spécialisés et partiellement instanciés, où les « *variables explicites* » var_i qui caractérisent les *templates* (voir TAB 1) ont été remplacées par des concepts (ou des individus) compatibles avec les

contraintes associées à ces variables dans les *templates* d'origine. Dans un modèle de recherche, les concepts insérés dans ces modèles sont utilisés ainsi comme des « *variables implicites* ». Quand les modèles p_i – construits manuellement par l'utilisateur ou générés automatiquement par les moteurs d'inférence – sont utilisés pour des opérations d'appariement avec les occurrences prédicatives c_j de la base, un concept de p_i peut unifier a) *tous les individus* qui correspondent à des instances du concept en question que l'on peut retrouver dans les différents c_j et, b) *tous le concepts* contenus dans ces c_j qui, en fonction de la structure de HClass, représentent des *spécialisations* (*subsumed*) par rapport à ce concept, ainsi que la totalité de leur instances.

Les procédures d'inférences de haut niveau de NKRL concernent principalement deux classes de règles, les « *transformations* » et les « *hypothèses* », voir Zarri (2009 : 201-234).

Les règles de « transformation » cherchent à *adapter, d'un point de vue sémantique*, un modèle de recherche p_i « *incorrect* » – dans le sens qu'il a été incapable de trouver une unification à l'intérieur de la base de connaissances – aux contenus réels de cette base en utilisant une sorte de *raisonnement par analogie*. Concrètement, les transformations cherchent à convertir de façon automatique p_i dans un ou plusieurs modèles différents p_1, p_2 … p_n à même de récupérer des données *qui ne sont pas des réponses exactes* à la question d'origine mais qui peuvent fournir des informations intéressantes à ce propos. De façon intuitive, et en revenant à TAB 1, une question d'utilisateur à propos de l'existence de livres d'art parmi les livres de Marie qui ne trouverait pas de réponse directe dans la base de connaissances car aucune de ses occurrences prédicatives serait en mesure de s'apparier directement à la question, pourrait obtenir une sorte de réponse « *indirecte* » par l'occurrence ex.c1 de la Table 1. Cela pourrait se concrétiser en utilisant une règle de transformation consistant à affirmer que, si quelqu'un a reçu un cadeau, il entre en possession de l'objet représenté par le cadeau en question. Le fait que les transformations produisent des informations « *utiles* » mais certainement pas des informations « *surement vraies* » est confirmé par le fait que, au moment où l'on pose la question d'origine, Marie aurait pu déjà égarer, vendre, détruire etc. le livre en question.

Du point de vue formel, une règle de transformation est composée d'une partie gauche, « *l'antécédent* » – c'est-à-dire la formulation, dans le format modèle de recherche, de la question qui n'a pas trouvé de réponse – et d'une ou plusieurs parties droites, le/les « *conséquent(s)* » qui fournissent l'énoncé d'un ou plusieurs modèles(s) de recherche à substituer au modèle d'origine, voir aussi l'exemple du TAB 4 dans la Section 4. En indiquant par A l'antécédant et par Cs_i tous les possible conséquents, les règles de transformations peuvent être modélisées comme indiqué dans l'Eq. 3 ; la restriction $var_i \subseteq var_j$ correspond à la *clause de sauvegarde habituelle* qui impose que toutes les variables déclarées dans l'antécédent A se retrouvent aussi dans le conséquent Cs_i accompagnées, le cas échéant, par des nouvelles variables.

$$A(var_i) \Rightarrow Cs_i(var_j), \quad var_i \subseteq var_j \ . \tag{3}$$

Les règles d'hypothèse permettent de construire automatiquement une sorte d'explication « *causale* » pour un événement (une occurrence prédicative) récupéré à l'intérieur d'une base de connaissances NKRL. Ces règles sont formalisées comme des "*biconditionals*" du type :

$$X \text{ iff } Y_1 \text{ and } Y_2 \ldots \text{ and } Y_n \ , \tag{4}$$

où la « tête » de la règle correspond à l'occurrence prédicative c_j à expliquer et les différentes étapes du raisonnement Y_i – appelées « *schéma des condition* » dans un contexte hypothèse – doivent toutes donner lieu à un résultat positif. Cela signifie que, pour chacune de ces étapes il doit être possible, pour le moteur d'inférence, de *créer automatiquement* au moins un modèle de recherche p_i « *réussi* » (*successful*), c.-à-d. capable, en utilisant *FUM*, de s'unifier avec succès à des informations présentes dans la base. Dans ce cas, l'ensemble des c_1, c_2 … c_n occurrences prédicatives récupérées par les schémas de condition Y_i grâce à leur traduction dans des p_i peuvent être interprétées en tant *qu'explication des causes* – ou, du moins, en tant qu'éclaircissement du contexte – de l'occurrence prédicative c_j d'origine.

Voir Zarri (2009 : 183-243) pour les détails algorithmiques et formels, et des exemples concrets d'utilisation, des procédures inférentielles de NKRL.

4. NKRL et la représentation des narratives iconographiques

Pour illustrer la façon dont l'on pourrait utiliser NKRL pour le traitement des narratives iconographique nous nous servirons d'un cas concret, l'encodage NKRL, voir TAB 2, de la scène centrale du tableau de Velázquez qui représente la « Reddition de Breda «. Cette scène montre Ambrosio Spinola, commandant en chef des troupes espagnoles pendant la deuxième phase de la Guerre des Quatre-Vingt Ans recevant, le 5 juin 1625, les clefs de Breda par Justinus van Nassau, gouverneur de la ville.

Toute représentation NKRL d'une narrative débute nécessairement par la création d'une *occurrence de liaison*, voir 3.1, qui spécifie sa structure générale. Dans le cas présent, breda.c1 inclut trois blocs *logiquement équivalents*, voir l'utilisation de l'opérateur de liaison COORD(ination). Le premier bloc, breda.c2, inclut quatre occurrences prédicatives coordonnées, dont #breda.c6 qui est introduite par breda.c5 en tant que *filler* du rôle CONTEXT (*construction complétive*, voir 3.1). De ce fait, breda.c5 et breda.c6 constituent un *ensemble cohérent* qui représente la formalisation de l'élément narratif le plus important du tableau : tandis que le vaincu lui donne les clefs de la ville (breda.c6), Ambrosio Spinola prévient (PRODUCE activity_blockage, breda.c5) une tentative de Justinus de s'agenouiller (genuflecting_) devant lui. activity_blockage est un concept HClass de type activity_ ; genuflecting_ est un concept spécifique, dans l'ordre, de negative_relationship et de relationship_. A noter, voir breda.c6, que genuflecting_a été *réifié* en le transformant dans un individu, GENUFLECTING_1, *de façon à pouvoir le référencier à l'intérieur d'autres occurrences (coréférence)*. Il apparaît ainsi dans breda.c7 où l'on indique que ce mouvement est, à la fois, amorcé (sketched_, spécifique de qualifier_ via general_characterising_property) et in_front_of (spécifique de binary_relational_property) de Spinola. Les structures du type « OWN OBJ property_ TOPIC… » sont utilisées pour décrire les propriétés associées à des entités *du type inanimé* représentées par le *filler* de SUBJ ; les propriétés des *entités animées* sont déclarées en utilisant un *template* du type BEHAVE.

Le codage du TAB 2 montre aussi l'importance de l'utilisation de listes SPECIF(ation) pour la construction *d'arguments structurés du prédicat (expansions)* ; ces listes, utilisables de façon récursive, permettent d'introduire des propriétés importantes à propos de l'élément, qui représente le premier argument de la liste. Par exemple, dans breda.c5, l'utilisation de SPECIF(ation) indique que la modalité d'interruption de la tentative de s'agenouiller, moral_suasion, est du type hand_gesture et que cette action concerne Van Nassau ; dans breda.c6, que les clés sont celle de la ville de Breda, etc. moral_suasion est un spécifique de suggestion_ qui renvoie à la hiérarchie mutual_relationship de HClass ; hand_gesture est un spécifique de gesture_ inclut dans la sous-arborescence physical_activity. A noter que SPECIF(ation) n'est pas le seul opérateur utilisable à l'intérieur d'expansions : les autres sont ALTERN1(ative), COORD1(ination), ENUM1(eration), caractérisés par des propriétés sémantiques analogues à celle des opérateurs de liaison ALTERN, COORD et ENUM correspondant, voir 3.1. Leur empilage à l'intérieur d'une expansion est réglé par la « *règle de priorité* » (Zarri, 2009 : 68-70). Par exemple, il n'est pas possible d'insérer des listes SPECIF à l'intérieur de COORD1 tandis que le contraire est acceptable, voir breda.c7.

breda.c1) (COORD breda.c2 breda.c3 breda.c4)

La formalisation de cette narrative iconographique est formée de trois composants.

breda.c2) (COORD breda.c5 #breda.c6 breda.c7 breda.c8)

Le premier composant est formé de quatre occurrences (# = construction complétive).

breda.c5)	PRODUCE	SUBJ	AMBROSIO_SPINOLA: (BREDA_)
		OBJ	activity_blockage
		MODAL	(SPECIF moral_suasion hand_gesture)
		TOPIC	(SPECIF GENUFLECTING_1 JUSTINUS_VAN_NASSAU)
		CONTEXT	#breda.c6
		date-1:	05/06/1625
		date-2:	

Produce:CreateCondition/Result (6.4)

(*Dans le cadre de* breda.c6), *Ambrosio Spinola arrête Justinus qui est en train de s'agenouiller.*

breda.c6)	RECEIVE	SUBJ	AMBROSIO_SPINOLA: (BREDA_)
		OBJ	(SPECIF key_to_the_city BREDA_)
		SOURCE	JUSTINUS_VAN_NASSAU
		CONTEXT	CELEBRATION_1
		date-1:	05/06/1625
		date-2:	

Receive:TangibleThing (7.1)

Ambrosio Spinola reçoit les clefs de Breda après la reddition de la ville.

breda.c7)	OWN	SUBJ	(SPECIF GENUFLECTING_1 JUSTINUS_VAN_NASSAU)
		OBJ	property_
		TOPIC	(COORD1 sketched_ (SPECIF in_front_of AMBROSIO_SPINOLA))
		date-1:	05/06/1625
		date-2:	

Own:CompoundProperty (5.42)

L'agenouillement est amorcé, et il a lieu devant Ambrosio Spinola.

breda.c8)	OWN	SUBJ	CELEBRATION_1: (BREDA_)
		OBJ	property_
		TOPIC	(SPECIF surrender_ BREDA_)
		date-1:	05/06/1625
		date-2:	

Own:CompoundProperty (5.42)

La célébration concerne la victoire d'Ambrosio Spinola.

breda.c3)	BEHAVE	SUBJ	AMBROSIO_SPINOLA
		MODAL	commander_in_chief
		TOPIC	SPANISH_ARMY
		CONTEXT	EIGHTY_YEARS_WAR
		{ obs }	
		date-1:	05/06/1625
		date-2:	

Behave:Role (1.11)

A la date du 5 juin 1625, Ambrosio Spinola est le commandant en chef de l'armée espagnole.

breda.c4)	BEHAVE	SUBJ	JUSTINUS_VAN_NASSAU
		MODAL	(SPECIF governor_ dutch_)
		TOPIC	BREDA_
		CONTEXT	EIGHTY_YEARS_WAR
		{ obs }	
		date-1:	05/06/1625
		date-2:	

Behave:Role (1.11)

A la même date, Justinus van Nassau est le gouverneur hollandais de la ville.

TAB 2 – *Représentation NKRL de la scène centrale de la « reddition de Breda ».*

Nous pouvons aussi remarquer que la structure logique d'une narrative comme celle de TAB 2 peut toujours être représentée *sous la forme d'un knowledge graph*, voir FIG. 1.

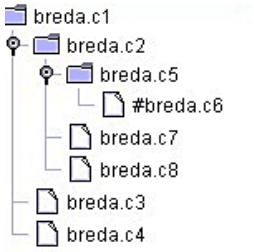

FIG. 1 – *Représentation arboriforme du formalisme de TAB 2*

La construction d'une représentation formelle détaillée de narratives iconographiques complexes n'aurait pas, en soi, un grand intérêt si cette représentation ne pouvait pas être utilisée *dans le cadre d'applications concrètes*. Pour cela, il faudrait toutefois disposer d'une base de connaissances bien plus large de celle représentée par les quelques occurrences de TAB 2. A titre d'exemple, on essayera toutefois d'utiliser ces données pour en déduire des indications à propos du comportement de Ambrosio Spinola vis-à-vis son adversaire. Pour simplifier, nous allons supposer que le modèle de recherche p_i à utiliser correspond à celui de TAB 3, qui provient d'une instanciation partielle du template Behave:ConcreteVersusHumanAttitude. L'utilisation de ce modèle correspond ainsi à demander si des informations décrivant une possible attitude positive de Ambrosio Spinola envers son adversaire sont présentes dans le système ; positive_attitude est un concept de la sous-arborescence personal_activity de HClass.

BEHAVE
SUBJ : AMBROSIO_SPINOLA :
OBJ : JUSTINUS_VAN_NASSAU :
MODAL : positive_attitude
{}
date1 : 05/06/1625
date2 :

TAB 3 – *Un modèle de recherche à propos des données de* TAB 2.

Le modèle de TAB 3 n'est pas en mesure, bien évidemment, de trouver une quelque *forme d'unification directe* avec les occurrences prédicatives de TAB 2. Dans la base de données des règles de transformation associées au système NKRL, l'on peut supposer, toutefois, de repérer une règle (de nature *suffisamment générale*) que l'on pourrait formuler simplement comme « le fait par une personne de mettre fin à une situation de soumission à son égard d'une deuxième personne implique une attitude favorable de la première personne envers la deuxième ». La formulation NKRL de cette est indiquée dans TAB 4.

Pour activer la transformation, il faudra vérifier que le modèle de recherche p_j à « transformer » est en mesure de s'apparier à l'antécédent de la règle : dans ce cas, *la règle sera activée* et les variables incluses dans l'antécédent assumeront les valeurs associées aux

rôles correspondant du modèle, dans notre exemple *var1* = AMBROSIO_SPINOLA, *var2* = JUSTINUS_VAN_NASSAU. Ces valeurs seront transférées au premier conséquent dans la partie droite de la règle ; celui-ci, *transformé en modèle de recherche et caractérisé par des nouvelles variables*, essayera à son tour de trouver des appariements à l'intérieur de la base en générant ainsi des nouvelles valeurs pour les nouvelles variables. Les valeurs en question seront transposées dans le deuxième conséquent, et ainsi de suite ; *la transformation sera validée si tous les conséquents trouveront un appariement valide avec les données de la base*. Dans notre cas, l'appariement du modèle de recherche dérivé de *conseq1* se fera avec breda.c5 de TAB 2, et la valeur GENUFLECTING_1 – instance de genuflecting_, qui est un spécifique de la contrainte negative_relationship – sera affectée à *var3* et transférée au modèle dérive de *conseq2*. Ce dernier modèle s'appariera à breda.c7 de TAB 2, et les deux occurrences prédicatives, breda.c5 et breda.c7 seront ainsi transmises à l'utilisateur en tant que « *possible réponse indirecte* » à la question d'origine. L'on notera que la transformation *t41* de TAB 4 satisfait à la *clause de sauvegarde* mentionnée dans la sous-section 3.2, voir Eq. 3, car l'on retrouve dans la partie droite de la règle la totalité des variables introduites par l'antécédent plus deux nouvelles variables, *var3* et *var4*.

t41: *"recovering from a submissive condition" transformation*

antecedent:

BEHAVE SUBJ *var1*
OBJ *var2*
MODAL positive_attitude

var1 = individual_person
var2 = individual_person
var1 ≠ *var2*

first consequent schema (*conseq1*):

PRODUCE SUBJ *var1*
OBJ activity_blockage
TOPIC (SPECIF *var3 var2*)

var3 = negative_relationship

second consequent schema (*conseq2*):

OWN SUBJ *var3*
OBJ property_
TOPIC (SPECIF *var4 var1*)

var4 = binary_relational_property

Le fait par une personne d'interrompre une manifestation de soumission à son égard d'une deuxième personne implique une attitude favorable de la première personne envers la deuxième.

TAB 4 – *Un exemple de règles de transformation.*

En élargissant considérablement l'embryon de base de connaissances de TAB 2 par l'ajout, en particulier, d'un (important) ensemble d'informations à propos du contexte historique de la reddition de Breda, l'on pourrait se servir des règles d'hypothèse de NKRL pour évaluer certaines « *raisons possibles* » (certaines « causes ») mises en avant pour expliquer l'attitude bienveillante (du moins, selon l'interprétation de Vélasquez) de Spinola

envers son adversaire, *décidément inhabituelle en temps de guerre au 17e siècle*. Parmi les explications avancées, l'on peut mentionner celle évoquant la mise en scène d'une astucieuse opération de propagande au bénéfice de la maison royale d'Espagne, le fait que l'armée espagnole avait effectivement beaucoup admiré le courage et l'efficacité de ses ennemis, la volonté de Vélasquez d'exalter une certaine « manière chrétienne » de faire la guerre, l'amitié de Vélasquez envers Spinola qu'il avait connu pendant un voyage en Italie, etc. D'autres directions d'investigation intéressantes pourraient concerner, par exemple, la confirmation de certaines conjectures à propos *d'influences possibles sur Vélasquez exercées par des chefs-d'œuvre bien connus par le peintre et traitant d'arguments similaires*, comme les ouvres de Rubens concernant « La rencontre du roi Ferdinand d'Hongrie et du Cardinal-Infante Ferdinand d'Espagne » ou la « Réconciliation de Jacob et d'Ésaü ».

Il est bien évident que ces types de traitements qui demandent/demanderaient l'utilisation de quantités importantes de « *connaissances du domaine* » ne pourront pas se faire sans une collaboration très stricte entre informaticiens/experts d'IA et expert d'Histoire et d'Histoire de l'Art. Par ailleurs, la nécessité d'une collaboration stricte avec les experts du domaine pour la mise à point des *procédures d'inférence de NKRL* est une contrainte bien connue pour une utilisation profitable de ce langage, voir à ce propos, par exemple, Zarri (2011).

5. Conclusions

Cet article préconise l'émergence d'une *nouvelle génération* de procédures de digitalisation dans le domaine de l'Héritage Culturel, moins préoccupées de reproduire exactement l'aspect purement « physique » (dimensions, techniques de production, origine, lieu de conservation, etc.) des « objets » appartenant à ce domaine et plus sensible à la possibilité de réaliser une description *la plus complète réalisable* du « *message* », implicite ou explicite, que ces objets charrient ainsi que de leur contexte historique, social et culturel. Cette possibilité est d'une importance particulière pour tous ces objets qui véhiculent un contenu du type « *iconographique* » – tableaux, dessins, fresques, mosaïques, sculptures, murales mais aussi, de façon plus générale posters, illustrations publicitaires, bandes dessinées, dessins animés, films... – et qui constituent une composante tout à fait fondamentale de l'héritage culturel. Ces objets relatent en effet des « histoires » (des « narratives ») sous une forme visuelle qui, pour pouvoir être décrites *de façon complète et réellement informative*, ne peuvent pas se limiter à une description sous forme de métadata ou en utilisant des instruments de représentation des connaissances inappropriés pour la représentation d'événement complexes, du type web sémantique dans sa forme « binaire » traditionnelle. Des exemples possibles du *manque « d'expressivité »* des outils traditionnels ont été fournis dans la Section 2.

L'article suggère ainsi de se servir pour une représentation informatique adéquate des narratives iconographiques complexes d'un *outil spécialisé* comme NKRL qui est, dans le même temps, un outil de haut niveau pour la représentation du « contenu sémantique » de narratives complexes et un environnement informatique complet pour l'exploitation « intelligente » de ces narratives. Ses mécanismes inférentiels utilisent, par exemple, des techniques de raisonnement *du type analogique* (« transformations ») pour trouver des réponses indirectes quand il n'est pas possible de répondre directement à une question donnée, et des techniques de *raisonnement hypothétique* (« hypothèses ») pour générer des

relations nouvelles (comme des relations de cause) entre informations qui sont a priori disjointes. La Section 4 de l'article décrit, à titre d'exemple, la représentation et l'exploitation dans un contexte NKRL d'une narrative iconographique complexe correspondant à la scène centrale du tableau de Velázquez à propos de la « Reddition de Breda ». L'utilisation d'une règle d'inférence du type « transformation » permet ainsi d'assimiler par inférence un geste, celui du vainqueur empêchant le vaincu à s'agenouiller, à une attitude bénévole du premier envers le second, c.-à-d., à un comportement réellement inhabituel du temps des guerres du 17e siècle. Voir aussi, dans Amelio et Zarri (2019), un deuxième exemple récent d'application de techniques du type NKRL à la description/élaboration d'une situation iconographique particulièrement complexe comme celle représenté par le tableau de « La Gioconde » (Mona Lisa).

Références

Amelio, A., and Zarri, G.P. (2019), Conceptual Encoding and Advanced Management of Leonardo da Vinci's Mona Lisa: Preliminary Results, Information MDPI 10(10), 321.

Bal, M. (1997), Narratology: Introduction to the Theory of Narrative, 2d ed. Toronto, University of Toronto Press.

Bernstein, A., Hendler, J., and Noy, N. (2016), A New Look at the Semantic Web. Communications of the ACM 59(9), 35-37.

Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., and Veninata, C. (2019), ArCo: The Italian Cultural Heritage Knowledge Graph, Proceedings of The Semantic Web Conference – ISWC 2019., LNCS vol 11779, pp. 36-52. Cham, Springer.

European Commission (2020), Horizon 2020, Work Programme 2018-2020: 13. Europe in a Changing World – Inclusive, Innovative and Reflective Societies. Luxembourg, Publication Office of the EU.

Isaac, A., *et al.*, eds. (2011), Europeana Data Model Primer. The Hague, Europeana Foundation (https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf, dernier accès le 18/12/2020).

Jahn, M. (2017), Narratology: A Guide to the Theory of Narrative (version 2.0). Cologne, English Department of the University (http://www.uni-koeln.de/~ame02/pppn.htm, dernier accès le 18/12/2020).

Le Boeuf, P., Doerr, M., Ore, C.E., and Stead, S., eds. (2018), Definition of the CIDOC Conceptual Reference Model (version 6.2.3). Heraklion, ICOM/CIDOC Documentation Standard Group.

Nilsson, M., Powell, A., Johnston, P., and Naeve, A. (2008), Expressing Dublin Core Metadata Using the Resource Description Framework, RDF (DCMI Recommendation 2008-01-04). Silver Spring (MD), Dublin Core Metadata Initiative (http://dublincore.org/documents/dc-rdf/, dernier accès le 18/12/2020).

Noy, N.F., Fergerson, R.W., Musen, M.A. (2000), The Knowledge Model of Protégé-2000: Combining Interoperability and Flexibility, Knowledge Acquisition, Modeling, and Management – Proceedings of EKAW 2000, LNCS 1937, pp. 17-32. Berlin, Springer.

Trame, J., Kessler, C., and Kuhn, W. (2013). Linked Data and Time – Modeling Researcher Life Lines by Events, Proceedings of the 11th International Conference on Spatial Information Theory, COSIT 2013, LNCS vol. 8116, pp. 205-223. Berlin, Springer.

Troncy, R., van Ossenbruggen, J., Pan, J.Z., and Stamou, G., eds., Halaschek-Wiener, C., Simou, N., and Tzouvaras, V., contributors (2007), Image Annotation on the Semantic Web, W3C Incubator Group Report 14 August 2007 (https://www.w3.org/2005/Incubator/mmsem/XGR-image-annotation/, dernier accès le 18/12/2020).

Zarri, G.P. (2009), Representation and Management of Narrative Information – Theoretical Principles and Implementation, London, Springer.

Zarri, G.P. (2010), A Conceptual Methodology for Dealing with Terrorism "Narratives", International Journal of Digital Crime and Forensics 2(2), 47-63.

Zarri, G.P. (2011), Knowledge Representation and Inference Techniques to Improve the Management of Gas and Oil Facilities. Knowledge-Based Systems 24(7), 989-1003.

Zarri, G.P. (2013), Generalized World Entities as a Unifying IoT Framework: A Case for the GENIUS Project, Internet of Things and Inter-Cooperative Computational Technologies for Collective Intelligence, Studies in Computational Intelligence vol. 460, pp. 345-367. Berlin, Springer.

Zarri, G.P. (2014a), Sentiments Analysis at Conceptual Level Making Use of the Narrative Knowledge Representation Language, Neural Networks 58(1), 82-97.

Zarri, G.P. (2019), Functional and Semantic Roles in a High-Level Knowledge Representation Language. Artificial Intelligence Review 51(4), 537-575.

Vers un système de crowdsourcing pour la transcription des cahiers de fouille d'archéologues

Christophe Tufféry*, Claudia Marinica**
Maximilien Rioult**, Yulin Xie**

* Inrap, 121 rue d'Alésia 75685 Paris Cedex, Doctorant CY Cergy Paris Université
christophe.tuffery@inrap.fr
**LS2N UMR 6004, équipe DUKe, Polytech Nantes, Université de Nantes, France
Claudia.Marinica@univ-nantes.fr, {maximilien.rioult, yulin.xie}@etu.univ-nantes.fr

1. Résumé

Ce papier se situe dans le cadre d'un travail en cours de thèse par le projet[1] qui se propose d'étudier les effets du numériques sur l'archéologie comme discipline depuis une quarantaine d'années ainsi que sur les pratiques et les identités des archéologues. Dans ce cadre, l'étude de de carnets de terrain d'archéologues et de comprendre le rôle du numérique dans cette évolution. Ce papier s'intéresse à explorer les possibilités de transcription de carnets de terrain des années 1970-1980 sur un chantier archéologique en France.

La documentation concernée se compose de cahiers au format A4, manuscrits, rédigés par plusieurs des fouilleurs du chantier. A côté de zones de texte manuscrits, concernant le déroulement de la fouille, les observations, les découvertes, les choix, les hypothèses d'interprétation, mais aussi des commentaires sur la météorologie, ou des blagues ou remarques humoristiques, se trouvent des croquis, des photographies annotées, et des éléments plus originaux comme des cartes postales ou des extraits de documents sur support papier découpés et collées.

Le projet de numérisation et transcription de ce type de documentation rejoint d'autres expérimentations ou projets en cours dans le domaine comme par exemple le projet *Bulliot, Bibracte et moi*[2], dirigé par l'EPCC de Bibracte, concernant les carnets de terrain de Gabriel Bulliot datant de la fin du XIXème siècle (Dépalle et Girard 2019).

Pour résoudre le problème de transcription des carnets de terrain concernés ici, une solution envisagée, et déjà mise en place dans certains projets, correspond au développement d'une plateforme de *crowdsourcing* qui permettra aux experts du domaine de transcrire les pages des carnets.

Le *crowdsourcing* (ou production participative) décrit une forme d'externalisation d'une activité composée d'un ensemble de tâches. Ce concept a été introduit par How (2008) qui le définit comme "*l'acte d'accepter un travail traditionnellement exercé par un agent désigné et de l'externaliser à un groupe de personnes non défini, généralement sous la forme d'un appel ouvert*". Ainsi, deux types d'acteurs participent dans ce processus : le demandeur qui organise les activités à externaliser, et les contributeurs qui proposent d'effectuer un ensemble des tâches des activités proposées par le demandeur.

---

[1] La thèse est réalisée à Cergy Paris Université, dans le cadre de l'EUR *Humanités, Création, Patrimoine*, et en partenariat avec l'Institut National du Patrimoine.
[2] https://bbm.hypotheses.org/

La mise en place d'une plateforme de *crowdsourcing* est organisée en plusieurs étapes : (1) l'organisation en tâches de l'activité proposée ; (2) l'évaluation des résultats ; (3) le développement d'un mécanisme d'incitation et (4) la gestion des contributeurs. Parmi ces 4 étapes, l'organisation des tâches est de loin la plus importante et celle qui impacte le plus les résultats obtenus. Ainsi, les tâches peuvent être de différents types (inventives, routinières, axées contenu), et organisées selon différentes stratégies (organisation séquentielle, parallèle ou *divide-and-conquer*).

Pour la transcription des carnets de terrain, la complexité de cette réalisation est amenée par la diversité du contenu de ces manuscrits :contexte de la fouille, météo, activités du jour, l'état d'avancement de la fouille, les noms des participants, émotions des fouilleurs, citations humoristiques, des croquis, des schémas, des dessins, etc. La grande divers de ces éléments permet de d'entrevoir la complexité du processus de crowdsourcing à mettre en place.

Dans la littérature, différentes plateformes pour la transcription de documents manuscrits ont été déjà mises en place. Ces plateformes sont basées sur le principe de *crowdsourcing*, ou bien elles intègrent des techniques de reconnaissances automatique. Les plateformes comme Transcrire[3], demandent aux contributeurs de prendre en charge un certain nombre de tâches. Le but de cette plateforme est de transcrire plusieurs collections de notes manuscrites, de carnets de terrain et de journaux. Elle permet aux contributeurs de transcrire du texte, croquis, photos, cartes, et sections illisibles, mais sans proposer une catégorisation spécifique de ces éléments. D'un autre côté, des plateformes comme eScriptorium[4] et Transkribus[5] (Dépalle et Girad, 2019) proposent l'automatisation du processus de transcription de documents manuscrits en utilisant des techniques d'apprentissage automatique, comme l'apprentissage profond.

Pour la mise en place du système de *crowdsourcing* pour la transcription envisagée ici, il s'avère impossible de réutiliser des plateformes déjà développées car elles ne permettent pas aujourd'hui la catégorisation des éléments retranscrits d'une manière suffisamment fine. De plus, la comparaison des deux outils de développement des plateformes de *crowdsourcing*, Pybossa[6] et Scribe[7], a permis de souligner que le second est plus adapté aux besoins du projet car il permet une meilleure configurabilité de la conception et de la gestion de tâches.

La plateforme est actuellement en cours de développement et devrait être finalisée au printemps 2021.

Références

J. Howe. Crowdsourcing: How the Power of the Crowd is Driving the Future of Business. Business & Internet. Random House Business, 2008.

C. Dépalle et J-P. Girard. "Bulliot, Bibracte et moi" une expérience de sciences participative en archéologie. Culture et Recherche, « Recherche culturelle et sciences participatives », n°140, hiver 2019-2020, p. 78. En ligne [http://www.culture.gouv.fr/Sites-thematiques/Enseignement-superieur-et-Recherche/La-revue-Culture-et-Recherche/Recherche-culturelle-et-sciences-participatives]

[3] http://transcrire.huma-num.fr/
[4] https://www.escriptorium.uk/
[5] https://readcoop.eu/transkribus/
[6] https://pybossa.com/
[7] https://scribeproject.github.io/