Atelier DAHLIA

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

Organisateurs :

Claudia Marinica (MIDI - ETIS, Université de Cergy-Pontoise)
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 (EGC2019)

le 22 janvier 2019 à Metz

Editeurs :

Claudia Marinica
Laboratoire ETIS - Université de Cergy-Pontoise, Institut IDHN
page web : https://perso-etis.ensea.fr//marinica/
email : claudia.marinica@u-cergy.fr

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

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

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

Accès en ligne :

Atelier DAHLIA : http://dahlia.egc.asso.fr/atelierDAHLIA-EGC2019.html
Groupe de travail DAHLIA : http://dahlia.egc.asso.fr

Membres du comité de programme

Catherine Faron, I3S, Université de Nice Sophia Antipolis

Jean-Gabriel Ganascia, LIP6, UPMC

Fabrice Guillet, LS2N, Polytech'Nantes

Pascale Kuntz, LS2N, Polytech’Nantes

Florent Laroche, LS2N, Ecole Centrale de Nantes

Julien Longhi, AGORA, IDHN, Université de Cergy-Pontoise

Sabine Loudcher, ERIC, Université de Lyon 2

Jean-Philippe Magué, ENS Lyon

Claudia Marinica, ETIS, IDHN, Université de Cergy-Pontoise

Françoise Paquienséguy, ELICO, Sciences po Lyon

Julien Velcin, ERIC, Université Lyon 2

Gilles Ventuniri, LIFAT, Université de Tours

Table de matières

Étude des quartiers : défis et pistes de recherche
Loic Bonneval, Fabien Duchateau, Franck Favetta, Aurélien Gentil, Mohamed Nader Jelassi, Maryvonne Miquel et Ludovic Moncla…… 1

Modélisation 3D urbaine et historique de la ville de Charleville
Sylvain Rassat et François-Joseph Ruggiu…….. 12

Remonter le temps pour comprendre le passé : l'immersion virtuelle au service des historiens
Paul François, Florent Laroche, Françoise Rubellin et Jeffrey Leichman…….. 15

Harmonisation de l'acquisition des données d'opérations d'archéologie préventive. Retours d'expériences et perspectives à partir de l'application EDArc
Christophe Tuffery et Stéphane Augry……. 21

Les événements dans l'ontologie CRMCR dédiée à la conservation et la restauration des oeuvres d'art
Claudia Marinica, Inès Bannour, Luc Bouiller et Olivier Malavergne…… 28

Enrichissement sémantique de données d'archives sonores d'ethnomusicologie par alignement
Nedra Mellouli et Aude Julien Da Cruz Lima…….. 35

Identification automatique des sources des notices zoologiques du Speculum naturale de Vincent de Beauvais
Etienne Cuvelier, Sébastien de Valeriola et Céline Engelbeen…… 40

Extraction et formalisation du savoir-faire industriel
Meriem Mejhed Mkhinini, Ouassila Narsis et Christophe Nicolle…….. 42

Apport du Text Mining pour l'exploration de relations dans les textes. Application à la découverte d'appariements entre descriptions textuelles d'objets d'intérêt et localisations dans la Tapisserie de Bayeux
David Condaminet, Antoine Widlöcher, Bruno Crémilleux, Pierre-Yves Buard et Julia Roger…… 44

Le repas gastronomique des Français comme patrimoine culturel immatériel de l'humanité : caractérisation et transmission à travers les tweets des chefs 2 et 3 étoiles au Guide Michelin
Julien Longhi, Zakarya Després, Claudia Marinica, Vincent Marcilhac et Felipe Diaz Marin……. 56

HeritaMus: a machine of representation of actors networks settling the "parliament of things" in a cultural heritage context
Nedra Mellouli et Pedro Felix……... 64

Projet Frénaud 2018 : constitution et exploitation d'une base de données
Marianne Froye et Zakarya Després……. 70

Valorisation de récits de vie de Républicains espagnols
Catherine Dominguès, Laurence Jolivet et Carmen Brando…… 83

Étude des quartiers : défis et pistes de recherche

Loïc Bonneval*, Fabien Duchateau**, Franck Favetta**, Aurélien Gentil*,
Mohamed Nader Jelassi**, Maryvonne Miquel**, Ludovic Moncla**

*Centre Max Weber, Université de Lyon, France
prénom.nom@univ-lyon2.fr
**LIRIS UMR5205, Université de Lyon, France
prénom.nom@liris.cnrs.fr

Résumé. Le projet Home In Love (HiL) s'intéresse à la recommandation de biens immobiliers, en particulier dans le cas où l'on ne connait pas sa future ville de résidence (e.g., mutation professionnelle). Si le choix d'un logement est facilité par les nombreuses ressources disponibles (e.g., sites web avec photos, visites virtuelles), cela reste compliqué de se faire une idée concrète des quartiers où se trouvent les logements disponibles. L'un des enjeux concerne donc la description et la comparaison de quartiers selon les domaines d'application (e.g., recherche immobilière, étude sociale, recensement du patrimoine). Cet article décrit les défis et les pistes de recherche (en informatique) liés à cette étude des quartiers.

1 Introduction

Le projet pluridisciplinaire Home In Love [1] (HiL) a pour objectif la recommandation de biens immobiliers pour des personnes ne connaissant pas ou peu leur future ville de résidence. Dans cette optique, le travail de caractérisation des quartiers potentiellement recommandés s'avère central.

Durant ces dernières décennies, les quartiers ont fait l'objet d'une attention particulière en sciences humaines et sociales (Authier et al. (2007)). Les cas d'utilisation liés au quartier occupent une place importante à l'ère des *"smart cities"*, avec des projets de simulation pour l'efficacité énergétique (Perez et al. (2016)), de gestion du patrimoine (Devernois et al. (2014)), d'étude d'impact de l'environnement sur la santé (Morland et al. (2002)) et la mobilité (Chaix (2014)), ou encore de reconstitution numérique du patrimoine (Pardoen (2015)). Par rapport au contexte du projet HiL, on peut citer des applications centrées sur la recommandation de quar-

Ce travail a été réalisé au sein du LABEX IMU (ANR-10-LABX-0088) de l'Université de Lyon, dans le cadre du programme "Investissements d'Avenir" (ANR-11-IDEX-0007) de l'Etat Français, géré par l'Agence Nationale de la Recherche (ANR).

1. http://imu.universite-lyon.fr/projet/hil

tiers comme Kelquartier [2], Cityzia [3], ville-ideale [4], vivroù [5] ou encore le site DataFrance [6] qui intègre des indicateurs fournis par l'INSEE et par d'autres sources (e.g., IGN, L'Express). Pour étudier ces quartiers, il est important de les décrire avec précision et de pouvoir les comparer.

La notion de quartier (i.e., définition, délimitations, perception) est plurielle et sujette à différentes interprétations (Authier et al. (2007)), ce qui rend complexe leur utilisation pour un domaine d'application donné. Des projets s'attachent à décrire les quartiers d'une ville selon un point de vue, par exemple celui de ses habitants (Authier (2008)). Généralement, ces descriptions de quartiers se concentrent sur quelques villes, et ne permettent pas un traitement automatique à plus grande échelle. De même, la délimitation des quartiers n'est pas toujours disponible, et de nombreux travaux s'intéressent à la détection automatique de ces « frontières à l'interprétation variable ». Un défi important concerne donc la méthodologie à adopter pour décrire des quartiers et leur délimitation. Dans de nombreux cas d'utilisation, comme le regroupement de quartiers similaires, la classification des quartiers selon leur fonction, ou la recommandation, il est utile de comparer des quartiers, i.e., de déterminer si deux quartiers possèdent des caractéristiques semblables. Ce défi repose généralement sur l'utilisation d'un processus ou d'un algorithme de comparaison qui exploite la description des quartiers. Enfin, il est crucial de vérifier, en particulier dans le cas des traitements automatisés, la qualité des résultats produits par rapport à la réalité. Par exemple, on souhaite contrôler qu'un algorithme de regroupement ait bien associé des quartiers similaires ou vérifier qu'un algorithme de recommandation ait proposé des quartiers pertinents à un utilisateur. Pour cela, il faut réfléchir à l'évaluation et notamment la manière de valider les résultats, ce qui peut se révéler très complexe au vu de la perception et du flou qui entourent ces quartiers.

Dans cet article, nous dressons un panorama de l'étude des quartiers sous un angle informatique, en identifiant et décrivant les pistes de recherche principales que sont la description, la délimitation, la comparaison et l'évaluation en lien avec les quartiers.

2 Pistes de recherche

Dans cette section, nous décrivons quatre pistes de recherche que nous considérons comme cruciales pour l'étude des quartiers : la description, la délimitation, la comparaison et l'évaluation.

2.1 Description de quartiers

Afin de caractériser les quartiers, une première étape consiste à déterminer les informations pertinentes permettant leur description (pour un domaine d'application). Cette sélection nécessite des échanges entre les différents acteurs (chercheurs en SHS, informaticiens, etc.) afin d'établir la liste des données utiles pour l'application. Par exemple, l'INSEE quadrille le territoire français en IRIS [7] (Ilots Regroupés pour l'Information Statistique). Il existe plusieurs types d'IRIS, ceux d'habitation (entre 2000 et 5000 résidents), ceux d'activités (e.g.,

2. http://www.kelquartier.com/
3. http://www.cityzia.fr/
4. http://www.ville-ideale.fr/
5. http://www.vivrou.com/
6. http://datafrance.info/
7. http://www.insee.fr/fr/metadonnees/definition/c1523

un campus universitaire) et ceux divers peu habités (e.g., un parc). L'INSEE fournit pour chacun des 50 000 IRIS de nombreux indicateurs (e.g. le nombre d'épiceries, la répartition par type de logement ou par catégorie socio-professionnelle). Certaines données seront agrégées (e.g., somme du nombre de boulangeries et du nombre d'épiceries pour représenter les "commerces de proximité"). Si l'on compare des quartiers très différents, il faut réfléchir à la normalisation des données (e.g., selon leur superficie, leur population, leur densité). Les données choisies peuvent donc être traitées de façon absolue (e.g., nombre de boulangeries dans le quartier) ou relatives (e.g., densité de commerces par rapport à celle du reste de l'agglomération). L'avantage de cette deuxième option est de rapprocher des quartiers d'agglomération différentes mais ayant une position comparable (e.g., un « quartier central »n'aura pas les mêmes caractéristiques à Lyon ou dans une ville moyenne, mais pourra avoir le même « rôle »ou fonction dans les deux cas). Une tendance récente concerne la collecte de données des réseaux sociaux, comme les *tweets*, les *likes* ou les *checkins* qui sont ou peuvent être localisés. Un corpus d'images habilement constitué peut également servir à représenter un quartier (Kennedy et Naaman (2008)). Enfin, le choix des informations pertinentes peut être actualisé et est dépendant de la disponibilité des données.

En effet, une fois les données identifiées, une deuxième difficulté est la recherche et la disponibilité des sources de données. Le web facilite aujourd'hui cette étape, et le mouvement "Données Ouvertes" (e.g., `data.gouv.fr`) permet de disposer de nombreux jeux de données. Cependant, il convient de vérifier la qualité de ces données (e.g., provenance, dates de mise à jour, utilisation par d'autres applications). Les jeux de données à caractère sensible ou ayant une potentielle valeur monétaire peuvent être plus difficiles à obtenir. Par exemple, pour la recherche immobilière, une information utile est le prix de vente (ou de location) d'un bien. Or cette donnée sur le prix est rarement disponible ou reste incomplète (e.g., uniquement fournie pour les plus grandes villes), ou elle n'a pas le bon niveau de granularité (e.g., au niveau d'un département, mais pas d'une ville ou du quartier). Sur les aspects techniques, certaines API peuvent être indisponibles à certaines périodes ou avoir des limitations en terme de requêtes. Enfin, les sources de données s'accompagnent d'une licence d'utilisation : des restrictions peuvent donc empêcher leur utilisation.

Le troisième problème est l'intégration de ces données hétérogènes pour faciliter leur exploitation (Halevy et al. (2006)). Les concepts qui décrivent les données sont habituellement comparés en utilisant des techniques d'appariement de schémas ou d'ontologies (Bellahsène et al. (2011)). Pour les données elles-mêmes, le *record linkage* ou appariement de données permet de détecter les informations équivalentes et donc d'éviter la redondance (Christen (2012)). Des données plus complexes (e.g., textuelles, multimédia) utiliseront des outils de détection d'entités nommées (Shen et al. (2015)), d'extraction d'informations (e.g., spatio-temporelles dans Strötgen et al. (2010)) ou des annotateurs automatiques (Zhang et al. (2012)). Dans le cas des données sur les IRIS de l'INSEE, les indicateurs sont fournis dans des dizaines de fichiers, qui ne sont pas tous organisés de la même manière (interprétation différente des concepts, hétérogénéité des libellés, regroupement ou division d'IRIS, etc.). Les solutions pour l'intégration peuvent être manuelles (e.g., saisie des données dans un tableur, codage d'un script) ou automatisées en utilisant des outils tels qu'OpenRefine [8], Talend [9], Karma ((Gupta et al., 2012)),

8. `http://openrefine.org/`
9. `http://fr.talend.com/products/data-integration/`

BigGorilla (Chen et al. (2018)). La description des quartiers à partir de différentes sources agrégées doit donc être stockée dans un format approprié (e.g., GeoJSON).

2.2 Délimitation des quartiers

Comme indiqué précédemment, la définition d'un quartier n'est pas fixée et elle dépend du contexte (e.g., économique, historique, politique) et de la perception (e.g., point de vue de l'administration, des habitants). La délimitation d'un quartier (e.g., liste de coordonnées géographiques formant un polygone) fait partie de sa description, mais nous la considérons à part car c'est un défi particulièrement complexe et qui apparaît comme optionnel dans certains cas d'utilisation. Selon les données disponibles, la délimitation d'un quartier consiste à identifier son contour en :

- Se basant sur les définitions de l'administration ;
- Regroupant des zones plus petites (e.g., les IRIS) pour former un quartier, ou en divisant une zone (e.g., un arrondissement, un grand IRIS) en plusieurs quartiers ;
- Exploitant des cartes ou des systèmes d'information géographique (SIG) ;
- Exploitant des sources semi-structurées, textes, ou images (aériennes) ;
- Exploitant les *likes* et *checkins* de réseaux sociaux comme Foursquare, ;
- Réalisant des enquêtes auprès des populations.

L'administration met aujourd'hui à disposition certaines informations sur les quartiers tel que le contour cartographique. C'est par exemple le cas dans le cadre de la définition des quartiers prioritaires de la politique de la ville [10]. Nous pouvons également citer des communes (e.g., Sèvres, Nanterre, Saint-Quentin) qui mettent à disposition les informations de délimitations de leurs quartiers sur `data.gouv.fr`. La méthodologie détaillant le découpage n'est pas toujours accessible, et ces données ne prennent en compte que le point de vue administratif. Mais ces informations peuvent servir de base à une étude ou à un comparatif par rapport à d'autres perceptions.

Un quartier peut résulter d'un regroupement ou d'une division d'autres unités géographiques. Dans de nombreux travaux, un IRIS est tout simplement assimilé au quartier. Par exemple, dans Préteceille (2009), l'IRIS est considéré comme l'unité spatiale la plus pertinente pour étudier la ségrégation car il correspond mieux au quartier vécu des habitants. Cependant, selon l'objectif de la recherche, cette simplification n'est pas toujours opératoire. En effet, cette vision s'oppose à celle des travaux de Maurin (2004), dans lesquels un découpage plus fin est préconisé (e.g., similaire à l'ancienne unité géographique appelée îlot [11] et définie par l'INSEE comme « *un pâté de maison en zone dense ou un ensemble limité par des voies en zone périphérique* »). Dans Barret et al. (2019), les auteurs utilisent également les IRIS mais prennent en compte le voisinage (IRIS adjacents à un IRIS) pour lisser les données, caractériser plus finement les espaces étudiés et s'approcher au mieux de la notion de quartier. Enfin, il est possible de regrouper ou de diviser des unités géographiques qui ne sont pas adaptés au domaine d'application. Dans Actif et al. (2013), des « grands quartiers »sont ainsi créés à partir des IRIS, mais la méthode n'est pas discutée. Le problème avec cette réorganisation des unités géographiques est la validité des données, qui ne sont donc plus au même niveau et peuvent éventuellement introduire des biais dans les analyses.

10. `http://www.data.gouv.fr/fr/datasets/5a561801c751df42d7fca9b6/`
11. `http://www.insee.fr/fr/metadonnees/definition/c1656`

Plusieurs fournisseurs géographiques proposent de visualiser les quartiers. Chez Google Maps, des noms de quartiers avec leur contour apparaissent dans les villes, mais la méthodologie utilisée pour les déterminer n'est pas décrite. Les noms de quartiers sont également présents sur Bing Maps ou Here Maps, mais seul un point représente le quartier, ce qui est insuffisant. OpenStreetMap [12] et Wikimapia [13] offrent pour certaines zones un découpage en quartiers. Comme ces sites sont collaboratifs, la délimitation est de fait subjective et potentiellement incomplète à l'échelle d'une ville. Dans la base de données géographique Geonames [14], les quartiers sont généralement présents à travers des unités géographiques plus importantes. Par exemple, le quartier « *Serin* »à Lyon est associé au quatrième arrondissement, car il est cité dans la page Wikipédia de cet arrondissement. De plus, la visualisation d'un quartier (sur un fond de carte Google Maps) est très approximative car représentée par un rectangle.

Dans les sources de données semi-structurées, Wikipédia figure parmi les sites les plus utilisés. On y trouve une page [15] « catégorie : quartier de ville en France ». Mais la délimitation du quartier accompagne rarement ces descriptions, qui de plus ne concernent que quelques dizaines de villes. Le « Linked Open Data » [16] relie sémantiquement les entités de nombreux jeux de données. Dans la base de connaissances Wikidata, les quartiers sont de type « district » [17], dont la définition trop générale (« type de division administrative existant dans certains pays, de tailles variables allant du quartier à la région ») illustre là encore la difficulté d'uniformiser cette notion de quartier et ne permet pas d'exploiter cette source de données sans traitement spécifique. Les sources de données textuelles sont largement disponibles sur le web, en particulier des sources touristiques ou patrimoniales. Par exemple, le site `http://www.patrimoine-lyon.org/` décrit les (vieux) quartiers de Lyon avec de nombreux détails spatiaux et cartographiques, mêlant textes, cartes et photographies. Une façon d'exploiter de telles sources est détaillée dans Brindley et al. (2014), où des noms de quartiers sont extraits d'un grand volume de documents, et reliés ensuite à des codes postaux. Cela permet d'étudier l'évolution du contour des quartiers au fil du temps. Pour exploiter des documents textuels, il est nécessaire d'appliquer des méthodes de traitement automatique du language (TAL), notamment pour l'extraction d'informations géographiques (e.g., lieux, bâtiments, monuments, rues), comme le soulignent Miller et Han (2009); Leidner et Lieberman (2011). Les annonces immobilières ainsi que les offres de location provenant de sites tels que AirBnB [18] comportent parfois des descriptions détaillées du quartier, avec le point de vue d'un expert ou d'un habitant (Guérois et Madelin (2017)). Dans le rapport de Tang et Sangani (2015), il est question de prédire le quartier et le prix d'une annonce (pour la ville de San Francisco), et il est envisageable de déterminer les limites d'un quartier en fonction de ces prédictions. Enfin, les images aérienne et/ou satellite sont l'objet de recherches pour la détection automatique d'objets (e.g., véhicules) ou de phénomènes (e.g., déforestation, glissement de terrain). Les bâtiments peuvent être détectés et classés par type en les comparant à des exemples annotés (Du et al. (2015)) ou en exploitant les images « Street View »(Kang et al. (2018)). La détection de quartiers à partir d'images reste un défi majeur, mais des progrès sont

12. `http://www.openstreetmap.org/`
13. `http://wikimapia.org/`
14. `http://www.geonames.org/`
15. `http://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Quartier_de_ville_en_France`
16. `http://linkeddata.org/`
17. `http://www.wikidata.org/wiki/Q149621`
18. `http://www.airbnb.fr`

réalisés, par exemple pour détecter des villages « absorbés »par l'expansion rapide des grandes villes chinoises (Huang et al. (2015)).

Une autre piste pour détecter les contours est l'exploitation des données de géolocalisation issues de réseaux sociaux, mouvement qui a connu un engouement certain depuis une quinzaine d'années. Avec Livehoods [19], la délimitation de quartiers est possible en analysant les *checkins* des réseaux sociaux (Cranshaw et al. (2012)). Dans Hoodsquare (Zhang et al. (2013)), les activités des habitants sont analysées et associées à l'une des 300 catégories de la hiérarchie de Foursquare pour en déduire le ou les raisons de fréquenter un quartier, et d'en déterminer les contours. Les aspects temporels (e.g., période la journée) ainsi que les activités des touristes sont également pris en compte. À Beijing, les fonctions d'une zone géographique (e.g., résidentiel, éducation, diplomatique, culturel ou historique) sont découvertes en considérant ce problème comme la découverte de thématiques pour un document (Yuan et al. (2012)). Les zones sont regroupées en fonction de la distribution des points d'intérêt (*via* un algorithme de *clustering*), et les activités humaines de chaque groupe servent à identifier sa ou ses fonctions principales. Comme l'indiquent les auteurs de ces travaux, les données peuvent avoir un biais (e.g., plus forte probabilité de *liker* un bar que son lieu de travail).

Les enquêtes auprès de populations sont la dernière piste pour identifier les contours de quartiers. Des travaux permettent à l'utilisateur de dessiner sur une carte interactive son quartier (ou « région ») d'intérêt, et différentes informations sont alors affichées comme les activités principales ou points d'intérêt représentatifs ((Kumar et al., 2015)). Les habitants peuvent aussi décrire les points d'intérêt représentatifs de leur quartier afin d'en déduire ses frontières (Berjawi et al. (2013)). Pour obtenir des résultats valides, et en particulier des délimitations « sans trous », il faut récolter l'avis de nombreuses personnes. Côté sciences humaines de sociales, il y a peu d'enquête systématique sur ce thème. Seuls les travaux de Pan Ké Shon (2005) exploitent une question de l'enquête permanente « *condition de vie de 2001* », menée par l'INSEE (la question étant "pouvez-vous me dire en quelques mots ce que représente votre quartier pour vous ?").

2.3 Comparaison de quartiers

Une troisième piste de recherche concerne la comparaison de quartiers. En effet, de nombreux cas d'application ont besoin d'établir la ressemblance entre quartiers. Ce processus repose essentiellement sur la description des quartiers, dont les données serviront à mesurer un degré de similarité.

Dans un premier temps, les données utiles à la comparaison doivent être sélectionnées parmi toutes celles qui caractérisent les quartiers. Le voisinage d'un quartier peut être exploité, par exemple pour estimer le climat urbain ou la co-occurrence d'activités dans une zone.

Différentes techniques permettent d'établir des comparaisons entre objets. La similarité cosine et la mesure de Jaccard sont les algorithmes les plus connus pour réaliser cette opération. Ils permettent de calculer directement le degré de ressemblance entre deux quartiers décrits comme des vecteurs de valeurs (Yu et al. (2016) et Zhang et al. (2013)). La distance *Earth mover* (EMD) mesure l'effort pour "transformer" un quartier en un autre (Le Falher et al. (2015)).

19. `http://livehoods.org/`

Pour comparer des quartiers, il est également possible de les regrouper, par exemple en utilisant des algorithmes de partitionnement ou de *clustering*. Les algorithmes de partitionnement comme KMeans, Affinity Propagation ou Spectral Clustering nécessitent de spécifier le nombre de groupes. Au contraire, les algorithmes de *clustering* comme DBSCAN et ses alternatives estiment automatiquement le nombre de groupes, mais requièrent un paramètre ϵ représentant la distance au-dessus de laquelle de nouveaux groupes sont créés. Les auteurs de Livehoods se basent sur l'algorithme *spectral clustering* pour comparer des quartiers (Cranshaw et al. (2012)).

Le *case-based reasoning* détecte des cas similaires pour rapprocher des quartiers. Par exemple, la situation d'une personne (e.g., composition du ménage, distance maison-travail) à la recherche d'une résidence est analysée pour proposer des quartiers où vivent des résidents dans une même situation (Yuan et al. (2013)).

Des algorithmes de classification peuvent être utilisés, en particulier avec des données issues des réseaux sociaux. Les travaux de Le Falher et al. (2015) utilisent *Information Theoretic Metric Learning* (ITML) et *Large Margin Nearest Neighbor* (LMNN) pour déterminer une matrice contenant les lieux les plus proches à partir des activités humaines réalisées dans ces lieux (vecteurs d'entrée) et des catégories de Foursquare (classes).

2.4 Évaluation

Pour certaines études, par exemple sociologiques, le processus informatique sert souvent d'outil dont le résultat (i.e., tableaux, visualisation, diagrammes) est ensuite analysé et interprété par des experts. Dans d'autres cas, le résultat produit par un processus informatique, en particulier la délimitation du quartier et la comparaison, a besoin d'être vérifié et validé, parfois de manière automatique.

Les enquêtes, déjà mentionnées en section 2.2, sont un moyen répandu pour l'évaluation, que ce soit sous forme de questionnaire ou d'entretien. Par exemple, pour mesurer l'impact d'un quartier sur l'activité physique, des enquêtes ont été envoyées à une centaine de résidents de deux quartiers de San Diego (Saelens et al. (2003)). Dans Lovejoy et al. (2010), les enquêtes permettent d'évaluer le degré de satisfaction des résidents vis à vis de leur quartier. Dans l'article de Yuan et al. (2012), des habitants de Beijing (au moins six ans de résidence) sont questionnés pour annoter des zones géographiques avec leur fonctions d'utilité. Cette expertise est ensuite comparée aux résultat de l'algorithme. Le web facilite aujourd'hui la création et le traitement d'enquêtes type questionnaire, et permet de toucher un plus grand nombre d'individus.

Pour une évaluation automatique, un résultat expertisé ou estimé est requis. Par exemple, pour évaluer la comparaison de quartiers, il est possible de disposer d'un jeu de données manuellement rempli qui liste les quartiers similaires. L'évaluation de la délimitation est particulièrement complexe du fait de la définition floue du quartier. Un cas d'utilisation plus précis peut permettre de trouver un mode d'évaluation adapté et fiable. L'évaluation automatique est aussi rendue possible par le micro-travail. En effet, on peut disposer de jeux de données annotés ou vérifiés par des personnes rémunérées grâce à des outils en ligne comme Amazon Mechanical Turk (Kittur et al. (2008)). Les villes développent parfois des planifications urbaines, et dont les informations peuvent être utiles pour l'évaluation. À Belo Horizonte, une méthodologie fournit des indicateurs sur la qualité de vie d'une dizaine de quartiers. Le nombre de services culturels, de santé, d'environnement, etc. est donc connu et mis à jour régulièrement.

Les travaux de Smarzaro et al. (2017) cherchent à vérifier si les fournisseurs cartographiques (Facebook, Foursquare, Google Places et Yelp) confirment ces statistiques expertisées. Dans les travaux de Yuan et al. (2012), l'algorithme d'annotation des zones géographiques est comparée à la planification urbaine de Beijing. Bien que les deux cartes se recoupent effectivement à certains endroits, une étude plus approfondie sur l'ensemble de la ville est nécessaire.

3 Conclusion et perspectives

Dans cet article, nous avons dressé un panorama des pistes de recherche informatique pour l'étude des quartiers en sciences humaines et sociales : description, délimitation, comparaison et évaluation. Ces grands axes sont utilisables dans de nombreux cas d'utilisation, que ce soit le regroupement ou la classification de quartiers, la recommandation, ou la prédiction de l'évolution d'un quartier.

Le projet HiL s'intéresse à plusieurs de ces pistes avec l'objectif de recommander le quartier idéal lors d'une recherche immobilière. Pour la description des quartiers, nous nous basons sur les données IRIS que nous agrégeons lorsque celles-ci renvoient à des découpages trop fins. La prise en compte du voisinage d'un IRIS permet d'étendre sa description en considérant les caractéristiques situées à proximité. Dans les grandes villes (où les IRIS sont souvent petits), le voisinage se rapproche d'une forme de quartier. En parallèle, nous étudions la possibilité de délimiter des quartiers en regroupant des IRIS adjacents partageant des descriptions similaires. Enfin, la comparaison de deux IRIS est essentielle pour la recommandation, et nous avons expérimenté quelques algorithmes (e.g., mesure cosine, DBSCAN) que nous devons affiner en exploitant les profils utilisateur. L'évaluation et la justification des recommandations sont des problèmes encore ouverts.

Références

Actif, N., A. Levet, S. Hoarau, H. Maillot, F. Andy, M. Boyer, C. Calteau, L. Trentin, et C. Ory (2013). Des quartiers inégaux face à la précarité. In *Cartographie sociale des territoires*. Insee.

Authier, J.-Y. (2008). Les citadins et leur quartier. *L'Année sociologique 58*(1), 21–46.

Authier, J.-Y., M.-H. Bacqué, et F. Guérin-Pace (2007). *Le quartier*. La Découverte.

Barret, N., F. Duchateau, F. Favetta, M. Miquel, A. Gentil, et L. Bonneval (2019). À la recherche du quartier idéal. In *EGC (à paraitre)*.

Bellahsène, Z., A. Bonifati, et E. Rahm (2011). *Schema matching and mapping*. Springer.

Berjawi, B., M. Colomb, T. Joliveau, F. Favetta, F. Duchateau, et M. Miquel (2013). Outil de repérage urbain à travers la prise de points de repère. Prototype, Laboratoires EVS et LIRIS.

Brindley, P., J. Goulding, et M. L. Wilson (2014). A data driven approach to mapping urban neighbourhoods. In *SIGSPATIAL*, pp. 437–440. ACM.

Chaix, B. (2014). Quartiers, mobilité et santé : l'étude record. *Les Cahiers de l'IAU*, 170–171.

Chen, C., B. Golshan, A. Y. Halevy, W.-C. Tan, et A. Doan (2018). BigGorilla : An Open-Source Ecosystem for Data Preparation and Integration. *IEEE Data Eng. Bull. 41*(2), 10–22.

Christen, P. (2012). *Data matching : concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

Cranshaw, J., R. Schwartz, J. Hong, et N. Sadeh (2012). The livehoods project : Utilizing social media to understand the dynamics of a city.

Devernois, N., S. Muller, et G. Le Bihan (2014). *Gestion du patrimoine urbain et revitalisation des quartiers anciens : l'éclairage de l'expérience française*. Agence française de développement (AFD).

Du, S., F. Zhang, et X. Zhang (2015). Semantic classification of urban buildings combining vhr image and gis data : An improved random forest approach. *ISPRS journal of photogrammetry and remote sensing 105*, 107–119.

Guérois, M. et M. Madelin (2017). Comment les hôtes et clients d'Airbnb parlent-ils des lieux ? Une analyse exploratoire à partir du cas parisien. In *EXCES-EXtraction de Connaissances à partir de donnéEs Spatialisées*.

Gupta, S., P. Szekely, C. A. Knoblock, A. Goel, M. Taheriyan, et M. Muslea (2012). Karma : A system for mapping structured sources into the semantic web. In *Extended Semantic Web Conference*, pp. 430–434. Springer.

Halevy, A., A. Rajaraman, et J. Ordille (2006). Data integration : the teenage years. In *VLDB '06 : Proceedings of the 32nd international conference on Very large data bases*, pp. 9–16. VLDB Endowment.

Huang, X., H. Liu, et L. Zhang (2015). Spatiotemporal detection and analysis of urban villages in mega city regions of china using high-resolution remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing 53*(7), 3639–3657.

Kang, J., M. Körner, Y. Wang, H. Taubenböck, et X. X. Zhu (2018). Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Kennedy, L. S. et M. Naaman (2008). Generating diverse and representative image search results for landmarks. In *Proceedings of the 17th international conference on World Wide Web*, pp. 297–306. ACM.

Kittur, A., E. H. Chi, et B. Suh (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 453–456. ACM.

Kumar, C., W. Heuten, et S. Boll (2015). Visual overlay on openstreetmap data to support spatial exploration of urban environments. *ISPRS International Journal of Geo-Information 4*(1), 87–104.

Le Falher, G., A. Gionis, et M. Mathioudakis (2015). Where Is the Soho of Rome ? Measures and Algorithms for Finding Similar Neighborhoods in Cities. *ICWSM 2*, 3–2.

Leidner, J. L. et M. D. Lieberman (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special 3*(2), 5–11.

Lovejoy, K., S. Handy, et P. Mokhtarian (2010). Neighborhood satisfaction in suburban versus traditional environments : An evaluation of contributing characteristics in eight California neighborhoods. *Landscape and Urban Planning 97*(1), 37–48.

Maurin, E. (2004). Le ghetto français. *Enquête sur le séparatisme social*.

Miller, H. J. et J. Han (2009). *Geographic data mining and knowledge discovery*. CRC Press.

Morland, K., S. Wing, A. D. Roux, et C. Poole (2002). Neighborhood characteristics associated with the location of food stores and food service places. *American journal of preventive medicine 22*(1), 23–29.

Pan Ké Shon, J.-L. (2005). La représentation des habitants de leur quartier : entre bien-être et repli. *Économie et statistique 386*(1), 3–35.

Pardoen, M. (2015). Les oreilles à l'affût ! restitution d'un paysage sonore : œuvre de l'imaginaire ou recherche d'authenticité. *Silences et bruits du Moyen Âge à nos jours : Perceptions, identités sonores et patrimonialisation. L'Harmattan*.

Perez, N., A. Mailhac, C. Inard, et P. Riederer (2016). Outil d'aide à la décision multicritère pour la conception de systèmes énergétiques à l'échelle du quartier. In *IBPSA France Conference*.

Préteceille, E. (2009). La ségrégation ethno-raciale a-t-elle augmenté dans la métropole parisienne ? *Revue française de sociologie 50*(3), 489–519.

Saelens, B. E., J. F. Sallis, J. B. Black, et D. Chen (2003). Neighborhood-based differences in physical activity : an environment scale evaluation. *American journal of public health 93*(9), 1552–1558.

Shen, W., J. Wang, et J. Han (2015). Entity linking with a knowledge base : Issues, techniques, and solutions. *Knowledge and Data Engineering, IEEE Transactions on 27*(2), 443–460.

Smarzaro, R., T. F. d. M. Lima, et C. A. Davis Jr (2017). Could data from location-based social networks be used to support urban planning ? In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 1463–1468. International World Wide Web Conferences Steering Committee.

Strötgen, J., M. Gertz, et P. Popov (2010). Extraction and exploration of spatio-temporal information in documents. In *Workshop on Geographic Information Retrieval*, pp. 16. ACM.

Tang, E. et K. Sangani (2015). Neighborhood and price prediction for san francisco airbnb listings.

Yu, M., G. Li, D. Deng, et J. Feng (2016). String similarity search and join : a survey. *Frontiers of Computer Science 10*(3), 399–417.

Yuan, J., Y. Zheng, et X. Xie (2012). Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 186–194. ACM.

Yuan, X., J.-H. Lee, S.-J. Kim, et Y.-H. Kim (2013). Toward a user-oriented recommendation system for real estate websites. *Information Systems 38*(2), 231 – 243.

Zhang, A. X., A. Noulas, S. Scellato, et C. Mascolo (2013). Hoodsquare : Modeling and recommending neighborhoods in location-based social networks. In *Social Computing*, pp. 69–74. IEEE.

Zhang, D., M. M. Islam, et G. Lu (2012). A review on automatic image annotation techniques. *Pattern Recognition 45*(1), 346–362.

Summary

The French project Home In Love (HiL) aims at recommending real estates (for buying or renting), especially when people move in a new city. With the growing amount of websites (including pictures, virtual tours), one can easily search for and select an ideal property. However, summarizing information about the neighborhood(s) remains an issue. In this paper, we identify and describe the main challenges related to the study of neighborhoods (description, border detection, comparison and evaluation).

Modélisation 3D urbaine et historique de la ville de Charleville.

Sylvain Rassat*, François-Joseph Ruggiu**

*7 rue Victor Schoelcher, 75014 Paris
sylvain.rassat@cnrs.fr
**7 rue Des Clos Saint Marcel, 92300 Sceaux
francois_joseph_ruggiu@paris-sorbonne.fr

Résumé. La base de données démographique et historique « CHARLEVILLE », forte d'environ 66000 personnes (soit plus d'un million d'entrées), a été constituée entre 2007 et 2011. La principale caractéristique de la ville de Charleville, concernant l'histoire de sa population, est la réalisation par les autorités municipales de la fin du XVIIe siècle, et jusqu'au XXe siècle, d'un recensement nominatif, spatialisé et annuel des habitants.
Grâce à à la qualité et au volume informels, il est possible de modéliser en 3D l'histoire urbaine en employant la technologie dite « BIM » (ou Building Information Model).

1. Charleville : une ville aux archives exceptionnelles

La ville de Charleville (devenue Charleville-Mézières en 1966) a été fondée en 1606 par Charles de Gonzague (1580-1637), duc de Nevers et de Mantoue.

Cette principauté, aux limites du royaume de France et des Pays-Bas Espagnols, puis Autrichiens, est longtemps restée une petite ville à l'échelle du royaume. Elle compte, environ, 4000 habitants au début du XVIII^{e} siècle, 8000 à la veille de la Révolution et 12000 en 1873.

Pour les historiens et les démographes, Charleville dispose d'une ressource démographique exceptionnelle grâce aux recensements annuels, nominatifs et exhaustifs mis en place par les autorités municipales de la fin du XVII^{e} siècle au début du XX^{e} siècle.

Ces listes nominatives permettent le suivi des trajectoires de vie des habitants de la fin du XVIIe siècle à la fin du XIXe siècle (Ruggiu, 2005).

Le programme ANR-06-CORP (Corpus et outils de la recherche en sciences humaines et sociales) -0005 : « Mobilités, populations, familles dans la France du Nord de la fin du XVII^{e} siècle à la fin du XIX^{e} siècle » ou « MPF » (2007-2011) a été lancé afin de numériser et d'exploiter ce fort potentiel historique et démographique (Rathier et Ruggiu, 2005).

2. Le Building Information Model et les sciences humaines et sociales

Comme prolongement au projet « MPF », fut créé un système d'information géographique alliant des données cartographiques à la base de données.

Dans un registre de 200 folios, allant de 1836 à 1843, chaque parcelle est numérotée en fonction du plan cadastral général levé en 1834. Chaque unité cadastrale possède un numéro unique, lien avec les sources manuscrites.

Grâce à ce S.I.G 2D, il est possible de modéliser en 3D la ville de 1836 avec la technologie Building Information Model (B.I.M). L'exploitation simultanée et diachronique de l'exhaustivité des entités spatiales et démographiques de chaque élément urbain est désormais possible. Le tissu urbain et certains bâtiments marquants ont été modélisés (format CityGml et IFC) selon plusieurs niveaux de définition (ou L.O.D, sur une échelle croissante de 1 à 4) grâce aux archives iconographiques très précises en présence (levers de coupe et d'élévation de 1764 par exemple).

Le B.I.M (Volk et al. 2014) permet la modélisation 3D de tous les éléments architecturaux, tout en leur associant des informations standardisées. Par exemple à l'échelle du bâtiment, un mur peut être défini par son matériau de construction et par son appartenance à un instant « T » à un logement précis et à un propriétaire donné.

Néanmoins, les différentes bibliothèques de formes géométriques standardisées du B.I.M sont, peu ou prou, exploitables pour des problématiques historiques ou encore archéologiques. Effectivement, presque tous les éléments doivent être redessinés et parfois de nouvelles familles d'entités architectoniques doivent créées (parements ou charpente par exemple).

Fig. 1 - Rendus B.I.M de la ville de Charleville en 1834, Rassat, Sylvain. 2017

Cette jonction entre les bases de données démographiques, cartographiques et le B.I.M ouvre la voie à une compréhension conjointe de la ville de Charleville et des populations l'occupant Charleville, une ville neuve créée au début du XVIIe siècle, permettra très rapidement un suivi spatial et démographique exhaustif sur près de trois siècles.

Références

Ruggiu F-J., (2005) Pour une histoire de Charleville et de sa population sous l'Ancien Régime », Revue Historique Ardennaise, tome XXXVII, 2005, pp 77-88.

Rathier R, Ruggiu F-J (2005) La population de Charleville de la fin du XVII^{e} siècle à la fin du XIX^{e} siècle, Histoire & Mesure, 2013, 2, pp 3-16.

Volk R., Stengel J., Schultmann F., (2014) Building Information Modeling (BIM) for existing buildings — Literature review and future needs. Automation in Construction. Vol. 38, 2014, pp 109-127.

Summary

The "CHARLEVILLE" demographic and historical database, with a population of around 66,000 (more than one million admissions), was created between 2007 and 2011. The main characteristic of the city of Charleville, concerning the history of its population, is the realization by the municipal authorities of the end of the seventeenth century, and until the twentieth century, a nominative, spatialized and annual census of inhabitants.

Thanks to the informal quality and volume, it is possible to model urban history in 3D using the so-called "BIM" (or Building Information Model) technology.

Remonter le temps pour comprendre le passé : l'immersion virtuelle au service des historiens

Paul François*, Florent Laroche**
Françoise Rubellin***, Jeffrey Leichman****

*LS2N, L'AMo
paul.francois@ls2n.fr
**LS2N, École Centrale de Nantes
florent.laroche@ls2n.fr
***L'AMo, CETHEFI
francoise.rubellin@univ-nantes.fr
****LSU, Louisiana State University
jleichman@lsu.edu

Résumé. Les dispositifs d'immersion en réalité virtuelle atteignent aujourd'hui un degré de réalisme permettant aux chercheurs d'expérimenter des situations du passé comme s'ils avaient emprunté une machine à remonter le temps. Dans cet article prospectif, nous proposons une ébauche de méthodologie pour permettre la captation et l'enregistrement des connaissances tacites des experts, révélées par l'émulation intellectuelle que provoque le sentiment de présence en réalité virtuelle. Cela nous permet d'envisager de nouvelles manières de capitaliser la connaissance des historiens autour d'une maquette numérique.

1 Introduction

Serions-nous capable de remonter le temps, l'étude des coutumes et des espaces disparus du passé en serait grandement facilitée. Nous pourrions accompagner un expert de la période considérée et le laisser nous expliquer ce qu'il voit, ce qu'il convient de faire en telle circonstance, ses doutes et ses surprises. Si le voyage dans le temps n'est bien sûr pas d'actualité, il est aujourd'hui possible avec les technologies d'immersion en réalité virtuelle d'avoir l'impression d'être dans des espaces-temps différents du nôtre. Ainsi la restitution archéologique s'est-elle depuis longtemps attachée à produire des immersions convaincantes dans des espaces disparus, en s'appuyant notamment sur les vestiges conservés. Mais plus que le produit de cette restitution, c'est alors le processus qui est source de découvertes.

Dans cet article prospectif, nous posons les bases d'un système permettant d'utiliser l'immersion d'experts dans des espaces historiques pour extraire leurs connaissances tacites concernant la période, les manières, les usages associés au lieu ou à l'objet restitué virtuellement. Nous nous intéresserons à la méthodologie à mettre en place et aux outils à convoquer pour ce type d'immersion virtuelle et présenteront succinctement un cas d'étude à la Foire Saint-Germain à Paris au XVIII^{e} siècle.

2 État de l'art

L'usage des technologies immersives pour l'étude ou la restitution du patrimoine de manière générale et de vestiges historiques ou archéologiques en particulier n'est pas nouvelle. Dès le début des années 2000, la question de l'immersion virtuelle apparaît dans la recherche, avec l'objectif premier d'offrir une meilleure contextualisation des données de fouille (Krasniewicz, 2000), avant même de pouvoir réaliser des analyses sur des modèles virtuels. Le développement de plateformes de réalité virtuelle de type CAVE, puis de casques de réalité virtuelle a permis une démocratisation de l'accès à des expériences d'immersions. Elles ont dès lors été utilisées dans l'archéologie, notamment pour la restitution de la Rome Antique par le CIREVE (Fleury et al., 2015), et plus généralement par la restitution de nombreux monuments antiques par Archéovision, par exemple.

L'intérêt du dispositif immersif pour le patrimoine de manière générale n'est plus à démontrer puisqu'il permet idéalement, selon Laroche (2017), de « capitaliser les connaissances et conceptualiser les informations ; démontrer la valeur patrimoniale de l'objet du passé ; piloter la visualisation et la manipulation de données par les connaissances et d'assurer la possibilité d'enrichissement des données directement en immersion ».

Malgré cela, le Consortium 3D-SHS, qui fédère les initiatives liant humanités numériques et technologies 3D, présente toujours les maquettes interactives ou la réalité augmentée comme un « livrable », c'est à dire une étape finale du processus de restitution (Vergnieux et al., 2017). Ce livrable est l'expression de la connaissance explicite, exprimée sous forme de modèle tridimensionnel et des données qui l'accompagnent, associée à un espace à restituer. La connaissance tacite, telle qu'exprimée par Takeuchi et Nonaka (2004) est celle des « idées, intuitions, inspirations » et des « croyances, perceptions, idéaux, valeurs, émotions et modèles mentaux » qui tous ensemble « forment la manière dont nous percevons le monde autour de nous ».

Cette dimension intrinsèquement sensible, à contre-courant *a priori* d'une recherche scientifique profondément factuelle, s'exprime déjà dans la recherche et la restitution des ambiances du passé. Les ambiances sonores font par exemple l'objet de travaux de recherche permettant d'ajouter un sens supplémentaire aux restitutions, et donc de transporter avec autant de force le visiteur dans ce qu'il voit et entend. Le son impose également de nouvelles modalités de compréhension, celles de l'instant fugace, aptes à modifier la manière dont chercheurs et historiens peuvent travailler (Pardoen, 2016).

3 Problématique

Une grande partie de la connaissance des historiens, archéologues et experts reste inexplorée à la suite d'un travail de restitution historique ou archéologique. Ces connaissances tacites permettent pourtant de renseigner tout un pan des modèles numériques créés, concernant leurs usages, les concepts et croyances auxquels ils renvoient, *etc*. Dans le même temps, les dispositifs numériques permettent une immersion de plus en plus convaincante, avec une expérience véritablement personnelle voire intime. À l'aide de ces outils, nous cherchons à reproduire l'émulation intellectuelle qui accompagnerait un historien si on avait le capacité de remonter le temps pour le faire voyager dans sa période de prédilection.

Nous formons l'hypothèse qu'une telle immersion permettrait d'accéder plus facilement aux connaissances tacites des experts. Ces connaissances peuvent prendre la forme de com-

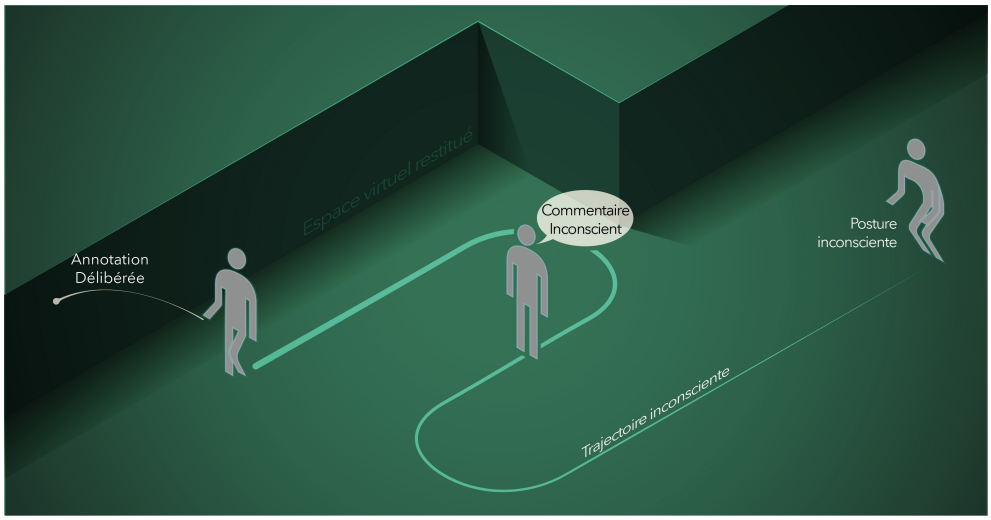

FIG. 1 – *Ensemble des connaissances tacites à capturer lors de l'immersion d'un expert.*

mentaires « conscients », bien sûr, mais ceux-ci s'accompagnent de gestes, postures, trajectoires plus ou moins inconscients et qui relèvent avant tout d'automatismes. La figure 1 représente les informations pertinentes qui nous souhaitons pouvoir enregistrer lors d'une utilisation de ce dispositif de capture de l'expérience.

Il convient donc de réussir à concevoir une expérience suffisamment convaincante pour qu'elle produise une émulation intellectuelle, sans toutefois introduire des hypothèses indémontrables, mais également d'enregistrer toutes ces données pour capturer les connaissances tacites ainsi révélées.

4 Concevoir l'environnement de capture de l'expérience

Le processus de conception de cette expérience virtuelle nécessite dans un premier temps de pouvoir rassembler et classer des sources, avant de concevoir un modèle numérique. L'ensemble s'inscrit dans une chaîne numérique mêlant solutions logicielles et matérielles.

4.1 Sources convoquées

Les sources convoquées pour la restitution des espaces et des usages sont diverses et reflètent l'éventail de documents dont se servent archéologues ou historiens. Nous proposons deux manières de distinguer ces sources, soit par leur support, soit par les éléments qu'elles renseignent. Le support graphique permet de regrouper l'ensemble des sources prenant la forme de plans, images, photos, peintures, nuages de points, *etc*. Le support textuel, quant à lui, regroupe les descriptions scientifiques, techniques ou journalistiques, qui se veulent factuelles, les monographiques et toutes les oeuvres littéraires qui par leur contexte, leurs descriptions,

peuvent évoquer ou renseigner un bâtiment et ses usages. On pourra par exemple se confronter à des romans, des poèmes, des pièces de théâtre, et ce bien que ces documents ne soient pas instinctivement convoqués dans un processus de restitution. L'indexation et la conservation de l'ensemble de ces sources doit également être envisagé afin de pouvoir maintenir un lien entre les interprétations de ces documents (sous la forme de modèles) et les documents eux-mêmes.

Peut-on considérer les experts comme des sources ? L'objectif de notre démarche étant de pouvoir capitaliser les connaissances tacites des experts, ceux-ci sont alors considérés *de facto* comme des sources d'information. Il convient néanmoins de limiter le recours à ce type de connaissances dans le processus préalable de conception de la maquette numérique servant de support à l'immersion, sans quoi l'expérience finale se limiterait, dans les cas les plus extrêmes, à faire parcourir par un expert des espaces uniquement issus de sa propre imagination.

4.2 Modèle conçu

Les outils de conception de maquettes numériques sont aujourd'hui bien connus et de nombreux logiciels permettent de réaliser des modèles architecturaux convaincants qui puissent être ensuite utilisés pour de l'immersion en réalité virtuelle. Nous souhaitons revenir plus en détail ici sur deux injonctions contradictoires de notre démarche de recherche. Dans un premier temps, nous souhaitons concevoir une restitution historique qui tienne compte de l'état des savoirs sur l'objet ou l'espace à restituer. Ainsi, chercher le réalisme de l'image ou du modèle produit semble contreproductif si l'on en croit Desjardin et al. (2012), puisqu'il « nous éloigne très souvent de la réalité de la connaissance que nous avons en ne permettant pas la perception par l'utilisateur de son imperfection ».

Dans le même temps, les études sur le sentiment de présence en immersion, qui traduit la sensation « d'être là » dans le monde virtuel, laissent plutôt penser que le réalisme a un impact fortement positif sur cette sensation (Hvass et al., 2017). Pour pouvoir résoudre ces injonctions contradictoires, il convient avant tout de produire un modèle présentant une cohérence globale pour cette expérience, et ainsi de garder une granularité constante dans l'ensemble du modèle et des textures, alors même que la connaissance sur certains détails de restitutions archéologiques ou historiques est très inégale. Nos premières expérimentations semblent également montrer que des modèles peu réalistes permettent tout de même d'obtenir les connaissances tacites que nous recherchons.

4.3 Chaîne numérique

Le matériel utilisé pour permettre l'immersion dans la réalité virtuelle des experts est composé d'un casque HTC Vive, relié à un ordinateur portable dimensionné pour permettre cet usage. Le casque Vive est accompagné de ses deux stations permettant le positionnement dans l'espace, ainsi que de deux manettes. Afin de diffuser le contenu dans le casque, nous utilisons la plateforme Steam VR, couplée au logiciel de conception de jeux vidéo Unity. L'usage de ces deux plateformes nous permet une relative indépendance vis-à-vis du matériel et laisse envisager d'utiliser d'autres dispositifs si de nouvelles technologies immersives venaient à se démocratiser.

L'immersion elle-même permet aux experts de parcourir les lieux à l'échelle 1 de deux manières différentes. Il est possible de se déplacer librement et naturellement dans un espace

de 3*3m environ, qui correspond à la limite des détecteurs de mouvement associés à la technologie Vive. Ce premier mode de déplacement permet une interaction très naturelle avec l'environnement représenté en trois dimensions. La seconde modalité de déplacement et l'usage des options de « téléportation » fournis par Steam VR dans Unity et qui permettent, grâce à l'usage des manettes, de choisir l'endroit de la scène dans lequel l'utilisateur veut se déplacer.

5 Capturer l'expérience

Notre premier prototype d'immersion d'expert a concerné la restitution d'un théâtre de marionnettes à la Foire Saint-Germain, à Paris, dans les années 1760. La Foire était un haut lieu de production artistique et théâtrale pendant tout le XVIII^{e} siècle et la restitution des espaces de spectacle présente donc un enjeu patrimonial majeur (Rubellin, 2018). Connue uniquement par une miniature réalisée par Louis-Nicolas von Blarenberghe, cette salle de marionnettes n'a laissé aucun vestige archéologique qui puisse être interprété. Sa restitution nécessite donc non seulement de réinterpréter la miniature en essayant de déjouer les pièges du peintre pour présenter sous son meilleur jour cet espace, mais également de résoudre les très nombreuses incertitudes quant à la morphologie de cet espace.

Puisque le document graphique sur lequel nous nous appuyons s'est particulièrement attaché à représenter l'ambiance du lieu, nous avons entrepris la restitution quasi photoréaliste de l'atmosphère de ce petit théâtre. L'ensemble des textures et des éclairages a donc également pu faire l'objet d'un travail avec les experts, qui a rapidement fait naître un problème de granularité puisqu'il était possible d'aller très en détails sur certains aspects. Nous avons néanmoins souhaité garder un niveau de détails cohérent dans l'ensemble du modèle.

Les experts convoqués pour ce projet – une spécialiste de l'histoire théâtrale de la Foire au XVIII^{e} siècle et un spécialiste de l'histoire culturelle française du XVIII^{e} siècle – ont été placés en autonomie dans l'espace virtuel, avec la possibilité de s'y déplacer librement. Notre dispositif de capture de l'expérience est pour le moment sommaire : une caméra enregistre les mouvements et les paroles de l'expert en réalité virtuelle dans le champ des capteurs, tandis que l'ordinateur enregistre simultanément ce qui est vu par l'expert à l'intérieur du casque de réalité virtuelle. Grâce au *timecode* des deux fichiers, il est possible de visualiser simultanément les commentaires d'un expert, son attitude, sa position et ce qu'il voyait à cet instant précis.

Cette première approche ne permet pas l'enregistrement des données dans un format adéquat pour une réutilisation ou pour une analyse. Pour cela, nous devons mettre en place un système qui permette à l'avenir :

— D'identifier les éléments de l'espace virtuel auxquels se rattachent les annotations conscientes ou inconscientes de l'expert ;
— De séparer et d'associer des échantillons sonores contenant des commentaires à des éléments de l'espace virtuel ;
— D'enregistrer de manière précise des postures ou des trajectoires et de les replacer dans la maquette tridimensionnelle ;
— De visualiser l'ensemble de ces éléments, comme à l'avenir le reste des sources, dans la réalité virtuelle.

6 Conclusion

Malgré l'état de prototype de notre dispositif d'expérimentation, nous avons pu vérifier que l'immersion virtuelle était suffisamment convaincante auprès des experts pour convoquer dans un premier temps des réponses de type « réflexe » avec l'environnement virtuel que nous avons restitué. Ces sessions permettent de manière globale de renseigner le modèle virtuel en lui apportant des connaissances concernant son usage, grâce à des données de trajectoire ou de posture. Celles-ci correspondent dans notre cas au chemin suivi par une personne pour entrer dans un théâtre et à la manière de se tenir en pareille circonstance à l'époque, par exemple. Ces connaissances tacites rendent compte d'une situation vécue par une personne avec un point de vue donné. Il nous semble que l'étude de la réponse d'un expert à un environnement virtuel s'apparente à de l'ergonomie archéologique virtuelle, définissant par la même occasion un nouveau champ de recherche. L'expérience d'un espace ne peut néanmoins se limiter à parcourir seul des lieux vides. Il convient donc d'imaginer pouvoir peupler ces espaces d'avatars reproduisants les usages qui y sont attendus, permettant à l'expert en immersion d'avoir des réponses à un environnement visuel, sonore mais aussi social.

Références

Desjardin, E., O. Nocent, et C. de Runz (2012). Prise en compte de l'imperfection des connaissances depuis la saisie des données jusqu'à la restitution 3d. *Archeolgie e Calcolatori*.

Fleury, P., S. Madeleine, et N. Lefèvre (2015). A roman street at the time of constantine : Interactive visit with access to ancient source materials. *Papers from the 41st Conference on Computer Applications and Quantitative Methods in Archaeology*.

Hvass, J. S., O. S. Larsen, K. B. Vandelbo, N. Nilsson, R. Nordahl, et S. Serafin (2017). Visual realism and presence in a virtual reality game. *3DTV-Conference*.

Krasniewicz, L. (2000). Immersive imaing technologies for archaeological research. *Virtual Reality in Archaeology, BAR International Series* (843).

Laroche, F. (2017). *KLM for Heritage*. Manuscrit d'habilitation à diriger des recherches, Université de Nantes.

Pardoen, M. (2016). Oyez, oyez ! le paysage sonore au service du passé : création ou travail scientifique ? *Musique et écologies du son. Propositions théoriques pour uen écoute du monde*.

Ravel, J. S. (2002). Le théâtre et ses publics : pratiques et représentations du parterre à paris au xviiie siècle. *Revue d'Histoire Moderne et Contemporaine 49-3*.

Rubellin, F. (2018). Historiographie des théâtres de la foire : pour en finir avec le populaire ? *Cahiers de l'Association Internationale des Études Françaises 70*.

Takeuchi, H. et I. Nonaka (2004). *Hitotsubashi on Knowledge Management*, Chapter 1. John Wiley and Sons.

Vergnieux, R., J.-F. Bernard, et M. Chayani (2017). *Livre Blanc - Consortium 3D SHS*. Huma-Num.

Harmonisation de l'acquisition des données d'opérations d'archéologie préventive. Retours d'expériences et perspectives à partir de l'application EDArc

Christophe Tufféry* et Stéphane Augry**

* Direction Scientifique et Technique
Inrap - 121 rue d'Alésia - CS 20007 - 75685 Paris Cedex 14
christophe.tuffery@inrap.fr
http://www.inrap.fr
** Inrap, 4, rue du tertre 44470 Carquefou
stephane.augry@inrap.fr
http://www.inrap.fr

Résumé. Les activités de l'Inrap en matière d'acquisition de données archéologiques ont conduit à concevoir et développer depuis 2015 une application qui permet de tendre vers l'harmonisation des pratiques d'enregistrement numérique sur le terrain. Baptisée EDArc, cette application utilise plusieurs normes et thésaurus qui permettent d'inscrire ce projet dans une démarche d'interopérabilité technique et sémantique. Le déploiement de cet outil est en cours depuis trois 2015, et uniquement sur la base du volontariat des utilisateurs. La mise en œuvre d'EDArc et d'une tablette numérique sur une opération de fouille au Mans a permis d'observer l'usage de ces outils numériques auprès d'une équipe d'agents opérationnels, plus ou moins à l'aise avec les outils numériques. Ainsi cette expérience et d'autres permettent de s'interroger sur l'impact de dispositifs numériques sur les pratiques et l'identité professionnelle de leurs utilisateurs et de ne pas se contenter d'une vision trop idyllique de l'utilisation d'outils informatiques.

1. Enregistrement des données archéologiques de terrain, pourquoi choisir EDArc ?

L'Inrap (Institut National de Recherches Archéologiques Préventives), est un établissement public à caractère administratif, sous la double tutelle du Ministère de la Culture et de celui de l'Enseignement Supérieur, de la Recherche et de l'Innovation. En application de la loi sur l'archéologie préventive de 2001 modifiée en 2003 puis en 2016, tout projet d'aménagement du territoire doit donner lieu à une phase préalable de diagnostic du potentiel archéologique qui peut être réalisé par l'Inrap ou les services compétents et habilités de collectivités territoriales. Si le diagnostic est positif, c'est-à-dire qu'il révèle la présence de vestiges archéologiques à l'emplacement du projet d'aménagement, une fouille prescrite par le Service Régional de l'Archéologie, qui dépend de la Direction des Affaires Culturelles peut ensuite avoir lieu, parfois dans le cadre d'une mise en concurrence entre l'Inrap et des opérateurs publics ou privés respectivement agréés (ou habilités) par l'Etat.

Lors de toute opération archéologique, l'ensemble des observations et des découvertes archéologiques sont enregistrées dans des fiches ou carnets papier et, depuis quelques

années, dans des systèmes d'enregistrement de terrain numériques, fonctionnant sur micro-ordinateurs et, plus récemment, sur tablettes.

Actuellement, de nombreux systèmes d'enregistrement sont utilisés par les archéologues de l'Inrap (Tufféry et al. 2017a, 2017b, Badey 2017). Une étude comparative a été réalisée entre ces différents systèmes. A quelques nuances près, ils utilisent tous les mêmes types d'entités puisque l'Archéologie est basée sur des protocoles communs : emprise, tranchée, sondage, unités stratigraphiques, faits archéologiques, ensembles, structures, mobiliers, prélèvements, documentation, etc.

Toutefois, le vocabulaire en usage peut varier d'une équipe à l'autre et chaque chantier peut développer des problématiques propres nécessitant un enregistrement particulier.

Afin de répondre à plusieurs enjeux liés à l'interopérabilité entre ces divers systèmes et aux principes FAIR (Facilement Trouvable, Accessible, Interopérable et Réutilisable) des données produites dès le terrain, la Direction Scientifique et Technique et la Direction des Systèmes d'Information de l'Inrap ont développé depuis 2015 l'application EDArc (Enregistrement de Données Archéologiques),

Cet article, à travers l'expérience d'un chantier emblématique, permet de faire un bilan critique de l'introduction de nouveaux outils informatiques dans la chaine opératoire de l'archéologie préventive.

2. Présentation de l'application

EDArc est une application logicielle fonctionnant en local avec un navigateur Web tel que Google Chrome, une base de données, un ensemble de pages.html, des fichiers JavaScript et CSS. EDArc n'a pas besoin d'une connexion Internet. Le navigateur Web sert à afficher les interfaces utilisateur et à saisir les données de terrain. Utilisé sur des tablettes durcies sous Windows lors des opérations archéologiques de l'Inrap, les données produites avec l'application EDArc sont enregistrées dans une base de données SQLite installée en local sur le poste de l'utilisateur. Cela permet, si besoin, d'afficher et d'interroger les données directement dans un logiciel SIG tel que QGis, en liaison avec des données géoréférencées.

A chaque opération l'archéologue produit une base de données terrain particulière. Néanmoins, il est possible d'embarquer dans EDArc pour une opération de fouille, les données préalablement saisies lors de l'opération de diagnostic, permettant ainsi de disposer dans un même outil de l'ensemble des données d'un site archéologique.

Le modèle conceptuel de données d'EDArc est relativement simple. Il s'appuie sur le modèle générique (Desachy 2008) qui distingue principalement les entités de contexte archéologique (US, faits, structures, etc.), les entités de documentation (photos, minutes, etc.), les entités d'élément mobilier recueilli (mobiliers, prélèvements, etc.).

Le modèle physique de données est constitué d'une série de tables de données (entre 7 et 13 selon les versions et les besoins des utilisateurs) qui comprennent des données générales administratives sur l'opération archéologique et des données brutes et interprétées sur les observations et découvertes archéologiques : Opération, Métadonnées, Unités Stratigraphiques (US), Fait, Ensemble, Tranchée, Sondage, Prélèvement, Mobilier, Minute et Photo, etc. A ces tables s'ajoutent plusieurs tables de jointures qui permettent d'interroger les liens (stratigraphiques et chronologiques) entre certaines tables de données et d'exploiter ainsi les avantages du modèle relationnel.

Plusieurs champs des tables s'appuient sur des listes de valeurs dont certaines sont issues de vocabulaires contrôlés (application Patriarche du Ministère de la Culture) ou de thesaurus

(micro-thésaurus Sujets des thesaurus PACTOLS[1]). En revanche, il n'existe pas de vocabulaires de référence ni *a fortiori* de thesaurus pour toutes les données archéologiques enregistrées dans EDArc. Quant aux métadonnées, elles décrivent la ressource documentaire qu'est la base de données EDArc de l'opération. Ces métadonnées s'appuient sur la norme internationale ISO 15836 dite Dublin Core. Elle comprend 15 descripteurs formels (titre, créateur, éditeur...), thématiques (description, langue...) et relatifs à la propriété intellectuelle (droit, auteur...). Ces descripteurs ont été retenus par l'Inrap pour permettre le versement de ses rapports d'opérations archéologiques auprès du CINES (Centre National Informatique de l'Enseignement Supérieur), qui en assure l'archivage pérenne.

A ce jour, EDArc est le seul système d'enregistrement de terrain à avoir pris en compte la nécessité d'utiliser des normes tels que le format XML pour importer et exporter des données, ce qui le rend directement interopérable et compatible avec de nombreux systèmes d'information respectant ces mêmes formats normés.

De plus, EDArc offre la possibilité d'exporter les données nécessaires pour *Le Stratifiant*, une application utilisant le logiciel Excel. (Desachy 2008). Grâce aux trois fichiers au format .xls que permet d'exporter EDArc vers *Le Stratifiant* (ExportUS, ExportRelations, ExportSynchros), cette application permet ensuite aux utilisateurs de contrôler la cohérence des données et de leurs relations de chronologie relative entre les US et de dessiner le diagramme stratigraphique en tenant compte des phases de regroupement des US associées et des éventuels éléments de datation absolue disponibles. Si le contrôle de cohérence révèle des erreurs dans la saisie des données sur les US, celles-ci sont identifiées précisément dans *Le Stratifiant*. L'utilisateur peut alors revenir dans EDArc, corriger toutes les erreurs révélées avant de réimporter dans *Le Stratifian*t les trois fichiers corrigés et de pouvoir construire le diagramme stratigraphique avec des données de meilleure de qualité.

L'interface d'EDArc et l'usage de la tablette présentent beaucoup de similitude avec les smartphones aujourd'hui dans les poches de nombreuses personnes y compris des plus réfractaires ou des moins habiles avec l'outil informatique. L'application est appréciée pour sa facilité d'utilisation, sa simplicité de mise en œuvre puisqu'elle ne nécessite l'installation d'aucun programme informatique particulier. L'utilisation possible de l'application sur divers types de tablettes sous Windows et prochainement sous Android est aussi appréciée.

La possibilité d'accéder directement dans un logiciel de SIG comme QGis aux données de la base de données SQLite dans laquelle les données d'EDArc sont enregistrées, est souvent souligné par les utilisateurs. Cela évite de devoir procéder par des exports et des imports entre EDArc et ces types de logiciels.

3. L'exemple du chantier des jardins de la cathédrale du Mans

A ce jour, une vingtaine d'opérations archéologiques de l'Inrap ont utilisé EDArc pour l'enregistrement des données de terrain.

Parmi les opérations importantes ayant adopté l'application, l'opération de fouille des jardins de la cathédrale du Mans dans la Sarthe apparaît emblématique à plus d'un titre. Cette fouille a été réalisée, de septembre 2017 à juillet 2018, aux abords du chevet gothique de la cathédrale Saint-Julien, au Mans, à l'occasion de l'aménagement des jardins, programmé par

[1] https://www.frantiq.fr/fr/thesaurus et https://pactols.frantiq.fr/opentheso/index.xhtml

la mairie[2]. La fouille a porté sur une surface de plus de 2 000 m² et concerne une stratigraphie pouvant atteindre plusieurs mètres d'épaisseur. Cette fouille est également caractérisée par les nombreux axes de recherches induits par les découvertes même si ces dernières se focalisent autour du fait urbain, l'étude de la ville, au sens général.

EDArc a été utilisé sur une tablette numérique. L'enregistrement se concentre autour de la notion d'unité stratigraphique (US) dans le sens de plus petite unité d'enregistrement de terrain. Il s'agit d'une notion primordiale au cœur de la démarche de terrain. Ici l'ensemble de l'équipe a procédé à l'enregistrement de ces US. Ce point mérite d'être souligné car le chantier réunissait plusieurs générations d'archéologues aux compétences variés. La possibilité de générer aisément les données au format XML garantit la disponibilité des données dans un formalisme normé, auquel se conforment de plus en plus de logiciels et d'applications utilisés par les agents de l'Inrap tout au long du cycle de vie des données archéologiques.

L'application permet également une veille concernant la complétude des fiches et un contrôle supplémentaire sur la réalisation des différentes étapes de la fouille (prise de vue, levers topographiques en particulier).

Pour la partie étude, aujourd'hui en cours, le lien avec chaque base de données propre aux différents spécialistes est très simple grâce aux possibilités d'export. C'est à souligner, d'autant que leur nombre est relativement important et qu'ils n'appartiennent pas tous forcément à l'Inrap. Le « phasage », (la mise en cohérence chronologique sur laquelle s'appuie le discours archéologique) des données géoréférencées est en cours de réalisation mais l'outil informatique permet d'affronter la masse de données générées. L'étape qui consistait à saisir les fiches papiers est supprimée. Le risque d'erreur de saisie est ainsi réduit et les manipulations en vue du travail de synthèse sont plus facilement accessibles.

Par ailleurs, EDArc permet aux responsables scientifiques d'opérations archéologiques de produire relativement simplement les inventaires scientifiques prévus en fin des rapports d'opérations (inventaires réglementaires de la section 3) tels qu'ils sont définis par l'arrêté du 27 septembre 2004 portant définition des normes de contenu et de présentation des rapports d'opérations archéologiques. Notons ici que la masse de données produites peut rapidement devenir exponentielle. A titre exemple, sur le chantier du Mans les unités stratigraphiques se comptent en milliers. Le mobilier archéologique est très varié et est contenu sur plusieurs mètres linéaires d'étagère. On dénombre 250 relevés manuels sur calque A3 et une quantité innombrable de clichés issus de quatre appareils photos numériques différents.

La réalisation des inventaires peut donc rapidement se relever fastidieuse.

En conclusion sur l'opération du Mans, les gains qualitatifs et quantitatifs ont été notables sans avoir pour autant mis de côté une partie de l'équipe, comme cela peut être le cas lors du déploiement de nouveaux protocoles ou de nouvelles technologies qui modifient parfois en profondeur les pratiques professionnelles. Ainsi une partie de l'enregistrement archéologique se retrouve « dématérialisé » ; les tâches laborieuses de saisie sont été écartées et l'édition des inventaires est grandement automatisée. La qualité de l'enregistrement (complétude et fiabilité des fiches) a également connu une amélioration même si ce volant de l'expérimentation n'a pas été confronté à une évaluation détaillée. Enfin le format numérique se prête bien à l'archivage même si des protocoles nouveaux sont probablement à mettre en place en lien avec une formation minimale des équipes archéologiques. En dernier lieu, il convient d'insister sur les facilités et les possibilités crées entre les données de terrain brutes

[2] https://www.inrap.fr/les-fouilles-des-jardins-de-la-cathedrale-du-mans-13569

et les données géoréférencées ou encore les données issues des différentes études spécialisées.

4. Suites envisagées et perspectives

Depuis 2015, des travaux d'appariement entre EDArc et le CIDOC CRM, norme ISO depuis 2006, ont été réalisés, d'abord dans le cadre du programme européen ARIADNE (Tufféry et al. 2016) puis poursuivis depuis trois ans (Tufféry et al. 2017c, Le Goff 2018). Ces résultats vont pouvoir contribuer prochainement au groupe de travail Web des données du Consortium MASA, soutenu par la Très Grande Infrastructure de Recherche HumaNum. L'intérêt de ces travaux est de faciliter la publication dans le Web Sémantique des données de terrain produites à l'aide d'EDArc, sous la forme de triplets sémantiques au format Resource Description Framework (RDF). Une fois l'appariement effectué, les données pourront être publiées sur le web sémantique depuis un End point SPARQL. Ces données seront alors des ressources documentaires que les machines pourront mieux exploiter à l'aide de langages de requêtes et de plateformes d'interrogation adaptées à l'exploitation des données archéologiques publiées dans le Web des données.

L'une des retombées de la mise en œuvre de l'application EDArc est le fait de démontrer la possibilité d'un ensemble de données minimales à enregistrer sur le terrain qui soit à la fois admissible par l'ensemble des archéologues de l'Inrap et en même temps qui respectent la diversité des pratiques actuelles et des problématiques scientifiques qui peuvent varier d'une opération à l'autre. De ce point de vue, EDArc présente une souplesse de modification et d'adaptation à des opérations, des contextes archéologiques et des types de données spécifiques à prendre en compte.

Il faut aussi souligner que l'introduction sur les opérations archéologiques d'une application comme EDArc s'inscrit, pour l'Inrap et au-delà pour la communauté archéologique, dans le cadre du déploiement de techniques et dispositifs numériques, initié depuis une dizaine d'années avec l'utilisation des premières tablettes numériques. A ce jour une centaine de ces équipements ont été acquis par l'Inrap et utilisés sur plusieurs centaines d'opérations.

EDArc modifie considérablement la répartition traditionnelle du temps consacré à la saisie numérique des données, habituellement réalisée après la phase terrain, à partir des enregistrements réalisés sur supports papier pendant la fouille. Ainsi l'utilisation d'EDArc sur le terrain laisse davantage de temps au traitement des données après la fouille puisque les données sont produites nativement sous forme numérique dès le terrain. Dès lors, les pratiques et les logiques de travail évoluent, de même que la répartition des tâches au sein des équipes ou encore les compétences numériques des agents. Cette évolution multiple se traduit par des situations très différentes selon les agents. Si pour une partie d'entre eux, l'introduction d'outils de travail numériques sur le terrain constitue une opportunité pour faire évoluer leurs pratiques et leurs compétences dans un sens positif, pour d'autres agents cela représente une évolution contrainte, qui modifie en profondeur et de façon très rapide leurs manières habituelles d'exercer leur activité et peut aller jusqu'à perturber leur identité sociale et professionnelle (Dubar 1991).

Depuis le début de l'introduction d'EDArc et de tablettes sur les opérations archéologiques de l'Inrap, nous avons veillé à prendre en compte la diversité des réactions, des capacités, et parfois de difficultés incontestables des agents de l'Inrap face à ce type de dispositif. Ainsi, des actions de formation et de tutorat ont été mises en œuvre. Sans ce genre d'accompagnement, le risque est grand de voir apparaître des fractures entre d'un côté les

agents les plus à l'aise, les primo-adoptants, et de l'autre côté ceux pour lesquels le recours au numérique est synonyme de difficulté, voire de souffrance au travail. Accompagner ainsi l'introduction de dispositifs numériques dans un champ professionnel comme celui de l'archéologie, permet de veiller à ne pas confondre « l'informatique comme pratique et comme croyance », comme cela a déjà été proposé depuis une vingtaine d'années (Gollac M. et al. 2000).

Actuellement plusieurs évolutions d'EDArc sont en cours pour répondre encore mieux aux besoins des utilisateurs et pour tenir compte des retours d'expériences.

EDArc est maintenant présentée dans les actions de formation à l'enregistrement de terrain qui ont lieu à raison de trois sessions par an, soit environ 25 agents formés par an, lesquels peuvent ensuite facilement transférer leur maîtrise de l'application à leurs collègues sur le terrain. Ainsi si les outils numériques entraînent une modification profonde de la nature du travail archéologique, les formations permettent, si ce n'est d'écarter, au moins d'atténuer les aspects négatifs. Des retours d'expérience et des bilans réguliers s'avèrent également nécessaires.

5. Références

Badey S. et Moreau A. (2017), Teaching Archaeology or Teaching Digital Archaeology: Do We Have to Choose ?, in M. Matsumoto, E. Uleberg (eds.). In CAA 2016. Proceedings of the 44th Annual Conference on Computer Applications and Quantitative Methods in Archaeology - Oceans of Data. Oslo, Museum of Cultural History, 30/08-03/09/2017, 10p.

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

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

Le Goff E. et Tufféry C. (2018), Inrap's experience in the digital acquisition of field archaeological data and associated metadata compliant with CIDOC CRM and Dublin Core standards. Conference CIDOC 2018 Heraklion, Crete, Grèce, 2 octobre 2018

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

Tufféry C., Felicetti A., Jard P., Holzem N., Guillemard T. (2016), An essay of mapping archaeological land-record systems used by Inrap with CIDOC-CRM and CIDOC-CRMarchaeo extension using 3M on-line tool. CAA 2016. Proceedings of the 44th Annual Conference on Computer Applications and Quantitative Methods in Archaeology - Oceans of Data. Oslo, Museum of Cultural History, 30/08-03/09/2017, Session #11: Supporting researchers in the use and re-use of archaeological data: continuing the ARIADNE thread

Tufféry C., Le Goff E., Boudry J., Nurra F. (2017a), Using CIDOC-CRM for exchanging field data: the point of view of the National Institute for Preventive Archaeological Research (Inrap), EAA, Session #251, Maastricht, 30/08-01/09/2017

Tufféry C., Le Goff E., Boudry J., Nurra F. (2017b), Recours au CIDOC-CRM pour évaluer l'interopérabilité de données archéologiques de terrain très variées: présentation des premiers résultats des tests effectués par l'Inrap, in Spatial Analysis and GEOmatics SAGEO, Rouen, 6-9 novembre 2017, 4 pages. https://hal.archives-ouvertes.fr/hal-01649750/document

Tufféry C. et Le Goff E. (2017c), Contribution de l'Inrap à l'usage du CIDOC-CRM pour les données archéologiques d'enregistrement de terrain. Participation à la table-ronde des Journées MASA Interopérabilités, 20-22 novembre 2017, Tours : http://masa.hypotheses.org/430

Summary

Inrap's activities in the field of archaeological data acquisition have led to the design and development since 2015 of an application that aims to harmonize digital recording practices in the field. Called EDArc, this application uses several standards and thesauri that allow this project to be part of a technical and semantic interoperability approach. The deployment of this tool has been ongoing since three years 2015, and only on a voluntary basis. The implementation of EDArc and a digital tablet on a search operation in Le Mans made it possible to observe the use of these digital tools with a team of operational agents, more or less at ease with digital tools. Thus, this and other experiences allow us to question the impact of digital devices on the practices and professional identity of their users and not to be satisfied with an overly idyllic vision of the use of computer tools.

Les événements dans l'ontologie CRMCR dédiée à la conservation et la restauration des oeuvres d'art

Claudia Marinica*, Inès Bannour*, Luc Bouiller**, Olivier Malavergne***

*ETIS UMR 8051, Université Paris-Seine, Université de Cergy-Pontoise, ENSEA, CNRS
{claudia.marinica, ines.bannour}@ensea.fr,
**Centre de Recherche et de Restauration des Musées de France, Paris
luc.bouiller@culture.gouv.fr
***Laboratoire de Recherche des Monuments Historiques, Champs-sur-Marne
olivier.malavergne@culture.gouv.fr

Résumé. Des institutions, avec différents coeurs de métier, participent dans le processus de conservation-restauration lors du cycle de vie de l'objet culturel et produisent des données diverses. Il est nécessaire aujourd'hui de pouvoir accéder à ces données à partir d'un seul point d'entrée. Dans ce papier, nous présentons l'aspect événementiel de l'ontologie de référence proposée, CRMCR, qui est dédiée à la conservation et la restauration des objets culturels.

1 Introduction

Le processus de conservation-restauration inclut les tâches qui permettent la préservation du patrimoine culturel tout en conservant sa signifiance et son état proche d'origine. Ce processus est associé aux collections d'art des musées et comprend des tâches comme la vérification, la documentation, le stockage, la conservation, et la restauration (Szczepanowska, 2012).

Les institutions culturelles prennent en charge des diverses tâches dans le processus de conservation-restauration des objets culturels, et, en conséquence, ils produisent des grandes masses de données qui devraient être accessibles et interopérables (Doerr, 2009). Afin de faciliter l'échange de connaissances, l'objectif n'est pas de déplacer toutes les données produites dans des silos centralisés, mais d'utiliser les technologies du web sémantique (Berners-Lee et al., 2001) et les approches d'intégration de données (Lenzerini, 2002) pour accéder aux données en utilisant un modèle conceptuel sémantique partagé.

Dans le contexte du patrimoine culturel, plusieurs modèles conceptuels ont été proposés dans la littérature, souvent sous forme d'ontologies. Mais, la plus part des ontologies proposées sont soit trop génériques pour représenter la spécificité du domaine de conservation-restauration, comme CIDOC-CRM (Doerr et al., 2007), soit trop spécifiques, comme les extensions de CIDOC-CRM ou le Monument Damage Ontology (Cacciotti et al., 2015).

Dans ce papier, nous présentons l'aspect événementiel de CRMCR, un modèle ontologique qui fournit une compréhension unifiée du processus de conservation et restauration des objets culturels. CRMCR est une extension de l'ontologie CIDOC-CRM et utilise des concepts et des relations de l'extension CRMSCI de CIDOC-CRM. CRMCR permet de décrire un objet culturel à travers ses caractéristiques (identification, location), mais, plus important, elle

permet de décrire un ensemble d'événements qui sont au coeur du processus de conservation-restauration, comme les événements dégradants et non-dégradants qui surviennent dans la vie d'un objet culturel. L'ontologie CRMCR a été présentée dans son intégralité dans (Bannour et al., 2018) ; dans ce papier, nous détaillons les événements en lien avec le processus de conservation-restauration.

L'ontologie CRMCR que nous avons proposée est actuellement utilisée dans le développement d'un système d'intégration ayant pour but de fournir un point d'accès unifié à deux bases de données : la base de données EROS du C2RMF (Centre de Recherche et de Restauration des Musées de France) et la base bibliographique CASTOR du LRMH (Laboratoire de Recherche des Monuments Historiques). L'objectif du système d'intégration développé est d'accueillir à terme des nouvelles bases et de devenir un outil de référence dans le domaine de la conservation et restauration des oeuvres d'art.

La papier présente dans la section 2 de manière succincte le processus de construction de l'ontologie CRMCR. Les sections 3 et 4 décrivent respectivement les événements dégradants et les événements non-dégradants. La dernière section conclut le papier.

2 Construction et description générale de CRMCR

Il est important de noter que la présentation des besoins des institutions culturelles est réalisée dans (Bannour et al., 2018). En nous basant sur ces conclusions, et étant donné que CIDOC-CRM est un modèle ontologique générique, nous avons développé le modèle ontologique CRMCR pour la conservation et la restauration des oeuvres comme une extension de CIDOC-CRM. Ainsi, nous étendons les objectifs de CIDOC-CRM en rajoutant des nouveaux concepts qui spécialisent ceux du CIDOC-CRM, et en mettant en place les relations nécessaires entre ces concepts. Plus précisément, nous construisons une architecture basée sur une seule ontologie, qui est présentée dans la figure 1 (nous précisons que les figures présentes dans ce papier ont été produites par les auteurs), et qui est composée de :

1. L'ontologie de haut niveau CIDOC-CRM à partir de laquelle nous utilisons l'infrastructure basée sur les événements, et l'identification (partie haute).
2. L'ontologie de domaine CRMSCI qui est une extension de CIDOC-CRM pour les observations scientifiques et pour les mesures (en bas à gauche).
3. L'ontologie du domaine CRMCR qui est une extension de CIDOC-CRM pour la conservation restauration que nous avons développée et qui spécialise des éléments de CRMSCI et CIDOC-CRM surtout concernant les événements dégradants ou non-dégradants qui affectent un objet culturel, et les résultats produits par les études scientifiques.
4. Un ensemble de thésauri du domaine pour la désambiguation de la terminologie.

La notation suivante est utilisée dans la suite : CIDOC-CRM - *E* pour les classes, *P* pour les relations, et la couleur bleu ; CRMSCI - *S* pour les classes, *O* pour les relations, et la couleur verte ; et CRMCR - *C* pour les classes, *R* pour les relations, et la couleur orange. Le modèle ontologique CRMCR est composé de : 93 concepts et 82 relations de CIDOC-CRM ; 22 concepts et 24 relations de CRMSCI ; 63 nouveaux concepts et 27 nouvelles relations.

Dans la suite, nous allons nous concentrer sur les événements en lien avec le domaine de la conservation-restauration : (1) les événements dégradants (altérations), et (2) les événements non-dégradants correspondants aux études scientifiques, les interventions, etc.

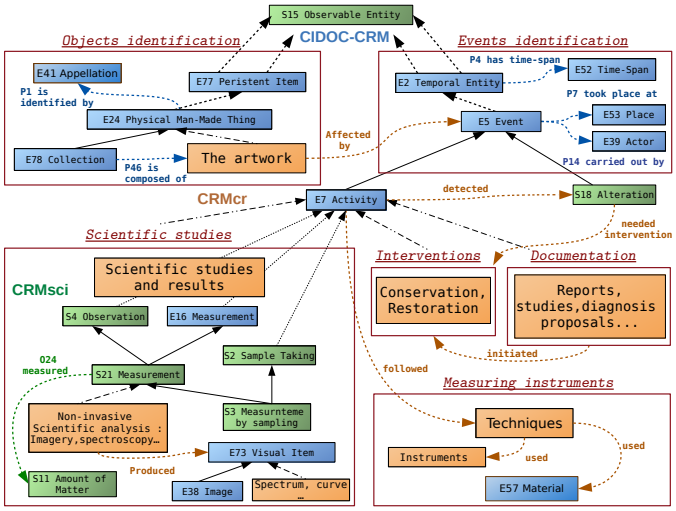

FIG. 1 – *CRMCR : éléments de CIDOC-CRM et CRMSCI et nouveaux.*

3 Les événements qui dégradent l'état des objets culturels

Un objet culturel subit, au cours de son cycle de vie, des événements dégradants qui sont des altération intentionnelles ou non-intentionnelles ; ces événements dégradants peuvent être des événements naturels ou des événements produits par l'homme qui créent, altèrent ou changent des objets en modifiant de manière permanente leur forme ou leur cohérence sans changer leur identité (Doerr et al., 2013).

CRMSCI définit le concept *S18 Altération* comme un type de phénomène qui peut être observé par le processus de *S21 Mesure* réalisé sur un *E18 Element Physique*. Une altération peut être soit une *S17 Genèse physique* ou une *E11 Modification* dans le *E18 Element Physique*. Par contre, CRMSCI ne définit pas ni la zone altérée ni le facteur qui a causé l'altération.

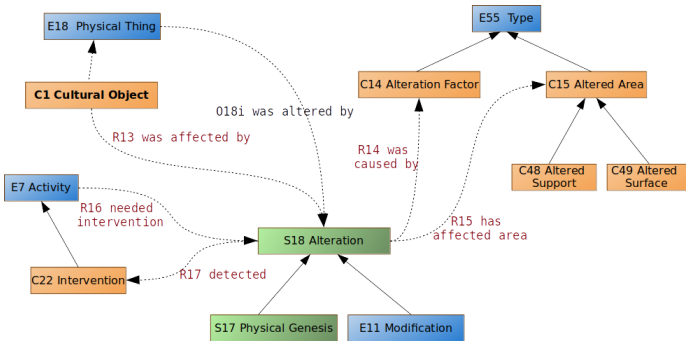

FIG. 2 – *Les événements dégradant l'état d'un objet culturel.*

Dans CRMCR, on intègre le concept *S18 Altération* de CRMSCI (figure 2) où l'objet cultu-

rel peut être affecté par un événement d'altération sur un zone et par une cause précise. L'événement d'altération *R15 a une zone affectée* une *C15 Zone Altérée*, qui peut être soit un *C48 Support Altéré* ou une *C49 Surface Altérée*. Ensuite, une altération peut être causée par un *C14 Facteur d'Altération* (des événements naturels ou produits par l'homme). De plus, une altération nécessite une intervention, mais auparavant elle devrait être analysée à l'aide d'études scientifiques afin d'avoir plus de détails sur le type d'intervention nécessaire.

4 Les événements qui ne dégradent pas l'état des objets culturels

Une *E7 Activité* est définie en CIDOC-CRM comme un événement composé d'actions intentionnelles réalisées par un *E39 Acteur*. Par rapport l'événement dégradant *S18 Altération*, une *E7 Activité* est un événement non-dégradant. Dans la conservation et restauration des oeuvres, les principales activités réalisées sur les objets culturels sont (1) les analyses, (2) les interventions et (3) la documentation de ces dernières activités. Dans la suite, nous nous concentrons sur les deux premières activités.

4.1 Les analyses réalisées sur les objets culturels

Les analyses ont un rôle principal, soit dans l'identification et la documentation d'un objet culturel, soit dans le déclanchement d'activités comme la réparation, la restauration et la conservation. Ces activités sont classées comme invasives ou non-invasives. Elles peuvent également être classées en fonction des techniques utilisées (chimiques, thermodynamiques, etc.).

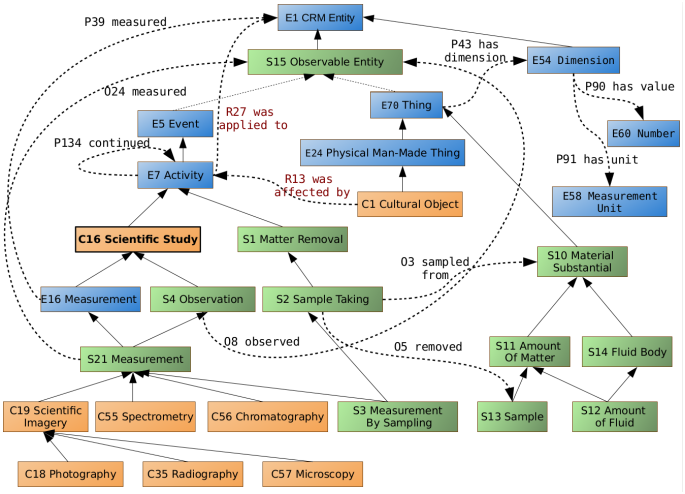

FIG. 3 – *Les analyses réalisées sur les objets culturels.*

L'ontologie CRMSCI décrit les observations et les études scientifiques qui peuvent être réalisées sur une *S15 Entité Observable*. Un *C1 Objet Culturel*, défini dans CRMCR, est une

sousClasseDe une *S15 Entité Observable*, ainsi, un ensemble d'analyses, héritées de CRMSCI, peuvent être réalisées sur un objet culturel : des *S4 Observation*s peuvent être réalisées (*S2 Echantillonnage*) afin de récolter un ensemble de *S21 Measures* et ensuite de produire un ensemble de résultats sur ces données. Plus précisément, un objet culturel peut être affecté par (*R13 A été affecté par*) plusieurs *C16 Etudes Scientiques* : *S21 Mesures* ou *S4 Observations*.

Dans CRMCR, nous intégrons deux types de mesures : (1) les mesures invasives réalisées par le *S3 Echantillonnage par Mesure* (de CRMSCI), et (2) nous avons rajouté un ensemble de mesures non-invasives. Ces dernières mesures sont intégrées par les concepts : *C19 Imagerie Scientifique*, *C55 Spectrométrie*, and *C56 Chromatographie*, comme le montre la figure 3.

4.2 Les résultats des analyses réalisées sur les objets culturels

Les résultats des analyses sont différents d'une activité de mesure à une autre, mais, le processus de récolte des résultats d'analyse peut être défini comme il suit. Une *E16 Mesure* peut mesurer (1) un ensemble de *C55 Conditions Expérimentales* comme la pression, température, (2) un concept de *S9 Type de Propriété* de CRMSCI, et (3) le *S11 Montant de Matière* qui compose l'objet observé ou échantillonné comme décrit par CRMSCI (*S3 Mesure par Echantillonnage*). Ces trois types de mesures sont présentés dans la figure 4 et ils font partie des résultats qui peuvent être conclus à partir des *S4 Observations* classiques ou des *S21 Mesures*.

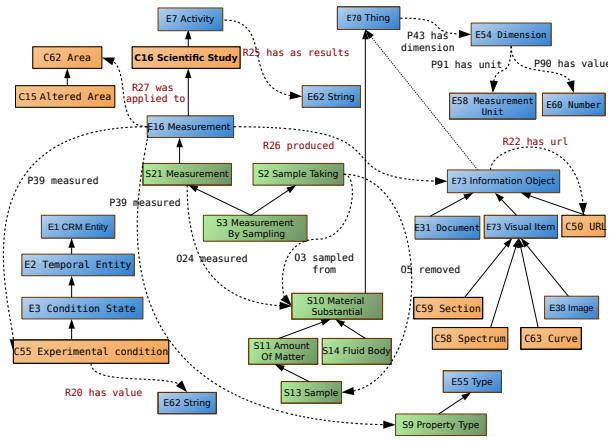

FIG. 4 – *Les résultats des analyses réalisées sur un objet culturel.*

Dans CIDOC-CRM, chaque objet de type *E70 Chose P43 a dimension* une *E54 Dimension*. Une *E54 Dimension P90 a comme valeur* une valeur (*E60 Nombre*), et *P91 a unité* une *E58 Unité de Mesure*. Ces valeurs mesurent les *C55 Conditions Expérimentales*, le *S11 Montant de Matière*s et le *S9 Type de Propriété*. Enfin, après une étude scientifique, les experts produisent des images et des rapports (*E73 Objet d'Information*).

4.3 Les interventions réalisées sur les objets culturels

Les *C16 Etudes Scientifiques R17 détectent* une *S18 Altération* qui *R16 nécessite une intervention*, et donc un type de *C22 Intervention* est lancé comme le montre la figure 5. Une intervention peut être une activité de conservation ou de restauration afin de réparer l'objet culturel et le préserver. Ainsi, dans CRMCR, nous avons ajouté les concepts *C22 Intervention*, *C23 Conservation*, *C24 Restauration* et une liste d'interventions spécifiques validées par les expert du domaine et définies à travers la terminologie spécifiées par le International Council of Museums - Committee for Conservation [1] (ICOM-CC).

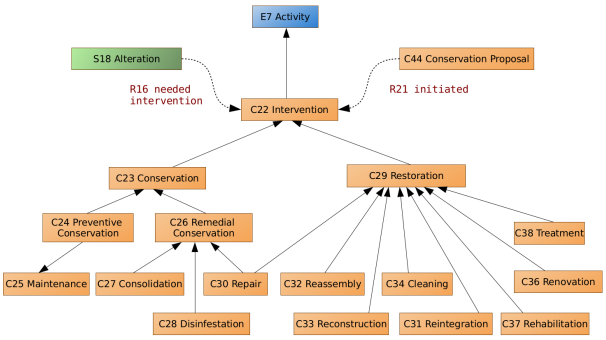

FIG. 5 – *Les interventions réalisées sur les objets culturels.*

Le concept *C23 Conservation* comprend toutes les activités qui protègent les objets culturels tout en assurant leur disponibilité. L'activité de conservation regroupe le *C24 Conservation Préventive* (éviter la détérioration) et la *C26 Conservation Réparatrice* (réduire le processus de détérioration déjà entamé). Le concept *C29 Restauration* concerne les actions réalisées quand l'objet a perdu une partie de sa signification ou sa fonction à cause des altérations. Ainsi, la restauration peut être une activité de *C30 Réparation*, *C31 Réintégration*, *C32 Réassemblage*, *C33 Reconstruction* or *C34 Nettoyage*, comme le montre la figure 5.

5 Conclusion

Dans ce papier, nous avons détaillé les événements intégrés dans l'ontologie CRMCR que nous avons proposée et qui est dédiée à la conservation et restauration des objets culturels. Cette ontologie a été développée comme une extension de l'ontologie CIDOC-CRM et elle intègre certains concepts et relations de l'ontologie CRMSCI. L'ontologie développée est basée sur les événements qui affectent une oeuvre d'art tout au long de sa vie ; par exemple, la création de l'oeuvre est un événement, mais également les interventions qui peuvent être réalisées sur l'oeuvre.

L'ontologie CRMCR que nous avons présenté est évolutive avec le domaine, et elle est actuellement utilisée dans le développement d'un système d'intégration ayant pour but de fournir un point d'accès unifié à deux bases de données : la base de données EROS du C2RMF

1. http ://www.icom-cc.org/

(Centre de Recherche et de Restauration des Musées de France) et la base bibliographique CASTOR du LRMH (Laboratoire de Recherche des Monuments Historiques). L'objectif du système d'intégration développé est d'accueillir à terme des nouvelles bases et de devenir un outil de référence dans le domaine de la conservation et restauration des oeuvres d'art.

Remerciements

Ce travail a été réalisé dans le cadre du projet PARCOURS (`projet-parcours.eu`) soutenu par la Fondation des Sciences du Patrimoine (EUR-17-EURE-0021).

Références

Bannour, I., C. Marinica, L. Bouiller, R. Pillay, C. Darrieumerlou, O. Malavergne, D. Kotzinos, et C. Niang (2018). CRMcr - a CIDOC-CRM extension for supporting semantic interoperability in the conservation and restoration domain. In *Digital Heritage 2018*, San Francisco, United States.

Berners-Lee, T., J. Hendler, et O. Lassila (2001). The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.

Cacciotti, R., M. Blasko, et J. Valach (2015). A diagnostic ontological model for damages to historical constructions. *Journal of Cultural Heritage 16*(1), 40 – 48.

Doerr, M. (2009). Ontologies for cultural heritage. In *Handbook on Ontologies*, pp. 463–486. Springer.

Doerr, M., G. Hiebel, et Y. Kritsotaki (2013). The scientific observation model an extension of cidoc-crm to support scientific observation. *Paper presented at the 29th CRM-SIG Meeting, Heraklion, Greece, October 21 - 25*.

Doerr, M., C.-E. Ore, et S. Stead (2007). The cidoc conceptual reference model : a new standard for knowledge sharing. In *Tutorials, posters, panels and industrial contributions at the 26th international conference on Conceptual modeling-Volume 83*, pp. 51–56. Australian Computer Society, Inc.

Lenzerini, M. (2002). Data integration : A theoretical perspective. In *PODS*, pp. 233–246.

Szczepanowska, H. (2012). *Conservation of Cultural Heritage : Key Principles and Approaches ?.(410 pages) Publisher ; Routledge, Classics, Archaeology and Museum Studies ; Taylor Francis Books, United Kingdom.*

Summary

Institutions with different specializations participate in the conservation-restoration process of a cultural object and produce a considerable amount of data. Nowadays, it is necessary to be able to access these data from a single entry point. In this paper, we present the event-based characteristic of the proposed ontology, CRMcr, which is dedicated to the conservation and restoration of cultural objects.

Enrichissement sémantique de données d'archives sonores d'ethnomusicologie par alignement

Nédra Mellouli*, Aude Julien Da Cruz Lima **

*LIASD(EA 4383), Université Paris8 Vincennes, Saint Denis
n.mellouli@iut.univ-paris8.fr,
**LESC (UMR 7186), CNRS, Université Paris Nanterre
aude.da-cruz-lima@cnrs.fr

Résumé. ATMAH : "Alignement Tool for Music Archive heritage" est un projet multidisciplinaire financé en 2016 par la COMUE [1] de l'Université Paris Lumières. Ses disciplines phares sont principalement le web des données et les sciences de l'information appliquées à la SHS. Son objectif principal est d'associer les partenaires culturels (BnF, Musée du quai branly, LESC et la Maison des Cultures du Monde) impliqués dans la diffusion d'archives sonores en ethnomusicologie, afin d'améliorer la gestion des vocabulaires d'indexation selon trois aspects : accès, enrichissement et interopérabilité. Il est donc fondamental de renforcer leur valorisation au niveau national et international dans le contexte des données ouvertes liées. Le défi de l'étude proposée est de démontrer l'intérêt de l'interconnexion et la faisabilité de sa réalisation entre plusieurs bases de données web et systèmes de plates-formes, avec des vocabulaires aussi bien internes spécifiques (structurés en SQL, MARC, etc.) qu'externes ("hub" de dimension internationale, comme langue française du consortium Bnf RAMEAU et MIMO de Europeana, conformes aux standards du web des données). A partir des cas d'utilisation dérivés des classifications d'instruments de musique et d'autres vocabulaires du domaine (tels que la description vocale, les supports audio, les genres musicaux traditionnels, les entités nommées, etc.), il sera nécessaire d'établir des scénarios d'utilisation et des spécifications à tester en fonction des besoins des professionnels de l'information et de la recherche (échange, alignement semi-automatisé, visualisation graphique et dynamique des données, annotation collaborative). Dans la première étape du projet, nous avons ciblé notre étude sur l'amélioration de la gestion des vocabulaires contrôlés dans la plateforme audio web des archives sonores utilisée par l'équipe du CREM et sur les outils et procédures d'alignement. Au delà des archives du CREM, l'outil visé par le projet ATMAH est étroitement lié aux besoins en cours des partenaires (nationaux et internationaux) en tant que consortiums (TGIR Huma-Num du CNRS, France), LabEx Past in Present cluster, Paris Lumières Université, Europeana Sounds.

1. ATMAH : Alignement Tool for Music Archive heritage s'inscrit dans le cadre de l'appel à projets COMUE 2016 de l'Université Paris Lumières

1 Vers une interopérabilité des vocabulaires d'indexation dans l'environnement numérique et du web

La diffusion et l'échange d'information dans un environnement numérique en ligne nécessite de partager des vocabulaires communs pour améliorer la mise en correspondance (interconnexion) des données tout en respectant la spécificité des terminologies métiers. A ce jour où des demandes de partage d'information entre des communautés nouvelles émergent, la plateforme d'archives sonores (Telemeta) [2] utilisée par le CREM ne permet pas seule de lever ce verrou technologique dès lors que les vocabulaires métiers restent spécifiques et dépendants des ressources de chaque équipe. Pour exploiter le potentiel de l'environnement numérique du web, il faut mettre en place un mécanisme d'interopérabilité sémantique à partir de vocabulaires pivots et d'outils de gestion et d'alignements permettant de définir les liaisons appropriées aux différents termes du vocabulaire de référence. Pour résoudre ces problèmes, ce travail contribue à l'étude des possibilités de représenter et de gérer ces vocabulaires dans des formats RDF (SKOS et OWL) permettant la construction de ressources liées à un niveau sémantique, et la recherche de solutions afin de les connecter à la plateforme Telemeta. Ensuite, face à la pluralité et la diversité de ces terminologies métier il est nécessaire de penser des solutions permettant d'améliorer l'accès et l'interopérabilité des vocabulaires contrôlés de la plateforme d'archives sonores du CREM afin de 1) renforcer les passerelles avec des vocabulaires externes de référence, 2) évaluer la stabilité et la pérennité des ressources, 3) valoriser les ressources des producteurs dans le contexte du Linked Open Data. Enfin ce travail vise à dynamiser et étendre les réseaux scientifiques aux niveaux national et international, notamment à travers l'élaboration de passerelles multilingues.

2 Méthodologie et application

Durant la première partie du projet qui s'est tenue en 2016 et 2017, nous avons structuré l'étude en deux phases. Lors de la première phase nous avons procédé à l'analyse des différents vocabulaires dont nous disposons et à établir les schémas de mapping les plus pertinents. Les formats SKOS et RDF ont été retenus pour la représentation des vocabulaires sources. Pour ce faire, nous avons utilisé OpenRefine(Mathieu, 2018) pour nettoyer, transformer le vocabulaire d'un format à un autre et l'étendre aux services web et aux données externes. Une fois le vocabulaire formaté, la seconde phase vient répondre à deux questions : 1) comment aligner ce vocabulaire source avec un vocabulaire cible ? 2) Quels vocabulaires cibles sont intéressants quant à l'enrichissement du vocabulaire source ?

RAMEAU [3] et MIMO(Manguinhas et al., 2016) sont deux vocabulaires cibles conformes aux standards du web de données que nous avons sélectionné pour la thématique instruments de musique. Enfin, la tâche d'enrichissement et de mise en correspondance des vocabulaires sources avec des données externes exige la mise en place d'une démarche qualité et d'évaluation d'un certain nombre d'outils d'alignement tels que ONAGUI [4], YAM++(Ngo et Bellah-

2. https://archives.crem-cnrs.fr/
3. http://rameau.bnf.fr/utilisation/liste.htm
4. ttps://github.com/lmazuel/onagui

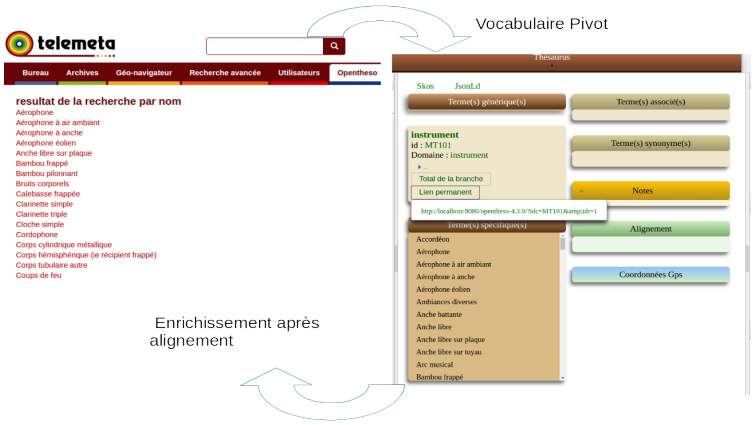

FIG. 1 – *Interface d'enrichissement du vocabulaire dans Telemeta par OpenTheso (Julien Da cruz Lima et Mellouli, 2017)*

sene, 2012), CultuurLink [5] ou OpenTheso (cf.Figure 1). Cette démarche cherche à répondre à trois critères : 1) optimiser le taux de couverture du vocabulaire source par les résultats d'alignement, 2) maximiser le matching positif et minimiser les contradictions, 3) écarter les ligneurs spécifiques. Pour cela, la première étape d'exploration du vocabulaire a été dédiée à l'extraction des termes utilisés par les données de Telemeta et qui concernent les instruments de musique. Cette extraction a permis dans un premier temps de constituer une base de vocabulaires bien identifiée. Lors de la seconde étape, nous avons cherché à enrichir ce vocabulaire par des méthodes d'alignement souvent utilisées pour trouver des correspondances sémantiques entre une ontologie source et une ontologie cible. Notre contribution a consisté à considérer et modéliser une ontologie spécifique aux vocabulaires extraits depuis les données de Telemeta et à l'aligner avec différentes ontologies cibles. Différents tests expérimentaux des procédures d'alignements ont été entrepris sur les archives sonores et audiovisuelles d'ethnomusicologie (du CREM) qui portent sur les référentiels d'instruments de musique et après avoir étudié les outils d'alignements disponibles dans la littérature. Cette étape a été très difficile à réaliser avec satisfaction puisque la plupart des outils d'alignements disponibles avaient été non concluants et souvent très spécifiques au domaine d'application. Néanmoins, nous avons retenu quatres outils qui se distinguent par leur interface graphique et par leur simplicité à savoir OnaGui, Cultuur Link et Yam++, ainsi que le gestionnaire OpenTheso incluant des fonctionnalités d'alignement. Grâce à ces outils nous envisageons de collaborer avec les développeurs afin de les adapter à notre besoin et de pouvoir comparer les résultats que nous obtiendrons sur nos vocabulaires, voire même sur de nouveaux vocabulaires.

5. http://cultuurlink.beeldengeluid.nl/app/

3 Conclusion

Le défi de cette étude et l'objectif de ATMAH est de démontrer l'intérêt de l'interconnexion et la faisabilité de sa réalisation entre plusieurs bases de données web et systèmes de plates-formes, avec des vocabulaires internes spécifiques et externes. A partir des cas d'utilisation dérivés des classifications d'instruments de musique et d'autres vocabulaires du domaine, nous avons expérimenté notre méthodologie sur un volume de données indexées avec les différents vocabulaires contrôlés. Ce volume correspond à environs 75000 fiches documentaires à ce jour avec une progression de 5 à 10 % par an. Les vocabulaires contrôlés utilisés peuvent contenir entre une centaine et plusieurs milliers de termes. En particulier, le vocabulaire des instruments de musique est très complexe. Il peut varier entre les formes génériques et vernaculaires couvrant des centaines de langues du monde entier, avec les diversités des translittérations en fonction des pratiques et des époques. Les résultats tangibles obtenus depuis le début du projet sont principalement d'ordre organisationnelle et méthodologique qui vise à mettre en place sur le long terme un réseau de partenaires interinstitutionnel et pluridisciplinaire et d'une coopération effective exploitant la procédure d'alignement visée pour les données du CREM. Les résultats tangibles visés en perspective sont d'ordre qualitatif avec l'enrichissement des données des différents partenaires et la contribution à la création de référentiels du domaine comme les expressions vocales, l'enrichissement des vocabulaires avec les résultats d'alignement, l'enrichissement sémantique des données indexées avec des référentiels pivots du LOD.

Références

Julien Da cruz Lima, A. et N. Mellouli (2017). Alignment tools for music archive heritage (atmah). *The International Association of Sound and Audiovisual Archives (IASA), Berlin*.

Manguinhas, H. ., V. Charles, A. Isaac, T. Miles, A. . Lima, A. Néroulidis, V. Ginouvès, D. Atsidis, M. Brinkerink, M. Hildebrand, et S. Gordea (2016). Yam++ : A multi-strategy based approach for ontology matching task. *The 16th European Networked Knowledge Organization Systems (NKOS16)*.

Mathieu, S. (2018). Nettoyer et préparer des données avec openrefine. Atelier pour les journées du consortium masa, (https://msaby.gitlab.io/atelier-openrefine-MASA/).

Ngo, D. et Z. Bellahsene (2012). Yam++ : A multi-strategy based approach for ontology matching task. *ten Teije A. et al. (eds) Knowledge Engineering and Knowledge Management. EKAW 2012. Lecture Notes in Computer Science, vol 7603. Springer, Berlin, Heidelberg 14*, 421–425.

Summary

ATMAH is a multidisciplinary study gathering computer science and information science applied to SHS. Its main purpose is to associate french cultural and scientific partners (BnF, Quai branly Museum, LESC, LIASD and the Maison des Cultures du Monde) involved in ethnomusicology audio archive dissemination, to improve the management of indexing vocabularies with regards to three aspects : access, enrichment and interoperability. The question

is how to enhances their valorisation at a national and international level in the context of the Linked Open Data. The challenge of the proposed study is to prove the interest of interconnection and the feasibility of its realization between several web data bases and platform systems, with specific internal vocabularies (structured in SQL, MARC, etc.) and external vocabularies ("hub" of international dimension, as french language of Bnf RAMEAU and MIMO consortium of Europeana, conform to the standards of the data web). Based on use cases derived from classifications of musical instruments, and other vocabularies in the domain (such as voice description, audio carriers, traditional music genres, named entities, etc.) it will be necessary to establish use scenari and specifications to be tested in accordance with the needs of information and research professionals (exchange, semi-automatised alignment, graphical and dynamical datavisualization, collaborative annotation). On a first step of the ATMAH project in 2016-2017, we started to work on both improving controled vocabularies management in Telemeta web audio plateform used by CREM team and testing alignment tools and procedures. ATMAH is a COMUE project [6] closely linked with the partners' ongoing programs (national and internation) as consortiums (TGIR Huma-Num from CNRS, France), LabEx Past in Present cluster, Paris Lumières University, Europeana Sounds.

6. ATMAH : Alignement Tool for Music Archive heritage s'inscrit dans le cadre de L'appel à projets COMUE 2016 de l?Université Paris Lumières

Identification automatique des sources des notices zoologiques du *Speculum naturale* de Vincent de Beauvais

Étienne Cuvelier*, Sébastien de Valeriolar* Céline Engelbeen*

*Laboratoire QUARESMI
ICHEC - Brussels Management School,
Boulevard Brand Whitlock, 2, 1150 Bruxelle
{etienne.cuvelier,sebastien.devaleriola,celine.engelbeen}@ichec.be
https://quaresmi.hypotheses.org/

Résumé.

Au XIII^e siècle, qui est parfois considéré comme « l'âge d'or » des encyclopédies médiévales, plusieurs auteurs regroupent et compilent l'ensemble des savoirs sous cette forme particulière. Ils agissent ainsi avec l'idée d'organiser l'ensemble des connaissances de l'époque (des textes antiques mais aussi des traductions de textes grecs ou arabes vers le latin, qui alors arrivent en Occident).

Une question naturelle se pose dès lors, celle de l'identification des sources utilisées par les encyclopédistes médiévaux. Ces encyclopédies formant un corpus très volumineux, cette démarche est longue et fastidieuse, raison pour laquelle elle n'a pas été entreprise par l'historiographie de manière complète et globale. C'est ce qui a motivé le développement de notre méthodologie d'identification automatique basée sur des outils de fouille de textes (*text mining*) que nous présentons dans cet exposé.

Parmi les encyclopédistes médiévaux les plus influents, nous avons choisi de travailler sur une partie de l'œuvre du dominicain Vincent de Beauvais. Décédé en 1264, ce dernier rédige le *Speculum maius*, qui forme l'une des encyclopédies les plus importantes du Moyen Âge. Parmi d'autres raisons qui ont guidé ce choix, signalons que cet auteur place en tête de chacune de ses notices un renseignement (généralement incomplet) sur la source utilisée. Cette information nous renseigne a priori sur la source effectivement utilisée par Vincent (et donc sur la qualité des identifications obtenues automatiquement par l'ordinateur), et nous place donc dans un cadre similaire à celui de l'apprentissage supervisé.

L'idée générale de notre méthodologie est de permettre de comparer une notice du texte de l'encyclopédie à un ensemble de sources potentielles sélectionnées a priori (par lesquelles on trouve Pline, Aristote, Ambroise de Milan, Isidore de Séville, Thomas de Cantimpré, etc.). Il s'agit donc de confronter chacune des notices considérées (soit 2 411 notices) à chacune des sources potentielles (soit 13 524 candidates sources). Avant de pouvoir effectuer cette comparaison, un ensemble d'activités de pré-traitement sont nécessaires, parmi lesquelles on peut citer la tokénisation et la lemmatisation. Ces opérations ne vont pas de soi, surtout lorsqu'il s'agit de traiter du texte en latin médiéval. Chacune de ces comparaisons produit une métrique de comparaison, puis, sur base de l'ensemble des résultats, l'ordinateur sélectionne la source potentielle la plus probable. Plusieurs métriques de comparaison ont été testées ; c'est la similarité *cosine* qui a finalement été retenue. Celle-ci nous permet d'atteindre un taux global

d'identifications correctes de 87,34%, c'est-à-dire que dans 87,34% des cas, la notice identifiée par l'ordinateur comme une source de l'encyclopédiste médiéval correspond au renseignement placé par celui-ci en tête de la rubrique correspondante. Ce résultat global suggère que notre méthodologie pourrait aider à identifier les sources d'autres encyclopédies qui, quant à elles, ne mentionnent pas de tel renseignement, et dont les sources ne sont donc a priori pas connues. Il nous semble de plus que l'intérêt de notre travail ne s'arrête pas là : l'analyse des sources qui n'ont pas été identifiées correctement par l'ordinateur nous permet de mieux comprendre comment Vincent de Beauvais utilisait les ouvrages qu'il considérait comme des références, et les conditions dans lesquelles il travaillait. Il est par exemple intéressant de remarquer que les auteurs utilisés par le dominicain ne sont pas tous logés à la même enseigne : certains sont très bien reconnus par la similarité *cosine*, d'autres beaucoup moins bien (voir le tableau 1). Nous avançons des hypothèses permettant d'expliquer ces différences.

	Ambroise	Aristote	Augustin	Isidore	Orosius	Palladius	Pierre	Pline	Solin	LDNR
Ambroise	76	2				1		1		36
Aristote	2	324						3		84
Augustin		1	6							3
Isidore	4	9	1	248	1		1	2	4	52
Orosius					1					
Palladius						58		1		
Pierre							9			
Pline		9		5		5	1	870	2	49
Solin	1	1						1	85	6
LDNR		13	1		1			1		430

TAB. 1 – *Distribution des attributions automatiques (colonnes) en fonction des auteurs indiqués par Vincent (lignes)*

Références

[1] Étienne Cuvelier, de Valeriola, Sébastien, and Céline Engelbeen. Identification automatique des sources des notices zoologiques du *Speculum naturale* de Vincent de Beauvais. *Soumis à la Revue d'intelligence artificielle*.

[2] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, 1986.

[3] B. Van den Abeele. Vincent de Beauvais naturaliste : les sources des livres d'animaux du *Speculum naturale*. In S. Lusignan, M. Paulmier-Foucart, and M.-C. Duchenne, editors, *Lector et compilator : Vincent de Beauvais, frère prêcheur : un intellectuel et son milieu au XIII^e siècle*, pages 127–151. Créaphis, Grâne, 1997.

Extraction et formalisation du savoir-faire industriel

Meriem MEJHED-MKHININI* Ouassila LABBANI NARSIS* Christophe NICOLLE*

* Laboratoire Le2i - Arts & Métiers, Univ. de Bourgogne Franche-Comté, Dijon, France.
{meriem_mejhed-mkhinini, ouassila.narsis,cnicolle}@u-bourgogne.fr

Résumé. L'expression et la formalisation du savoir-faire dans le milieu industriel sont limitées par la culture du secret et la nature implicite de l'expérience associée. Pour transmettre ce dernier, il faut le montrer en utilisant des termes spécifiques dont le sens est hypercontextualisé. Pour répondre à ces contraintes, nous proposons une démarche, qui, de la conduite d'interviews des experts métiers à la spécification formelle des connaissances dans une ontologie permet de construire des raisonnements numériques sur des savoir-faire industriel.

Au sein des sociétés industrielles, l'expression orale et la démonstration sont les véhicules principaux de transmission des savoir-faire. Les sachants expliquent et montrent aux apprenants sans que ces informations apparaissent dans les documentations techniques. Le savoir-faire associe des connaissances, des contraintes et usages et des raisonnements. Avec l'évolution de la pyramide des âges, de nos jours, beaucoup de sachants quittent leur entreprise avant d'avoir rencontré leurs successeurs. Il y a une perte importante de savoir-faire, une augmentation des pannes machines, parfois une perte de qualité de production et surtout une incapacité pour l'entreprise à anticiper des coûts lors de la budgétisation de nouveaux contrats clients. À l'heure de l'industrie 4.0 où la robotisation semble le nouvel eldorado pour les industriels, l'importance de la valeur humaine de la connaissance semble être, enfin, dans toutes les têtes.

Pour répondre à cet enjeu capital pour l'avenir de l'industrie, nous proposons une approche pour capter cette connaissance souvent implicite, de la qualifier pour en déterminer la valeur et la vérité, de la formaliser dans des systèmes numériques de raisonnement explicable. Car le savoir-faire est une connaissance impliquant un raisonnement. Dans notre contexte, Ce raisonnement doit être transmis ensuite aux nouvelles générations et notre ambition est que l'IA enseigne à l'apprenant ce qu'elle a capté des sachants.

Notre approche est née de la combinaison de deux champs théoriques, la psychologie du développement et l'ingénierie des connaissances. Pour extraire le savoir-faire des experts, nous avons utilisé la didactique professionnelle qui a pour but l'analyse du travail pour transférer les compétences dans le cadre d'une formation (Pastré (2002)). Son principe est basé sur une conduite d'entretien individuel à l'aide d'une grille de questions. Ces entretiens sont ensuite retranscrits pour obtenir une expression textuelle du savoir-faire. Pour qualifier la valeur de ces retranscriptions, un travail commun entre l'interviewer et le service informatique est mené. L'objectif est de passer d'une représentation informelle exprimée en langage naturel en expression homogène et semi-formelle sous la forme de diagrammes UML [1]. Ce standard de modélisation d'un domaine selon le paradigme objet propose un ensemble de diagrammes permettant l'expression des différentes facettes d'un savoir-faire : la définition des éléments statiques de la connaissance, la définition des éléments dynamiques associés (les usages) et une représentation des contraintes simples. Cette approche permet aux informaticiens de construire un espace numérique d'expression des savoirs pour l'entreprise. Ce logiciel répond aux besoins d'expression, mais pas à la nécessité d'une compréhension numérique. UML ne permet

1. https ://www.omg.org/

qu'une représentation semi-formelle de la connaissance. Cela limite la représentation des raisonnements associés au savoir-faire. Pour répondre à cette limitation, nous devons traduire cette représentation UML dans une équivalence sémantique formelle. Dans notre étude, nous avons identifié les ontologies (Studer et al. (1998)) et le domaine du linked data (Bizer et al. (2009)) pour permettre cette représentation sémantique formelle de la connaissance. Plusieurs travaux ont étudié la relation entre UML et les langages d'ontologie. Certaines approches (Cranefield (2001)), utilisent des règles de transformation définies en XSTL [2]. D'autres approches sont basées sur des métamodèles tels que l'ODM (Ontology Definition Meta-model) ((OMG) (2014)). La spécification ODM présente un métamodèle pour le langage d'ontologie OWL ainsi qu'un profil UML pour modéliser des ontologies sans préciser la mise en œuvre d'outils pour la traduction d'UML vers OWL.

D'une manière globale, les travaux de transformations d'UML vers une ontologie se restreignent à l'aspect syntaxique des éléments en UML. Pour répondre au mieux à la représentation du savoir-faire, il faut non plus travailler au niveau de la structure, mais au niveau du sens. C'est l'ambition de notre recherche, une transformation guidée par le sens, limitant la perte sémantique. Pour cela, nous avons développé un ensemble de mesures de similarité sémantique en liaison avec la base linked open vocabularies (lov) [3]. Cette base décrit plus de 650 ontologies. Cette approche nous permet d'identifier les ensembles sémantiques correspondant à la description des connaissances exprimées en UML. Nous travaillons actuellement sur la sélection des constructeurs sémantiques les plus pertinents et leur alignement. Le résultat formera une proposition d'ontologie qui sera amendée par les experts métiers à l'issue de notre démarche de didactique professionnelle. Une première ontologie construite manuellement servira de référence pour étudier les écarts potentiels. Cette ontologie, déjà en service, a permis de démontrer la faisabilité industrielle du raisonnement sur l'espace d'expression numérique des connaissances développé au sein de l'entreprise partenaire.

Références

Bizer, C., T. Heath, et T. Berners-Lee (2009). Linked data - the story so far. *Int. J. Semantic Web Inf. Syst. 5*(3), 1–22.

Cranefield, S. (2001). Uml and the semantic web. In *Proceedings of the International Semantic Web Working Symposium, Palo Alto*.

(OMG), O. M. G. (2014). Ontology Definition Metamodel. Number : formal/14-09-02.

Pastré, P. (2002). L'analyse du travail en didactique professionnelle. *Revue française de pédagogie* (138), 9–17.

Studer, R., V. R. Benjamins, et D. Fensel (1998). Knowledge engineering : Principles and methods. *Data Knowl. Eng. 25*(1-2), 161–197.

Summary

The expression and formalization of know-how in industry is limited by its confidentiality and the implicit nature of the associated experience. To pass it on, we must show it using specific terms, whose meaning is hyper-contextualized. To answer these constraints, we propose an approach, which, from conducting interviews of business experts to the formal specification of knowledge in an ontology can build numerical reasoning on industrial know-how.

2. eXtensible Style Sheet Language Transformations
3. https ://lov.linkeddata.es/dataset/lov

Apport du Text Mining pour l'exploration de relations dans les textes. Application à la découverte d'appariements entre objets d'intérêt et localisations dans la Tapisserie de Bayeux

David Condaminet*, Antoine Widlöcher*, Pierre-Yves Buard**
Bruno Crémilleux*, Julia Roger**

*Normandie Univ., UNICAEN, ENSICAEN, CNRS – UMR GREYC, Caen, France
prenom.nom@unicaen.fr
**MRSH, UNICAEN, CNRS – USR 3486, Caen, France
prenom.nom@unicaen.fr

Résumé. Nous présentons dans ce travail une approche fondée sur l'utilisation de méthodes de *Text Mining* pour l'exploration de relations dans des textes issus des Humanités, et plus précisément la découverte d'appariements entre descriptions textuelles d'objets d'intérêt (personnages, lieux, événements) et localisations dans la *Tapisserie de Bayeux*. Cette approche repose sur une phase d'amorçage permettant d'obtenir un premier appariement minimal mais fiable entre des objets d'intérêt et leurs localisations, le corpus contenant de tels appariements devant ensuite être exploré par une étape de fouille de données textuelles focalisée sur les mécanismes de mise en relation. Nous présentons ici notre démarche globale et détaillons le processus d'amorçage, qui repose notamment sur la mise en place d'un environnement d'observation mettant l'humain au centre du processus d'analyse et de contrôle.

1 Introduction

Les textes procèdent par nature massivement à des mises en relation, sur différents plans (syntaxique, sémantique...), à différentes échelles (propositions, phrases, discours, texte...) et selon différents paradigmes (relations de dépendance, causales, rhétoriques, temporelles, coréférentielles ...). L'accès aux informations véhiculées par ces relations constitue un enjeu majeur pour l'interprétation des données textuelles, et en particulier pour des tâches d'extraction d'information où il ne s'agit pas seulement d'identifier dans les textes des éléments de sens isolés (par exemple de entités nommées, des personnes, des lieux...), mais des dispositifs au sein desquels différents éléments interagissent (présence d'une personne en un certain lieu, à un certain moment... par exemple).

Dans la perspective d'exploiter automatiquement ces relations portées par les textes, la description manuelle des configurations linguistiques correspondantes, en vue de leur projection sur corpus, s'avère souvent si coûteuse qu'on souhaite au moins partiellement l'automatiser, et il n'est donc pas surprenant que différents travaux proposent des méthodes permettant l'extraction automatique des relations dans les textes, parmi lesquels on peut citer par exemple Mé-

tivier et al. (2015) dans le domaine de la recherche des relations entre gènes et maladies ou encore Tikk et al. (2013) et Cellier et al. (2015) pour la recherche d'interactions entre protéines ou entre gènes. Ces travaux sont fondés sur l'exploitation de méthodes de fouille de données ou d'apprentissage automatique.

Mais l'automatisation de l'extraction des règles suppose la disponibilité de données d'apprentissage en quantité suffisante, données au sein desquelles les objets liés sont pré-annotés. Faute d'en disposer, le recours à des ressources externes, telles que des terminologies ou ontologies de domaine porteuses de relations connues entre formes connues, peut permettre de sortir de l'impasse, par projection sur corpus de ces connaissances disponibles, en vue d'apprendre ensuite les modalités de leur mise en relation textuelle, modalités qui pourront ensuite être utilisées pour découvrir des configurations où s'articulent de manière similaire, au sein de relations inconnues, des formes connues ou elles-mêmes inconnues.

Dans le domaine des Humanités qui nous occupe ici, les mises en relation portées par les textes constituent aussi évidemment un enjeu d'importance. L'exploitation de méthodes automatisées du type de celles indiquées ci-dessus se heurte néanmoins à une double difficulté : (a) les données annotées sont encore globalement rares ; (b) les ressources externes projetables sur corpus pour procéder à une annotation automatique ne le sont pas moins.

Il s'avère donc nécessaire d'imaginer des méthodes pouvant être amorcées sur des données peu ou pas enrichies. C'est l'objet du travail que nous présentons ici.

Avant d'aborder la présentation de cette contribution, il convient de s'arrêter sur une difficulté inhérente aux phénomènes de mise en relation, difficulté qui n'est pas propre aux relations rencontrées dans les textes issus des Humanités, difficulté déterminante pour le choix de la méthode et suffisamment contre-intuitive pour être présentée ici, dès l'introduction. Une première approche des phénomènes de mise en relation pourrait en effet laisser penser qu'il s'agit là d'un phénomène pouvant être abordé de façon purement compositionnelle et ascendante, en considérant que l'analyse des relations suppose, dans cet ordre strict (1) la découverte des éléments isolés et (2) leur mise en relation. Il convient au contraire d'insister sur le fait que l'articulation entre éléments et relations s'avère en réalité beaucoup plus dialectique, l'élément isolé ne pouvant devenir significatif que parce qu'il entre dans une relation donnée. Si l'on vise par exemple les relations causales entre deux faits, il serait vain de penser que la cause ou la conséquence existent isolément et préalablement à cette relation.

Notre travail vise l'application de méthodes de *Text Mining* pour l'exploration de relations dans des textes issus des Humanités, et plus précisément la découverte d'appariements entre descriptions textuelles d'objets d'intérêt (personnages, lieux, événements) et localisations dans (c'est-à-dire positionnement sur) la Tapisserie de Bayeux, appariements décrits par des textes.

La méthode que nous présentons ci-après possède les propriétés suivantes, que nous souhaiterions mettre en emphase :

— elle est le fruit d'une étroite collaboration avec des chercheurs en SHS ;
— elle n'est pas limitée dans son principe au domaine spécifique dont elle relève ici ;
— elle met l'humain au centre du processus d'analyse et de contrôle ;
— elle permet de compenser partiellement la faible disponibilité des données enrichies.

La section 2 présente le contexte et les objectifs globaux de notre travail. La section 3 présente de manière détaillée la phase d'amorçage devant aboutir à la constitution d'un corpus d'apprentissage porteur d'appariements fiables entre localisations et objets d'intérêt. Nous y mettons notamment en évidence la nécessité d'interactions avec l'opérateur humain et présen-

tons, à la section 3.4, l'interface d'observation combinée des connaissances déjà acquises et des données textuelles.

2 Contexte et présentation schématique de notre méthode

Nous précisons dans cette section le contexte et les objectifs de notre travail. Nous commençons par donner quelques définitions.

2.1 Définitions

La Tapisserie est considérée, par simplification, comme une suite d'événements graphiquement décrits ;

Un objet d'intérêt (parfois noté OI ci-après) désigne la description textuelle d'un épisode, d'un personnage, d'un lieu ou de toute autre entité ou réalité prenant part à un événement décrit par la Tapisserie ;

Une scène est l'un des 58 segments conventionnellement délimités sur la Tapisserie ;

Une figure désigne une illustration de la Tapisserie présente dans notre corpus textuel et faisant référence, par sa légende, à une ou plusieurs scènes de la Tapisserie ;

Une localisation désigne la mention faite par le texte d'une séquence de la Tapisserie, celle-ci pouvant désigner une scène ou une figure – cette dernière étant indirectement et implicitement liée à une ou plusieurs scènes.

2.2 Contexte : événements et localisations dans la Tapisserie

Le travail que nous présentons ici s'inscrit dans un projet plus vaste visant à repenser un système documentaire donnant accès à l'abondante littérature relative à la Tapisserie de Bayeux. Dans ce cadre, notre travail doit plus particulièrement contribuer à l'effort de mise en relation des documents ou des parties de documents avec les zones de la Tapisserie qu'ils décrivent ou qu'ils commentent.

Notre corpus de travail est constitué des actes d'un colloque international (Bouet et al. (2004)) consacré à la Tapisserie de Bayeux, dont les contributions sont supposées représentatives des mécanismes par lesquels les chercheurs font référence aux différentes parties de l'œuvre brodée. Les événements rapportés par la Tapisserie peuvent ainsi faire l'objet, dans ce corpus, de descriptions textuelles désignant des objets d'intérêt (« l'envoi du messager », « l'arrivée de Guillaume en Angleterre », « la mort d'Edouard »...), ainsi que d'une localisation sur la Tapisserie (par exemple par l'indication d'un numéro de scène). L'ensemble documentaire est donc porteur d'une collection de références aux parties de l'ouvrage pouvant soit décrire textuellement les objets d'intérêt, soit donner des localisations, soit, dans certains cas, fournir à la fois la description textuelle et la localisation. Ainsi, on trouve par exemple : « Conan s'enfuit de Dol (scène 18) » (p. 191), « La scène 10 montre la rencontre des envoyés de Guillaume » (p. 191), « La scène 32, celle de la comète » (p. 193).

D'une manière générale, il est utile de disposer d'un appariement entre localisation et descriptions textuelles, pour que des localisations puissent être automatiquement rapportées

aux objets d'intérêt qui les concernent et que, inversement, les descriptions textuelles d'objets d'intérêt puissent être localisées sur la Tapisserie. Les avantages qui en résulteraient en matière d'aide à la navigation sont assez évidents : depuis la description textuelle d'un objet d'intérêt, on pourrait par exemple automatiquement rediriger le lecteur vers la ou les zones correspondantes sur la Tapisserie. Mais il deviendrait dès lors également possible d'enrichir significativement les modes d'indexation des contenus associés à la Tapisserie. Par exemple, les documents privilégiant le référencement par localisation pourraient bénéficier d'une indexation s'appuyant sur les descriptions textuelles issues du *mapping*, pour autoriser ainsi notamment une recherche par mots-clés, là où les seules localisations sont pourtant explicitement indiquées dans le texte.

2.3 Objectifs

L'objectif du travail présenté ici est la mise en place d'une solution permettant l'extraction semi-automatisée de ces appariements entre descriptions textuelles et localisations. Plus précisément, il s'agit d'extraire le *mapping* à partir de passages où coïncident une localisation et une description textuelle, passages pouvant ressembler aux exemples présentés ci-dessus. Pour y parvenir, il convient de localiser dans le corpus les passages porteurs de cette double identification. Pour cela, l'idée directrice de notre méthode est d'exploiter des techniques de fouille de texte pour la découverte de patrons linguistiques caractéristiques de la double identification, motifs généralisant des énoncés de mise en relation. La section suivante présente la méthode d'ensemble que nous proposons, méthode qui repose a) sur la construction d'un premier appariement minimal mais fiable (phase qui sera détaillée ensuite dans la section 3) puis b) son enrichissement grâce à la fouille de données séquentielles.

2.4 Présentation schématique de la méthode

La méthode proposée repose sur les deux phases complémentaires présentées ci-après, chacune pouvant être répétée itérativement.

2.4.1 Phase d'amorçage

L'objectif de la phase d'amorçage (cf. figure 1), phase qui pourra faire l'objet de plusieurs itérations, est d'obtenir un premier appariement minimal mais fiable, appariement associant des représentations textuelles d'OI et des localisations. La méthode retenue consiste à :

1. adopter un formalisme permettant l'expression de règles décrivant des patrons linguistiques susceptibles de capter un appariement entre description et localisation. Ce formalisme doit : a) être exploitable pour la définition manuelle de règles ; b) être utilisable lors des étapes ultérieures du projet pour encoder les motifs appris lors du processus de fouille (cf. phase suivante) ;
2. utiliser ce formalisme pour décrire manuellement les patrons correspondant à certains motifs triviaux immédiatement repérables (notamment dans les légendes des figures) ;
3. projeter ces patrons sur corpus pour extraire l'information utile des passages correspondants ;
4. capitaliser les informations d'appariement fiables ainsi extraites.

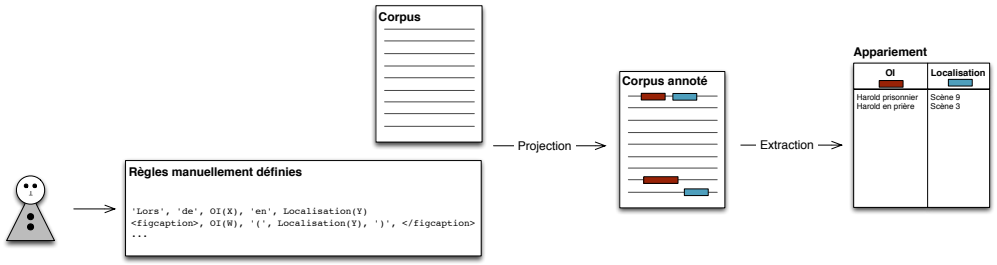

FIG. 1 – *Processus d'amorçage*

2.4.2 Phase d'enrichissement par exploration des relations

La phase d'enrichissement (cf. figure 2), qui pourra elle aussi faire l'objet de plusieurs itérations, consiste à exploiter l'appariement fiable pour isoler des énoncés où apparaissent OI et localisations que l'on sait fortement corrélés, pour découvrir automatiquement dans ces énoncés des structures textuelles caractéristiques de la mise en relation entre OI et localisations. La méthode retenue repose sur l'enchaînement suivant :

1. on dispose d'un corpus enrichi au sein duquel sont déjà annotées des informations relatives aux OI et aux localisations que l'on sait fortement corrélés ;
2. on sélectionne des énoncés comportant un OI et une localisation ;
3. on applique sur les passages sélectionnés une méthode de fouille séquentielle, pour apprendre des règles de mise en relation ;
4. ces règles sont reformulées selon le formalisme exploité à la phase 1 ;
5. ces règles reformulées sont projetées sur le reste du corpus, conformément au schéma présenté en figure 1, en vue d'enrichir nos connaissances sur l'appariement.

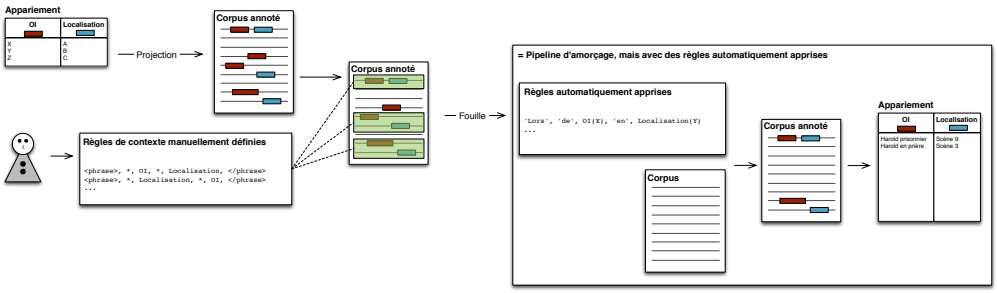

FIG. 2 – *Processus d'enrichissement*

Notons que l'application des règles apprises doit permettre non seulement de localiser de nouveaux appariements, mais aussi de déterminer les contours des OI (et dans une moindre mesure des localisations), ce qui fait écho au principe dialectique évoqué ci-dessus : un OI ne devient tel ici que dans la mesure où il intervient dans une relation. Intuitivement, l'OI est ce dont on parle et ce que l'on situe.

2.5 Avancement actuel de nos travaux

À l'heure où nous écrivons ces lignes, la phase 1, qui sera présentée en détail dans la section suivante, est toujours en cours. On le verra, l'amorçage, loin d'être réalisable naïvement, suppose la mise en place d'une interaction avec l'utilisateur, en vue d'obtenir un premier appariement suffisamment large pour permettre la constitution d'un corpus d'apprentissage suffisant pour la phase 2.

3 Extraction automatisée d'appariements entre descriptions textuelles et localisations

La première phase de notre travail consiste à mettre en place une méthode d'extraction semi-automatisée d'appariements fiables entre descriptions textuelles d'objets d'intérêt et leurs localisations à partir d'un corpus de textes d'étude, structuré au format XML, mais sans annotation préalable des localisations et objets d'intérêt. Ces appariements serviront de base pour la seconde phase du travail, phase présentée sommairement dans la section précédente mais qui ne sera pas détaillée dans cet article.

3.1 Initialisation de l'amorçage

Une lecture rapide du corpus permet d'identifier des énoncés d'appariement évidents dans les légendes de figures, ainsi que dans les titres de sections. Par exemple, on trouve dans Bouet et al. (2004) la légende "Fig. 3 : Tapisserie de Bayeux : Harold quitte l'Angleterre (scène 4)" (p. 200) ou le titre "La fuite de Conan (scène 18)" (p. 190).

On remarque également que, en présence d'illustrations, le numéro de figure est souvent utilisé dans le texte, pour mettre en correspondance indirectement les OI et les scènes décrites par les figures, sans mentionner explicitement les scènes. Il s'agit donc là également d'appariements fiables dont la détection est triviale et nous avons choisi d'utiliser les numéros de figure comme marqueurs de localisation : un énoncé mettant en relation un OI et une figure sera jugé aussi pertinent pour la suite qu'un énoncé établissant une relation entre un OI est un numéro de scène.

À partir de règles très simples exprimées avec des expressions régulières, nous avons extrait dans cette première phase d'amorçage 117 relations réputées certaines, i.e. 117 triplets distincts de type (*OI*, *scène*, *figure*) renvoyant globalement à 56 scènes – sur les 58 existantes –, 76 figures et 82 objets d'intérêt différents.

Sur la base de ces premiers appariements réputés fiables, une première projection sur le reste du corpus doit permettre de sélectionner un premier ensemble de phrases (ensemble devant être *in fine* utilisé comme corpus d'apprentissage pour la phase 2 d'exploration des relations) porteuses à la fois d'une localisation et d'un objet d'intérêt. Pour clarifier ce point, 3 ensembles de contextes doivent être distingués :

— CS : phrases contenant au moins une expression de *scène* ;
— CF : phrases contenant au moins une expression de *figure* ;
— COI : phrases contenant au moins une expression d'un *objet d'intérêt*.

Les contextes dits de localisation correspondent à l'ensemble $(CS \cup CF)$. D'une manière générale, la sélection d'énoncés devant constituer le corpus d'apprentissage $\mathcal{C}$ sera un sous-ensemble de $C = (CS \cup CF) \cap COI$.

Si nous limitons notre corpus d'apprentissage aux phrases (ensembles de mots, S ci-après) présentant une localisation (l ci-après) et un OI (o ci-après) ayant un lien fiable selon l'appariement $\mathcal{M}$ obtenu ci-dessus, le corpus d'apprentissage correspond alors à l'ensemble de phrases suivant :

$$\mathcal{C} = \{S \mid S \in C, \exists o \in S, \exists l \in S, (o, l) \in \mathcal{M}\}$$

Ainsi définie, la sélection d'énoncés devant constituer le corpus d'apprentissage pour la phase ultérieure serait de taille dramatiquement réduite. En effet, sur les données issues de Bouet et al. (2004) : $|\mathcal{C}| = 23$.

Notons que, même en relaxant la contrainte sur la fiabilité de l'appariement, le volume d'énoncés n'augmente pas de manière très importante, ce qui renvoie au fait que les OI et les localisations issues de cette toute première phase sont simplement en nombre trop réduit. En effet : $|C| = 56$.

3.2 Enrichissement de l'amorçage

Il est dès lors évidemment nécessaire de procéder à un enrichissement des données d'amorçage. Pour cela, il est notamment indispensable de pouvoir :

— contrôler l'état actuel des connaissances acquises, c'est-à-dire les appariements fiables ;
— prévisualiser l'état actuel du corpus $\mathcal{C}$ en cours de constitution ;
— identifier des configurations proches de celles qui sont actuellement sélectionnées et qui mériteraient une formalisation explicite dès la phase d'amorçage, en vertu de leur fiabilité.

De cette triple exigence ont résulté les étapes suivantes de notre travail, étapes ayant principalement consisté jusqu'ici à :

1. mettre en place un environnement permettant à l'opérateur informaticien et/ou humaniste une navigation efficace dans les connaissances acquises et les contextes d'acquisition (voir section 3.4) ;
2. proposer différentes méthodes d'élargissement des données visualisables via cette interface, pour permettre à l'opérateur d'identifier les configurations non encore prises en charge mais devant être traitées en priorité.

3.3 Élargissement des données observables

L'élargissement des données observables, qui vise à placer sous le regard de l'expert des configurations probablement remarquables, en vue d'améliorer l'amorçage, est réalisé selon deux axes complémentaires :

1. à un niveau local, un élargissement de la présentation des OI ;
2. à un niveau global, un élargissement de la présentation des contextes de mise en relation entre OI et localisation.

3.3.1 Élargissement au niveau des OI

Ce premier élargissement, au niveau local des OI, vise à permettre à l'opérateur de clarifier la nature des OI. Pour cela, la démarche retenue consiste à analyser les OI déjà sélectionnés, en quête de régularités exprimées par des motifs séquentiels, puis de projeter ces motifs sur le reste du corpus, pour voir s'ils correspondent à une formalisation acceptable de la notion d'OI.

Pour cela, nous nous appuyons sur :

— la bibliothèque Python *Natural Language Toolkit* (NLTK) Bird et al. (2009) pour la segmentation en phrases puis la projection des motifs ;

— *TreeTagger* de Schmid (1994) [1] pour la tokenisation, l'étiquetage morpho-syntaxique et la lemmatisation ;

— l'outil SDMC (*Sequential Data Mining under Constraints*) Béchet et al. (2015) pour la fouille de séquences.

Chacune des descriptions textuelles d'OI extraites précédemment est ici une transaction. Pour chaque transaction, chacun de ses mots, représenté par son étiquette, est un item. Nous extrayons alors les motifs séquentiels fréquents avec SDMC sur les OI préalablement étiquetés avec comme paramètres pour SDMC : $mingap = maxgap = 0$, $minsize = 3$, $maxsize = 5$ et $minsup = 10$. Après examen des motifs extraits, nous pouvons retenir 4 motifs, qui correspondent en fait, simplement, à des syntagmes nominaux :

```
{NOM_COMMUN} {PREPOSITION} {NOM_COMMUN}
{NOM_COMMUN} {PREPOSITION} {NOM_PROPRE}
{PREPOSITION} {DETERMINANT} {NOM_COMMUN}
{DETERMINANT} {NOM_COMMUN} {PREPOSITION} {NOM_PROPRE}
```

Nous pouvions évidemment largement subodorer ce fait, à la simple lecture des descriptions des OI. Toutefois, nous souhaitions mettre en place un *pipeline* réutilisable ensuite à chaque itération, en y incluant également la seconde phase non encore abordée (celle de l'exploration des relations) : à chaque nouvel élargissement de l'appariement, des nouveaux OI seront retenus, dont nous souhaitons pouvoir identifier les éventuelles structures communes, en permettant toujours à l'opérateur d'élargir aussi son horizon, en quête de nouvelles régularités significatives pouvant donner lieu à l'établissement manuel de règles jugées essentielles et fiables.

La projection des motifs sur le reste du corpus est réalisée à l'aide du module *RegexpParser chunker* de NLTK, dont l'expressivité est en adéquation avec notre représentation des données textuelles. Auparavant, l'ensemble du corpus est tokenisé à l'aide du TreeTagger et segmenté en phrases à l'aide du module *PunktSentenceTokenizer* de NLTK.

3.3.2 Élargissement au niveau des contextes de mise en relation entre OI et localisation

L'élargissement au niveau des contextes de mise en relation repose lui-même sur deux mesures complémentaires :

1. la relaxation de la contrainte d'appartenance du couple (o, l) à l'appariement fiable $\mathcal{M}$, avec l'hypothèse, acceptable pour les besoins de l'observation, que la co-présence d'un OI et d'une localisation dans le même contexte (ici la phrase) résulte souvent d'une mise en relation. Cette relaxation revient, pour reprendre la notation utilisée ci-dessus, à passer de $\mathcal{C}$ à C ;

1. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

2. la relaxation de la contrainte d'appartenance des OI à l'ensemble $\{o \mid \exists l, (o, l) \in \mathcal{M}\}$, c'est-à-dire l'élargissement des contextes à toute combinaison d'une localisation d'une part et, soit d'un OI, soit d'une instance quelconque de l'un des motifs extraits de la fouille d'autre part.

3.3.3 Effets du double élargissement

Sur la base de ce double élargissement, nous pouvons présenter à l'opérateur un ensemble de 151 contextes porteurs d'OI et de localisations, valeur à rapporter d'une part à $|\mathcal{C}| = 23$ et d'autre part à $|C| = 56$. Notons que, sans l'assimilation des mentions de figures à des localisations (assimilation scène/figure évoquée ci-dessus), le nombre de contextes observables tomberait de 151 à 77.

Reste alors à observer lesdits contextes, et à y repérer des structures fiables pouvant permettre d'enrichir le jeu de règles exploitables pour cette phase d'amorçage. Pour cela, nous plaçons entre les mains de l'opérateur une interface de navigation qui sera l'objet de la prochaine section.

3.4 Interface d'exploration des données

L'interface que nous présentons ici est un outil destiné à la fois au spécialiste de l'analyse des données, en charge du processus d'extraction automatisé, et au spécialiste de ces données spécifiques elles-mêmes, acteur des SHS.

Cette interface doit tout d'abord permettre l'exploration des données textuelles pour la découverte d'appariements fiables et la construction d'un corpus d'apprentissage utilisable lors de la phase d'analyse des relations. Elle sera également utile lors de cette seconde phase pour contrôler la progression de l'évolution des connaissances acquises.

Cet outil, dont le fonctionnement accorde une place privilégiée à l'utilisateur (cf. section 3.4.2), permet de guider le travail de construction incrémentale des connaissances en mettant l'humain au centre du processus d'analyse et de contrôle (cf. section 3.4.1).

3.4.1 Intérêt et rôle de l'interface

Cette interface web, dont la figure 3 fournit une première illustration, permet au spécialiste en sciences des données de *contrôler* l'évolution du processus d'extraction de connaissances, d'avoir une meilleure visualisation des contextes textuels dont les connaissances sont extraites et de prévisualiser le corpus d'apprentissage qui pourra résulter de l'exploitation des connaissances acquises. De surcroît, elle permet de confronter ce que le dispositif actuel capte déjà, et ce qu'il devrait idéalement capter, conformément aux mesures d'enrichissement présentées dans les sections précédentes. En conséquence de ces mesures, il pourra en effet observer des configurations proches de celles qui ont été déjà validées, y découvrant de nouvelles formes de mise en relation entre localisation et objets d'intérêt, dont certaines, suffisamment régulières et fiables, mériteront d'être explicitement formalisées et ajoutées au dispositif d'amorçage.

Pour un usager moins impliqué dans le processus d'extraction, mais spécialiste ou simple usager du corpus, par exemple documentaliste ou historien, cette interface offre un environnement efficace pour naviguer à la fois sur la *Tapisserie de Bayeux* et dans un corpus qui lui est dédié, au travers de différents critères de sélection : les scènes, les figures de l'ouvrage, ou

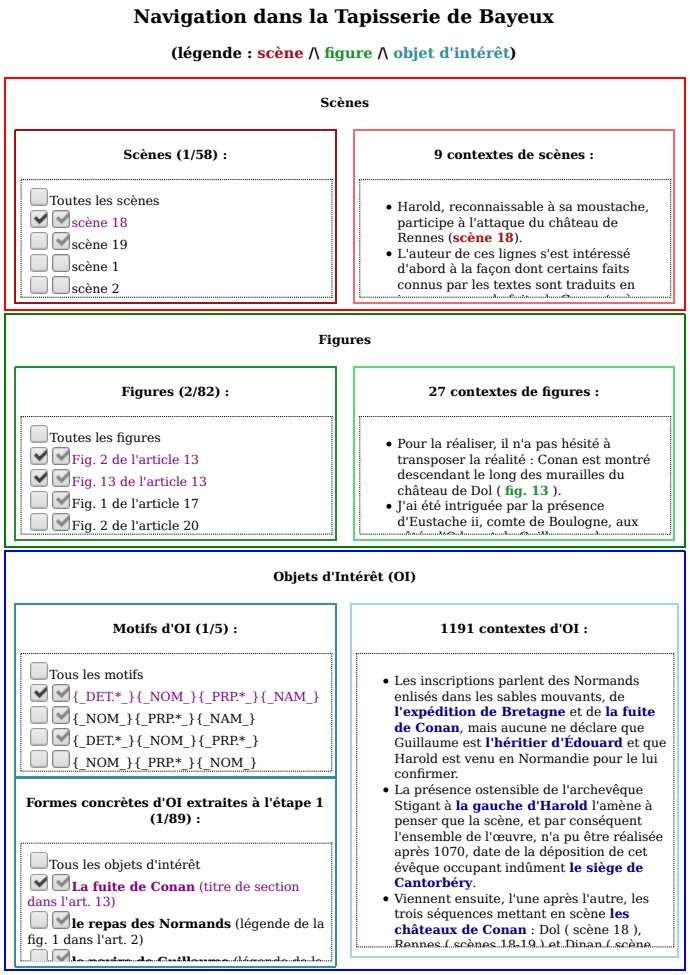

FIG. 3 – *Interface d'observation*

encore les objets d'intérêt, ces derniers pouvant être distingués par le degré de leur fiabilité, selon qu'ils sont extraits d'appariements jugés certains, ou de généralisations résultant de la projection de motifs.

3.4.2 Usage de l'interface

L'utilisateur est à l'initiative sur la construction des différents ensembles de contextes : CS, CF et COI (placés respectivement de haut en bas en colonne de droite), selon ce qu'il choisit de sélectionner respectivement comme scènes, figures, motifs et OI (de haut en bas en colonne de gauche). L'usage des quatre listes de sélection d'items se traduit par la mise à

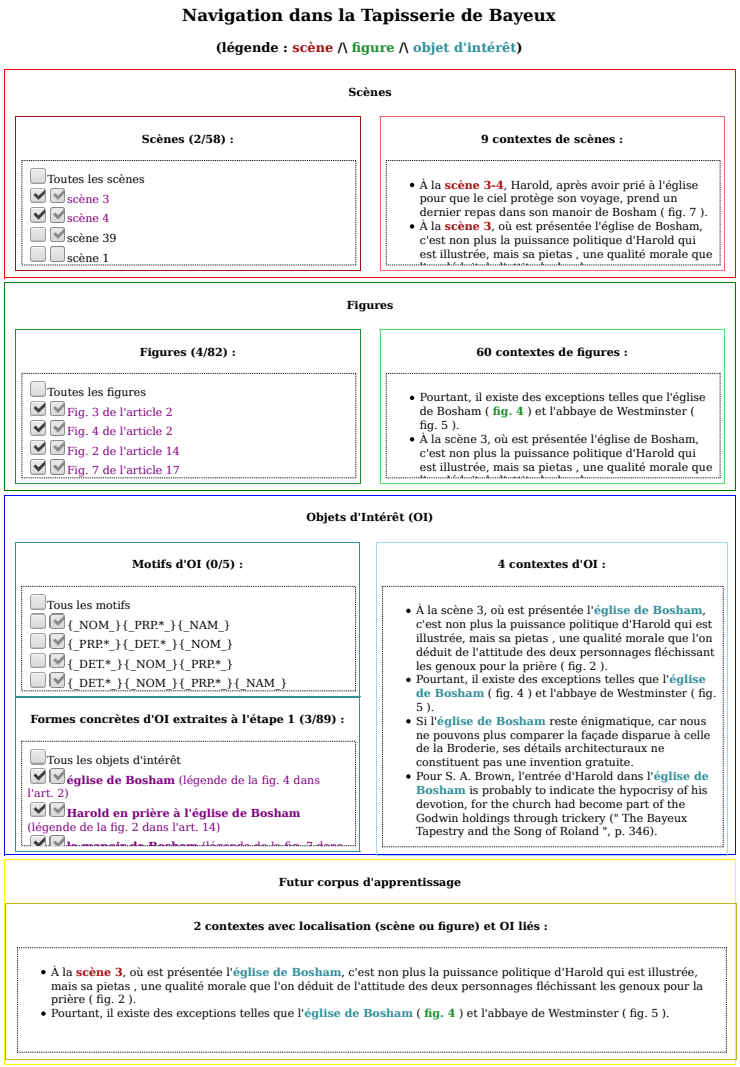

FIG. 4 – *Interface d'observation*

jour des contextes correspondants (à droite), et par la représentation conséquente du corpus d'apprentissage en cours de construction (en bas de la figure 4).

Lorsqu'il sélectionne un item d'une des catégories exposées ci-dessus, les items des 3 autres catégories étant liés à cet item, d'après l'état actuel des connaissances, sont pré-sélectionnés (via colonne des *checkboxes* située à droite). L'utilisateur peut alors, s'il le souhaite, poursuivre ce chemin de propagation des connaissances, en sélectionnant les items adéquats (via colonne des *checkboxes* située à gauche).

Le corpus d'apprentissage, pour sa part, est automatiquement déduit des ensembles de

contextes configurés par l'utilisateur, conformément aux mesures présentées en section 3.1. L'utilisateur contrôle ainsi, indirectement, la construction du corpus d'apprentissage.

La figure 4 illustre l'état de l'interface pour un cas d'utilisation lors duquel l'utilisateur recherche des événements qui ont eu lieu dans la ville de Bosham. On peut remarquer, au bas de cette figure, la construction résultante du corpus d'apprentissage $\mathcal{C}$.

4 Conclusion et perspectives

Dans cet article, nous avons présenté une démarche globale fondée sur l'utilisation de méthodes de fouille de données textuelles pour l'exploration de relations dans des textes issus des Humanités. Nous avons détaillé la première étape consistant à produire un premier appariement minimal mais fiable entre des objets d'intérêt et leurs localisations. Cette étape place l'humain au centre du processus, celui-ci intervenant via une interface de navigation et d'exploration des données. La poursuite de ce travail consistera désormais en l'exploitation de méthodes de fouille de données pour apprendre des règles de mise en relation.

De façon plus générale, la recherche de relations entre entités, par exemple entre personnes et localisations, est un sujet d'étude intéressant dans de nombreux autres corpus issus des Humanités. Nous projetons ainsi par exemple d'utiliser ce type de démarche pour l'étude du corpus des *Minutes du Procès Nuremberg* dont la MRSH de Caen a récemment mis une version numérisée à disposition de la communauté.

Références

Béchet, N., P. Cellier, T. Charnois, et B. Crémilleux (2015). Sequence mining under multiple constraints. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain, April 13-17, 2015*, pp. 908–914.

Bird, S., E. Loper, et E. Klein (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Bouet, P., B. Lévy, et F. Neveux (2004). *La Tapisserie de Bayeux : l'art de broder l'Histoire*. Office universitaire d'études normandes (ouen), Presses Universitaires de Caen.

Cellier, P., T. Charnois, M. Plantevit, C. Rigotti, B. Crémilleux, O. Gandrillon, J. Kléma, et J. Manguin (2015). Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. *J. Biomedical Semantics 6*(27).

Métivier, J.-P., L. Serrano, T. Charnois, B. Cuissart, et A. Widlöcher (2015). Automatic Symptom Extraction from Texts to Enhance Knowledge Discovery on Rare Diseases. In J. H. Holmes, R. Bellazzi, L. Sacchi, et N. Peek (Eds.), *Proceeding of the 15th Conference on Artificial Intelligence in Medicine, AIME 2015*, Pavia, Italy, pp. 249–254.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Tikk, D., I. Solt, P. E. Thomas, et U. Leser (2013). A detailed error analysis of 13 kernel methods for protein-protein interaction extraction. *BMC Bioinformatics 14*(12).

Le repas gastronomique des Français comme patrimoine culturel immatériel de l'humanité : caractérisation et transmission à travers les tweets des chefs 2 et 3 étoiles au *Guide Michelin*

Julien Longhi*,**,******, Zakarya Després**, Claudia Marinica**,***,
Vincent Marcilhac****,*****, Felipe Diaz Marin*****

*AGORA, EA7392, Université Paris-Seine
**Institut des Humanités Numériques de l'Université de Cergy-Pontoise
***ETIS UMR 8051, Université Paris-Seine, Université de Cergy-Pontoise, ENSEA, CNRS
**** Espaces Nature et Culture, ENeC - UMR8185, CNRS
*****Pôle de Gastronomie de l'Université de Cergy-Pontoise
****** Institut Universitaire de France (IUF)

{julien.longhi, claudia.marinica, zakarya.despres, vincent.marcilhac,
felipe.diaz-marin}@u-cergy.fr

Résumé. Depuis 2010, le "repas gastronomique des Français" a été inscrit sur la liste représentative du patrimoine culturel immatériel de l'humanité. Cette appartenance amène des questions liées à la préservation et la transmission de ce patrimoine. Ces tâches sont actuellement (parfois inconsciemment) réalisées en partie par ces chefs de cuisine reconnus et starisés, même si leur image élitiste est bien éloignée des valeurs de partage et de convivialité qui étaient au cœur du dossier de candidature à l'Unesco. Dans ce travail, nous nous intéressons à la transmission de ce patrimoine gastronomique réalisée par les chefs étoilés via les messages envoyés sur Twitter. Ainsi, nous proposons d'analyser les tweets des chefs 2 et 3 étoiles au *Guide Michelin* afin de caractériser le processus de transmission.

1. Introduction

Depuis 2010, le "repas gastronomique des Français" a été inscrit sur la liste représentative du patrimoine culturel immatériel de l'humanité[1] : parmi ses composantes importantes figurent "le choix attentif des mets parmi un corpus de recettes qui ne cesse de s'enrichir ; l'achat de bons produits, de préférence locaux, dont les saveurs s'accordent bien ensemble ; le mariage entre mets et vins ; la décoration de la table ; et une gestuelle spécifique pendant la dégustation (humer et goûter ce qui est servi à table)".

L'appartenance du repas gastronomique des Français au patrimoine de l'humanité, en tant que patrimoine culturel immatériel de l'humanité, amène à s'interroger sur la gastronomie française en tant que patrimoine, notamment du point de vue des données produites autour d'elle. En effet, le site de l'Unesco précise aussi que "des personnes reconnues comme étant des gastronomes, qui possèdent une connaissance approfondie de la tradition et en préservent

[1] https://ich.unesco.org/fr/RL/le-repas-gastronomique-des-francais-00437

la mémoire, veillent à la pratique vivante des rites et contribuent ainsi à leur transmission orale et/ou écrite, aux jeunes générations en particulier". Donc, quand on traite du repas gastronomique des Français en tant que patrimoine immatériel, on ne s'intéresse pas seulement aux plats qui font partie de la gastronomie française, mais aussi aux personnes qui permettent la préservation et la transmission de ce patrimoine qui, dans le cas d'œuvres d'art matérielles, peuvent être comparées aux conservateurs ou restaurateurs. A ce sujet, le ministère de l'agriculture[2] précise également qu'il s'agit d'un patrimoine à transmettre, et qu'il est important de sauvegarder le repas gastronomique des Français.

Dans ce travail, notre objectif est de chercher à caractériser les discours des "personnes reconnues" en matière de gastronomie, afin de voir si elles contribuent à la préservation de ce patrimoine, et, si tel est le cas, de chercher à voir comment les sciences des données peuvent aider à appréhender cette patrimonialisation. Pour atteindre cet objectif, nous proposons d'analyser les tweets des chefs deux et trois étoiles *au Guide Michelin* et/ou de leurs restaurants à l'aide d'analyses statistiques. Dans la littérature, des travaux autour des tweets gastronomiques permettent d'analyser les tweets envoyés par les clients et non par les chefs ou les restaurants (Mouritsen et al., 2017), avec souvent des objectifs autour de la santé (Harris et al., 2017, Nguyen et al., 2017). A notre connaissance, il n'y a pas des travaux qui proposent d'analyser les tweets des chefs/restaurant afin de comprendre la transmission de ce patrimoine.

La suite de l'article est organisée comme il suit : la section 2 présente les données que nous avons traitées. La section 3 présente la méthodologie mise en place et la section 4 présente les résultats préliminaires. La dernière section conclut l'article.

2. Description des données

Pour prendre en comptes des discours faisant autorité en matière de gastronomie, nous avons choisi de nous intéresser aux chefs qui ont obtenu deux ou trois étoiles au *Guide Michelin* dont l'attribution des étoiles repose sur des critères identiques afin de garantir la cohérence de la sélection. Ces critères sont au nombre de cinq : qualité des produits, maîtrise des cuissons et des saveurs, personnalité du chef dans sa cuisine, rapport qualité/prix et régularité dans le temps et sur l'ensemble de la carte. Les étoiles ne jugent que « ce qui est dans l'assiette » ; elles viennent uniquement récompenser la qualité de la cuisine. 3 étoiles indiquent qu'il s'agit « d'une cuisine remarquable, valant le voyage » et 2 étoiles désignent « une table excellente méritant un détour »[3]. Le choix du *Guide Michelin* et des chefs récompensés par 2 ou 3 étoiles comme échantillon de notre étude se justifie par le fait que le *Guide Michelin* reste en France le guide gastronomique de référence dont le classement fait autorité et par le fait que l'attribution de 2 ou 3 étoiles apporte une grande notoriété aux restaurants et aux chefs distingués, dont le discours est relayé par les médias.

Parmi cette liste des restaurants primés par le *Guide Michelin*, nous avons sélectionné ceux qui possédaient un compte Twitter ou dont le chef gère son propre compte Twitter, voire les deux. Nous avons pu extraire 47 923 tweets de ces 61 comptes, ce qui représente

[2] http://agriculture.gouv.fr/le-repas-gastronomique-des-francais-un-patrimoine-culturel-immateriel-de-lhumanite

[3] https://fr.wikipedia.org/wiki/Liste_des_restaurants_deux_et_trois_étoiles_du_Guide_Michelin

775 754 mots. Nous nous sommes servis d'un script Python basé sur la librairie Tweepy, qui permet de faire appel à l'API de Twitter pour récolter la plupart des tweets à partir d'un nom de compte. Certains comptes sont beaucoup plus prolifiques que d'autres : seuls 7% des comptes représentent 25% de notre corpus, c'est à dire 4 comptes sur 60.

3. Méthodologie

Du point de vue méthodologique, nous cherchons à utiliser un outil qui prenne en compte la matérialité linguistique du corpus : nous nous attachons donc aux formes linguistiques (lemmes notamment), au moins dans un premier temps. Ainsi, si l'usage d'outils ou de mesures issus de la fouille de textes peut être envisagé dans un premier temps, une analyse préalable du point de vue textométrique nous semble être une démarche intéressante. La textométrie « (ou statistique textuelle, lexicométrie, logométrie) propose une approche instrumentée des corpus, articulant synthèses quantitatives et analyses à même le texte » (Lebart et Salem, 1994). En plus de fournir des procédures de tri et de calculs statistiques pour l'étude de corpus numériques, elle « établit une modélisation contextuelle et contrastive : le texte est caractérisé par ses mots par rapport à leur usage dans le corpus, le mot est caractérisé par ses cooccurrents, etc. » (Pincemin, 2012).

La spécificité du corpus doit aussi être prise en compte, afin que chaque tweet soit considéré comme une unité textuelle (un texte). Parmi les différents logiciels existants, le choix de *Iramuteq* (http://www.iramuteq.org) pour ce travail exploratoire est motivé par notre double intérêt conceptuel pour les liens entre formes et thèmes d'un côté, et entre formes et profilages d'un autre côté. Ceci s'inscrit dans le cadre de la *Théorie des objets discursifs* (Longhi, 2015), qui s'inspire, à partir du concept d'*objet discursif* (Longhi, 2008), de la *Théorie des formes sémantiques* (Cadiot et Visetti, 2001), qui mobilise les concepts de *motifs*, *profils* et *thèmes*. Aussi, nous retenons la classification lexicale (Ratinaud et Marchand, 2015) implémentée dans *Iramuteq* car elle permet de faire ressortir les thématiques propres aux chefs est de « regrouper des *mondes lexicaux* et de mettre en évidence les thématiques générales du corpus », la méthode cherchant à « rendre compte de l'ordre interne d'un discours, à mettre en évidence ses mondes lexicaux » (https://datahist.hypotheses.org/11). Elle découle de l'analyse factorielle des correspondances.

L'autre fonctionnalité, l'Analyse de similitudes (ADS), utilisée sur la partie de corpus pertinente pour l'étude (dans la première partie de la section 4), établit un calcul sur la base d'« un indice de co-occurrence (combien de fois les éléments vont apparaitre en même temps) » pour donner un résultat visuel « où la taille des mots est proportionnelle à la fréquence et où la taille des arêtes et proportionnelle à la force ». Ceci permet de rendre compte visuellement de la fréquence des mots en lien avec les associations spécifiques. Cette fonctionnalité représente d'une certaine manière le « profilage » des unités, c'est-à-dire qu'elle rend compte de leur stabilisation dans le corpus à travers des associations fréquentes qui « profilent » les usages des formes dans tel ou tel domaine ou pratique.

Pour la classification Reinert, en considérant des corpus parfois de taille considérable plusieurs dizaines de classes lexicales peuvent être produites. Dans notre cas le corpus est de taille raisonnable, et le logiciel propose trois grands ensembles, comme le montre la figure 1.

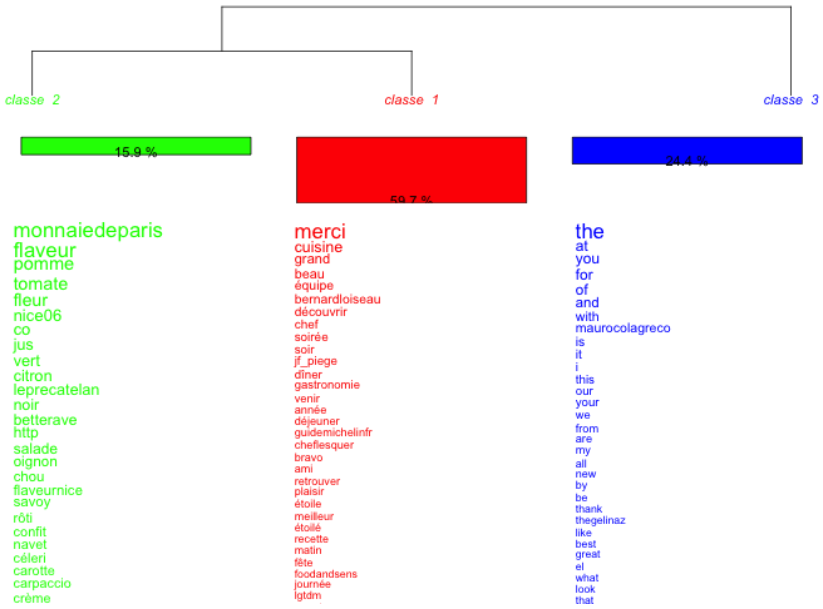

FIG. 1 - *Classes obtenues après l'utilisation de la classification Reinert sur le corpus.*

On trouve donc trois classes, ou univers dans ce corpus. Une partie importante du corpus, représentée par la classe centrale ayant près de 60% des tweets, est en lien avec de la "promotion", et une autre partie, représentée par la classe de droite et par 24% des tweets environ, s'y apparente aussi, mais en anglais. Un exemple de tweets de ce type est donné dans la figure 2.

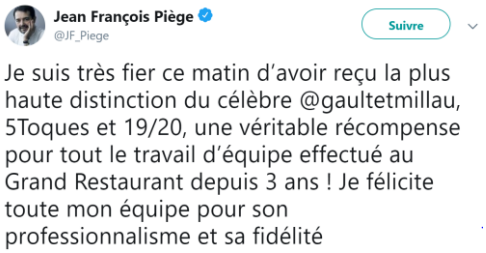

FIG. 2 – *Tweet de promotion envoyé par le chef Jean François Piège.*

Cependant, près de 16% (la classe de gauche) sont néanmoins des messages contenant des informations culinaires à proprement parler, avec des termes concernant des ingrédients ou des techniques/recettes.

En filtrant le corpus afin de constituer un sous-corpus spécifique à cette classe, on peut avoir une vision moins bruitée de ce que pourrait être le "patrimoine culinaire" à travers les discours numériques des grands chefs. Pour cela, on utilise la fonction "Segments de texte

(ST) caractéristiques" de la classe considérée, puis on exporte tous les ST présents dans la classe, afin de constituer, avec leur regroupement, un nouveau corpus.

L'analyse de ce nouveau corpus permet ensuite d'utiliser d'autres fonctionnalités, comme l'analyse de similitudes, qui produit des graphes à partir de la librairie « igraph » de R. Le tableau en entrée est un tableau de présence / absence. La matrice de similitude est calculée à partir de l'un des indices proposés. La plupart des indices proposés sont issus de la librairie « proxy » de R.

4. Résultats

L'utilisation de l'analyse de similitudes produit le résultat présenté dans la figure 3. Dans cette figure, nous avons un ensemble de clusters qui permettent de trouver des ingrédients et leurs associations, des parties de recettes perceptibles à travers des désignations de plats, ou encore des assortiments ou des spécifications.

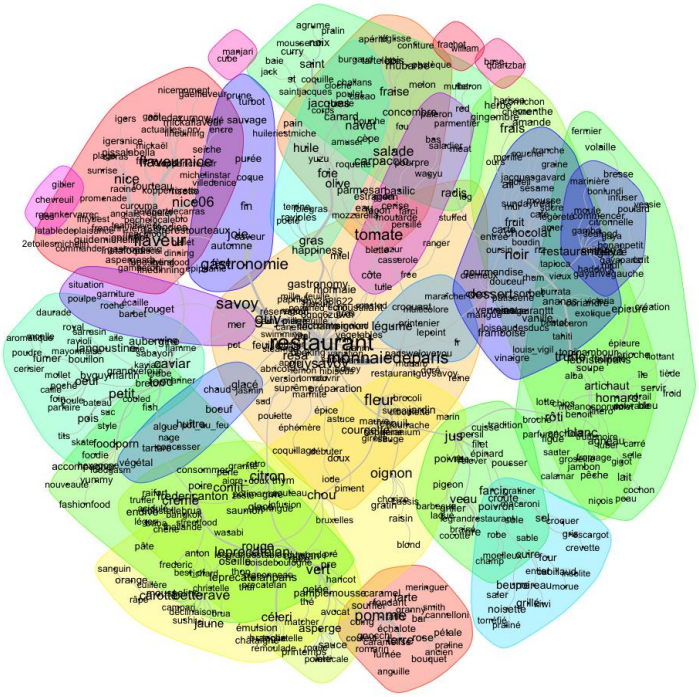

FIG. 3 - *Analyse de similitudes du sous-corpus constitué.*

Pour aller plus loin dans l'analyse, nous utilisation l'outil Gephi (https://gephi.org/) qui permet de mettre ces conclusions visuelles en valeur, comme l'indique la figure 4 avec par exemple, à droite, des termes tels que "crabe", "carpaccio", "cassis", "raisin", etc.

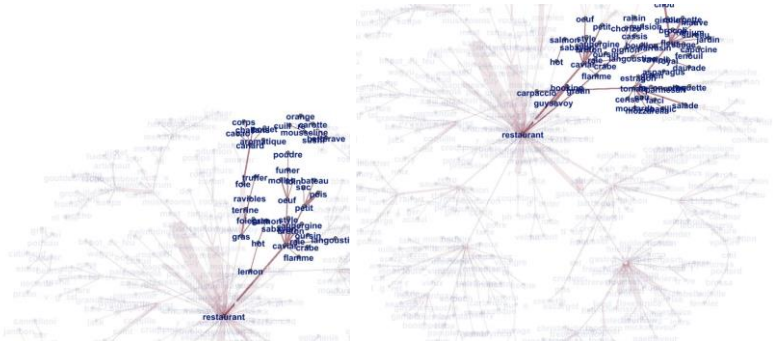

FIG. 4 - *Gephi comme outil de visualisation de "zones" liées dans le cluster général*

La recherche devra donc se poursuivre pour analyser plus précisément les liens entre ces termes, la manière dont on peut utiliser ces données pour appréhender le patrimoine gastronomique, mais il apparaît clairement que ce corpus peut fournir des indications intéressantes pour cartographier les différents aspects de ce patrimoine : ingrédients, techniques, associations, mise en scène du produit.

Une piste intéressante pour expérimenter cela consistera notamment à utiliser la fonction POS Tagging (Part of speech) afin de regarder la catégorie grammaticale des mots utilisés, et dans chaque catégorie les plus fréquents (figure 5).

Forme	Freq.	Types
farcir	137	ver
rire	98	ver
cuire	97	ver
fumer	89	ver
souffler	72	ver
accompagner	67	ver
griller	65	ver
découvrir	61	ver
servir	55	ver
consommer	49	ver
commencer	41	ver
saler	39	ver
pousser	39	ver
déguster	34	ver
infuser	30	ver
mettre	29	ver
croquer	28	ver
parfaire	25	ver
concasser	25	ver
retrouver	22	ver
matcher	21	ver
goûter	21	ver
cuisiner	21	ver
truffer	20	ver
sauter	20	ver
régaler	19	ver

Forme	Freq.	Types
tomate	349	nom
fleur	329	nom
gastronomie	321	nom
pomme	320	nom
citron	270	nom
jus	244	nom
truffe	242	nom
nice	238	nom
caviar	236	nom
oignon	232	nom
salade	227	nom
dessert	227	nom
betterave	225	nom
chou	215	nom
crème	202	nom
joie	198	nom
terre	189	nom
carpaccio	182	nom
céleri	177	nom
homard	175	nom
navet	174	nom
carotte	173	nom
veau	169	nom
foie	164	nom
chocolat	163	nom
fraise	162	nom
légume	158	nom
ail	155	nom
oeuf	151	nom
tarte	149	nom
huile	148	nom

Forme	Freq.	Types
noir	261	adj
vert	254	adj
petit	223	adj
rôti	192	adj
frais	190	adj
confit	182	adj
blanc	173	adj
gras	162	adj
rouge	160	adj
jaune	138	adj
nouveau	117	adj
grillé	100	adj
glacé	96	adj
cuit	77	adj
croustillant	67	adj
sauvage	65	adj
bleu	63	adj
premier	62	adj
doux	62	adj
beau	60	adj
végétal	56	adj
gourmand	51	adj
tartare	50	adj
nouvelle	48	adj
chaud	45	adj
pourpre	44	adj
sablé	43	adj
mini	43	adj
feuilleté	43	adj
suprême	41	adj
gris	41	adj
mauve	40	adj
velouté	39	adj
poché	38	adj

FIG. 5 - *Les termes les plus fréquents des catégories grammaticales verbes, noms et adjectifs.*

On remarque qu'avec les verbes on peut extraire des techniques culinaires ("farcir", "souffler", "infuser") mais aussi des manières de vivre l'expérience gastronomique ("découvrir", "déguster") ce qui nous conforte dans l'idée que ces données ont un intérêt patrimonial. Par exemple, les tweets dans la figure 6 présentent des plats mettant en œuvre des techniques et des produits qui s'intègrent au processus patrimonial présenté.

FIG. 6 – *Tweets impliquant des verbes et des noms en lien avec des techniques et des produits.*

La catégorie des noms est intéressante d'un point de vue patrimonial également, car malgré le corpus de chefs 2 et 3 étoiles, les ingrédients mentionnés ne sont pas forcément luxueux et on peut donc penser que le patrimoine culinaire français se fonde sur certains produits traditionnels du terroir ("pomme", "oignon", "chou"). Ce lien entre patrimoine et terroir sera à approfondir. Par ailleurs, il serait intéressant de comparer ces résultats avec les travaux menés en sciences sociales (notamment le sociologue Claude Fischler) sur les ingrédients mentionnés dans les appellations culinaires des spécialités mentionnées par les chefs triplement étoilés au *Guide Michelin* : la truffe, le homard, le caviar ou le chocolat figurent parmi les ingrédients les plus mentionnés ; mais on observe que de plus en plus de légumes sont mentionnés dans les spécialités, ce qui témoigne d'un nouveau statut de ces produits dans la haute-cuisine. Enfin, beaucoup d'adjectifs de couleur sont mentionnés, ce qui indique que les noms utilisés sont souvent spécifiés, soit parce que l'ingrédient lui-même contient un adjectif de couleur ("haricot vert"), soit parce qu'il s'agit d'une spécificité de la recette. D'autres adjectifs concernent davantage la dimension patrimoniale, à travers des qualités attribuées via la préparation ("confit", "croustillant", "gourmand"). L'utilisation de ces adjectifs montre que la description sensorielle des plats est au cœur de la rhétorique culinaire des chefs. Un travail plus spécifique des associations entre ces catégories reste à faire, afin de saisir la complexité combinatoire de ces parties du discours.

5. Conclusion

Dans ce papier nous présentons un travail en cours avec des résultats préliminaires autours du patrimoine gastronomique français. A cette fin, nous avons analysé les tweets

envoyés par les chefs ou les restaurants 2 ou 3 étoiles au *Guide Michelin*. Dans la suite des travaux, une extension du corpus aux chefs 1 étoile sera envisageable. Il faudra néanmoins vérifier les conséquences d'une telle extension en termes quantitatifs (nombre de comptes et quantité des données) et qualitatifs (particularités de ce corpus et pertinence pour la problématique générale).

Références

Cadiot P. et Visetti Y.-M. (2001), Pour une théorie des formes sémantiques, PUF.

Harris J. K., Hawkins J. B., Nguyen L., Nsoesie E. O., Tuli G., Mansour R., et Brownstein J. S. (2017), Using Twitter to Identify and Respond to Food Poisoning: The Food Safety STL Project, Journal of public health management and practice: JPHMP, 23(6), pp 577-580.

Lebart L. et Salem A. (1994), Statistique textuelle, Dunod.

Longhi J. (2008), Objets discursifs et doxa. Essai de sémantique discursive. l'Harmattan, coll. « Sémantiques ».

Longhi J. (2015), La Théorie des objets discursifs : concepts, méthodes, contributions. Mémoire d'HDR, Université de Cergy-Pontoise.

Mouritsen O. G., Edwards-Stuart R., Ahn Y. Y., et Ahnert S. E. (2017), Data-driven Methods for the study of Food Perception, Preparation, consumption, and culture, Frontiers in ICT, 4.15.

Nguyen Q. C., Meng H., Li D., Kath S., McCullough M., Paul D., ... et Li F. (2017), Social media indicators of the food environment and state health outcomes, Public health, 148, pp 120-128.

Pincemin B. (2011), Sémantique interprétative et textométrie – Version abrégée, Corpus, 10 , pp 259-269.

Ratinaud P. et Marchand P. (2015), Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014), Mots – Les langages du politique, 108, pp 57-77.

Summary

Since 2010, the French Gastronomic Meal has been registered in the representative list of the immaterial cultural heritage of the humanity. This decision brings questions related to the tasks of preservation and transmission of this heritage. These tasks are currently (most often without intention) undertaken partially by chefs recognized and lionized, even if their elitist reputation is far from the values of sharing and conviviality that were at the heart of the application submitted to Unesco. In this work, we are interested in the French gastronomy transmission carried out by Michelin-starred chefs by means of Twitter messages. Thus, we propose to analyze the tweets of 2- and 3-stars Michelin-starred chefs in order to characterize the transmission process.

HeritaMus: a machine of representation of actors' networks concerning cultural (tangible and intangible) heritage

Nédra Mellouli*, Pedro Félix**

*LIASD(EA 4383), Université Paris8 Vincennes, Saint Denis
n.mellouli@iut.univ-paris8.fr
**Instituto de Etnomusicologia INET_MD, Universidade Nova de Lisboa
lx.felix@gmail.com

Résumé. Le patrimoine culturel (matériel et immatériel) fait l'object des nouvelles technologies numériques depuis ces 5 dernières années. La conjoncture favorable de la digitalisation et le retour en force de l'intélligence articifielle, ont permis le développement de multiples plateformes et outils dédiés à la valorisation du patrimoine, autrefois méconnu au grand public et accessible seulement à une minorité de personnes détenteur du savoir. L'archive sonore numerisée du Fado ne peut pas se limiter seulement à la conservation et l'exploitation simple des enregistrements sonores. Au contraire, le Fado comme le Flamenco, est un univers de pratique qui doit être décrit comme une aggrégation d'entités dynamiques liées les unes au autres dans le temps et dans l'espace. Les liens entre les entités constituent un réseau dynamique reflétant la perception du point de vue de chacun sur cette univers de pratique musicalle des ethnomusicologues mais aussi à les communautés de praticiens eux même. Le Réseau est en réalité une multitude de réseaux individuels/ parciels issus de cartes mentales représentant différents perspectives autour d'une même "réalité".

1 Digital tool research on the relationship between tangible and intangible heritage

Heritage, musical practices, and musical industries are seldom associated concepts in a research project. "Heritage" as been associated to traditional practices, and the musical industries to popular music international repertoire. But in both cases, the associations are factually erroneous (as showed by authors like (Tschmuck, 2012) or (Manuel, 2013)) because since the start of their commercial activities, international companies always acted internationally ; publishing art music recordings and promoting international "stars" and locally publishing recordings of local musical genres, repertoire and musicians to feed local gramophone markets that are seen (by the community of practitioners, stakeholders and some researchers) as "authentic" representations of "old" traditional musical forms, as "heritage". It's fundamental to carefully document the interconnections between those topics. With the current proposal, by focusing specifically on the relationship between heritage practices (practiced by different types of actors, from musicians to researchers, institutions and companies), historical sound documents

like historical recordings and current uses of community's history (by the cultural manifestation stakeholders), we aimed at developing an innovative open web API of cooperative research programme with the stakeholders of Fado and Flamenco from Spain) supported by the development of a new management and research tool. Heritamus (http://heritamus.fcsh.unl.pt/) is an open source web tool focused on deepening the intricate relationship between intangible and tangible heritage, gathering all kinds of actors (human and non-human, intangible and tangible) in a map that best displays the networks through new tools for visualization of complex data. The devlopment of Heritamus is achieved as part of the European HERITAGE PLUS project JPI-CH program. Furthermore, this project will develop an innovative research approach in order to bridge the gap of articulation between tangible and intangible heritage, patrimony that has been treated differently by different institutions. Its proposal congregates knowledge from different areas of study like anthropology and ethnomusicology, museum and heritage studies and skills ethnography, computer programming, recorded sound analysis, sound archive management, audio restoring, database design, visual representation, augmented reality, museum and display technologies. It is achieved through a transdiciplinary research team with multi-faceted researchers (social scientists with database programming experience), stakeholders and community practitioners whose knowledge is integrated and being valued, with associated partners as mediators and beta-testers.

2 Context

Fado and Flamenco were the chosen cultural practices for this research as they both have a history of commercial sound recording, most of them unknown to the actual community of practitioners. In both cultural universes the knowledge transmission is based on an organic, lifelong process of acculturation and apprenticeship of a long and complex network of actors, from teachers to recordings, from recording publishers to museums, from academia to private archives and collections. Both submitted successful applications for the UNESCO Intangible Cultural Heritage list (2011 and 2010, respectively). Since then, the reinforced national and international interest on both practices called for the access to historical sound recordings and some resources where made available to the general public (like, for example, Fado Digital Sound Archive at http://arquivosonoro.museudofado.pt).

3 Theoretical context : ANT theory, Category theory, Computation

HeritaMus is an acronym that brings together **Herita**ge and **Mus**ic, echoing Latour *Cogitamus* (Cotelette, 2010)). Like "cogitamus", or "we think" ? no longer the Cartesian *cogito* or "I think" ? HeritaMus highlights the multiple acts, performed by multiple actors, of production of heritage by collecting, safeguarding, giving access, and researching it. Conventionally, at academic, museological, archival and heritage circles, the heritage act has been assigned to designated specialists, precisely the ones that have had a more facilitated access to historical sources, namely sound recordings, and whose narrative has been dominant to the point of erosion of alternative or controversial narratives. At HeritaMus there is no determinant, absolute and definitive truth, no complete and isolated actor, but dynamic actors, each one with it ?s

own trajectory, with its reasons to act, with their personal diplomacies. The project focus on bringing the communities of practice to the process of creating heritage by registering their own narrative, even if controversial. Previously conducted fieldwork made clear that data and the work built upon it has been, in fact, affected by the politics of prestige of the stakeholders and the consolidation (or solidification in ANT terminology) of narratives. So, at HeritaMus, we wanted to give voice to multiple actors, each one with their own reason to act. We wanted to follow the associations established by each actor in order to define itself and the others, being them human or non-human. In a word, the project looks for the democratization of data access and curation. The conceptual starting point is the definition of "actor". An actor is any entity, human or non-human, with a reason to act, whose action transforms the conditions. This transformation (or translation) results in the establishment of relationships between two actors. From this very simple movement, multiplied across time, space, and mobilized actors, we are entitled to trace a network of relationships among actors ("nodes"). Each actor is defined by the network it constructs with other actors, just like the network is defined by the actors that participate on it. Our purpose was to design an tool to register those movements. Starting from this simple principle, the project adds another dimension to it by considering the one who records, the one who speaks, who trace the network also an actor in itself. And this is not a minor issue. All actors, since they act in any way, are elements of the network, not transcendental like entities. Through this, it is possible to trace diplomacies and controversies. So, it is fair to say that, at the core of the project lays the politics (or diplomacies) of actors, considering their reasons to represent, write, speak, catalog, display, describe, register, and so on. The poetics and politics of social sciences has been already intensely debated (since, at least Clifford and Marcus 1986), but always in a "after" writing moment. In order to be truly and effectively democratic, the narrative production has to have user friendly tools to free actors from mediation, producing their own narratives. To summarize, HeritaMus application is based on Latour ((Latour, 2005)) and Actor-Network Theory (ANT) conceptual toolbox ((Law et Hassard, 1999)) that provides a user-friendly (in order to guarantee a democratic access) way to register and curate data on tangible assets and intangible knowledge, building heritage. The project focused on deepening the intricate relationship between intangible (knowledge, memory, and identity) and tangible heritage (namely, historical sound recordings). The team started from a specific dataset for the test phase : the relationships between practices (by different types of actors, from musicians to institutions), sound documents (historical recordings) and current uses and re-uses of community's history and knowledge. During the test phase we established a cooperative research laboratory among the Fado community in Portugal (that should be replicated in Spain with Flamenco community) on the uses of historical commercial recordings that are now understood by the communities as historical documents of past practices, to understand the genre's memory shaping processes and heritage configuration. The team wanted to research the intense dialogue between human actors (like musicians) with non-human actors (like the 78 rpm shellac recordings), knowing, from the start, that both were essential actors in the process of shaping the musical genre, its community, history, music elements, practice, and identity.

4 The research process

The project is structured in three phases : 1) Preliminary inventory of materials. During this phase, we putted together an experimental laboratory, centered on a collaborative interac-

tion between community members, researchers and stakeholders to analyze data, promoting the construction of fluid assemblages of improbable and unexpected objects, places, people, ideas, values, techniques, gathered to depict the community's knowledge and assets, taking special care of the difficulty of representing ICH. The laboratory sessions were based on open discussion on display strategies, usability and concept validation. The main focus was how to depict tangible assets and intangible knowledge, human and non-human actors, in an articulated manner and on the same plane, knowing that agency is fluid, based on unexpected and ever-changing alliances and oppositions. 2) Design of digital tool. Being the technical core of HeritaMus, with the gathered ethnographic and historical information, the tool was designed to register the associations (relationships) between the items (nodes), tracing the network of actors. The graph database was programmed according to the needs of research and community of practice. The interface is addressed to the community of stakeholders and practitioners. The theoretical issues informed the structure and the interface designed for a user-friendly digital environment. The software is a complex, multi-level, graph database for the management of fundamental ethnographic raw data, that supports future safeguard programs of both tangible and intangible heritage, and registers any actions over time (namely technical actions, registration of new data, and trace more complex processes of cultural practices change). The graphs are used for data input, retrieval and visualization. It supports systematic analysis of complex cultural manifestations from different kinds of materials usually dispersed unrelated in hierarchical and atomistic databases. With Heritamus, users can easily represent complex networks of human and non-human actors promoting new interpretations and knowledge based on unexpected relationships. 3) Fieldwork. The ethnography produced was based on audio-elicitation sessions of free association with some of the community activists, chosen on the basis of previous ethnographic knowledge of the community, taking into account their knowledge of the manifestation in question and the community's recognition. The sessions focused on the ways the community of practitioners organize their world-view and their current use of historical sound recordings. By listening to those recordings with the practitioners, a "facilitator" registered comments, descriptions and reactions to those historical sound documents, the ways community members and stakeholders classify and perceive their heritage, and how this affects their present action and perception (aesthetic and technically assessing their peers and their performance). In a second, and longer, period of an year long fieldwork, a prototype of the tool was tested in order to achieve a final version of the software (released April 2018).

4.1 How it works

Tangible and intangible data are described by a predefined collection of metadata. Metadata are the different types of nodes in the graph. Connections between nodes are described by a list of generic relationship types. As nodes or relationships are described by a closed set of features and free comment's feature. Each new graph created relating to a mental representation of a user's knowledge is faithfully stored in a graphical database (Neo4J). Thanks to Neo4J's architecture, we have for each node all the nodes connected to it and the users who made these connections. The development of the Heritamus application allows to answer several challenges by proposing a simple and intuitive collaborative Web interface. We first recall the main application challenges. Indeed, the first challenge consists in generating in an inductive way the representation of knowledge from a user point of view. Through a mental representation of entities, each user is able to simply model the semantic links between enti-

ties through a graph enriched with personal information. This representation is simply the best model that reflects the vision of each user and that is how each of us perceives the world. The second Challenge consists in extracting different representation patterns of knowledge or the world, and to be able to confront them, project them, compare them to standard gold representations. The third challenge consists in enriching each representation model with connections between the different graphs. In other words, we seek first of all to understand why two entities are always in relations whatever the space, and how these relations evolve in time. These Functional Challenges related to this graph model, Heritamus lift very interesting theoretical locks and little studied in the literature from the point of view of computer modeling. Based mainly on graph theory, theoretical problems remain open. For example, the generalization of a spatio-temporal graph from a network of graphs is a functional need that can correspond to the detection of a standard and common model of a knowledge world. Another functional problem as interesting as the previous one is to determine controversial points of view and to know how to explain them. It is very important here to emphasize the advantages of these graph-based representation models. The first advantage is the simplicity of the model since no formalization effort is necessary for each person to semantically represent his own world. The second advantage lies in the use of this same graphic model to visualize and explore knowledge on several facets. Graph theory allows, with the same tool used for modelling, the visualisation of knowledge at several scales. Indeed, given that the graph is at the same time spatial, temporal, semantic, historical, political, etc. Heritamus is quite an open system allowing to elarchir and to enrich one's own knowledge by relying on open linked Data. Open data can enrich existing knowledge on the one hand, and fill knowledge gaps that may exist because of rights on the other. Finally, Heritamus could be used for educational purposes without prior parameterization, allowing to extract relevant graphs for a precise world. Since it could be applied to areas other than music, such as learning computer programming, history teaching, etc., it could be applied to other fields.

5 Conclusion and futur works

Heritamus proposal congregates knowledge from different areas of study like anthropology and ethnomusicology, museum and heritage studies and skills ethnography, computer programming, recorded sound analysis, sound archive management, audio restoring, database design, visual representation, augmented reality, museum and display technologies. Through its graphical model we envisage studying its potential to be able to represent a universal knowledge that would be the maximum concept graph obtained by the intersection of all graphs dealing with the same context. If this graph exists and if it is unique, can it be considered as a domain ontology of as long as the graph is a faithful representation of knowledge ? Through mining of different graphs by applying spatial and temporal queries, we will be able to highlight the evolution of knowledge in terms of the emergence of new connections and/or the vanishing of others. For the interrogation and mining of these graphs we will investigate to use natural language queries for large mount of data. Heritamus will be used in other application context when a collective and shared knowledge are built by each person contributing to semantically represent his own world. Fianlly we evaluate Heritamus knowledge representation and perception by measuring ints impact on heritage management.

Références

Cotelette, P. (2010). Criativity and innovation in the music industry. *http://journals.openedition.org/lectures/1208*.

Latour, B.; Weibel, P. (2005). *Making Things Public. Atmospheres of democracy*. Cambridge: Cambridge University Press: Cambridge: MIT Press.

Law, J. et J. Hassard (1999). Criativity and innovation in the music industry. *Sociological Review Monograph Series: Actor Network Theory and after 47, Issue S1*, 1–251.

Manuel, P. (2013). *"Music cultures of mechanical reproduction" in Bohlman*. Cambridge: Cambridge University Press: he Cambridge History of World Music.

Tschmuck, P. (2012). Bruno latour, cogitamus. six lettres sur les humanités scientifiques. *Heidelberg: Springer*.

Summary

The cultural heritage (tangible and intangible) has been under the spotlight of new digital technologies. The favourable context of digitalization and the re-emergence of artificial intelligence have enabled the development of multiple platforms and resources dedicated to the enhancement of heritage, once unknown to the general public and accessible only to a minority of knowledge holders. The Fado Digital Sound Archive cannot be limited solely to the preservation and simple exploitation of sound recordings. On the contrary, Fado, like Flamenco, is a universe of practice that should be described as an aggregation of dynamic entities linked to each others in time and space. The links between the entities constitute a dynamic network reflecting the perception of everyone's point of view on this musical universe of practice and its culture, from ethnomusicologists to the community of practitioners themselves. The Network is in fact a multitude of individual/ partial networks resulting from mental maps representing different perspectives upon the same "reality".

Frénaud numérique : des prémices du projet à sa concrétisation

Marianne Froye*,**, Zakarya Després*,**

*Université de Cergy-Pontoise
**Institut des Humanités Numériques de l'Université de Cergy-Pontoise
marianne.froye@u-cergy.fr
zakarya.despres@gmail.com

Résumé. Le projet de numérisation a permis au laboratoire LT2D de l'Université de Cergy-Pontoise de devenir membre du consortium « Cahier » de la TGIR Huma-Num. Il vise la préservation et la valorisation d'un patrimoine littéraire : 2 ensembles majeurs de manuscrits d'André Frénaud.

1. Objectifs et contraintes de départ

Après avoir explicité le déroulement de la collecte des données[1] de notre corpus, nous envisagerons les mises en œuvre et les questionnements techniques et scientifiques qu'il a suscités. La présentation aura pour but d'expliquer les enjeux épistémologiques de l'encodage en XML-TEI et les exploitations théoriques et herméneutiques à court et moyen termes.

1.1 Présentation de l'auteur

Le corpus de notre étude porte sur deux ensembles de manuscrits *La Sorcière de Rome*[2] et *Gloses à la Sorcière*[3] d'un poète français du XX° siècle : André Frénaud (1907-1993). Poète reconnu de son vivant et par ses pairs, il reste relativement discret dans l'histoire littéraire. Il occupe pourtant une place primordiale dans les échanges des intellectuels durant la Guerre Froide et participe activement par son œuvre à l'évolution de l'écriture poétique au siècle dernier. C'est un poète majeur d'après-guerre. Les considérations métriques, stylistiques, lexicologiques pourraient être davantage exploitées par le biais d'une approche numérique de son œuvre. Pourquoi commencer un projet de numérisation sur ce poète et sur ce corpus ? Pour une première raison qui serait valable pour bon nombre d'auteurs : la conservation de documents uniques, mais qui l'est d'autant plus pour l'œuvre de ce poète. André Frénaud a écrit essentiellement au crayon de papier et même si les manuscrits datent du XX° siècle, les

[1] Le projet a reçu un soutien financier du consortium Cahier de la TGIR Huma-Num et de l'IDHN de l'Université de Cergy-Pontoise.
[2] André Frénaud, *La Sorcière de Rome*, Gallimard, 1973.
[3] André Frénaud, *Gloses à la Sorcière*, Gallimard, 1995

feuillets commencent à s'effacer. La numérisation permet donc de préserver ses manuscrits. Son œuvre est également très intéressante par la multiplicité de lectures qu'elle propose. Il a eu soin pour quasiment la totalité de son œuvre poétique d'écrire des commentaires en prose. La diversité générique était un défi à l'encodage. Jusqu'à présent les études en humanités numériques chez les littéraires ont davantage versé vers des textes soit d'un unique genre littéraire : prose théâtre…, soit en prose. La gageure résidait dans la transcription de la poésie libre de l'auteur et de ses commentaires en prose conjoints à l'écriture du poème, dans sa façon de créer un poème : le décompte du nombre de syllabes et du nombre de vers est capital.

1.2 Présentation des données[1]

Il peut paraître étrange de ne pas avoir choisi de numériser, transcrire et encoder ses premiers poèmes, pourtant le choix de ces données s'est imposée très rapidement pour plusieurs raisons. La première d'entre elles est certainement liée au fait que ces deux ensembles de manuscrits fonctionnent de manière close et sont très interactifs entre eux. Par ailleurs, nous ne disposions dans un premier temps que d'un budget limité et ce que nous numérisions et transcrivions, nous voulions pouvoir le proposer comme modèle pour poursuivre ensuite une recherche de plus grande ampleur. Notre but ne devait pas se réduire à une simple étape de conservation. Or, les premiers poèmes auraient été plus difficiles à travailler pour élaborer un modèle de données. D'une part, ils ont été écrits pendant la Seconde Guerre Mondiale, tous n'ont pas été gardés ou ne nous sont pas parvenus. André Frénaud a été mobilisé et a été prisonnier des Allemands, il est fort probable que lors de ses nombreux déplacements, des manuscrits se sont perdus. D'autre part, si nous voulions avoir des données significatives, il aurait fallu numériser un ensemble très important de manuscrits, ce qui aurait dépassé la somme allouée. Finalement, et c'est certainement scientifiquement la meilleure raison : *La Sorcière de Rome*[2] est le poème majeur d'André Frénaud. Composé de 15 amples mouvements, il concentre l'essentiel des interrogations existentielles, poétiques et esthétiques de l'auteur. Ce long poème est l'œuvre centralisatrice et totale de Frénaud. *Gloses à la Sorcière*[3] correspond au commentaire du poème. Si la pratique du commentaire est quasi systématique lors de l'écriture poétique, cette expérience-ci est la plus caractéristique, car c'est la seule qui ait donné lieu à la publication du commentaire du poème. D'autres poèmes ont fait l'objet de commentaires, mais ces derniers n'ont pas été eux-mêmes publiés en livre. Les gloses se déclinent telle une explication linéaire détaillée pour les premiers mouvements et avec une inflation croissante d'informations pour le reste des mouvements. Des considérations plus générales sont venues complétées les remarques très analytiques de ses premiers écrits. Le tout

[1] L'ensemble choisi dans le premier volet de ce projet représente environ 650 feuillets sur un total de 1430. Ils sont répertoriés sous les cotes FND Ms (97), FND Ms (98), FND Ms (99), FND Ms (100) et FND Ms (101).
[2] André Frénaud, *La Sorcière de Rome*, Gallimard, 1973.
[3] André Frénaud, *Gloses à la Sorcière*, Gallimard, 1995.

forme donc un ensemble clos, dialogique, Les manuscrits montrent que leur écriture est parfaitement imbriquée et quasi simultanée puisque bon nombre de tâtonnements versifiés sont accompagnés de commentaires et l'écriture des gloses a suscité la réécriture de certains vers. *Gloses à la Sorcière* a été publié de façon posthume.

1.3 Présentation du contexte

Le projet de numérisation et d'exploitation de ces deux ensembles a permis au laboratoire LT2D de devenir membre du consortium « Cahier »[1] en 2016, puis d'obtenir auprès de la TGIR Huma-Num par le biais de Cahier un financement pour le mener à bien l'année suivante. Plusieurs critères étaient exigés dont la mise à disposition à l'ensemble de la communauté scientifique des données. Monique Frénaud[2] a donné son accord pour une diffusion totale de ce corpus. Ce dernier est déposé à la bibliothèque littéraire Jacques Doucet[3], dépositaire de plusieurs legs du poète. Le plus récent date de 2016. Cet ensemble-ci de manuscrits a fait l'objet d'un legs spécifique en 1997. La bibliothèque est elle-même impliquée dans différents programmes de bibliothèques numériques et d'expositions virtuelles dont Frénaud a également fait l'objet à l'automne dernier et pour laquelle nous étions également commissaire scientifique[4].

2. Élaboration du corpus

2.1 Récolement

La numérisation du corpus a été préparée par un récolement[5] qui a anticipé des questions quant à l'exploitation des données. Le récolement qui pourrait apparaître comme une simple formalité pose d'emblée les questions des perspectives scientifiques que nous voulons donner au projet, notamment à travers les données que nous voulions conserver absolument lors du passage au format numérique. La reproduction numérique identique à l'objet physique n'est pas un détail et n'a pas pour seul enjeu la conservation exacte du *princeps*. Les chercheurs avec les conservateurs de bibliothèque oscillent entre la volonté de tout conserver dans un esprit de fidélité absolue et les limites techniques de pouvoir le faire, et surtout sa pertinence. En l'occurrence, pour ce corpus, il existe une très grande variété de formats : le poète pouvait écrire sur un coin d'enveloppe déchirée comme sur le verso de revues qu'il recevait ou encore sur des copies doubles qu'il s'appropriait dans toute leur largeur. Si les formats inférieurs au

[1] Cahier pour « Corpus d'Auteurs pour les Humanités : Informatisation, Édition, Recherche ». Consortium appartenant à la TGIR Huma-Num.
[2] Veuve et ayant droit d'André Frénaud.
[3] Paris, Place du Panthéon.
[4] http://bljd.sorbonne.fr/exhibit/36
[5] Assuré par M. Froye.

format A4 ne posaient pas véritablement de problème technique puisqu'ils ne nécessitaient pas de calibrage nouveau, les copies doubles avaient des incidences sur la numérisation et sur le mode de visualisation. Le récolement avait pour but initialement de préparer le travail de numérisation. Pour finir, récoler a aussi anticipé l'encodage. Le choix du support participait à la création, la gestion physique du papier et de son format expliquait le processus créateur du poète. Il devenait donc évident que ces données devaient être transcrites. Par exemple, l'utilisation des supports dans plusieurs sens vertical, puis horizontal puis de nouveau vertical avec des dessins ou encore avec des renvois ou des commentaires était le cœur même de l'explication de l'écriture conjointe des commentaires et du poème et montrait le processus créateur de Frénaud.

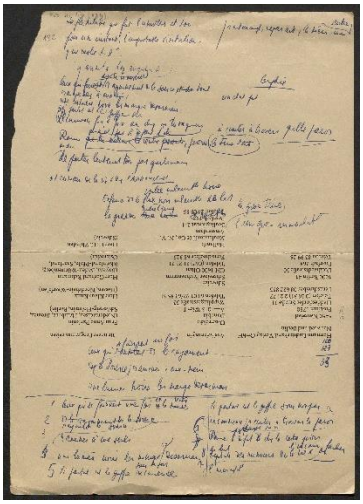

FIG. 1 - *FND Ms 100 (134) : exemple de support et d'organisation spatiale du manuscrit.*

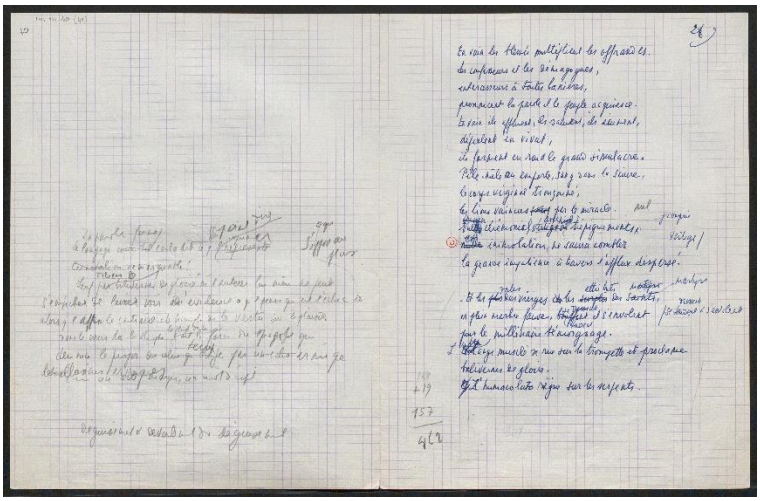

FIG. 2 - *manuscrit FND Ms 100 (41) : Quelques vers et leur commentaire.*

Ce détail physique a des incidences herméneutiques fondamentales pour expliquer la genèse de l'œuvre. En agissant ainsi, le poète réfléchit et formalise la portée et la signification de son œuvre. Il espère lui donner un certain sens et réfléchit donc principalement aux meilleurs choix linguistiques, lexicaux et métriques pour y parvenir. Il devient de ce fait à son tour lecteur et commentateur de sa propre œuvre.

2.2 Numérisation

La numérisation était externalisée auprès d'un prestataire avec lequel la bibliothèque a l'habitude de travailler. La société Arkhênum est coutumière des spécificités des manuscrits littéraires. La numérisation a dû être préparée par quelques réunions de travail avec le prestataire de service et les conservateurs de bibliothèque pour s'assurer du résultat que tous, nous voulions atteindre. Elles ont notamment levé les ambiguïtés du mode de visualisation des doubles pages et des supports particuliers type carton épais. Les supports étant des « documents sales », il n'était pas envisageable d'associer à la numérisation une reconnaissance automatique de caractères. Les fichiers ont donc été livrés aux formats JPEG et TIFF.

2.3 Transcription[1]

Le choix pour effectuer la transcription s'est porté sur Transkribus après l'étude de quelques options. Le travail préparatoire sur les manuscrits et les analyses génétiques que nous voulions mener à bien nécessitaient une transcription très fine. Mais elle était rendue délicate par plusieurs phénomènes. Tout d'abord, physique : la diversité des papiers devait être rendue. La transcription devait nécessairement faire état d'une description physique de l'objet transcrit. Le support choisi donnait des indications sur l'univers intellectuel de l'auteur durant sa création. Lorsque ce sont des versos d'affiche d'exposition ou des cahiers de l'ORTF, le chercheur a des indications importantes sur ses nourritures intellectuelles. Quelle signification lui donner ? Est-ce qu'il a écrit au verso de l'article parce qu'il ne lui prêtait pas davantage d'importance et montrait par là même un désaccord avec l'auteur de l'article ou au contraire doit-on y lire une forme de dialogue entre son œuvre et l'article critique présent au recto ? Est-ce que l'exposition a retenu toute son attention ? Par ailleurs, si la technique offre de nombreuses possibilités, tout devient plus délicat lorsque les lignes d'écriture ne sont plus droites et horizontales, lorsque le poète crée un axe paradigmatique sur son manuscrit pour le mot d'un vers. Or, ce dernier devait lui aussi être transcrit pour que les études lexicologiques soient de plus grande ampleur. Les ratures étaient aussi une gageure à transcrire. Il nous faudra encore attribuer par des critères les degrés de raturage : parfois le poète a volontairement noirci et recouvert son texte jusqu'à le rendre illisible.

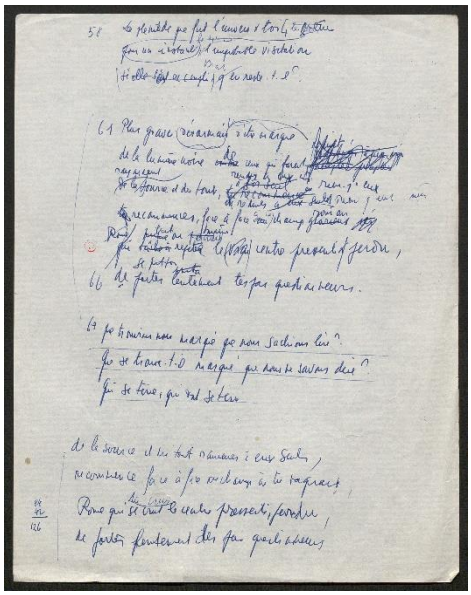

FIG. 3 - *Exemples de quelques ratures.*

[1] Assuré par M. Froye et R. Bouroubi.

D’autres fois, il ne s’agit que d’un mot et il est utile de signaler la rature et le mot raturé. Les renvois sont également très nombreux et très divers. Ce peut être un mot d’un vers qui suscite une explication formalisée par un trait pour signaler le lien entre le mot précis et le commentaire. Ce peut être aussi pour signaler une accumulation de commentaires. Le premier commentaire appelant une nouvelle série de commentaires. Dans ce cas de figure, on se doit de conserver la disposition sur la page et en même temps de déterminer la chronologie des commentaires.

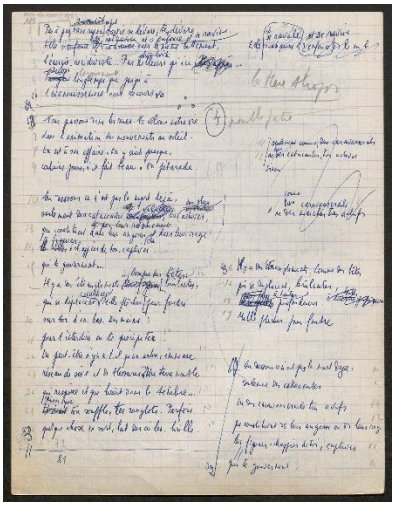

FIG. 4 - *Les différentes strates de création.*

Pour finir, le poète s’est fait dessinateur sur quelques manuscrits. Quand on connaît ses nombreuses amitiés et collaborations avec des peintres très connus du XX° siècle comme Mirò, il est impensable de perdre cette création non littéraire lors de la transcription et de l’encodage.

2.4 Encodage[1]

Transkribus permet de préparer l’encodage XML-TEI qui se fera sous Oxygen notamment lorsque plusieurs morceaux de textes figurent sur une même page.

[1] Assuré par Z. Després

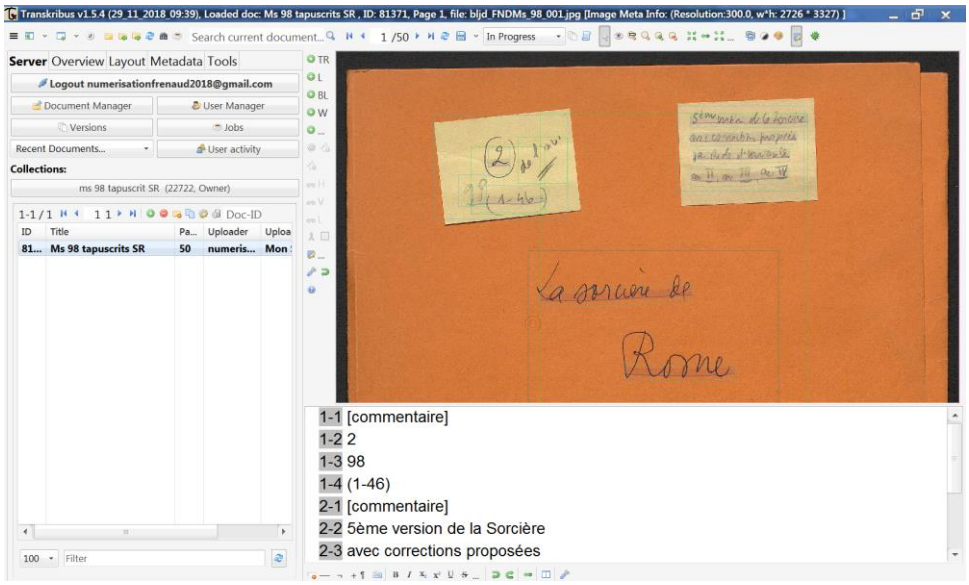

FIG. 5 - *Transkribus.*

Transkribus a l'avantage contrairement à un simple logiciel de traitement de texte de formaliser les zones de texte. Il est possible de transcrire le texte, certes, mais aussi un certain nombre d'éléments de la somme des informations annexes et paratextuelles tout aussi importantes et dont nous avions besoin. Ces zones de texte déterminées sous Transkribus permettent d'anticiper l'encodage. Le choix du principe d'encodage s'est assez naturellement porté sur la XML-TEI[1]. La TEI n'est pas le seul langage de balisage du texte, mais elle offre plusieurs avantages, le premier et pas des moindres, elle est très prisée par les littéraires qui s'intéressent aux humanités numériques. Gravite autour d'elle une communauté active de chercheurs qui ont levé déjà quelques difficultés de balisage. Surtout elle se construit en permanence par les apports des discussions au sein de différents forums.

[1] TEI : Text Encoding Initiative.

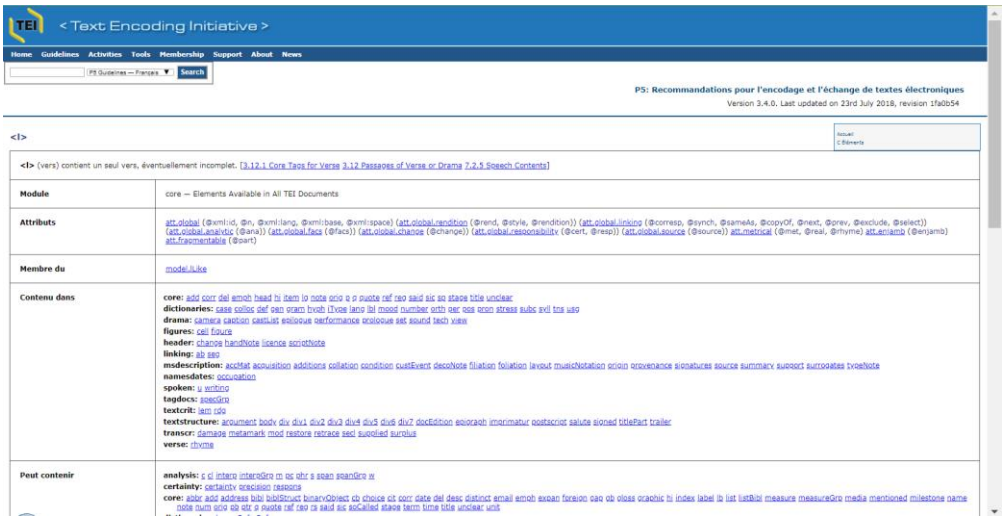

FIG. 6 - *Exemples de guidelines TEI.*

Autre avantage : la TEI permet la description du document textuel dans toutes ses dimensions. Rendre la disposition spatiale était une gageure de l'encodage, la TEI nous a semblé plus approprié que l'EAD fort utile par ailleurs pour l'indexation des entités nommées. Voici quelques éléments de l'encodage et du modèle de données :

Groupe de textes	<l>[VERS]</l>...	@type : poésie, commentaire... ex : <lg type="commentaire">...	(structure créée par Transkribus)
Vers	<l>[VERS]</l>		(structure créée par Transkribus)
Groupe de vers	<lg><l>[VERS]</l>...</lg>	@type : quatrain, tercet… ex : <lg type="quatrain">...	

Les différentes expériences d'encodage en XML-TEI de textes littéraires se sont davantage focalisées sur des textes en prose, les balises pour la poésie existent, mais elles

demandent à être affinées et interrogent les frontières génériques des textes littéraires. Baliser un texte demande au transcripteur de faire un choix interprétatif en amont.

3. Exploitation / exploration des données

3.1 Le public

Nous avons obtenu l'accord des ayants droits et de la bibliothèque pour la diffusion libre de ces manuscrits. Les images numérisées figureront sur le site de la bibliothèque littéraire Jacques Doucet. L'ensemble des données (numérisations, transcriptions et encodage) seront en libre accès sur le site internet du projet www.frenaudnumerique.fr et sont hébergées par ETIS. Elles figureront également sur le site du consortium Cahier. La visibilité est ainsi renforcée et la mise à disposition des chercheurs facilitée.

3.2 Nos hypothèses, nos perspectives et leur application

Notre perspective de travail dès l'origine du projet est de conjuguer une triple approche : littéraire, linguistique et informatique. Notre postulat était de participer à la communauté TEI en réfléchissant à de nouvelles balises à travers l'encodage de ce corpus pour développer à plus long terme des projets plus importants en humanités numériques sur des textes poétiques. L'application est multiple : ce sera tout d'abord en lexicologie. L'encodage permettra des études fines sur les champs dérivationnels utilisés par l'auteur, sur les fréquences de termes. Il favorisera également une approche génétique pour le lexique. À de nombreuses reprises, le poète bute, hésite sur un mot. Pouvoir mettre en perspective l'ensemble des choix paradigmatiques de l'auteur favorisera la compréhension de son œuvre. Les changements lexicaux tiennent parfois de la sonorité du mot. La facilité pour comparer les différents états d'un mouvement ou d'un vers apportée par le balisage redonnera la pleine dimension sonore de la poésie frénaldienne. Comme le montre également D. Mayaffre, la technique permettra également de faire émerger des objets d'étude qu'empiriquement nous ne soupçonnons pas.

Par ailleurs, André Frénaud commence à écrire des poèmes à un moment important de l'histoire de la versification : l'abandon de l'alexandrin à la suite des Surréalistes avant-guerre. Or, le décompte des vers très fréquent de Frénaud et la fascination - répulsion du chiffre « 12 » interrogent le rapport des poètes de sa génération au dodécasyllabe. En stylistique, l'encodage permettra d'approfondir la récurrence de certaines figures comme l'oxymore et de la lier au cheminement intellectuel de Frénaud. L'incomplétude de la négation, signe stylistique de Frénaud pourra faire l'objet également d'une étude plus précise. Pour mener à bien ces différentes perspectives d'étude, les données seront analysées avec TXM. La plateforme modulaire dont l'utilisation est possible sur poste de travail comme sur serveur est compatible avec le balisage en XML-TEI et assure une opération patrimoniale. Le principe est de mettre

à disposition des corpus en ligne pour mener des études en lexicométrie en s'appuyant sur des statistiques fondées.

Ultérieurement, les données pourront être exploitées selon d'autres perspectives patrimoniales. Le renvoi par des hyperliens aux monuments que le poète mentionne dans son poème, aux artistes et intellectuels qu'il convoque dans ses commentaires permettra de donner un caractère visuel à l'immatérialité du texte. Cartographier ses aires artistique, intellectuelle et linguistique donnera la possibilité de mieux appréhender la place de Frénaud dans l'histoire littéraire et ses connexions avec ses contemporains.

3.3 Nos pistes

Groupe de Travail « encodage des genres »

À partir de ce projet, notre volonté est de poursuivre et de participer à la réflexion tridisciplinaire au sein du groupe de travail « typologie textuelle » du consortium Cahier. La mise à disposition de ce patrimoine alimentera la constitution du thesaurus avec OpenTheso et les données pourront être versées dans la base de données Isidore.

4. Conclusion

Pour conclure, l'ambition de ce projet est de participer activement à la construction d'un patrimoine culturel en humanités numérique, de développer le réseau et les pratiques de travail en humanités numériques, et plus particulièrement l'approche des textes poétiques.

5. Références

Doueihi Milad, 2011, *Pour un humanisme numérique*, Paris, Éditions du Seuil.

Dufournaud Nicole, Gratsac Legendre Valérie. Manuel d'encodage XML-TEI - édition numérique de manuscrits baroques : Recommandations pour une application TEI. 2012. Consultable en ligne https://hal.archives-ouvertes.fr/hal-00718043

Gefen Alexandre, « Les enjeux épistémologiques des humanités numériques », Socio [En ligne], 4 | 2015, mis en ligne le 28 mai 2015, consulté le 11 janvier 2019. URL : http://socio.revues.org/1296

Gefen Alexandre, « Des humanités numériques en 2017 », *Mélanges de la Casa de Velázquez*, 47-2, 2017, p. 315-318

Heiden Serge, Magué Jean-philippe, Pincemin Bénédicte, TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. https://halshs.archives-ouvertes.fr/halshs-00549779

Mayaffre Damon. L'Herméneutique numérique. L'Astrolabe. Recherche littéraire et Informatique, 2002, pp.1-11. https://hal.archives-ouvertes.fr/hal-00586512/document
Mayaffre Damon, Philologie et / ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ? http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Mayaffre.pdf
Moretti, Franco, 2005, Graphs, Maps, Trees: Abstract Models for a Literary History, Londres et New York, Verso.
Rastier, François, 2001, *Arts et sciences du texte*, Paris, Presses universitaires de France.

6. Sitographie

http://www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_lite_fr.doc.html
https://github.com/oeuvres/poesie
https://groupes.renater.fr/wiki/cahier/typologie_textuelle

Valorisation de récits de vie de Républicains espagnols

Catherine Dominguès*, Laurence Jolivet*, Carmen Brando**

* Univ. Paris-Est, LASTIG COGIT, IGN, ENSG, Saint-Mandé
catherine.domingues@ign.fr, laurence.jolivet@ign.fr
** Centre de recherches historiques (UMR 8558 CNRS - EHESS), Paris
carmen.brando@gmail.com

Résumé. Ce travail vise à analyser des récits de vie de Républicains espagnols sous l'angle des lieux désignés et d'en proposer une représentation cartographique. Cette analyse combinée montre l'interdépendance des questions linguistiques et cartographiques soulevées.

Les corpus oraux nourrissent le « travail de la mémoire, du processus de ressouvenance et de la mise en paroles du passé » (dans Descamps (2008)) poursuivi par les historiens et les géographes de l'immigration. Dans ce contexte, le Réseau aquitain pour l'histoire et la mémoire de l'immigration a lancé en 2008 un programme expérimental de collecte pour recueillir la mémoire oubliée de populations immigrées engagées dans la vie de la région. Le projet MATRICIEL (PEPS CNRS UPE 2016) souhaite contribuer à la sauvegarde de ces témoignages oraux sur un support standard et compatible aux outils du traitement automatique des langues afin de permettre une exploitation systématique et ultérieure (croisements textométriques, cartographies, analyse des réseaux de personnes). Le travail présenté ici s'attache à localiser les récits de vie d'un groupe de Républicains espagnols exilés en France entre 1936 et 1939, et ayant participé à la résistance française. Notre proposition vise à analyser ces récits sous l'angle des lieux désignés afin de les valoriser sous forme cartographique.

Les récits de vie constituent un corpus de 18h30 dont l'observation montre que les lieux désignés et qui permettent de construire le récit relèvent des noms propres (NPr) : *Gurs*, mais aussi des noms communs (Nc) : *camp de regroupement*, *camp de triage*, *front nord*. L'identification automatique des NPr de lieux met en oeuvre des gazetiers – 2683 occurrences ont été détectées automatiquement et 2966 manuellement, correspondant à 351 NPr distincts –, celle des Nc s'appuie sur l'outil d'apprentissage Stanford-NamedEntityRecognition [1] entraîné sur des récits extraits du corpus – 2540 occurrences détectées automatiquement. Les performances du processus global donne une F-mesure de 75%, la reconnaissance des Nc de lieux obtenant de meilleurs résultats que celle des NPr. Les NPr sont associé à des coordonnées géographiques grâce à des gazetiers. Des informations attributaires sont aussi issues des gazetiers (la catégorie du lieu : pays, ville, relief, etc.) ou calculées (nombre d'occurrences, nombre de récits concernés, etc.) et valorisées grâce à la propriété synoptique de la carte (cf. figure 1).

Dans ce travail, toujours en cours, seuls les lieux NPr ont été localisés (comme dans Caquard et Cartwright (2014)). La question de la localisation des lieux Nc n'a pas été traitée et paraît, d'un point de vue géographique, indissociable de celle de l'échelle de la représentation.

1. http ://nlp.stanford.edu/software/CRF-NER.html

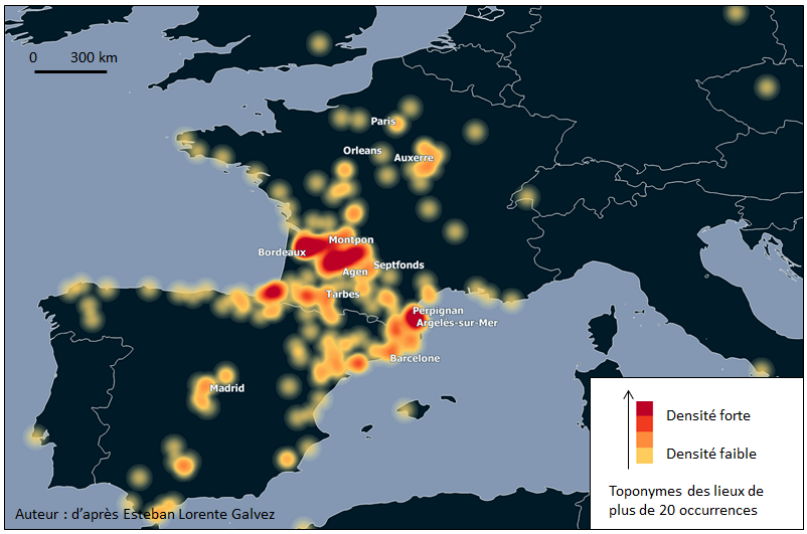

FIG. 1 – *Carte de chaleur de la localisation des NPr de communes (n = 255)*

D'un point de vue linguistique, elle relève à la fois des questions de métonymie : *Gurs* pour désigner le *camp de Gurs*, et de coréférences : le *camp* pour le *camp de Gurs*. La métonymie dans la désignation des lieux a des conséquences sur leur visualisation et impose de répondre à la question suivante : faut-il cartographier le texte (*Gurs*) ou l'interpréter auparavant (*camp de Gurs*) ? et se prolonge par la mise en place d'une mesure de rappel qui prendrait en compte la dimension spatiale des erreurs d'annotations. Une autre perspective concerne l'interprétation spatiale (sous forme de sur-ensemble, sous-ensemble, etc. du lieu désigné) des localisations relatives et des termes spatialement vagues (*banlieue de*, *alentours de*...). Enfin, la spatialisation des récits à travers les cartes peut accompagner ou susciter des questions de recherche :

— la répartition spatiale, par récit et pour le corpus, des lieux (de passage, de séjour) désignés et la comparaison avec les lieux emblématiques pour les Républicains espagnols. La figure 1 rend compte de fortes densités dans certaines zones, par exemple dans la région de Bordeaux, dans le Lot-et-Garonne, à la frontière franco-espagnole ;
— les lieux les plus désignés sont-ils emblématiques de la période ? de la résistance française ? de la migration espagnole ?
— la dimension spatiale des valeurs attributaires : des modalités identiques correspondent-elles à des lieux proches ?

Ces récits constituent un patrimoine immatériel (au sens de la convention UNESCO de 2003 [2]) et leur analyse contribue à construire la mémoire des lieux importants des Républicains espagnols installés en France. Ce travail donne l'occasion de rencontres entre les associations, le monde universitaire et les professionnels des archives. Enfin, la valorisation sous forme cartographique des récits illustre les questions récurrentes de l'annotation : qu'est-ce qu'un lieu (voir Purves et Derungs (2015)) ? comment prendre en compte les désignations métonymiques et les reprises anaphoriques des noms de lieux ?

2. http ://portal.unesco.org/fr/ev.php-URL_ID=17716&URL_DO=DO_TOPIC&URL_SECTION=201.html

Références

Caquard, S. et W. Cartwright (2014). Narrative cartography : From mapping stories to the narrative of maps and mapping. *The Cartographic Journal 51*(2), 101–106.

Descamps, F. (2008). Éditorial. *Bulletin de l'AFAS [En ligne] 44*.

Purves, R. S. et C. Derungs (2015). From space to place : Place-based explorations of text. *International Journal of Humanities and Arts Computing 9*(1), 74–94.