L'apport du web sémantique à la sauvegarde du patrimoine immatériel. Les cas du textile, de la musique et de la mine

Stéphane Chaudiron, Bernard Jacquemin, Eric Kergosien

L'apport du web sémantique à la sauvegarde du patrimoine immatériel. Les cas du textile, de la musique et de la mine*

Stéphane Chaudiron, Bernard Jacquemin, and Éric Kergosien

Univ. Lille, EA 4073 – GERiiCO, F-59000 Lille, France.
{Prenom.Nom}@univ-lille.fr

Résumé

L'article présente une méthodologie composite permettant de créer des ontologies de domaines dans le cadre de la préservation et la valorisation du patrimoine culturel et industriel. Fondée sur la reprise d'outils documentaires divers, l'analyse terminologique, des entretiens ethnographiques, la fouille de textes et l'analyse cartographique du web, cette approche est expérimentée dans trois univers distincts : la musique, le textile et les mines de charbon.

Mots-clefs : ontologie, musique, textile, mines de charbon.

Abstract

The article presents a methodology aiming at building domain ontologies in order to preserving and valuing cultural and industrial heritage. This approach is based on (i) merging different documentary tools,(ii) exploiting specialized terminologies, (iii) ethnographic interviews, (iv) text mining and (v) web analysis. Experimentations of this methodology have been made in three different domains: musical heritage, textile industry and coal mining industry.

Keywords: ontology, music, textile, coal mining.

1 Introduction

Les technologies du web sémantique sont désormais couramment utilisées par les internautes pour consulter des données institutionnelles (data.bnf.fr), administratives publiques (data.gouv.fr), encyclopédiques (DBpedia, Wikidata) ou privées (data.ratp.fr, data.sncf.com). Plus récemment, ces technologies sont également utilisées dans le cadre de la sauvegarde du patrimoine culturel et industriel (Chowdhury & Ruthven, 2015 ; Hastings, 2014) et peuvent participer à la préservation et la valorisation du patrimoine,

*Chaudiron, Stéphane, Jacquemin, Bernard et Kergosien, Éric, 2019. L'apport du web sémantique à la sauvegarde du patrimoine immatériel. Les cas du textile, de la musique et de la mine. In: Roxin, Ioan, Tajariol, Federico, Hosu, Ioan et Pélissier, Nicolas (éd.), *Information, communication et humanités numériques. Enjeux et défis pour un enrichissement épistémologique. Actes du 23e colloque bilatéral franco-roumain en Sciences de l'information et de la communication*, HumaNum, Cluj-Napoca (Roumanie), 18-20 octobre 2018. Cluj-Napoca (Roumanie): Accent, p. 311-328.

qu'il soit actuel et partagé comme le patrimoine musical, ou proche de la disparition (*endangered cultural heritage* [1]) comme les industries textiles et minières dans certaines régions du monde. Dans cet article, nous présentons une approche méthodologique visant à représenter les connaissances de trois domaines (la musique, le textile et les mines), au travers d'ontologies de domaine. Ces approches ont été développées dans le cadre des projets ANR Dorémus [2] (musique), DENIM [3] (textile) et ANR Mémo-Mines [4].

Les fondements des projets sont différents. Pour le patrimoine musical, il s'agit de définir un système de représentation de connaissances qui sont actuelles, évolutives et partagées entre différents acteurs. La musique enregistrée est omniprésente et les modalités d'écoute très nombreuses (*streaming* sur Deezer ou Qobuz, *playlists* sur les *smartphones*...), mais leur description demeure très pauvre et leur accessibilité problématique. Comme le soulignent Choffé et Leresche (2016, p. 1), « [i]l est en effet très difficile de trouver sur le web l'histoire d'une œuvre musicale, son origine culturelle, son compositeur, son parolier, ses influences, ses reprises, ses interprétations... Pourtant ces connaissances existent et sont décrites finement par les institutions culturelles et les médias, mais elles sont en majorité cachées dans leurs systèmes d'information qui constituent autant de silos. Pour relier les ressources musicales et l'information descriptive existante, il faut sortir cette information des silos et la rendre disponible sur le web de données ».

Le contexte des domaines du textile et de la mine est différent. Malgré les destructions induites par les deux grands conflits mondiaux de 1914-1918 et 1939-1945, les secteurs du textile et de l'industrie charbonnière constituaient dans les années 1950 deux des trois piliers industriels du Nord-Pas-de-Calais avec la sidérurgie. Si la dernière mine de charbon a fermé en 1991, l'industrie textile a su s'adapter grâce à un redéploiement qui s'est opéré autour de nouveaux débouchés dans les années 1980.

Si la richesse et la diversité des lieux de mémoire (musées, bibliothèques, associations, entreprises, etc.) liés à cette histoire industrielle sont un atout pour le territoire, la diversité et la dispersion des ressources (fonds, collections d'objets et de machines, témoignages) les rendent peu visibles. Malgré des initiatives en faveur de leur valorisation depuis la fin des années 1970 et l'inscription du bassin minier au patrimoine mondial de l'UNESCO en 2012, le constat s'inscrit dans celui, plus général, d'un manque de lisibilité du domaine de la culture scientifique, technique et industrielle en région Hauts-de-France.

Les lieux dans lesquels sont stockées et/ou disponibles ces ressources ne sont pas toujours clairement identifiés. Les moyens humains et matériels dans les différents centres documentaires sont souvent limités et ne permettent pas de valoriser de façon cohérente les fonds en respectant des règles communes de catalogage. L'un des objectifs des projets DENIM et Mémo-Mines est précisément d'utiliser les technologies du web sémantique afin de préserver le patrimoine industriel du textile et de la mine.

L'article est structuré en trois parties. Nous présentons tout d'abord le contexte et les objectifs des projets avant de décrire la méthodologie qui a été développée pour la

1. *Cf.* UNESCO : `http://whc.unesco.org/en/danger/`.

2. DOnnées en RÉutilisation pour la Musique en fonction des USages (projet ANR-2014-CE24-0020). Voir `http://www.doremus.org/`.

3. Données numériques, langages et représentations du patrimoine textile en région Nord-Picardie : quelles compréhensions réciproques ? (Projet Ministère de la Culture). Voir `https://reccits.hypotheses.org/753`.

4. `https://memomines.hypotheses.org/`.

construction des ontologies puis de présenter les modèles choisis (FRBRoo et CIDOC CRM).

2 Objectifs des projets

2.1 Le projet Dorémus

Le projet Dorémus est fondé sur le constat paradoxal que l'information descriptive de la musique est à la fois particulièrement riche, abondante et très utilisée, mais également peu normée, très disparate et donc difficilement exploitable. En conséquence, l'accès à une œuvre précise ou à une information portant sur une œuvre précise est souvent malaisé voire impossible. Ce projet se donne dès lors pour objectif de proposer un modèle descriptif de la musique qui prenne en compte la variabilité importante des données musicales décrivant les œuvres, les enregistrements, les compositions et toutes les réalités du domaine de la musique, de manière à rendre accessibles et interopérables les grandes bases d'information existantes qui décrivent les données musicales tout en conservant à la fois l'intégrité des données et leur niveau de granularité (Lisena et al., 2018). Il faut également que le modèle proposé apporte l'expressivité informationnelle nécessaire aux spécialistes de l'information et de la musique tout en restant accessible aux connaissances et aux pratiques du grand public (Cotte, Despres-Lonnet, Vandiedonck, Heizmann, & Jacquemin, 2015 ; Debruyne, 2012).

Pour répondre à ces aspirations, un consortium de chercheurs, développeurs et professionnels de l'information musicale a été rassemblé. Il comprend des membres de trois institutions gestionnaires de l'information musicale que Dorémus entend rendre compatible, exploitable et accessible : la Bibliothèque nationale de France [5] (BnF), la Philharmonie de Paris [6] et Radio France [7] (Paris). Plusieurs catalogues existent en effet, voire coexistent, au sein de ces établissements, qui décrivent les documents musicaux réalisés ou conservés en leur sein. Ces institutions apportent également leurs compétences et leur savoir-faire en termes de représentation et de manipulation de l'information descriptive de la musique. Deux laboratoires spécialisés dans la modélisation des données pour le web sémantique (EURECOM [8], Nice ; LIRMM [9], Montpellier) apportent leur expertise dans le développement du modèle informationnel, d'outils permettant le maniement du modèle et des données, et dans la réalisation des alignements nécessaires (référentiels, données descriptives). Le laboratoire GERiiCO [10] (Lille), spécialisé dans l'analyse des usages et pratiques informationnelles en particulier en contexte numérique, se charge d'étudier et de formaliser les demandes, les besoins et les envies des publics concernés pour orienter la définition du modèle ontologique et surtout les fonctionnalités auxquelles il ouvre (Cotte, 2011). Une société spécialisée dans la gestion de l'information musicale (Meaning Engines [11], Paris) apporte la dimension opérationnelle nécessaire au fonctionnement des outils imaginés dans le cadre du projet et assure la portabilité des outils et des données à l'échelle du

5. http://www.bnf.fr/.
6. https://philharmoniedeparis.fr/.
7. http://www.radiofrance.fr/.
8. http://www.eurecom.fr/.
9. http://www.lirmm.fr/.
10. https://geriico-recherche.univ-lille3.fr/.
11. http://www.meaningengines.com/.

web sémantique. Le cabinet Ourouk [12] (Paris), spécialisé dans la consultance sur des projets de management de l'information, est chargé de la gestion.

2.2 Les projets DENIM et Mémo-Mines

Les projets DENIM (financé par le ministère de la Culture) et Mémo-Mines [13] (financé par l'ANR) partagent l'objectif d'améliorer la visibilité des patrimoines textile et minier des Hauts-de-France. Pour ces deux secteurs industriels, le constat est celui d'une immense variété et richesse du patrimoine, mais d'une faible visibilité. L'objectif de chacun des projets est de proposer un modèle descriptif qui prenne en compte l'hétérogénéité des données.

Afin de borner le périmètre des ontologies à construire, nous nous appuyons sur la définition de la notion de patrimoine donnée par l'UNESCO (1954, 1970, 1982). En 1982, lors de la *Déclaration de Mexico sur les politiques culturelles*, l'UNESCO a reprécisé la définition en déclarant que le patrimoine culturel d'un peuple « s'étend aux œuvres de ses artistes, de ses architectes, de ses musiciens, de ses écrivains, de ses savants, aussi bien qu'aux créations anonymes, surgies de l'âme populaire, et à l'ensemble des valeurs qui donnent un sens à la vie. Il comprend les œuvres matérielles et non matérielles qui expriment la créativité de ce peuple : langue, rites, croyances, lieux et monuments historiques, littérature, œuvres d'art, archives et bibliothèques » (p. 3).

Comme le soulignent Babelon et Chastel (1994), l'intérêt pour le patrimoine industriel est assez récent. C'est en effet dans les années 1970 que l'on commence à comprendre que les vieux bâtiments méritaient mieux que la casse, qu'un paysage devait se protéger et que les « Gueules Noires » (nom donné aux mineurs de charbon), comme tous les ouvriers, qui vieillissaient, qui disparaissent peu à peu, ne devaient pas être gommées de la mémoire collective. Le « patrimoine industriel » s'impose alors dans les discours et devient un objet d'études. La multiplication des friches industrielles et l'épineuse question de leur devenir contribue à stimuler la réflexion et à susciter les débats. Le Comité international pour la conservation du patrimoine industriel en propose ainsi une définition plus précise (TICCIH, 2003, p. 1) : « Le patrimoine industriel comprend les vestiges de la culture industrielle qui sont de valeur historique, sociale, architecturale ou scientifique. Ces vestiges englobent : des bâtiments et des machines, des ateliers, des moulins et des usines, des mines et des sites de traitement et de raffinage, des entrepôts et des magasins, des centres de production, de transmission et d'utilisation de l'énergie, des structures et infrastructures de transport aussi bien que des lieux utilisés pour des activités sociales en rapport avec l'industrie (habitations, lieux de culte ou d'éducation)... ».

Ce sont ces définitions qui ont orienté notre analyse des documents afin d'en extraire les éléments caractéristiques du patrimoine.

3 Méthodes de conception

La méthodologie générale de conception des ontologies s'appuie sur une approche hybride qui prend en compte (1) l'utilisation de lexiques et/ou d'outils documentaires

12. `http://www.ourouk.fr/`.

13. Conversion des traces mémorielles en médiations numériques : le cas de la mémoire minière (projet ANR-16-CE38-0001-02).

comme des thésaurus ou des plans de classement (quand ils existent), (2) une cartographie des acteurs du domaine, (3) la collecte de la mémoire (captation vidéo de témoignages, entretiens ethnographiques, analyse de fonds presse...) et (4) l'analyse de fonds documentaires des domaines concernés. Nous précisons dans les trois sections suivantes les techniques effectivement mobilisées dans le cadre des différents projets.

3.1 Le projet Dorémus

La méthode mise en place au cours du projet Dorémus s'appuie essentiellement sur les outils documentaires propres aux institutions partenaires du projet et sur l'analyse des fonds documentaires de ces mêmes institutions. Elle tient compte à la fois des qualités informationnelles des catalogues hétérogènes à traiter et également de leur grande diversité. Ainsi, la Bibliothèque nationale de France (BnF) apporte des données catalographiques au format INTERMARC portant sur plus de 135 000 œuvres et sur près de 90 000 partitions ; Radio France exploite un format XML (*Extensible Markup Language*) pour décrire plus de 62 000 œuvres et plus de 9 000 partitions, mais aussi 9 500 concerts et 340 000 fichiers musicaux ; la Philharmonie exploite tantôt un format UNIMARC (près de 7 000 œuvres et plus de 30 000 partitions) et tantôt une structure XML (plus de 5 000 concerts et plus de 8 600 enregistrements musicaux).

Au-delà de l'élaboration de ce modèle conceptuel compatible et modulaire, la mise en œuvre de l'alignement d'informations concurrentes ou complémentaires, mais hétérogènes, au sein de ce modèle constitue un défi complexe. Il ne s'agit pas en effet de dupliquer Mozart (1756-1791), qu'il s'appelle Wolfgangus Theophilus, Wolfgang Amadeus ou simplement Amadeus – tout en le distinguant de son père Léopold –, pas plus qu'*Une petite musique de nuit* ne doit être différenciée d'*Eine kleine Nachtmusik* ou de la *Sérénade n° 13 en sol majeur*. Plusieurs vocabulaires issus des différentes institutions, et variant en granularité, format, structure, voire langue, doivent dès lors être alignés en amont de l'élaboration et du peuplement du modèle ontologique : personnes (auteurs, compositeurs, interprètes, etc.), éléments musicaux (instruments, périodes, clés musicales, voix, genres, formes, etc.) et descripteurs (thésaurus, lexiques).

Cependant, le caractère structuré de l'information originale (XML ou un format MARC) nous a amenés à convertir systématiquement les contenus initiaux en une structure SKOS [14] permettant de représenter sous forme de graphes les liens entre les informations présentes dans les divers vocabulaires utilisés. Cette structure SKOS présente l'avantage de rendre aisée une conversion vers d'autres langages de représentation du sens dans le web sémantique (OWL [15] ou RDF [16]). Dans le cadre de l'alignement des vocabulaires, elle permet également d'effectuer une comparaison systématique et une validation semi-automatique des graphes issus des différentes ressources de manière à associer les graphes identiques, à dissocier les graphes incompatibles et à compléter les graphes complètement ou partiellement compatibles, mais comportant une information permettant la désambiguïsation ou l'association. Deux graphes décrivant le compositeur Mozart, dont le prénom diffère, mais qui comportent la même date de naissance seront unifiés et bénéficieront de compléments d'information n'apparaissant que dans l'un des deux, comme un lieu de naissance ou une fonction par exemple. Cette structure permet enfin l'interopérabilité entre les vocabulaires ainsi générés et d'autres sources informationnelles accessibles via le web de données (telles que data.bnf.fr ou DBpedia) pour enrichir encore l'information disponible sur ces diverses réalités.

14. *Simple Knowledge Organization System*.
15. *Web Ontology Language*.
16. *Resource Description Framework*.

3.2 Le projet DENIM

Pour le domaine du textile, nous avons opté pour une méthode hybride qui combine une approche qualitative à une approche semi-automatisée consistant à dresser une cartographie des acteurs du patrimoine afin d'identifier les sources de données numériques existantes. La méthode cartographique permet en effet de construire des connaissances tant sur le réseau d'acteurs que sur les sources d'informations existantes.

La cartographie du web montre la présence de chaque acteur ou catégorie d'acteurs sur Internet en utilisant les hyperliens qui représentent les liens sociaux entre acteurs (Severo, 2012). Elle s'appuie sur des outils d'exploration du web pour l'identification des acteurs et sur des outils de représentation pour l'analyse des résultats identifiés. En ce qui concerne l'exploration, les outils les plus employés sont les *crawlers* (robots d'indexation).

Les *crawlers* peuvent être semi-automatiques (issueCrawler, Hyphe, etc.) ou, plus rarement, manuels (Navicrawler, etc.). Un crawler semi-automatique utilise un script qui suit et répertorie tous les hyperliens depuis un site donné, puis tous les hyperliens depuis les sites qu'il rencontre, et ainsi de suite. Un tel outil présente les avantages d'être généralement gratuit et simple à utiliser : il suffit de donner une liste de liens, de spécifier le type et la profondeur du *crawl* et l'outil fournit en réponse un réseau. Une limite importante est que, seul, il converge rapidement vers une petite minorité de sites (couche haute du web) qui constituent la cible de la large majorité des liens hypertextes (Barabási, Albert, & Jeong, 2000). Un *crawler* manuel permet de préciser clairement les limites d'un réseau en indiquant site par site s'il doit ou non intégrer le réseau.

La méthode hybride que nous avons utilisée combine une approche qualitative sous forme d'entretiens semi-directifs permettant de dresser un premier réseau des acteurs (et des ressources disponibles) du territoire d'étude, et une application quali-quantitative de la cartographie du web. En effet, dans le cas d'une étude d'un domaine tel que celui du patrimoine de l'industrie textile sur une échelle géographique limitée, la combinaison de deux types de *crawlers* est intéressante.

Une analyse manuelle avec Navicrawler à partir de la liste des sites dressée via les entretiens permet d'identifier de manière exploratoire les principaux acteurs du web dans le domaine considéré. Ensuite, l'emploi du *crawler* automatique Hyphe permet d'une part de repérer certains sites pertinents qui peuvent avoir échappé à l'observation manuelle, et d'autre part de reconstruire tous les hyperliens internes au corpus sélectionné. En ce qui concerne la représentation, les graphes se sont imposés comme la forme de visualisation pertinente pour représenter les liens entre acteurs. La topologie des réseaux peut ensuite être visualisée à l'aide d'un logiciel tel que Gephi dans lequel chaque nœud représente un site web et les arcs entre les nœuds visualisent les liens entre les sites. Les cartographies obtenues sont ensuite complétées par une analyse spatiale réalisée par un démonstrateur développé en nous appuyant sur Google Maps (Berthelot, Severo, & Kergosien, 2016).

À partir de 9 entretiens réalisés auprès d'acteurs du domaine dans la métropole lilloise, la méthode a permis d'identifier un réseau (non exhaustif) de 169 acteurs. À partir de cette liste, nous avons collecté un premier ensemble de 6 000 documents hétérogènes (images avec notices descriptives XML de la bibliothèque Georges Lefebvre – université de Lille, articles de presse de *La Voix du Nord* et de *Nord Éclair*, témoignages retranscrits, documents du Service Commun de la Documentation de l'université de Lille, notices et Plan local d'Urbanisme disponibles sur le portail de la Métropole européenne de Lille (MEL), notices de l'Inventaire de la Région). Nous nous sommes

alors concentrés sur trois types de documents pour la phase d'extraction et de structuration des connaissances :

— 142 articles de presse (*La voix du Nord*, *Nord Éclair*) collectés entre 2004 et 2016 par l'AASPT (Association des Anciens Salariés du Peignage de la Tossée) et numérisés via l'Agence Nationale de Reproduction des Thèses (ANRT, Lille) ;

— 59 témoignages retranscrits auprès des anciens acteurs du domaine par des étudiants de l'Institut Social de Lille en 2012 ;

— ouvrages anciens sur l'industrie textile et sur l'exposition internationale de Roubaix de 1911 (bibliothèque Georges Lefebvre, université de Lille) numérisés et OCRisés via l'ANRT. Ces ouvrages ont été mis à disposition du public sur le portail NordNum (bibliothèque numérique de l'université de Lille consacrée à l'histoire du Nord et du Pas-de-Calais).

Parmi l'ensemble des descripteurs caractérisant le domaine du patrimoine de l'industrie textile, nous nous sommes concentrés sur les informations de type *acteur*, *lieu* et *thématique* qui sont présentes dans le corpus analysé. Les lexiques des acteurs et des lieux sont respectivement constitués des 169 acteurs identifiés et de l'ensemble des communes de la région Hauts-de-France. Concernant le lexique thématique, comme il n'existe pas de ressource lexicale caractérisant le domaine de l'industrie textile de façon précise et avec une couverture importante, le choix a été de partir de la version XMLisée du *wiktionary* français, GLAWI, qui est une ressource grand public, libre et facile d'utilisation.

GLAWI est ainsi à notre connaissance le lexique disponible en ligne le plus représentatif du domaine (300 termes) parmi les lexiques existants (Rameau, Joconde, etc.) qui sont soit trop généraux, soit trop spécifiques. Il a donc été sélectionné comme ressource pivot pour la construction du lexique du domaine. À partir des 13 lexiques identifiés sur le web, divers traitements ont permis d'aboutir à un lexique enrichi de plus de 2 000 termes.

L'étape suivante a consisté à extraire de nouveaux termes à partir du corpus documentaire identifié lors de la cartographie du web. Pour chacun des 2 000 termes du lexique, l'algorithme *word2vec* a été utilisé pour rechercher l'ensemble des termes apparaissant régulièrement à proximité de ceux-ci. À l'issue de cette phase, et après validation manuelle, le lexique initial a été enrichi d'un peu plus de 48 termes.

Une fois les données extraites, nous avons construit une première base de connaissances permettant de structurer les informations liées au patrimoine industriel textile implicitement décrit dans les documents du corpus. Le travail a consisté (1) à formaliser chacune des informations extraites (acteurs, lieux, thématiques entités temporelles et références aux documents) dans une même base formalisée en OWL CIDOC CRM, puis (2) à les analyser/valider via un logiciel permettant de visualiser l'ontologie produite. Le modèle est présenté dans la section 4 de l'article.

3.3 Le projet Mémo-Mines

La méthode utilisée dans le cadre du projet Mémo-Mines est très similaire à celle présentée pour le domaine textile. Elle se décompose ainsi : (1) réalisation d'une cartographie, (2) collecte et analyse de ressources lexicales sur la mine (33 ressources), (3) analyse de différents outils documentaires (vocabulaires contrôlés, plans de classement, thésaurus…), (4) numérisation et analyse d'un fonds d'archives de presse (env. 1 000 articles, *La Voix du Nord*, *Le Nord*, 1990-2018), (5) analyse d'un fonds d'archives

vidéo (25 heures de captation de témoignages d'anciens mineurs) et (6) captation de nouvelles vidéos (démarche ethnographique en cours).

La cartographie des acteurs a été réalisée selon la même approche utilisée pour le textile, permettant d'identifier 51 acteurs (musées, associations…). L'étape de collecte et d'analyse lexicale de la mine (Daloz, 2018) a permis de construire un corpus de référence (Rastier, 2004) de 33 ressources (dictionnaires, lexiques, glossaires…) représentatives du Nord-Pas-de-Calais. Une analyse semi-automatique de ce corpus a ensuite permis de définir la nature des termes et de leurs relations, et d'étudier les particularités de la terminologie minière. Cinq critères ont été retenus : la volumétrie, la couverture informationnelle, la dialectologie, la spécialisation des ressources et la granularité sémantique. Une analyse détaillée des ressources a abouti à un lexique final de 2 400 termes intégrant les variantes dialectales (Daloz, 2018).

Le lexique a été structuré hiérarchiquement en domaines et sous-domaines à partir d'un index classant le lexique en six grands domaines : le travail du mineur, la mine, le transport, l'air et l'eau, le mineur, la productivité et appointements (Turpin, 2004). La plupart des termes se situent dans les deux premiers thèmes. Le domaine du travail du mineur est le plus spécialisé, avec sept sous-thèmes : l'abattage à l'explosif, le boisage et le soutènement, le gros matériel d'abattage et ses accessoires, les actions, les cadres et piles de soutènement, les métiers et tâches, les outils et pièces diverses. Le travail de structuration hiérarchique est poursuivi afin d'obtenir un thésaurus du domaine. L'enjeu concernant la granularité concerne le niveau sémantique de description des termes descripteurs. Celle-ci peut s'obtenir en étudiant les définitions des termes contenus dans le corpus. Ainsi, le terme générique d'ABANDAGE (ou son équivalent *abandache* en dialecte) sera le terme boisage selon trois lexiques dans lesquels ces termes apparaissent [17]. Un même terme peut être défini par plusieurs termes génériques différents : ceux-ci sont alors considérés comme synonymes.

Lorsque l'on remonte la chaîne hiérarchique dans la définition du terme générique boisage, deux sens se distinguent ; le premier est celui de « consolidation des galeries » ou « pose d'un soutènement » ou encore « opération d'étaiement », le deuxième désigne « l'ensemble des bois de soutènement/des étais ». C'est le premier sens qui est utilisé dans la définition d'abandage. Se dessine alors une hiérarchie que nous représentons dans le micro-thésaurus [18] suivant :

TG Consolidation
TS Consolidation des galeries
    EP Pose de soutènement
    EP Opération d'étaiement
TS Boisage
    EP Étaiement
TS Abandage (fr.)
    EP Abandache (dial.)

Il est également possible de choisir les termes génériques en étudiant les collocations et leur base. Par exemple, les termes *porion à terre*, *porion d'about*, *porion de coupe*, *porion de guides* ou *monte ed porion* ont pour point commun d'être formés sur la base *porion* qui sera alors considéré comme générique tandis que les autres seront spécifiques. Ces résultats montrent que l'étude des définitions et des unités lexicales

17. Abandage : *Boisage provisoire destiné à maintenir la paroi et le toit au début d'un chantier avant le déhouillement* (Turpin, 2004).
18. TG = terme générique, TS = Terme spécifique, EP = Employé pour.

composées du corpus permet d'accéder à une granularité sémantique très fine. À ce stade du projet, l'établissement des relations hiérarchiques entre les termes et leur analyse sémantique a permis de construire une première version d'un thésaurus de la mine, qui est également enrichi à partir d'une comparaison systématique avec d'autres outils documentaires identifiés (classifications, plans de classement, vocabulaires contrôlés).

Les étapes 4 à 6 de la méthodologie sont en cours de réalisation. L'analyse du corpus de presse se fera selon la méthode décrite pour le fonds presse du domaine textile.

4 Les modèles

Le web sémantique intéresse les organismes de grande taille confrontés à la multitude et à la diversité des données car il permet de lier et de faire communiquer des données d'origines très différentes. Celles-ci doivent être traduites dans un langage informatique spécifique propre aux données ouvertes afin de répondre aux exigences techniques exprimées par les besoins d'accès automatique au sens des informations plutôt qu'à leur forme : XML, RDF, OWL ou SPARQL en sont quelques exemples emblématiques. Mais si ces langages offrent bien les fonctionnalités nécessaires pour la mise en œuvre d'outils de traitement du sens, ils ne préjugent cependant pas de l'angle sous lequel sont abordées les données et leur signification, et ils laissent toute latitude pour en évoquer la logique.

En effet, un seul mode de représentation du sens ne peut à ce jour – et ne pourra sans doute jamais – prendre en charge la description universelle des données dans toutes leurs dimensions. Si aucune structure informationnelle spécifique ne semble avoir été conçue pour décrire le patrimoine culturel ou industriel, il existe cependant plusieurs exemples de formalismes créés pour décrire les objets culturels tout en exprimant les relations pouvant exister entre eux soit explicitement, soit en facilitant l'utilisation d'outils du web sémantique pour dépasser l'implicite. Il s'agit des modèles FRBR (Le Bœuf, 2013), CIDOC CRM (Doerr, 2003), et FRBRoo (Doerr, Le Bœuf, & Bekiari, 2008).

FRBR (*Functional Requirements for Bibliographic Records*) est un modèle de description qui distingue quatre niveaux d'information portant sur un même objet (initialement bibliographique), depuis ses caractéristiques physiques qui doivent être distinguées pour chaque exemplaire (« item ») jusqu'aux spécificités les plus abstraites de sa conception (« œuvre ») en passant par les spécifications de sa mise à disposition d'un public (« manifestation ») et celles de son contenu intellectuel (« expression »). À chaque niveau de description, le renseignement des champs informationnels n'est pas forcément opéré par une explicitation locale mais, dans la mesure du possible, par une référence au modèle FRAD (pour les personnes physiques et morales) ou au modèle FRSAD (pour les lieux, événements, concepts et objets). Un réseau dense de relations se construit dès lors entre les œuvres, entre les autorités et entre les descripteurs qui y sont attachés, sortant des limites classiques de la fiche descriptive.

Le modèle conceptuel de référence (*Conceptual Reference Model*) CIDOC CRM est un modèle de représentation de données conçu par le Comité International pour la DOCumentation du Conseil International des Musées pour permettre l'interopérabilité des référencements des objets de musées puis, par extension, de tout objet du patrimoine culturel physique ou non, selon la définition proposée par l'UNESCO. Il vise à dépasser les incompatibilités sémantiques et structurales des nombreuses sources d'informations hétérogènes portant sur des réalités patrimoniales et culturelles, afin de faci-

liter l'échange de documentations et la recherche dans ces documentations. La version actuelle (ISO 21127:2014) intègre 86 classes (acteurs, lieux, événements ou entités temporelles…) qui sont reliées entre elles par 137 propriétés distinctes. Le modèle est assorti de plusieurs outils, dont des implémentations OWL et RDF et des utilitaires de *mapping* avec d'autres formalismes (UNIMARC, EDM…).

FRBRoo est une évolution « orientée objet » imaginée à partir de FRBR et de CIDOC CRM. Très ambitieuse, l'ontologie FRBRoo est conçue pour prendre en charge, décrire et mettre en relation toute réalité de l'univers culturel. Le modèle dans son état actuel n'est pas encore stabilisé, et toutes les questions conceptuelles qu'il soulève n'ont pas encore obtenu de réponse. Son développement est néanmoins organisé de manière à ce qu'il puisse être instancié automatiquement par des données issues de ses modèles « parents », CIDOC CRM et FRBR.

4.1 Le projet Dorémus

Pour assurer la navigation sémantique dans l'information musicale, les partenaires de Dorémus ont choisi d'utiliser l'ontologie FRBRoo (Riva, Doerr, & Žumer, 2008 ; Smiraglia, Riva, & Žumer, 2013) qui bénéficie des avantages cumulés du web sémantique (unification sémantique de données hétérogènes distribuées, pérennité et ouverture des formats) et des modèles ontologiques (approche conceptuelle et explicitation des relations de sens) tout en implémentant les structures nécessaires pour le catalogage de l'information culturelle patrimoniale. En effet, FRBRoo est issu d'une harmonisation entre le modèle CIDOC CRM (Doerr, 2003) destiné aux musées et à l'information patrimoniale, et le modèle FRBR (Le Bœuf, 2013) dévolu à l'information bibliographique. En particulier, FRBRoo permet de décrire un objet culturel tel qu'un document musical en le caractérisant selon les entités catalographiques de FRBR que sont l'*Œuvre* (création intellectuelle), l'*Expression* (réalisation de l'*Œuvre*), la *Manifestation* (matérialisation de l'*Expression*) et l'*Item* (exemplaire « physique » et particulier de la *Manifestation*), ce qui permet de distinguer les caractéristiques permanentes d'une œuvre (compositeur, numéro d'opus, titre) et celles qui sont plus ponctuelles (interprètes lors d'un concert, enregistrement particulier, variations d'orchestrations). En outre, FRBRoo autorise la création et la publication d'extensions lorsque l'expressivité du modèle original est insuffisante au regard des besoins de modélisation d'un domaine.

Le modèle informationnel proposé par le consortium Dorémus place l'*Expression* au centre de la modélisation de l'information musicale, car elle correspond à la réalisation instinctive et directement compréhensible d'une *Œuvre*, quant à elle très abstraite et désincarnée. C'est en effet seulement à partir du niveau de l'entité *Expression* qu'une *Œuvre* peut être réalisée, et qu'une description – donc une association d'éléments d'information – peut être entreprise. L'essentiel de cette information catalographique sera donc associé soit à l'*Expression* (quand c'est l'œuvre elle-même qui est décrite), soit à la *Manifestation* (quand le descriptif porte sur une interprétation particulière ou un enregistrement spécifique de l'œuvre), soit à l'*Item* (quand il s'agit de données relatives à un fichier, un support enregistré ou un document particulier).

En outre, le modèle FRBR, et par conséquent FRBRoo, s'appuie également sur des entités d'autorités (personnes, collectivités) et d'indexation (événement, lieu…) qui permettent d'intégrer l'information issue de référentiels stables aux divers niveaux informationnels (instruments, orchestres, interprètes, labels, etc.). L'intégration ontologique de l'information strictement musicale issue des catalogues avec des données biographiques, géographiques, historiques… provenant des référentiels permet dès lors

une navigation pratiquement illimitée (Nentwig, Hartung, Ngonga Ngomo, & Rahm, 2017) à travers les graphes sémantiques ainsi créés, et rend également disponible tout type d'information encodée à la demande (Ferrara, Lorusso, Montanelli, & Varese, 2008 ; Shvaiko & Euzenat, 2013).

L'implémentation du modèle a donc été pensée et réalisée de manière à répondre non seulement aux exigences de la description documentaire par et pour les professionnels de la documentation musicale (Choffé & Leresche, 2016), mais également aux demandes des autres usagers passionnés de musique, quel que soit leur niveau de connaissances musicologiques. Le modèle ainsi que les données sont diffusés à travers une interface de consultation [19] élaborée à partir de demandes formulées par un échantillon de différents utilisateurs de l'information musicale interrogés au cours des phases successives du projet.

4.2 Les projets DENIM et Mémo-Mines

Du fait de son niveau élevé de maturité et de sa stabilité, nous avons choisi de mettre en œuvre l'ontologie CIDOC CRM dans les deux domaines du textile et des mines. Les outils déjà proposés ainsi que son interopérabilité planifiée avec FRBRoo 324ont conforté ce choix. Pour illustrer le résultat obtenu dans le domaine du textile, nous décrivons ici un exemple représentant l'ontologie produite pour quatre des documents du corpus documentaire. L'objectif étant de traiter sémantiquement l'information issue de sources hétérogènes, nous avons conservé des documents hétérogènes.

Il s'agit de quatre documents très distincts tant par leur forme que par leur contenu, ainsi que par leur producteur. Deux d'entre eux sont des descriptifs d'objets du patrimoine issus d'acteurs institutionnels (une fiche extraite de l'Inventaire général et une autre issue de la base de photographies du laboratoire de recherche IRHiS, Université de Lille), se présentant sous la forme de fichiers XML, mais dont la structure informationnelle est radicalement différente ; un autre est un article de presse de la *Voix du Nord*, en texte brut non structuré, et sans vocation descriptive spécifique ;le dernier est un document PDF de la MEL. Tous les quatre répondent à la requête géographique, et deux d'entre eux à une répondent au lexique du bâti industriel (usine textile, filature, lainerie, etc.). L'ensemble des informations pertinentes collectées dans le corpus de test a été intégré au modèle comme instances de classes, et les propriétés qui les relient ont été générées soit directement par le modèle, soit par un moteur d'inférences intégré au modèle.

La Figure 1 est une projection d'un extrait de l'ontologie peuplée par notre corpus de test, visualisée via le logiciel *Protégé* (Musen, Wieckert, Miller, Campbell, & Fagan, 1995).

On notera que les quatre documents-tests traités dans cet exemple sont bien mis en relation dans le modèle, et que de nouvelles propriétés, absentes des sources d'information originelles, leurs sont associées, soit par la puissance du modèle (dans la classe *E53 Place*, Roubaix est une ville du Grand Lille, lui-même agglomération du Nord, etc.), soit grâce au moteur d'inférences qui crée de nouvelles relations (un événement *E5 Event* tel que l'*Exposition internationale de Roubaix* est une entité temporelle – *E2 Temporal entity* – qui a forcément un début et une fin). Dans cet exemple, un premier document IRHIS_FL1269145.xml relate la participation du président de la République de l'époque à l'exposition internationale du Textile en 1911, et un second document

19. `https://overture.doremus.org/`.

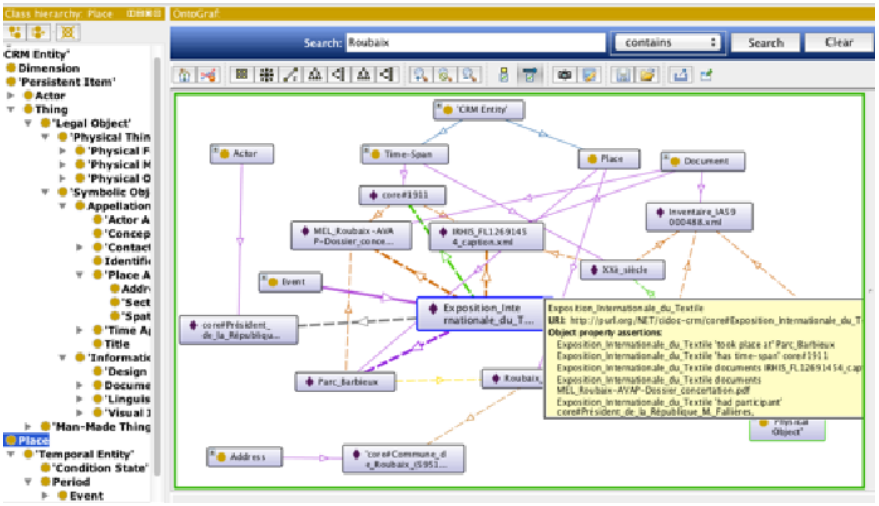

FIGURE 1 – Extrait de l'ontologie produite

MEL_Roubaix_AVA.pdf précise que l'événement a eu lieu le long du Parc Barbieux à Roubaix, commune du nord de la France.

5 Conclusion et perspectives

Cet article a présenté une méthodologie hybride permettant de concevoir des ontologies dans trois domaines culturels et industriels, destinées à préserver et valoriser le patrimoine. Alors que les contextes et les objectifs des trois projets sont en partie différents, l'article a mis en évidence des étapes communes dans la démarche générale, en particulier lors de la phase de collecte et de traitement des ressources hétérogènes. La plus grande proximité entre les patrimoines textile et minier a également validé la démarche quali-quantitative pour les phases de cartographie des acteurs à partir des ressources du Web et d'exploitation semi-automatique des corpus documentaires (fouille de textes et validation par les experts).

Le degré d'avancement différent des trois projets implique des perspectives différentes. Le projet ANR Dorémus est officiellement achevé mais le travail se poursuit, notamment pour enrichir l'ontologie et la valider auprès d'autres acteurs. Le projet DENIM est également terminé, mais nous souhaitons étendre la phase de construction d'une base de connaissances OWL CIDOC CRM à un volume plus important de documents. Une seconde action sera de proposer un moteur de recherche permettant de naviguer à travers le corpus indexé en s'appuyant sur la base de connaissances produite.

À ce jour, le projet Mémo-Mines est à mi-parcours. La ressource terminologique est en grande partie construite, le thésaurus est partiellement réalisé et l'ontologie du domaine minier est à construire.

Les recherches présentées dans cet article ont été partiellement financées par les projets ANR-2014-CE24-0020 « DOREMUS », ANR-16-CE38-0001 « MEMO-MINES » et Ministère de la culture « DENIM ».

Références

Babelon, J.-P., & Chastel, A. (1994). *La notion de patrimoine*. Paris : L. Levi.

Barabási, A.-L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks : The topology of the world-wide web. *Physica A : Statistical Mechanics and its Applications*, *281*(1–4), 69-77. doi: 10.1016/S0378-4371(00)00018-2

Berthelot, M.-A., Severo, M., & Kergosien, E. (2016). Cartographier les acteurs d'un territoire : une approche appliquée au patrimoine industriel textile du Nord-Pas-de-Calais. In *CIST. 3e colloque international. En quête de territoire(s) ? Looking for territorie(s) ? Proceedings* (p. 66-72). Grenoble : Collège international des sciences du territoire (CIST).

Choffé, P., & Leresche, F. (2016). DOREMUS : Connecting Sources, Enriching Catalogues and User Experience. In *Proceedings of IFLA WLIC 2016 - Connections. Collaboration. Community in Session 93 – Cataloguing and Information Technology.* Columbus, OH.

Chowdhury, G., & Ruthven, I. (2015). Managing digital cultural heritage information. In *Cultural heritage information : Access and management* (p. 1-12). London : Facet.

Cotte, D. (2011). *Émergences et transformations des formes médiatiques*. Paris : Hermès science-Lavoisier.

Cotte, D., Despres-Lonnet, M., Vandiedonck, D., Heizmann, M., & Jacquemin, B. (2015). Entre appréciation et description, les goûts musicaux à l'épreuve de la « data ». In C. Paganelli, S. Chaudiron, & K. Zreik (Eds.), *Documents et dispositifs à l'ère post-numérique. Actes du 18e Colloque international sur le Document Électronique (CiDE.18)* (p. 159-170). Montpellier : Europia.

Daloz, A. (2018). Vers la représentation terminologique d'un patrimoine culturel immatériel menacé de disparition : le cas du patrimoine minier. In *Toth 2018. Terminologie & ontologie : théories et applications.* Le Bourget du Lac.

Debruyne, F. (2012). Le disquaire et ses usagers. Du magasin au site Web. *Communication & langages*, *173*, 49-65. doi: 10.4074/S033615001201304x

Doerr, M. (2003). The CIDOC Conceptual Reference Module : An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, *24*(3), 75–92.

Doerr, M., Le Bœuf, P., & Bekiari, C. (2008). FRBRoo, a conceptual model for performing arts. In *Conference proceedings "The Digital Curation of Cultural Heritage"* (pp. 06–18). Athens. doi: http://www.cidoc2008.gr/cidoc/Documents/papers/drfile

Ferrara, A., Lorusso, D., Montanelli, S., & Varese, G. (2008). Towards a Benchmark for Instance Matching. In P. Shvaiko, J. Euzenat, F. Giunchiglia, & H. Stuckenschmidt (Eds.), *Proceedings of the 3rd International Conference on Ontology Matching* (Vol. 431, pp. 37–48). Karlsruhe , Germany : CEUR-WS.org.

Hastings, D. L. (2014). *Combating Visitor Pressure : Impact of Tourism on the Conservation of World Heritage Sites* (Master of Arts Thesis). University of Washington, Seattle.

Le Bœuf, P. (Ed.). (2013). *Functional Requirements for Bibliographic Records (FRBR) : Hype Or Cure-All ?* New York, London : Routledge.

Lisena, P., Achichi, M., Choffé, P., Cecconi, C., Todorov, K., Jacquemin, B., & Troncy, R. (2018). Improving (Re-)Usability of Musical Datasets : An Overview of the DOREMUS Project. *Bibliothek Forschung und Praxis*, *42*(2), 194-205. ((référence internationale WorldCat)) doi: 10.1515/bfp-2018-0023

Musen, M. A., Wieckert, K. E., Miller, E. T., Campbell, K. E., & Fagan, L. M. (1995). Development of a controlled medical terminology : Knowledge acquisition and knowledge representation. *Methods of Information in Medicine*, *34*(1-2), 85-95.

Nentwig, M., Hartung, M., Ngonga Ngomo, A.-C., & Rahm, E. (2017). A survey of current Link Discovery frameworks. *Semantic Web*, *8*(3), 419-436. doi: 10.3233/SW-150210

Rastier, F. (2004). Ontologie(s). *Revue d'intelligence artificielle*, *18*(1), 15-40.

Riva, P., Doerr, M., & Žumer, M. (2008). FRBRoo : Enabling a common view of information from memory institutions. In *World Library and Information Congress. 74th IFLA General Conference and Council.*

Severo, M. (2012). La cartographie du Web : le lien social sur le Net. In *Séminaire du Groupe f.m.r. (flux, matrices, réseaux).* Paris : UMR Géographie-Cité.

Shvaiko, P., & Euzenat, J. (2013). Ontology Matching : State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering*, *25*(1), 158-176. doi: 10.1109/TKDE.2011.253

Smiraglia, R. P., Riva, P., & Žumer, M. (Eds.). (2013). *The FRBR family of conceptual models : Toward a linked bibliographic future*. London ; New York : Routledge.

TICCIH. (2003). *Charte Nizhny Tagil pour Le Patrimoine Industriel*. Nizhny Tagil, Russie : The International Committee for the Conservation of the Industrial Heritage.

Turpin, B. (2004). *Les mots de la mine*. Paris : Maisonneuve et Larose.

UNESCO. (1954). *Convention pour la protection des biens culturels en cas de conflit armé, avec Règlement d'exécution*. La Haye : Centre du patrimoine mondial de l'UNESCO.

UNESCO. (1970). *Convention concernant les mesures à prendre pour interdire et empêcher l'importation, l'exportation et le transfert de propriété illicites des biens culturels*. Paris : Centre du patrimoine mondial de l'UNESCO.

UNESCO. (1982). *Déclaration de Mexico sur les politiques culturelles*. Mexico : Centre du patrimoine mondial de l'UNESCO.