Vers l'interopérabilité des données hétérogènes liées au patrimoine industriel textile

Towards interoperability of geographical data related to textile industry heritage

Éric KERGOSIEN, Bernard JACQUEMIN,
Marta SEVERO, Stéphane CHAUDIRON

GERiiCO, Université de Lille (France)

{Eric.Kergosien, Bernard.Jacquemin,
Marta.Severo, Stephane.Chaudiron}@univ-lille3.fr

Résumé

Le projet TECTONIQ étudie les dispositifs numériques mis en place par les différents acteurs impliqués pour gérer, diffuser et échanger les informations relatives au Patrimoine Industriel Textile (PIT) sur le territoire du Nord – Pas-de-Calais. Dans cet article, nous définissons tout d'abord notre domaine d'étude, à savoir le patrimoine numérique lié à l'industrie textile. Nous proposons ensuite une méthode hybride, c'est-à-dire une approche qualitative combinée à une approche quantitative semi-automatisée, afin de dresser une cartographie des acteurs du patrimoine nous permettant d'identifier les sources existantes de documents numériques hétérogènes. L'objectif du projet à terme étant de construire une base de connaissances qui structure et relie entre elles l'ensemble de ces données en respectant les normes définies pour le Web sémantique, nous justifions notre choix d'utiliser le modèle sémantique CIDOC CRM et nous présentons un extrait d'une première ontologie produite manuellement à partir d'un extrait du jeu de documents collecté.
Mots-clefs : Patrimoine de l'industrie textile, Territoire, Cartographie du Web, Organisation des connaissances, Interopérabilité, documents numériques hétérogènes, CIDOC CRM.

Abstract

The TECTONIQ project studies digital systems implemented by the different actors involved to manage, disseminate and exchange information about the Textile Industrial Heritage (TIH) on the territory of Nord of France. The objective is to provide a knowledge representation that interconnects all of these data using the semantic web structure technologies in order to assist the domain experts in producing and prividing digital content. The originality of the proposed project is to be part of a multidisciplinary approach to provide stakeholders, experts and non-experts, help in the discovery of knowledge specific to their heritage,

thanks to the extraction, structuration and visualization of knowledge from heterogeneous digital corpora. The studied territory is Nord-Pas-de-Calais in the north of France. In this paper, we present the methodology defined to get a stakeholder mapping of textile industrial heritage. Then, we present a first ontology built manually in OWL CIDOC CRM to merge data from heterogeneous numeric documents related to TIH.
Keywords: Textile industrial. Heritage, Territory, Web Mapping, Knowledge management, Interoperability, Heterogeneous numeric documents, CIDOC CRM.

1 Introduction

Riche d'une histoire de plus de dix siècles, le Nord – Pas-de-Calais (NPDC) est jalonné de bâtiments industriels, devenus monuments ou encore réaménagés en zones commerciales ainsi qu'en regroupements d'entreprises, témoins de ses influences historiques successives. Les documents numériques concernant le patrimoine industriel textile (PIT) sont variées (descriptifs d'objets, textes, images fixes et animées, sons, etc.) et portent sur des bâtiments, tissus, machines, techniques, acteurs, et plus généralement sur l'évolution dans le temps et dans l'espace de l'industrie textile propre à un territoire. Les acteurs institutionnels et associatifs notamment aux niveaux local et national produisent et enrichissent régulièrement ces connaissances. Pour sauvegarder et valoriser ce patrimoine, la Métropole Européenne de Lille (MEL) développe une politique de restauration ambitieuse liée à l'urbanisme, formalisée dans le Plan Local d'Urbanisme notamment. Le quartier de Moulins, Euratechnologie, la zone commerciale l'Usine de Roubaix et la Plaine Images à Tourcoing en sont autant d'exemples qui rendent le patrimoine toujours vivant, animé par des visites guidées, des expositions et des manifestations ouvertes à tous. De leur côté, les institutions expertes, notamment la DRAC et l'Inventaire général du patrimoine culturel de la Région, inventorient minutieusement l'évolution de ces sites dans le temps et dans l'espace, et travaillent ainsi à la sauvegarde de la mémoire de ce patrimoine bâti. Les archivistes, bibliothécaires et documentalistes, experts dans la conservation, le signalement et l'enrichissement des données numériques sous forme de métadonnées, participent également à la pérennisation de la mémoire relative au patrimoine matériel (objets de type tissu, machine…) et immatériel (méthodes et techniques, événements, mémoires…). Cependant, l'ensemble des contenus produits est hétérogène et sans liens explicites.

Dans cet article, nous présentons les premiers résultats de nos travaux menés dans le cadre du projet TECTONIQ [1] qui vise à l'identification et à la mise à disposition pour le plus grand nombre de la connaissance relative au PIT (matériel et immatériel) diffusée via les dispositifs numériques hétérogènes présents sur le territoire du NPDC. Une première difficulté consiste à identifier les contenus numériques relatifs au domaine étudié en prenant en compte la quantité et la qualité de ces contenus lors de la phase de collecte. Dans l'objectif de cadrer le domaine d'étude, nous précisons d'abord ce que nous entendons par Patrimoine Industriel Textile (PIT)

1. http://tectoniq.meshs.fr/

en nous appuyant notamment sur les définitions de l'UNESCO et du comité TICCIH (UNESCO, 2008 ; TICCIH, 2003). Nous présentons ensuite une méthode semi-automatisée pour identifier les acteurs diffusant de l'information du PIT afin ensuite de collecter les corpus numériques existants.

Des verrous sont notamment associés à l'hétérogénéité des documents traités aussi bien dans leur structure que dans leur contenu (notices descriptives, rapports techniques, compte-rendu de réunions publiques, articles de journaux, blogs, interviews retranscrites…). Une tâche importante à ce niveau consiste à construire un vocabulaire contrôlé de type ontologie décrivant le patrimoine industriel textile présent dans les documents, sur lequel nous pourrons nous appuyer pour distinguer et organiser le contenu relatif au patrimoine. Parmi les standards définis pour le Web sémantique, nous justifions notre choix d'utiliser le modèle sémantique CIDOC CRM (Doerr, 2003) qui offre (1) la possibilité de décrire de façon précise les différents aspects du patrimoine, et (2) d'homogénéiser la description du domaine d'étude présente dans les documents. Nous présentons une ébauche d'ontologie produite manuellement à partir d'un extrait réduit du jeu de données collecté.

Cet article est organisé de la façon suivante. La section 2 fait tout d'abord un état des définitions proposées par les acteurs institutionnels internationaux. Différents formalismes standardisés pour la représentation des connaissances liées au patrimoine sont ensuite présentés et comparés. La section 3 décrit notre méthodologie générique pour identifier, cartographier, et collecter acteurs et les données relatives au patrimoine. Une première ontologie minimale définie manuellement sur la base d'un jeu de tests est présentée et discutée.

2 État des lieux

2.1 Définition du domaine : le patrimoine industriel textile

L'UNESCO[2] occupe une place centrale sur la scène internationale culturelle et a beaucoup contribué à la définition de la notion de patrimoine (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 ».

L'intérêt pour le **patrimoine industriel** est un assez récent comme en témoigne cette citation de Jean-Pierre Babelon et André Chastel (1980) : « la patrimoine français s'est constitué par la conjonction de cinq patrimoines : la religion, la monarchie, la nation, le fait administratif et le fait technique. Le dernier porte en lui la notion

2. http://fr.unesco.org/

de patrimoine industriel dont l'émergence s'est accélérée à partir des années 1970, même si, auparavant, des historiens avaient attiré l'attention sur cette notion désormais importante ». 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'imposa dans le discours, se généralisa dans les ouvrages et les articles. La multiplication des friches industrielles sur notre territoire et l'épineuse question de leur devenir contribuèrent à stimuler la réflexion et à susciter les débats. Le Comité international pour la conservation du patrimoine industriel en propose ensuite une définition plus précise (TICCIH, 2003) : « 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)… ».

Le textile est un des champs du patrimoine industriel au même titre que d'autres activités industrielles telles la métallurgie, la chimie, la papeterie… L'historien Laurent Marty (1984), ajoute « le textile a produit des fils, des tissus, des usines, des maisons et des quartiers, mais surtout au centre de tout cela des hommes, avec leur travail, leurs loisirs, leur vie quotidienne ». Secteur industriel majeur en France pendant de nombreuses décades, le domaine du textile implique de nombreux acteurs, qui ont produit énormément de documents numérisés, renfermant des connaissances s'étendant sur plusieurs siècles.

2.2 Méthodes pour l'identification des acteurs et des sources de données du domaine

La projection de données collectées sous forme d'une cartographie est une méthode qui peut s'avérer propice à dégager des connaissances tant sur le réseau d'acteurs et son dynamisme que sur les sources d'informations existantes, leurs échanges et partages. C'est pourquoi nous avons choisi cette technique pour identifier les acteurs du PIT et les liens qu'il y a entre ces acteurs, afin d'identifier l'ensemble des ressources documentaires numériques disponibles.

Pour ce faire, les méthodes de collecte d'information à cartographier existantes se répartissent en deux catégories. Une première famille, celle des méthodes qualitatives, permet de réaliser une étude fine et précise. Parmi les méthodes existantes, nous pouvons notamment citer les questionnaires, les entretiens, ou encore les observations *in situ*. Bien qu'intéressantes, ces méthodes ne garantissent pas d'identifier l'ensemble des acteurs du domaine sur le territoire étudié. Une seconde famille de méthodes, dites quantitatives, s'appuie sur des techniques informatiques

afin d'étendre l'analyse, d'identifier et de cartographier des sources de données de façon (semi-)automatique. Parmi ces méthodes, la scientométrie et la cartographie du Web sont celles qui sont les plus utilisées. La scientométrie (Godin, 2005) consiste à analyser les citations bibliographiques dans les publications scientifiques propres à un domaine, un laboratoire, un chercheur… pour étudier comment les publications font référence les unes aux autres, ce qui permet d'évaluer et de visualiser le dynamisme et l'influence de chaque acteur scientifique, en fonction par exemple du nombre de fois où il est cité par ses confrères. La **cartographie du Web** montre la présence de chaque acteur ou catégorie d'acteurs au sein du réseau Internet constitué des sites d'acteurs et d'hyperliens entre eux 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) [3]. Les *crawlers* peuvent être semi-automatiques (*issueCrawler*[4], *Hyphe*[5], etc.) ou, plus rarement, manuels (*Navicrawler*[6], 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 *crawler* semi-automatique 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 (Barabasi *et al.*, 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. Dans le cadre de notre étude qui vise à identifier les acteurs du PIT et de leurs relations, nous souhaitons définir une cartographie du Web qui décrit le découpage du réseau Internet français sur la thématique de l'industrie textile en prenant comme territoire d'étude initial la région NPDC. Nous proposons une méthode hybride combinant une approche qualitative sous forme d'entretiens semi-directifs permettant de dresser un premier réseau précis des acteurs (et de leurs données numériques disponibles) du territoire d'étude, à une application quali-quantitative de la cartographie du web. En effet, dans le cas d'une étude d'un domaine tel que celui du PIT sur une échelle géographique limitée, la combinaison de deux types de *crawlers* peut s'avérer 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 (via un algorithme du type inter-acteur).

En ce qui concerne la représentation, les graphes se sont imposés comme la forme de visualisation classique pour représenter les liens entre acteurs. La topologie des

3. Logiciel qui permet de naviguer dans un ensemble de pages web et de tracer tous leurs liens hypertextes.

4. https://www.issuecrawler.net/

5. http://hyphe.medialab.sciences-po.fr

6. httpÂă://webatlas.fr/wp/navicrawler/

réseaux construits en s'appuyant sur un *crawler* peut ensuite être visualisée à l'aide de logiciels de manipulation de graphes tel que *Gephi*[7]. Le principe à la source de ce type de visualisation est que chaque site web constitue un nœud du graphe et que chaque lien hypertexte depuis ce site vers un autre est un arc depuis un nœud vers un autre sur le graphe.

2.3 Formalismes du Web sémantique

L'émergence du **Web sémantique** depuis une quinzaine d'années a mis en évidence divers langages informatiques pour 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 n'est pas capable à ce jour – et ne sera sans doute jamais capable – de 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 PIT, il existe cependant plusieurs exemples de **formalismes** créés pour décrire les objets culturels. Les principaux formalismes sont des modèles complexes qui permettent de 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 Boeuf, 2013), CIDOC CRM (Doerr, 2003), et FRBRoo (Doerr *et al.*, 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 – du plus matériel au plus conceptuel –, le renseignement des champs informationnels n'est pas forcément opéré par une explicitation locale, mais autant que faire se peut par une référence au modèle FRAD (pour les personnes physiques et morales) ou au modèle FRSAD (pour les lieux, événements, concepts et objets). Un dense réseau 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 de patrimoine culturel physique ou non, selon la définition proposée par l'UNESCO.

7. http://gephi.github.io/

Il vise à dépasser les incompatibilités sémantiques et structurales des nombreuses sources d'informations hétérogènes portant sur des réalités patrimoniales et culturelles pour faciliter 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 des 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. Reprenant les quatre niveaux de description de FRBR, il fait des entités originelles des conteneurs chargés d'intégrer les classes CIDOC CRM pour assurer l'interdépendance entre les richesses des deux modèles. 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. Du fait de son niveau élevé de maturité et de sa stabilité, de son adéquation avec les données du projet ainsi qu'avec son objectif d'agrégation de données hétérogènes, nous avons choisi de mettre en œuvre l'ontologie CRM CIDOC. Bien entendu, les outils déjà proposés, de même que son interopérabilité planifiée avec son évolution que constitue FRBRoo, nous ont également guidés dans ce choix.

3 Contributions scientifiques

3.1 Définition du patrimoine industriel textile

Dans un objectif de valorisation de l'ensemble de ces données, le projet TECTONIQ a pour objectif d'identifier, cartographier, mutualiser les données stockées dans différents formats pour les rendre interopérables. Pour ce faire, et sur la base de l'ensemble des définitions présentées ci-dessus, nous proposons tout d'abord une première explicitation de ce que nous entendons par patrimoine industriel textile afin de préciser les éléments caractéristiques du domaine que nous cherchons à identifier dans les documents textuels :

— les biens matériels : bâtiments, machines, équipements, ateliers, usines, sites de traitements et de raffinage, magasins, centres de productions ainsi que des activités sociales en rapport avec l'industrie textile (habitations, lieux de culte ou d'éducation) ;
— les biens immatériels : souvenirs, événements, fêtes, image collective, production intellectuelle transmise par le savoir faire qui peut être une succession de gestes dictés et montrés dans les centres de production.

Nous nous intéressons à la fois au patrimoine industriel patrimonialisé et au domaine de la filière textile d'aujourd'hui qui constitue le patrimoine vivant.

3.2 Notre méthode pour l'identification des acteurs et des sources de données du domaine

Dans l'objectif d'identifier les acteurs du patrimoine produisant et/ou possédant des données numériques sur le thème du PIT tel que défini section 3.1, nous proposons une méthodologie semi-automatique composée de trois étapes : (1) identification des principaux acteurs du patrimoine sur le territoire NPDC via des entretiens semi-directifs, (2) identification du réseau numérique d'acteurs du PIT à travers la cartographie quali-quantitative du Web via *Navicrawler* et (3) *Hyphe*). L'analyse du réseau obtenu est réalisée en nous appuyant d'abord sur les cartographies obtenues via l'outil *Gephi*, complétées ensuite par une analyse spatiale réalisée par un démonstrateur développé en nous appuyant sur *Google Maps*[8].

1. **Identification du noyau d'acteurs :** sur la base d'une première veille sur le Web, nous avons identifié 60 acteurs du patrimoine. Nous avons réalisé des entretiens semi-directifs auprès de 9 de ces acteurs présents dans le NPDC et disposant de données expertes sur ce patrimoine, parmi lesquels nous pouvons citer l'Inventaire de la Région, Musée d'art et d'industrie André Diligent (La Piscine), le service de l'urbanisme de la MEL, ou encore l'association PROSCITEC[9]. L'ensemble des informations collectées comprend à la fois des informations sur les acteurs eux-mêmes (coordonnées, statut, type de patrimoine à disposition, etc.), les données dont ils disposent (quantité, format, etc.), et les éventuels échanges/collaborations avec d'autres autres acteurs du domaine. Ces entretiens nous ont permis d'identifier un premier noyau de 118 acteurs.
2. **Construction du réseau d'acteurs via *Navicrawler*:** pour approfondir ce premier listing d'acteurs du domaine, nous avons utilisé le *crawler Navicrawler* qui offre la possibilité de valider uniquement les sites que nous identifions comme liés au domaine d'étude. En suivant cette procédure, nous avons défini un premier réseau de 160 acteurs.
3. **Enrichissement automatique du réseau d'acteurs via *Hyphe* :** *Hyphe* utilise un script qui suit et répertorie tous les liens d'un site puis tous les liens des sites qu'il rencontre et ainsi de suite selon une variable profondeur renseignée par l'utilisateur. Sur la base des 160 sites web donnés en entrée, *Hyphe* a fait ressortir une multitude de nouveaux sites web en précisant à chaque fois le nombre de liens entrants et sortants, respectivement pour chaque site les hyperliens pointant vers celui-ci et les hyperliens présents sur ce site et pointant vers d'autres. Un dépouillement des résultats nous a permis d'identifier 9 acteurs supplémentaires, que nous avons intégrés au réseau thématique constitué donc de 169 acteurs.

Nous avons ensuite organisé en catégories les sites du corpus selon les critères suivants : type de patrimoine (patrimonialisé ou vivant), rayonnement géographique, statuts et localisation. Enfin, nous avons analysé la distribution de ces catégories à

8. https://developers.google.com/maps/documentation/javascript/
9. http://www.proscitec.asso.fr/

travers des représentations sous forme de graphe par *Gephi*. Nous en présentons un premier bilan succinct ici. Tout d'abord, nous avons réalisé une première cartographie des 169 acteurs en les classant par type (voir Figure 1).

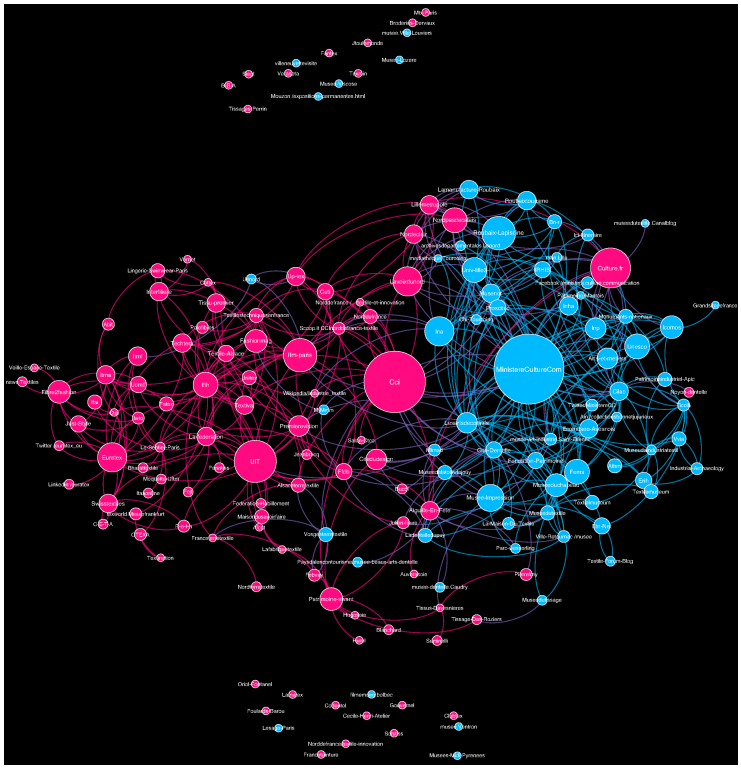

FIGURE 1 – Cartographie du Web du réseau d'acteurs du patrimoine de l'industrie textile organisé par type (patrimoine patrimonialisé en rouge et patrimoine vivant en bleu). Graphe réalisé avec *Gephi*. Algorithme *force-atlas 2*.

Les acteurs du patrimoine patrimonialisé (en bleu) représentent 40,83% du corpus et les acteurs du patrimoine vivant 59,17% (en rouge). La taille des nœuds est proportionnelle à l'*in-degree* (degré entrant), soit le nombre de liens pointant vers un site web. L'*in-degree* peut donc être considéré comme un indice de l'autorité sur le Web : plus un nœud est cité, plus les autres nœuds reconnaissent son intérêt et son importance (Severo et Venturini, 2015). Seuls quelques acteurs font le « pont » entre les deux ensembles, parmi lesquels nous pouvons citer en acteurs du patrimoine vivant la Chambre de Commerce et d'Industrie (CCI, 25 liens entrants), L'Union des Industries Textiles (UIT, 16 liens entrants), la Voix du Nord (12 liens entrants), le portail de la Région NPDC (10 liens entrants), le Ministère de la Culture et de la Communication (29 liens) ou encore le musée de la Piscine de Roubaix (13 liens) pour le patrimoine patrimonialisé. Ces six acteurs représentent près de 20% du total des liens. Certains acteurs sont isolés et ne sont donc cités par aucune autre acteur du

corpus. Les acteurs de la filière textile isolés sont des sites vitrines à vocation commerciale uniquement, et ceux du patrimoine patrimonialisé sont pour la plupart des sites de type blog créés non pas par des institutions mais par des particuliers avec un rayonnement très local (sur une commune). À noter que parmi les 169 acteurs étudiés, 29% ont un rayonnement international, 30,7% un rayonnement national et 40,3% un rayonnement régional.

La Figure 2 met en avant les acteurs présents sur le territoire français. Bien que le point de départ de l'étude soit la région NPDC avec 37,76% des acteurs du réseau, deux autres pôles régionaux ressortent : Rhône-Alpes (23,47%) et l'Alsace-Lorraine (9,18%). Cela s'explique par l'identité industrielle textile forte dans le passé de ces régions. L'Île-de-France ressort également, du fait de la présence de nombreux sièges sociaux d'entreprises et d'institutions. Les territoires du NPDC et de Rhône-Alpes ont un nombre quasiment similaire d'acteurs du patrimoine patrimonialisé que d'acteurs du patrimoine vivant alors que l'Alsace-Lorraine n'est représenté que par des acteurs du patrimoine vivant.

FIGURE 2 – Localisation des acteurs présents sur le territoire français (Métropole).

Enfin, nous focalisons notre étude sur le territoire de la région NPDC représenté par 41 acteurs (voir Figure 3). Les icônes représentent le type de structure de l'acteur à savoir si il s'agit d'une institution (organisme qui dépend de l'état ou d'une collectivité publique), une entreprise ou une association. Cette dernière catégorie comprend toutes les organisations professionnelles telles que les fédérations, syndicats, groupement, Union nationale professionnelle et pôles de compétitivité.

Les structures institutionnelles sont considérées comme disposant de patrimoine patrimonialisé. Dans la région NPDC, ces acteurs représentent plus de la moitié des 41 acteurs au total soit 53,6%. De plus la majorité de ces acteurs est concentrée sur la Métropole Européenne de Lille (MEL). Il s'agit pour la majorité de musées comme

FIGURE 3 – Localisation des acteurs présents dans le Nord – Pas-de-Calais

la Manufacture des Flandres et le musée de la Piscine à Roubaix, la Cité de la dentelle à Calais mais aussi de bibliothèques telles que la bibliothèque numérique de Roubaix, le service commun de documentation de l'université de Lille 3, d'archives comme les Archives Nationales du Monde du Travail à Roubaix, les Archives Départementales du Nord ou encore du laboratoire de recherche IRHIS de Lille 3, d'établissements éducatifs comme l'université de Lille 3, l'École Nationale des Arts et Industries Textiles de Roubaix ou bien de structures municipales ou régionales comme la direction de l'urbanisme de la MEL, la Direction Régionale des Affaires Culturelles ou le service de l'inventaire de la région NPDC, la Chambre de Commerce et d'industrie Nord de France, etc.

Les associations et entreprises forment l'ensemble d'acteurs de type patrimoine vivant, c'est-à-dire des acteurs de l'industrie textile d'aujourd'hui. Dans le NPDC, elles représentent 46,3% des acteurs. Parmi les associations, nous pouvons notamment citer des organisations professionnelles comme l'Union des Industries Textile du Nord, CLUBTEX (une association spécialisée pour la promotion des entreprises du textile technique) mais également des associations à vocation culturelle comme PROSCITEC ou encore MUSENOR (regroupe les musées de la région NPDC), et des pôles de compétitivité tel que Up-tex. Concernant les entreprises, celles qui figurent sur la carte proviennent pour la quasi totalité de la liste des entreprises labellisées « entreprise du patrimoine vivant » par l'État comme les entreprises Codentel, Jean Bracq ou Noyon Dentelle. Les autres sont des entreprises ayant pour activité majeure de fournir du contenu informationnel comme les journaux *La Voix du Nord* et *Nord Eclair*.

À partir de cette liste d'acteurs produisant et/ou diffusant du contenu sur le thème du PIT, nous avons collecté un premier corpus constitué de 1600 documents hétérogènes (images avec notices descriptives XML du laboratoire IRHIS, articles de presse en XML de la Voix du Nord, documents du SCD de Lille 3, notices et Plan lo-

cal d'Urbanisme disponibles sur le portail de la MEL, Notices de l'Inventaire de la Région). Nous travaillons actuellement à l'extraction automatique de descripteurs patrimoniaux à partir de méthodes mises en place par les différents partenaires du projet (Tahrat *et al.*, 2012 ; Kergosien *et al.*, 2011 ; 2014). L'objectif est de mettre en relation de façon semi-automatisée des données hétérogènes en nous appuyant sur la norme CIDOC CRM formalisée en OWL.

3.3 Interopérabilité des données : une première preuve de concept

Pour mener cette première expérimentation, nous avons choisi de traiter quelques uns parmi les documents de ce corpus. Ce test consiste à peupler une ontologie CIDOC CRM [10] à partir des informations identifiées dans ce corpus restreint. Nous avons choisi arbitrairement deux points d'entrée dans notre univers documentaire : un accès thématique patrimonial, avec le bâti du domaine textile, et un accès géographique, avec l'entrée « Roubaix ». Notre objectif étant de traiter sémantiquement l'information issue de sources hétérogènes, nous avons conservé des documents hétérogènes, issus de sources distinctes. 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 photos* du laboratoire de recherche IRHIS), 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.).

Le travail d'instanciation de l'ontologie CIDOC CRM a été effectué manuellement grâce au logiciel Protégé [11] (Musen *et al.*, 1995). L'ensemble des informations pertinentes collectée dans le corpus de test à 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 4 est une projection de l'ontologie peuplée par notre corpus de test. Comme nous l'avions prévu, la structure informationnelle de l'ontologie est bien adaptée à la description des objets de patrimoine que nous cherchons à mettre en évidence. On notera également et surtout que les quatre documents-tests de notre expérimentation 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 pré-

10. Proposition d'encodage du modèle conceptuel de référence CIDOC en OWL 2 par Simon Reinhardt, basée sur la version 5.0.1 de mai 2009 (http://www.cidoc-crm.org/official_release_cidoc.html).
11. Version Desktop 5.0 beta (http://protege.stanford.edu/).

sident de la République de l'époque à l'exposition internationale du Textile en 1911, et un second document MEL_Roubaix_AVA.pdf précise que l'événement a eu lieu le long du Parc barbieux à Roubaix, commune du nord de la France.

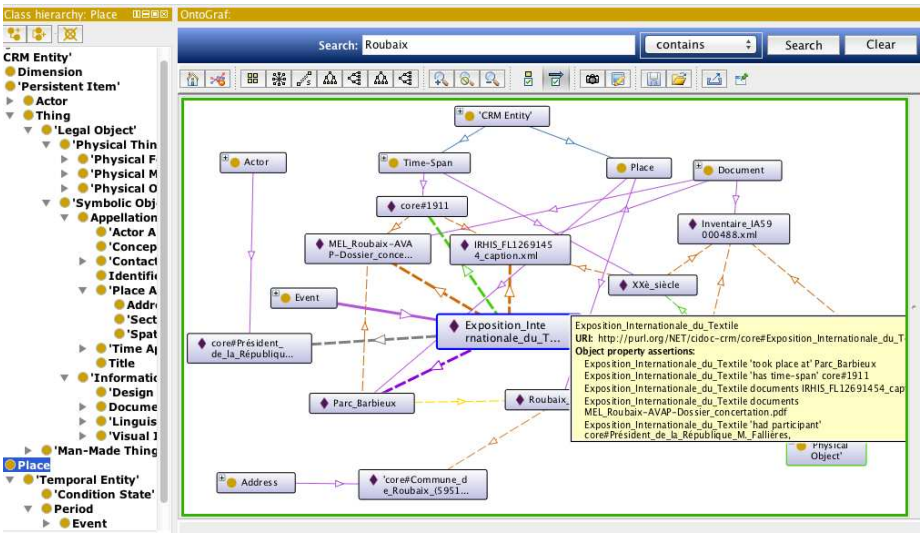

FIGURE 4 – Extrait de l'ontologie produite sur la base de quatre documents.

Au-delà de ce test nécessairement réduit, nous pouvons envisager raisonnablement une démarche de peuplement automatique, ou au moins semi-automatique, à partir d'informations identifiées dans la totalité des documents hétérogènes que nous avons à notre disposition. Cette identification de l'information sera assurée tantôt par la structure interne des documents décrivant le patrimoine industriel, tantôt par les stratégies d'identification d'entités nommées (Tahrat *et al.*, 2012 ; Kergosien *et al.*, 2014), exploitations de lexiques et thésaurus, démarche d'extraction d'information que nous comptons appliquer (Kergosien *et al.*, 2011). Un simple *mapping* entre l'information identifiée dans les documents et le modèle informationnel permettra non seulement d'enrichir cette information des propriétés associées au modèle et de la rendre accessible, mais également de mettre à profit les inférences qui en seront issues pour une valorisation plus efficace de ces masses d'informations patrimoniales.

4 Conclusion

Cet article présente les premiers résultats du projet TECTONIQ pour l'interopérabilité de contenus numériques sur le thème du patrimoine industriel textile (PIT) à partir de sources de données hétérogènes. Nous cadrons tout d'abord le sujet en délimitant le champ thématique qu'est le PIT sur la base des définitions de l'UNESCO et du groupe TICCIH. Nous proposons ensuite une méthode hybride combinant

des entretiens semi-directifs et une cartographie quali-quantitative du Web semi-automatique en nous appuyant sur les logiciels *Navicrawler* et *Hyphe* afin d'identifier les acteurs produisant et/ou diffusant de l'information sur le domaine étudié. Nous présentons enfin une ébauche d'ontologie construite manuellement au format OWL CIDOC CRM à partir d'un extrait réduit du jeu de données collecté, et permettant de mettre en relation des documents numériques hétérogènes en nous appuyant sur leur contenu.

En perspective à ces travaux, nous souhaitons automatiser la phase de construction d'une base de connaissances OWL CIDOC CRM en nous appuyant sur les listes de descripteurs patrimoniaux Lieux, événements, thématiques, entités temporelles et acteurs) produites par les chaînes de traitement des partenaires du projet TECTONIQ sur le corpus de 1600 documents.

Remerciements

Cet article a été rédigé dans le cadre du projet TECTONIQ (PEPS CNRS - InterMSH) hébergé à Maison européenne des sciences de l'homme et de la société (MESHS - USR 3185). Nous remercions nos partenaires institutionnels et notamment la MEL, l'Inventaire de la région NPDC, le laboratoire IRHIS et le SCD Lille 3. Ces travaux sont menés en utilisant le logiciel Protégé (*National Institute of General Medical Sciences of the United States National Institutes of Health*).

5 Références

Barabási, A.-L., Albert, R., Hawoong, J. (2000). Scale-free characteristics of random networks : The topology of the world-wide web. *Physica A*, vol. 281, num. 1, 69-77.

Severo, M. (2012). *La cartographie du Web : le lien social sur le Net.* Document de travail, GIS CIST – Collège International des Sciences du Territoire, disponible à https://halshs.archives-ouvertes.fr/halshs-00678768.

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

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

Doerr, M., Le Boeuf, P. et Bekiari, C. (2008). FRBRoo, a conceptual model for performing arts. In *Conference proceedings « The Digital Curation of Cultural Heritage ». ICOM-CIDOC Annual Meeting*, Athens, 618.

Godin B. (2005). *La science sous observation. Cent ans de mesure sur les scientifiques 1906-2006*, Laval, Presses de l'Université Laval, 81 p.

Kergosien E., Bessagnet M.-N. et Gaio M. (2011), Exploitation d'une cartographie sémantique à des fins de validation : application à l'indexation experte de corpus documentaires. *Documentation et Bibliothèques*, vol. 57, 19-28.

Kergosien E., Laval B., Roche M., Teisseire M. (2014). Are opinions expressed in land-use planning documents ? *International Journal of Geographical Information Science*, vol. 28, num. 4, 739-762.

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*, vol. 34, num. 1-2, 85-95.

Severo M., Venturini T. (2015, à paraître). Intangible cultural heritage webs : Comparing national networks with digital methods. *New Media & Society*.

Tahrat S., Kergosien E., Bringay S., Roche M., and Teisseire M. (2013). Text2Geo : from textual data to geospatial information. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, WIMS'13*, Madrid, 23 :1-23 :4.

UNESCO (2008). *Orientations devant guider la mise en œuvre de la Convention du patrimoine mondial*, Centre du patrimoine mondial de l'UNESCO, disponible à : http://whc.unesco.org/archive/opguide08-fr.