Département d'histoire
Faculté des lettres et sciences humaines
Université de Sherbrooke

Vers une nouvelle architecture de l'information historique : L'impact du Web sémantique sur l'organisation du *Répertoire du patrimoine culturel du Québec*

par Philippe Michon
Mémoire présenté pour obtenir
la Maîtrise ès arts (Histoire)

Université de Sherbrooke
Avril 2016

RÉSUMÉ

Le *Plan culturel numérique du Québec* (PCNQ) souligne l'importance pour le domaine culturel québécois, auquel participe étroitement les historiens, de s'intéresser aux possibilités du Web sémantique. Dans cette idée, ce mémoire étudie les avantages et les inconvénients de l'association entre le Web sémantique et l'histoire. D'un côté, on retrouve une nouvelle configuration du Web sous forme de données liées qui tente de s'inscrire dans un cadre pratique et, de l'autre, une discipline qui souhaite comprendre et préserver les faits passés. La réunion des deux concepts nécessite une implication interdisciplinaire entre programmeurs, professionnels en sciences de l'information et historiens. Face à ce travail interdisciplinaire, quels sont les enjeux et le rôle de l'historien dans le développement d'une plate-forme sémantique sur le patrimoine québécois?

Pour répondre à cette question, ce mémoire explique les liens étroits qui existent entre la discipline historique et les données liées. Après avoir défini un ensemble de concepts fondateurs tels que le *Resource Description Framework* (RDF), l'*Uniform Resource Identifier* (URI), les fichiers d'autorité et les ontologies, ce mémoire associe un corpus de personnes du *Répertoire du patrimoine culturel du Québec* (RPCQ) avec *DBpedia*, un joueur majeur du Web sémantique. Cette démonstration explique comment le patrimoine québécois s'articule dans le nuage des données liées. De cette expérimentation découle deux constats qui démontrent l'importance de l'implication historienne dans une structure sémantique. Le Québec n'a pas d'autorité sur ses propres données et on ne retrace actuellement que la grande histoire du Québec sans entrer dans ses particularités.

Mots-clés : Web sémantique, Données liées, *Resource Description Framework* (RDF), *Uniform Resource Identifier* (URI), fichiers d'autorité, ontologies, patrimoine québécois, *Plan culturel numérique du Québec* (PCNQ)

REMERCIEMENTS

La rédaction d'un mémoire sur le Web sémantique et l'histoire rappelle constamment l'importance de lier les bonnes personnes avec les bons événements. À ce titre, je me dois de remercier un réseau de personnes qui ont permis la réalisation de ce document.

Merci à Léon Robichaud, mon directeur et plus grand allié, qui a cru en mon projet et a accepté de m'accompagner dans ce périple qui débuta avec des agents intelligents pour se terminer dans une mer de triplets;

Merci à Harold Bérubé, lecteur aguerri, qui a toujours veillé à ce que je reste un historien avant tout;

Merci à Mathieu Rocheleau, celui à qui je dois tout, qui fut le premier à croire en moi il y a déjà 3 ans et qui m'offre encore aujourd'hui la chance d'œuvrer dans un domaine passionnant;

Merci à Philippe Dubé, ami inestimable, qui m'a embarqué dans de magnifiques projets et réflexions qui furent le premier canevas de ce mémoire;

Merci au *Ministère de la Culture et des Communications du Québec* qui m'a offert un lot impressionnant de données qui ont permis à ce mémoire d'exister;

Merci à Thomas Francart, Madeleine Lafaille, Guy Lapalme, Michel Gagnon et Hugues Boily, tous des piliers, qui contribuèrent à peaufiner ma réflexion et ma compréhension de mon sujet de recherche;

Merci à Joanne Burgess, directrice du *Laboratoire d'histoire et de patrimoine de Montréal*, d'avoir osé plonger dans le monde des données liées et de m'avoir permis d'assister au LODLAM2015, qui fut une expérience plus qu'enrichissante;

Merci à Tim Sherratt, historien australien, pour son aura et pour m'avoir fait comprendre que les historiens ont le droit de s'intéresser au Web sémantique;

Merci à Dhyana Robert et Rose Pelletier Lewis, la L-Team, pour votre humour particulier qui a rendu cette maîtrise plus qu'agréable;

Merci à Olivier Ariel, ami et colocataire, qui était toujours partant pour une activité hors mémoire et qui a toujours bien voulu m'écouter raconter mes histoires d'historien;

Merci à Francine Campbell et Sylvain Michon, mes parents, qui, malgré leur difficulté à cerner pleinement mon sujet de mémoire, ont toujours été des leviers me permettant d'accomplir mes ambitions;

Finalement un merci tout spécial à Sarah Bellefleur, mon amour que j'aime, qui, en plus d'avoir corrigé chaque ligne de ce mémoire et alimenté de manière soutenue ma réflexion, me prouve de jour en jour que notre futur sera merveilleux.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABBREVIATIONS

AAT: Art & Architecture Thesaurus
BAC: Bibliothèque et Archives Canada
BAnQ: Bibliothèque et Archives nationales du Québec
BMO: British Museum Ontology
CED: Corps expéditionnaire canadien
CHIA: Collaborative for Historical Information and Analysis
CHO: Cultural Heritage Object
CIDOC CRM: Conceptual Reference Model du Comité international pour la documentation
CIP: Calcul informatique de pointe
CNP: Classification nationale des professions
CONA: Cultural Objects Name Authority
DBC: Dictionnaire biographique du Canada
DHP: Data Hoover Project
EDM: Europeana Data Model
FoaF: Friend of a Friend
FRBR: Functional Requirements for Bibliographic
ICOM: International Council of Museums
LAM: Libraries, Archives and Museums
LCSH: Library of Congress Subject Headings
LOD: Linked Open Data
LODLAM: Linked Open Data in Libraries, Archives and Museums
MBAM: Musée des Beaux-Arts de Montréal
MCC: Ministère de la Culture et des Communications
NER: Named-Entity Recognition
OCDE: Organisation de coopération et de développement économiques
OCLC: Online Computer Library Center
ODC-PDDL: Open Data Commons Public Domain Dedication and License
OeCR: Oxford e-Research Centre
OWL: Web Ontology Language
PC: Ordinateur personnel
PCNQ: Plan culturel numérique du Québec
PURL: Persistent URL
RCIP: Réseau canadien d'information sur le patrimoine
RDA: Ressource: Description et Accès
RDF: Resource Description Framework
RDFS: RDF Schema
ROC: Reconnaissance optique de caractères
RPCPD: Réseau pancanadien du patrimoine documentaire
RPCQ: Répertoire du patrimoine culturel du Québec
SIG: Système d'information géographique
SKOS: Simple Knowledge Organization System
SMQ: Société des musées québécois
SSSO: Simple Service Status Ontology

SyMoGIH: Système modulaire de gestion de l'information historique
TGIR: Très grande infrastructure de recherche
TGM: Thesaurus for Graphic Materials
TGN: Thesaurus of Geographic Names
ULAN: Union List of Artist Names
URI: Uniform Resource Identifier
W3C: World Wide Web Consortium

INTRODUCTION

Les nouveaux rôles de l'historien 3.0

L'évolution des technologies de l'information a considérablement modifié la méthodologie de l'historien au cours des cinquante dernières années. De celles-ci, Internet et le Web sont celles qui ont eu un impact majeur sur l'accès à l'information et la diffusion de contenus. Outils facilitateurs, ils ont rapidement été adoptés. Aujourd'hui, il devient difficile de concevoir une recherche historique sans utiliser les outils du Web. Le travail de classification du bibliothécaire professionnel, de l'archiviste et du muséologue, jumelé aux possibilités du Web, ouvre un grand terrain exploratoire pour l'historien. Ce dernier devient alors un utilisateur de ces différents systèmes de bases de données institutionnelles.

L'historien a développé, avec le temps, une méthodologie de travail pour repérer, le plus efficacement possible, les sources et études susceptibles de répondre à ses questions de recherche. L'adaptation de l'historien peut être perçue comme un comportement passif face à celui des professionnels en sciences de l'information qui modélisent et normalisent des systèmes organisationnels. Cependant, ce rôle passif semble se transformer tranquillement en un rôle actif. En effet, de plus en plus d'historiens s'intéressent aux différents outils informatiques, mais surtout les critiquent de manière constructive par rapport à leurs propres objectifs disciplinaires. Cette mouvance porte à croire que plusieurs historiens migreront vers des rôles d'architectes de l'information historique.

Cette hypothèse émerge de deux constats qui mèneront vers une réorganisation des systèmes d'informations culturelles. Premièrement, le désir de contextualisation et de fédération à grande échelle est nécessaire pour permettre l'inscription du patrimoine québécois dans un réseau international. Deuxièmement, le Web est actuellement dans une phase de transition entre le Web social (2.0) et le Web sémantique (3.0), un processus qui peut être imperceptible pour l'usager, mais qui transforme les structures informationnelles du Web.

L'intérêt émergent du Web sémantique au Québec

Le Web sémantique n'est pas un concept nouveau. Dès 2001, Tim Berners-Lee, reconnu comme l'inventeur du Web, affirme déjà que le Web de documents tel qu'il existe ne permet pas une gestion efficace de l'information puisque les ordinateurs ne comprennent pas le contenu des pages web. Il faudrait donc mettre en place un Web sémantique qui donne sens à la donnée en désambiguïsant celle-ci pour ainsi permettre un décloisonnement de l'information actuellement coincée dans des sites web[1].

Depuis ce temps, plusieurs projets culturels basés sur cette idée du Web sémantique ont vu le jour. Aujourd'hui, il est indéniable que l'effort fourni par des acteurs tel *Europeana*, cette bibliothèque européenne qui regroupe plus de 1 500 institutions culturelles, n'a pas été vain puisque cette structure fédératrice permet un meilleur partage, mais surtout une meilleure contextualisation de leurs données[2].

[1] Tim Berners-Lee, James Hendler et Ora Lassila, « The Semantic Web », *Scientific American*, 17 mai 2001,http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb .html, consulté le 27 août 2015.
[2] Europeana, *Europeana - About us*, http://www.europeana.eu/portal/aboutus.html, consulté le 27 août 2015.

Les institutions patrimoniales québécoises, quant à elles, n'ont pas encore investi le Web sémantique. Par contre, conscient des changements qu'entraîne le numérique dans le traitement de l'information culturelle, le gouvernement du Québec amorce en 2010 une vaste consultation auprès de ses principaux partenaires pour déterminer les actions qui faciliteront ce virage numérique. De cette démarche découle le *Plan culturel numérique du Québec*. On annonce pour ce dernier, le 4 juin 2014, un budget de 110 M$ sur 7 ans. Ce plan se divise en trois axes : *Créer des contenus culturels numériques*, *Innover pour s'adapter à la culture numérique* et *Diffuser des contenus culturels numériques afin d'assurer leur accessibilité*[3]. Ces énoncés se concrétisent au sein de 50 mesures dont l'une qui s'intéresse à l'utilisation du Web sémantique pour traiter et diffuser le patrimoine, intitulée *Aider le réseau de la culture à s'approprier les technologies du Web sémantique afin de maximiser la présence des données culturelles québécoises dans le Web*[4]. Le *Répertoire du patrimoine culturel du Québec* (RPCQ) sera l'une des bases de données qui sera utilisée à titre d'exemple pour alimenter la réflexion autour de cette mesure. En proposant un modèle s'inspirant du projet du *Ministère de la Culture et des Communications* (MCC), ce mémoire s'intéresse au rôle que devrait jouer l'historien au sein de cette nouvelle structure des données culturelles numériques québécoises.

---

[3] Ministère de la Culture et des Communications du Québec, *À propos : Plan culturel numérique*, http://culturenumerique.mcc.gouv.qc.ca/a-propos/, 2015, consulté le 26 mars 2015.

[4] Ministère de la Culture et des Communications du Québec, *06 – Aider le réseau de la culture à s'approprier les technologies du Web sémantique afin de maximiser la présence des données culturelles québécoises dans le Web : Plan culturel numérique*, http://culturenumerique.mcc.gouv.qc.ca/aider-le-reseau-de-la-culture-a-sapproprier-les-technologies-du-web-semantique-afin-de-maximiser-la-presence-des-donnees-culturelles-quebecoises-dans-le-web-banq/, 2015, consulté le 26 mars 2015.

Le Web sémantique: prolongement de la méthodologie historienne

Aux premiers abords, il peut sembler que le maillage entre l'historien et le Web sémantique requiert une rupture avec la méthodologie classique. Le même constat peut être fait face à l'émergence des *humanités numériques* de manière large. Par contre, l'historiographie démontre l'effet contraire puisque certains historiens, dès l'arrivée de l'ordinateur, se sont intéressés à ces possibilités de traitement pour la discipline. L'histoire quantitative des années 1960 s'intéressait particulièrement aux calculs statistiques plus poussés qui devenaient alors possibles. C'est le cas de Robert Fogel et Stanley Engerman, économistes spécialisés en histoire économique qui publient en 1974 *Time on the cross : The Economics of American Negro Slavery*[5]. À partir d'une analyse statistique, ils ont conclu que le système esclavagiste est une institution économiquement viable, contredisant l'hypothèse voulant qu'il s'agisse plutôt d'un système anachronique auquel les Sudistes s'accrochaient pour des raisons socio-culturelles. Leur approche purement quantitative les mène aussi à conclure que le système profite même aux Afro-américains, ce qui leur vaudra de sévères critiques, pour ne pas avoir tenu compte des autres impacts de ce système sur les esclaves. Malgré ses imperfections, cette approche a démontré que l'ordinateur permet une analyse de masse qu'on ne pouvait imaginer quelques années auparavant.

D'abord limitée aux grands projets ayant un accès aux ordinateurs centraux, la création de bases de données se démocratise avec l'arrivée des ordinateurs personnels (PC) au début des années 1980. En plus de cette démocratisation, l'ordinateur devient

[5] Robert Fogel et Stanley Engerman, *Time on the Cross: The Economics of American Negro Slavery*, Boston, Little Brown, 1974, 336 p.

plus performant en permettant, comme le souligne Lou Burnard, de faire du traitement de texte, une caractéristique indispensable pour le travail historique[6]. L'autre volet fondamental de la recherche historique, la gestion de la recherche, est facilité par des bases de données personnelles comme *FileMaker*, logiciel qui apparaît pratiquement au même moment que le PC[7].

Une multitude d'outils commerciaux se retrouvent donc sur le marché, ce qui amène l'historien quantitatif Manfred Thaller à vouloir développer le rôle de l'historien dans la création de programmes adaptés aux objectifs de la discipline historique. Selon Thaller, la complexité des liaisons entre les unités d'information de nature historique nécessite une approche informatique qui lui est propre et Thaller n'hésite pas à confronter les historiens qui ne font qu'utiliser les programmes disponibles sur le marché[8]. Il a toujours maintenu cette position puisqu'il écrit en 2012 : « The Digital Humanities may have to take a much stronger part in the development, not only the reception, of technology[9]. »

L'appel de Thaller eut peu de résonance dans la communauté historienne puisque la démocratisation rapide des outils numériques additionnée aux possibilités du Web

---

[6] Lou Burnard, « The Historian and the Database » dans Evan Mawdsley (ed.), *History and computing III: historians, computers, and data: applications in research and teaching*, Manchester, Manchester University Press : Distributed exclusively in the United States and Canada by St. Martin's Press, 1990, p. 6.

[7] Glenn Koening, *FileMaker Early History*, http://www.dancing-data.com/filemakerhist.html, 2004, consulté le 30 octobre 2014.

[8] Matthew Woollard et Peter Denley, *A Tutorial for Kleio*, St. Katharinen, Max-Planck-Institut für Geschichte, http://www.hki.unikoeln.de/kleio/old.website/tutorial/intro.htm, 1993, consulté le 1er avril 2014.

[9] Manfred Thaller, « Controversies around the Digital Humanities: An Agenda », *Kontroversen um die Digitalen Geisteswissenschaften: Ein Arbeitsplan*, septembre 2012, vol. 37, no 3, p. 7.

mène les historiens à être plus passifs devant ceux-ci. Les programmeurs fournissent maintenant des plates-formes variées pour la gestion de contenus. Par exemple, *Wordpress* évite à ses utilisateurs de devoir programmer pour mettre en ligne des articles de blogues. Cette facilité d'accès et la passivité des chercheurs en sciences historiques face aux développements technologiques depuis l'apparition du Web apporte aujourd'hui son lot de problèmes au niveau des bases de données qui, rappelons-le, soutiennent tout le système informationnel numérique. La majorité de ces problèmes s'expliquent par un manque d'uniformité. Chaque particulier peut structurer son information sur un site web, ce qui entraîne la naissance de plusieurs systèmes qui fonctionnent en vase clos. Le Web sémantique tente de relier l'ensemble de ces données actuellement camouflées au sein d'interfaces diverses. L'historien obtient donc une nouvelle chance de s'intéresser à la structure afin de créer un système de données adapté à ses objectifs.

Le rôle de l'historien au sein de l'interdisciplinarité

Le Web sémantique est un concept difficile à cerner au premier abord. On peut rapidement se demander si la présence de l'historien est pertinente dans l'élaboration d'une plate-forme sémantique qui semble ne pouvoir être créée que par des programmeurs d'expérience. L'un des deux objectifs de ce mémoire est de démontrer que l'historien peut s'insérer dans le processus de développement autour de certains axes particuliers. Nous appuierons notre démonstration par l'exploration d'un terrain de recherche, soit le RPCQ. Ce sera alors l'occasion de s'intéresser aux éventuels possibilités et problèmes auxquels pourrait faire face le ministère lors du passage de leurs données classiques vers des données liées.

Le Web est souvent perçu comme un outil permettant la collaboration de diverses disciplines et le Web sémantique n'échappe pas à cette idée. Il est donc tout à fait possible que l'historien puisse contribuer à l'avancement de la réflexion autour des questions du Web sémantique. De plus, la complexité de sa discipline exige la conception d'un système extrêmement réfléchi et organisé. Si les bibliothécaires professionnels, les archivistes et les muséologues participent activement à l'amélioration des concepts entourant le Web sémantique, on peut prétendre que l'historien pourra, tout autant, bénéficier d'une participation à ce processus grâce à la semi-autonomisation de certaines étapes de recherche et de gestion de l'information. Du même coup, une approche interdisciplinaire oblige un fort maillage entre les bibliothèques, les centres d'archives et les musées (en anglais, les LAM pour *Libraries, Archives and Museums*). Pour ce qui est du renouvellement du RPCQ, on peut penser qu'il prendra la forme d'une interface fédératrice et collaborative puisqu'il s'agit d'un modèle que l'on retrouve abondamment en Europe et en Australie, et qui a fait ses preuves.

Une méta-analyse sur la longue durée

En moyenne 24 678 visiteurs uniques par mois utilisent le RPCQ pour consulter une quantité souvent limitée de fiches[10]. Dans le cadre de ce mémoire, le RPCQ sera analysé dans un ensemble de fiches plus volumineux pour déceler les incohérences impossibles à constater lors d'une consultation ponctuelle. Il faut comprendre que le Web sémantique s'appuie sur des identificateurs uniques de concepts et des propriétés qui permettent d'établir des équivalences ou de classifier les données. Ainsi, il sera possible de prévoir les changements que devra subir le RPCQ avant de prétendre

[10] Isabelle Jacques, *Statistiques de fréquentation du Répertoire du patrimoine culturel du Québec*, Québec, Ministère de la Culture et des Communications du Québec, 2015.

pouvoir s'adapter aux exigences du Web sémantique. Ensuite, l'interrogation directe des données permettra de saisir jusqu'où peut aller l'historien sans l'aide des autres disciplines.

Cette étude s'appuie aussi sur un concept historique qui refait surface grâce à l'ouvrage de Jo Guldi et David Armitage intitulé *The History Manifesto*, soit la longue durée. L'histoire de longue durée est un concept apparu en 1958 sous la plume de Fernand Braudel. L'idée est d'étudier des périodes historiques plus longues et ainsi mieux développer sa pensée critique face aux enjeux actuels. L'histoire de la longue durée prône une plus large contextualisation des concepts dans le temps pour en dégager les transformations. Dans *The History Manifesto*, les auteurs affirment que le numérique ouvre la voie à une très large contextualisation inconcevable au niveau de la pensée humaine. Cette affirmation correspond parfaitement aux possibilités qu'offre le Web sémantique. Étudier ce dernier sous un angle de longue durée change complètement le paradigme de classification actuel qui mise sur la contextualisation par la description et non la contextualisation par les transformations temporelles.

De la documentation technique vers un répertoire culturel québécois

Le RPCQ s'est avéré être la plate-forme la plus appropriée pour étudier le rôle des historiens dans le développement du Web sémantique. En effet, il s'agit d'une base de données extrêmement complète et diversifiée. Il s'agit du seul répertoire de cette envergure à associer différentes composantes du patrimoine culturel québécois, que ce soient les biens matériels, le patrimoine immatériel, les personnages ou les événements. De plus, étant supporté par le MCC, le RPCQ fait figure d'autorité comme base d'un

système consolidé en patrimoine québécois. Une fois que sera instaurée une démarche fonctionnelle applicable à l'ensemble de ses données, le MCC pourra mettre à profit ses infrastructures afin d'aider les institutions culturelles à s'intégrer à ce nouvel écosystème informationnel et ainsi assurer la pérennité de celui-ci. Cette méthode permettra une uniformité essentielle au bon déroulement d'un projet en Web sémantique puisque le MCC fournira aux petites, moyennes et grandes institutions, une plate-forme fonctionnelle et une documentation riche.

Présentement, plusieurs projets fonctionnent selon le principe de la base vers le haut, c'est-à-dire que des institutions migrent vers des systèmes sémantiques et éventuellement tentent d'aligner leurs données avec des partenaires. Cette méthodologie était idéale dans un contexte de défrichage. Face au peu d'information disponible concernant le Web sémantique, il fallait expérimenter à la base. Avec les multiples cas de figures à partir desquels on documente la démarche d'implantation de ce type d'environnements, il est maintenant envisageable de miser sur une approche inversée fédérée par une seule entité qui chapeaute ses partenaires. Il faut qu'une ou des institutions phares amorcent la réflexion afin d'inciter les particuliers à collaborer aux nuages des données liées. D'ailleurs, la documentation technique du *World Wide Web Consortium* (W3C) permet de se familiariser avec différents concepts fondamentaux du Web sémantique, mais les documents les plus utiles sont ceux des projets qui allient culture et Web sémantique. *Europeana*, nommé précédemment, en est un bon exemple, mais on retrouve aussi des projets comme CLAROS, la *British Museum Semantic Web Collection Online*, *Sémanticpédia* et *Au-delà des tranchées*. Tous ces projets ont produit de la documentation expliquant leur méthodologie qui sera fort utile pour le

développement d'une plate-forme québécoise, en particulier concernant les vocabulaires et les outils informatiques utilisés.

Le présent mémoire débute en positionnant la méthodologie historienne comme étant une préfiguration du Web sémantique. Évidemment, cet amalgame entre cette technologie et la discipline historique apporte des changements à la pratique historienne qui se présenteront sous trois grands paradigmes. Ces derniers, comme nous le verrons, ont un impact notoire sur la méthodologie classique, mais s'arriment parfaitement avec les visées et les concepts clés de l'histoire. Ensuite le lecteur sera amené à comprendre le fonctionnement du Web sémantique en y dégageant ses principaux concepts et mécanismes. Après avoir étudié divers projets associant Web sémantique et histoire, cette démonstration se termine avec un cas pratique mettant en scène les personnes répertoriées dans la base de données du RPCQ. Ce passage de la théorie à la pratique permet de mieux situer le rôle et les limites de l'historien face au Web sémantique et au développement d'une plate-forme collaborative.

CHAPITRE I : LES TROIS NOUVEAUX PARADIGMES QU'APPORTE LE WEB SEMANTIQUE DANS LA PRATIQUE DE L'HISTORIEN

Le présent chapitre identifie les objectifs d'une démarche historique classique et les adapte aux possibilités du Web sémantique. Cet angle permettra de comprendre les transformations des pratiques sous trois paradigmes : l'historien qui devient architecte de l'information historique, l'automatisation des processus de liaison des données et l'émergence de nouvelles littératies.

D'historien à architecte de l'information historique

Construire l'histoire par la liaison de faits

L'histoire totale, concept élaboré par l'*École des Annales*, a permis de faire éclater les frontières de la discipline historique au courant du 20e siècle. Les limites de l'historien ne sont plus la politique, le militaire et la diplomatie. L'histoire avait dorénavant pour mandat d'expliquer et non de réciter, de condamner ou de couvrir d'éloges un événement du passé. Ce discours narratif s'inscrit dans une « histoire-problème », c'est-à-dire une méthodologie qui tente de répondre à des problématiques qui remettent continuellement en question les postulats et les méthodes de la discipline historique[1]. Selon Marc Bloch, co-fondateur avec Lucien Febvre de la revue *Annales d'histoire économique et sociale* dont la pensée a renouvelé l'historiographie de son époque, il faut se distancer du relatif en s'intéressant à la globalité de l'histoire. Cette totalité serait le concept qui faciliterait l'explication objective du passé[2]. Il faut donc s'intéresser à d'autres corpus que les écrits et aux autres

[1] François Furet, « De l'histoire-récit à l'histoire-problème », *Diogène*, 1 janvier 1975, no 89, p. 116.
[2] Marc Bloch, *Apologie pour l'histoire ou métier d'historien*, Paris, Colin, 2011 (1949), p. 70.

disciplines des sciences humaines. L'historien couvre alors un terrain épistémologique très vaste qui requiert une actualisation constante de la méthodologie. L'histoire quantitative française est issue de cet élargissement des approches visant notamment à mieux connaître les caractères de la société et de l'économie. Cependant, la totalité en histoire ne concerne pas uniquement la totalité des faits, mais aussi la totalité des regards et des méthodologies. L'histoire totale ne pouvant, selon sa conception originale, répondre à toutes les questions, la discipline a connu un éclatement au milieu des années 1980 menant à la multiplication des objets d'étude et des méthodes.

Malgré l'abandon du projet de l'*École des Annales*, il peut être intéressant de voir si le Web sémantique pourrait contribuer à l'aboutissement de certains de leurs idéaux de fédération des faits. Il ne s'agit pas ici de proposer un retour intégral aux objectifs de l'histoire totale, qu'on a depuis laissé de côté comme étant irréalisable, mais bien de s'appuyer sur les éléments fonctionnels du concept pour faciliter la mise en commun de connaissances et ainsi rassembler les composantes d'une discipline très éclatée.

Un intérêt marqué pour le facteur social amène les historiens à s'intéresser, au courant du 20e siècle, au fait comme unité d'analyse. Le chercheur se trouve alors devant un problème dichotomique majeur. Il doit bien comprendre un événement dans sa singularité tout en l'insérant dans une trame temporelle moins limitative. L'historien devient alors un artisan[3] qui associe différentes archives pour en dégager un récit qui répondra à ses objectifs de recherche. Ce travail de liaison est plus que nécessaire

[3] François-Xavier Petit, *Qu'est-ce qu'être historien?*, http://www.histoire-pour-tous.fr/education/179-metiers-histoire/3416-quest-ce-quetre-historien-.html, 21 décembre 2010, consulté le 6 octobre 2015.

puisqu'un fait, selon Paul Veyne, ne peut être compris sans un maillage étroit avec d'autres :

> Les faits n'existent pas isolément, en ce sens que le tissu de l'histoire est ce que nous appellerons une intrigue, un mélange très humain et très peu « scientifique » de causes matérielles, de fins et de hasards; une tranche de vie, en un mot, que l'historien découpe à son gré et où les faits ont leurs liaisons objectives et leur importance relative: la genèse de la société féodale, la politique méditerranéenne de Philippe II ou un épisode seulement de cette politique, la révolution galiléenne[4].

Un fait doit être mis en résonance avec d'autres pour exister. Ainsi on s'éloigne de l'opinion puisqu'un ensemble de composantes permet de valider l'information.

Il existe par ailleurs une imbrication de faits. Sous forme hiérarchique, un fait peut se décomposer en une séries d'autres qui permettent de comprendre certaines particularités tout en contextualisant le fait général. Jacques Cartier plantant sa croix à Gaspé est un fait historique qui s'inscrit dans la découverte du Canada par l'explorateur. Ce maillage de faits est indispensable à la compréhension de l'histoire. La complexité des associations est différente selon la nature des faits. Il faut moins de faits pour déterminer le nom d'une personne que pour énumérer les causes du déclenchement de la Première Guerre mondiale.

Évidemment, cette quête de récit par ramifications oblige de donner sens à l'information recueillie. Le rôle social de l'historien n'est pas de simplement fournir une réponse à un questionnement, mais de présenter toutes les problématiques entourant le sujet pour montrer clairement la complexité du processus historique[5]. Les différents

[4] Paul Veyne, *Comment on écrit l'histoire*, Éditions du Seuil, Paris, 1971, p. 46.
[5] François-Xavier Petit, *loc. cit.*

supports informationnels contiennent une série de faits qui devront être analysés et contextualisés. Le processus historique est donc de déconstruire des récits pour en reconstruire d'autres, comme le rappelle Tim Sherratt :

> Let's think for a moment about the work of a historian — identifying actors, defining relationships, documenting the complex networks that bring together people, places and events over time. It's painstaking, exhilarating [sic] and potentially soul-destroying work. It's also an exercise in data modelling. Whether the results are preserved in a triplestore, a spreadsheet, or on a drawer full of index cards — it's nodes and edges, it's entities and relationships, it's data[6].

Entre la phase de déconstruction et de reconstruction du récit, l'historien compile ou consulte un volume variable de données extirpées en partie de son récit original. C'est le cas d'un livre ancien que l'on consulte à la collection nationale. Du même coup, ce livre est aussi un fait en soi de par son existence. Il y a donc une similitude entre un fait et une donnée. De plus, il devient primordial d'avoir un minimum d'autres faits entourant ce livre pour éviter une totale décontextualisation. À titre de précision, un fait existe tandis que la donnée est construite. Malgré leur grande similitude, il ne faut pas omettre que la donnée capte un fait, dans sa totalité ou en partie comme le rappelle Johanna Drucker[7].

Pour faciliter la compréhension du récit, deux balisages doivent être faits par un historien. Le premier est conceptuel puisqu'il faut éviter les termes à multiples sens. Le second est analytique, l'historien dévoile la méthode qu'il utilisera pour analyser son corpus. L'étape finale est évidemment la mise en place du récit qui prend la forme d'un texte. Il s'agit donc d'une mise à plat de ce maillage complexe de sources pour en

[6] Tim Sherratt, *Stories for machines, data for humans*, http://discontents.com.au/stories-for-machines-data-for-humans/, 10 avril 2015, consulté le 7 octobre 2015.
[7] Johanna Drucker, « Humanities Approaches to Graphical Display ». *DHQ:Digital Humanities Quarterly*, volume 5, no. 1, 2011, p. 3.

présenter un filon compréhensible et linéaire. L'historien ajoute donc un document à l'historiographie existante pour développer une nouvelle piste de réflexion.

Les limites de la démarche historienne et la publication inachevée

La question est maintenant de savoir si la méthodologie historienne répond bien aux objectifs de la discipline. Rappelons ici l'intérêt soutenu des historiens du 20e siècle à comprendre l'histoire dans son ensemble, c'est-à-dire par l'analyse de ses singularités politiques, économiques, diplomatiques, sociales, culturelles et environnementales. L'historien use alors de différents stratagèmes pour organiser l'histoire. C'est le cas de la périodisation historique qui correspond, selon Jacques Le Goff, à une action humaine sur le temps qui ne peut être neutre[8]. L'historien, comme tout individu, participe à cette construction de faits. Ce constat l'oblige donc à repenser son positionnement pour rendre sa démarche plus efficace.

Le numérique est sans contredit un atout majeur, mais la solution épistémologique ne se résout pas à la simple utilisation des nouveaux outils technologiques. Partant du postulat que l'histoire visant à présenter le tableau le plus complet possible ne peut se comprendre à l'échelle humaine, l'historien se doit d'incorporer son travail à un plus grand ensemble capable d'agréger les faits historiques et les liens les unissant. Évidemment, ces limites humaines sont déjà connues et l'historien balise son travail en conséquence. Toutefois, le numérique permet de compiler les données afin d'accélérer la recherche et le Web sémantique facilite la mise en commun de l'information. En effet, l'historien ne pouvant explorer avec la même

[8] Jacques Le Goff, *Faut-il vraiment découper l'histoire en tranches?*, Paris, Seuil, 2014, p. 12.

acuité l'ensemble des sources utiles à la compréhension de sa question de recherche utilise les écrits de ses prédécesseurs. Cependant, cette approche se limite encore une fois aux capacités humaines et peut rarement remonter la chaîne de faits, par exemple, jusqu'au balbutiement d'un concept et de l'exprimer dans ses multiples contextualisations à travers le temps. La question est alors de savoir si la forme narrative est l'outil idéal pour contribuer à la compréhension de l'histoire puisqu'il s'agit, encore aujourd'hui, du seul rendu permettant d'évaluer les compétences demandées par cette discipline.

Sachant que l'historien ne peut comprendre seul l'histoire dans toute sa complexité, est-il possible de favoriser l'écriture collaborative et inachevée via des plates-formes Web? Une œuvre collective en continuel développement facilite l'ajout ponctuel d'informations sans devoir nécessairement réécrire une contextualisation déjà proposée par d'autres professionnels. On évite alors d'ajouter un volume considérable d'écrits épars pouvant contenir un nombre variable de problématiques et qui devra être remis en question dans un autre essai. Ce regard sur l'œuvre collective permet d'entrevoir le changement de position que le Web sémantique impose aux historiens souhaitant y contribuer. Ces derniers passent du rôle d'auteur à celui de participant d'un plus grand ensemble. L'historien n'a plus à réécrire une série d'informations pour exposer un nouvel angle d'approche, mais peut simplement modifier ou interroger directement le travail d'un autre collègue.

Le projet *The Historian's Macroscope: Big Digital History* de Shawn Graham, Ian Milligan et Scott Weingart, utilise cette philosophie en présentant une version

préliminaire ouverte (en anglais: *Open Draft Version*) de leur prochain ouvrage[9]. Les auteurs soumettent leur travail au public afin que celui-ci contribue à la vérification et à la complétion des faits évoqués. Les auteurs souhaitent une rétroaction du public pour bien structurer leur ouvrage : « [It] does give us an opportunity to rethink how the overall structure of this work will come together, the things that we are taking for granted or haven't made explicit enough. It is rare I think for academics at any stage to read a book from start to finish, so perhaps this lesson will help us craft a better book that allows for dipping in to suit the reader's needs[10]. » À voir les statistiques de réutilisation de l'ouvrage dans les premiers jours du projet[11], il est indéniable que les auteurs produiront un livre mieux réfléchi.

Cette approche annonce un possible changement dans la manière de construire les récits historiques par le biais des technologies numériques. La méthodologie historienne est efficiente puisqu'elle souhaite associer un ensemble de faits pour en dégager un récit qui fait sens et qui rappelle fortement, comme le souligne Tim Sherratt, le fonctionnement du Web sémantique[12]. Cependant l'angle applicatif actuel est coercitif. Ce travail manuel associé avec justesse à l'artisanat doit s'inscrire dans une nouvelle structure favorisant un partage direct de l'information et une construction automatisée des liens entre les faits. L'historien doit s'intéresser à la structure des données et non plus simplement à la livraison de celles-ci.

---

[9] Shawn Graham, Ian Milligan et Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope*, http://www.themacroscope.org/2.0/, 2015, consulté le 7 octobre 2015.
[10] Shawn Graham, *Reflecting on our process*, http://www.themacroscope.org/?page_id=303, 11 septembre 2013, consulté le 7 octobre 2015.
[11] *Ibid.*
[12] Tim Sherratt, *Every story has a beginning*, http://discontents.com.au/every-story-has-a-beginning/, 4 octobre 2011, consulté le 31 août 2015.

Mettre de l'avant une structure réfléchie pour la discipline

L'historien qui souhaite participer pleinement au Web sémantique devra se transformer pour se rapprocher du rôle d'architecte de l'information tel que défini par Benoît Habert, Jean-Michel Salaün et Jean-Philippe Magué:

> Un architecte conçoit un habitat pour qu'il soit approprié aux besoins spécifiques (logement, bureau, commerce…) des personnes qui y vivront ou qui en seront les utilisateurs. L'architecte de l'information structure les contenus et leur accès (navigation, recherche) pour qu'ils soient le mieux adaptés aux tâches des utilisateurs effectifs. Au centre de son raisonnement se trouve la détection (findability). Les utilisateurs doivent trouver aisément, à point nommé, sous la forme requise, l'information précise qui leur est nécessaire. L'architecte de l'information doit tout à la fois être un spécialiste de l'organisation et du repérage des contenus et un spécialiste de l'expérience utilisateur ou utilisabilité (UX–user experience)[13].

Cette définition sous-entend, encore plus directement dans un contexte historique, la mise en valeur des contenus. Si la démarche s'oriente en partie sur l'expérience utilisateur, les contenus seront consultés plus massivement. L'historien intéressé par ce type de démarche devra acquérir cinq nouvelles compétences répertoriées dans le *Référentiel de compétences en Architecture de l'information*[14].

Premièrement, l'historien doit maîtriser la gestion dynamique des projets. Succinctement, il s'agit de préparer les assises du projet, soit définir les objectifs et découper le projet en étapes claires et précises. Dans le cas présent, il faut présenter les phases d'action permettant de mettre en place une structure capable d'accueillir tous sujets historiques concernant le Québec et établir les mécanismes de collaboration

---

[13] Benoit Habert, Jean-Michel Salaün et Jean-Philippe Magué, « Architecte de l'Information: Un métier », *ADBS*, 2012, vol. 49, n° 1, (coll. « Documentaliste - Sciences de l'Information »), p. 4-5.

[14] Jean-Michel Salaün, *Référentiel de compétences en Architecture de l'information*, http://archinfo01.hypotheses.org/453, 7 octobre 2013, consulté le 7 octobre 2015.

puisqu'il s'agira d'une plate-forme qui permettra à tous les professionnels d'y inclure des données ou des liens entre les données.

Deuxièmement, l'historien doit savoir coopérer et dialoguer avec les métiers connexes qui travailleront au développement de la plate-forme. Évidemment il faut être à l'écoute des obligations de chaque discipline historique, mais aussi des partenaires techniques. L'historien doit savoir où il doit contribuer pour déployer au maximum son potentiel. Cette frontière sera définie à l'aide d'un cas de figure au dernier chapitre. Cependant, on peut déjà affirmer que la mise en place de la structure technologique passera par des programmeurs et des informaticiens. Pour cette raison, la démarche conceptuelle doit obligatoirement s'arrimer aux exigences techniques. Ce dialogue transdisciplinaire s'appuie nécessairement sur la compréhension des différents langages qui seront présentés dans le prochain chapitre.

Troisièmement, la réflexion doit se faire autour de la notion d'expérience utilisateur qui correspond selon la norme ISO 9241-210 « aux réponses et aux perceptions d'une personne qui résultent de l'usage ou de l'anticipation de l'usage d'un […] système.[15] » Dans le contexte d'une plate-forme collaborative répondant aux logiques historiques, il faut analyser deux types d'utilisateurs : les professionnels et le grand public. D'un côté, il faut que la saisie de l'information reste simple malgré une structure qui doit pouvoir intégrer différents type de données. De l'autre, la consultation

[15] Yannick Grenzinger, *Qu'est-ce que l'expérience utilisateur ? | Ergonomie, Expérience Utilisateur, Design Thinking*, http://ux-fr.com/experience-utilisateur-definition/, consulté le 7 octobre 2015.

doit être fluide et permettre la création d'interfaces personnalisées selon les questions imprévisibles des utilisateurs[16].

Quatrièmement, l'historien doit s'intéresser aux modes de structuration. Il doit se familiariser avec les représentations de données à sa disposition et savoir lesquels sont susceptibles de résoudre ses problématiques. Dans le contexte du Web sémantique, il doit s'approprier les concepts techniques et les schémas conceptuels structurés que nous verrons plus tard. Cette étape est cruciale pour la passation d'une discipline organique, c'est-à-dire basée sur des processus humains, vers une discipline structurée, suivant des normes et pratiques internationales. Cette étape inclut aussi de s'investir dans la conception d'un moteur de recherche plus près de ses objectifs et de s'éloigner, par le fait même, des moteurs généraux qui ne focalisent pas sur une contextualisation de la donnée. Le moteur de recherche Google, par exemple, incorpore dans sa méthode de tri un algorithme qui tient compte de la popularité des sites web. Cette approche peut très fortement biaiser une recherche en sciences humaines, sachant que peu d'utilisateurs vont au-delà de la troisième page de résultats fournis par le moteur de recherche[17].

Cinquièmement, l'historien doit être en mesure de concevoir des prototypes. Dans le contexte actuel de financement de la recherche, il faut développer des projets ayant des résultats rapides. On ne peut penser structurer un projet de plate-forme collaborative sans passer par des projets-pilotes. Évidemment, l'historien se doit de

---

[16] Réseau pancanadien du patrimoine documentaire, *« Démonstration de faisabilité » de la Visualisation des Données ouvertes liées (LOD)*, Canada, 2012, http://www.canadiana.ca/sites/pub.canadiana.ca/files/PCDHN%20Proof-of-concept_Final-Report-FRA.pdf, consulté le 27 août 2015, p. 8.

[17] Chitika, *The Value of Google Result Positioning*, http://chitika.com/google-positioning-value, 7 juin 2013, consulté le 14 janvier 2016.

prouver que cette structure permettra des avancées considérables pour la discipline. Le meilleur moyen, dans ce cas de figure, est de s'intéresser à un sujet précis comme le *Réseau pancanadien du patrimoine documentaire* (RPCPD) l'a fait avec le projet *Au-delà des tranchées*[18] qui sera analysé au prochain chapitre. Une fois que le projet a fait ses preuves, il suffit d'agrandir le corpus et de continuer les rétroactions sur le modèle.

Cet argumentaire répond à la déclaration d'Emmanuel Leroy Ladurie qui affirmait en 1967 lors d'une conférence sur l'histoire quantitative : « …l'historien de demain sera programmeur ou il ne sera plus…[19] » Le débat concernant les capacités de programmation en humanités numériques fait toujours rage[20]. L'historien n'a pas l'obligation encore aujourd'hui de savoir coder, mais, paraphrasant Olivier Le Deuff lors du *Congrès Humanités Numériques 2015*, il devra sans doute savoir encoder et décoder[21]. L'historien doit pouvoir ajouter du contenu de différents formats (encoder) et être capable d'interpréter des interfaces présentant les données (décoder). Avec ces deux compétences, il sera en mesure de dialoguer avec les techniciens et de fournir un contenu pertinent et qui respecte les normes ontologiques.

---

[18] Réseau pancanadien du patrimoine documentaire*, loc. cit.,* 86 p.
[19] Émilien Ruiz, *Les historiens seront-ils finalement programmeurs ?*, http://www.boiteaoutils.info/2011/09/les-historiens-seront-ils-finalement/, 22 septembre 2011, consulté le 8 octobre 2015.
[20] *Ibid.*
[21] Olivier Le Deuff, Quelles littératies et formations pour les humanités digitales ?, Humanités numériques 2015, http://hn2015.org/programme/, Montréal, 12 août 2015.

Automatisation du processus historique

Longue durée et reconnaissance optique de caractères

Le second paradigme concerne l'automatisation du processus historique puisque la masse informationnelle disponible peut difficilement être analysée manuellement. Cette idée d'automatisation prend naissance en partie dans le concept d'histoire de longue durée. Celui-ci refait surface en 2015 avec l'ouvrage *History Manifesto* de Jo Guldi et David Armitage. Reprenant les écrits de Fernand Braudel, les auteurs s'inquiètent de la présence quasi unilatérale du « short-termism[22] » en histoire. *History Manifesto* affirme que la dernière décennie marque un retour à ces études plus étendues grâce au développement des outils informatiques.

Avec la mise en place d'outils de reconnaissance optique de caractères (ROC), il est possible d'effectuer des recherches par fouilles de textes automatisées[23]. Ces outils permettent de visualiser l'utilisation d'un terme dans un corpus prédéterminé. Évidemment, l'outil développé par *Google*, le *Google Books Ngram Viewer* est un système phare de cette méthodologie. L'inconvénient de cet outil est le corpus qui se limite aux livres numérisés et disponibles sur *Google Books*. Cette contrainte favorise l'histoire anglo-saxonne qui s'impose alors comme norme internationale. D'autres erreurs peuvent survenir lors de la reconnaissance des caractères comme le fait de lire le terme « fuck » au lieu de « suck » puisque le l'algorithme reconnait le « s » long du

[22] Jo Guldi et David Armitage, *The History Manifesto*, Cambridge University Press, Cambridge, 2014, p. 2.
[23] Jo Guldi et David Armitage, *op. cit*., p. 90.

19e siècle comme étant un « f »[24]. Il faut donc avoir un esprit très critique face aux outils d'automatisation à grande échelle dont l'algorithme et le corpus sont hermétiques.

Pour étudier d'autres corpus bibliographiques, Jo Guldi et Chris Johnson-Roberson ont développé une extension *Zotero* du nom de *Paper Machines* : « Paper Machines is [a] bibliographic management software that makes cutting-edge topic-modeling analysis in Computer Science accessible to humanities researchers without requiring extensive computational resources or technical knowledge. It synthesizes several approaches to visualization within a highly accessible user interface[25]. » Ce type d'application permet de gérer plus rapidement et plus efficacement un volume de données beaucoup plus grand qu'avec une méthodologie classique. Cette simple gestion de mots permet de poser un lot d'hypothèse sur l'impact de la longue durée sur l'utilisation d'un terme ou le développement d'une idée[26].

Cependant, une telle approche demande un investissement intellectuel et économique de la part des universités pour rendre possible ce type d'études[27]. Cet investissement matériel sera vain si une équipe spécialisée n'est pas instaurée pour encoder et décoder les données recueillies. La question est maintenant de savoir comment on obtient ces corpus. L'analyse est automatisée, mais la constitution et l'organisation du répertoire restent manuelles, conséquence du fonctionnement en silos

[24] Jakub Marian, *A curiosity about the F-word in Google Ngram Viewer*, https://jakubmarian.com/a-curiosity-about-the-f-word-in-google-ngram-viewer/, consulté le 14 janvier 2016.
[25] Jo Guldi et Chris Johnson-Roberson, *Paper Machines | Visualize Your Zotero Collections*, http://papermachines.org/, 2015, consulté le 8 octobre 2015.
[26] Jo Guldi et David Armitage, *op. cit*., p. 91.
[27] *Ibid*., p. 105.

des systèmes de recherches en ligne. La solution se trouve probablement dans la mise en place d'une plate-forme participative en données liées.

Le Web sémantique pour interroger la datamasse dans un récit textuel

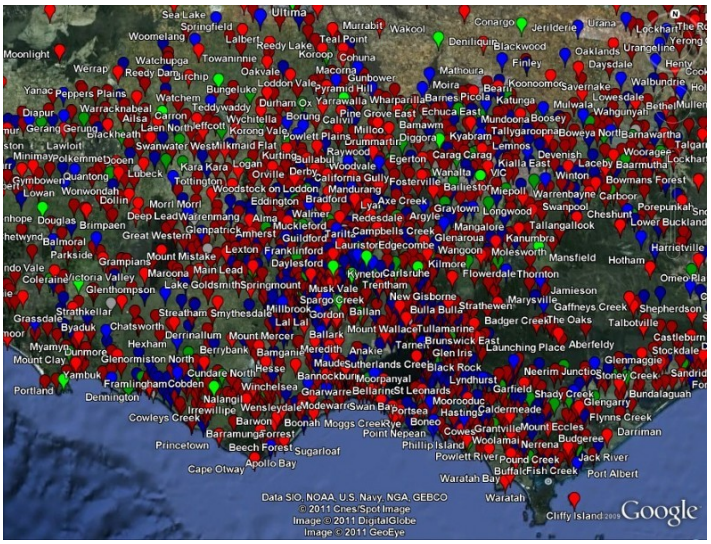
Dans une perspective de portail collaboratif basé sur le Web sémantique, l'analyse de ce volume de données risque de se complexifier. La quantité d'informations emmagasinée dans des bases de données ne cesse de croître, au point où il existe maintenant des études spécialisées dans la gestion de ce qu'on appelle la datamasse (en anglais: *Big data*). Un des joueurs majeurs de ce domaine dans la province est *Calcul Québec*. Ce « regroupement d'universités québécoises réunies autour du calcul informatique de pointe (CIP)[28] » s'allie aux experts disciplinaires pour contribuer à l'économie du savoir du Québec. Le potentiel calculateur de leurs superordinateurs est idéal pour accueillir la masse de données culturelles du RPCQ et de ses futurs partenaires. Toutefois, quels types d'interfaces faudra-t-il générer pour obtenir une information claire et qui structure efficacement les multiples récits historiques? Cette interrogation a été posée par Tim Sherratt dans son article « Every story has a beginning »[29].

La réflexion de cet historien australien s'arrime avec les notions de longue durée et celle d'histoire totale, qui demandent une compréhension de l'événementiel dans un ensemble plus large. À titre d'exemple, Sherratt raconte la vie d'Alexander Dewar Kelley, soldat australien qui a participé à la Première Guerre mondiale. On comprend

[28] Calcul Québec, *À propos de Calcul Québec*, http://www.calculquebec.ca/fr/a-propos-de-cq/calcul-quebec, 2015, consulté le 8 octobre 2015.
[29] Tim Sherratt, *loc. cit.*

facilement l'impact de la guerre sur cet individu et sur sa famille de par ses écrits personnels et les écrits professionnels le concernant. Cependant, si on s'intéresse à l'impact de la guerre sur les familles de soldats australiens, il faudra traverser un corpus d'enregistrements qui dépasse les 375 000 entrées[30]. Pour évaluer l'ampleur du corpus, Sherratt utilise des marqueurs sur *Google Maps* pour localiser seulement le lieu de naissance de ces soldats ou leur lieu d'enrôlement :

Figure 1.1: Lieux de naissance ou d'enrôlement des soldats australiens ayant participé à la Première Guerre mondiale



Source : Tim Sherratt, *Every story has a beginning*, http://discontents.com.au/every-story-has-a-beginning/, 4 octobre 2011, consulté le 31 août 2015.

Malgré un recadrage sur l'état de Victoria, la représentation graphique atteint une limite dans son utilité puisque la masse de données comble totalement l'espace. La représentation nous permet de constater que les soldats provenaient de partout en Australie, mais sans pouvoir dégager d'autres constats. L'implication historienne est donc nécessaire afin de réfléchir à des interfaces capables de répondre à des problématiques de recherche puisque le numérique et l'automatisation des processus risquent de s'imposer de plus en plus.

[30] *Ibid.*

En effet, cette méthodologie orientée vers l'automatisation est, selon Dan Cohen, la voie d'exploration des sources de la prochaine décennie : « These computational methods which allow us to find patterns, determine relationships, categorize documents, and extract information from massive corpuses, will form the basis for new tools for research in the humanities and other disciplines in the coming decade.[31] » En théorie, plus la quantité de données accessible en ligne augmente, plus il est possible d'en dégager des généralités. Il s'agit donc d'une méthodologie inductive, soit de partir des données pour en extraire des constats généraux[32]. Devant un nouvel outil comme le Web sémantique, il devient difficile, voire inapproprié, de développer une hypothèse de recherche qui découlerait d'une méthodologie classique de l'histoire, souvent déductive, que les chercheurs en humanités numériques interrogent et transforment. Cette idée s'exprime dans la définition même des humanités numériques proposée dans le *Manifeste des Digital Humanities*, qui rappelle que cette transdiscipline repose sur des perspectives heuristiques[33]. Tim Sherratt croit qu'il faut un amalgame entre automatisation à grande échelle et la conception humaine, voire sensible, de la discipline historique. Il faut conserver une part d'expressivité provenant de l'écrit. Il faut éviter de masquer les histoires individuelles au profit d'un ensemble statistique trop grand qui dilue les nuances. Le texte exprimant avec éloquence les faits historiques, comment peut-on en optimiser la réutilisation dans le cadre des données liées sans en morceler la narration?

---

[31] Daniel J. Cohen, « From Babel to Knowledge: Data Mining Large Digital Collections », mars 2006, http://www.dlib.org/dlib/march06/cohen/03cohen.html, vol. 12, n° 3.
[32] Mireille Blais et Stéphane Martineau, « L'analyse inductive générale: description d'une démarche visant à donner un sens à des données brutes », *Recherches qualitatives*, 2006, vol. 26, n° 2, p. 3.
[33] Marin Dacos, *Manifeste des Digital humanities*, http://tcp.hypotheses.org/318, 26 mars 2011, consulté le 27 février 2014.

Sherratt croit qu'il faut conserver le texte comme source première de la recherche historique, mais lui donner une forme par laquelle un ordinateur aura la capacité de repérer certains éléments particuliers et ce, grâce au Web sémantique. Cette intelligibilité par l'ordinateur permettra la navigation entre les différents médias d'où proviennent ces textes. Cette idée s'inscrit dans le prolongement de celles d'Edward L. Ayers qui écrivait en 1999: « May we now be able to, need to, write a new kind of history, a history that can be arrayed and understood in multiple sequences and layers, a history that involves and rewards more engagement on the part of the reader than a book requires or permits? We might call that history, for convenience, "hypertextual," since it would involve linked text in a manipulable electronic environment[34]. » À cette époque, Ayers ne pouvait qu'utiliser les liens hypertextes pour présenter sa vision de l'histoire non-linéaire. L'avantage des liens sémantiques est qu'à partir d'une base d'entrées, un nombre de liens se crée automatiquement sans devoir entreprendre soi-même une recherche des composantes connexes.

Pour illustrer son idée, Tim Sherratt propose un prototype de livre utilisant des données liées pour enrichir le contenu intitulé *Inigo Jones – The weather prophet*[35]. Ce projet montre le réel potentiel des données liées lorsqu'on les ajoute à un texte long. Le programme derrière le texte permet de déterminer une série de liens entre ce dernier et des vocabulaires externes que l'on verra dans le prochain chapitre. Cette extraction automatique permet d'enrichir l'écrit autant au niveau du contenu que du contenant.

[34] Edward L. Ayers, *History in Hypertext*, http://www.vcdh.virginia.edu/Ayers.OAH.html, 1999, consulté le 27 février 2014.
[35] Tim Sherratt, *Inigo Jones - The weather prophet*, http://lodbookdev.herokuapp.com/#!/text/1/, consulté le 14 janvier 2016.

Pendant la rédaction, l'auteur a créé, sans peut-être le remarquer, une base de données qui peut dorénavant servir à créer différentes visualisation du texte. Dans le prototype, on peut lister les personnages, les organisations, les sources, les endroits et les événements présents dans le texte[36]. Ces deux dernières composantes sont aussi présentées sur une carte du monde et une ligne du temps délimitant avec précision le cadre spatio-temporel. À l'échelle des institutions culturelles, il serait possible pour un centre d'histoire locale d'utiliser des données provenant d'un musée national pour peaufiner son offre informationnelle.

Ce livre lié proposé par Tim Sherratt lui permet d'enrichir facilement son travail tout en l'incorporant au nuage de données liées existant. Cet enrichissement mutuel facilite la lecture du texte et la réutilisation d'un grand pourcentage des données puisqu'elles sont extraites du bloc texte. Toutefois, les compétences en informatique de Tim Sherratt dépassent celle de la grande majorité des historiens. Si l'on souhaite une implication massive des historiens dans cette nouvelle forme de narration automatisée, il est primordial d'adapter les compétences historiennes face à cette nouvelle possibilité.

Adapter la méthodologie historienne classique aux possibilités numériques

Le coffre à outils du chercheur débutant : Guide d'initiation au travail intellectuel dirigé par Jocelyn Létourneau s'adresse à un public collégial et universitaire pour faciliter la réalisation des travaux écrits. Il s'agit d'un ouvrage généraliste pour les chercheurs débutants en sciences humaines, mais co-écrit par plusieurs historiens qui en font un incontournable pour la rédaction en sciences historiques. Reprenant la réédition

[36] *Ibid.*

de 2006 de l'ouvrage datant de 1989, la démonstration suivante analyse l'impact du Web sémantique sur différentes étapes de la méthodologie historienne identifiées dans les chapitres « Comment se documenter à l'ère électronique », « Comment élaborer une stratégie de recherche » et « Savoir communiquer sa pensée par écrit ».

« Comment se documenter à l'ère électronique » met en garde le jeune chercheur puisque: « La cybernavigation demande en effet une cybercompétence. Elle exige de l'utilisateur une conscience aiguë des avantages et des inconvénients du média.[37] » Misant énormément sur la consultation de catalogues de bibliothèque, notamment celui de *Bibliothèque et Archives nationales du Québec* (BAnQ), on souligne que les ouvrages référencés ne présentent pas leurs contenus de manière exhaustive[38]. C'est devant ce manque de précision que les auteurs listent 26 conseils pour une meilleure recherche à l'ère numérique comme « connaître les moteurs de recherche qu'on emploie[39] ». Difficile de connaître les habitudes de travail des historiens, mais fort est à parier que peu d'utilisateurs des systèmes de recherche s'intéressent aux limites ontologiques de ceux-ci. Avec le Web sémantique, l'utilisateur est confronté aux ontologies puisque celles-ci sont au cœur de la structure du Web sémantique comme nous le verrons au prochain chapitre. Il est alors plus facile de concevoir les limites d'un outil. De plus, l'utilisation des données liées favorise une unification des différents systèmes réduisant le nombre d'outils à connaître et à surveiller.

[37] Jocelyn Létourneau, *Le coffre à outils du chercheur débutant: guide d'initiation au travail intellectuel*, Montréal, Boréal, 2006, p. 35.
[38] *Ibid.*, p. 43
[39] *Ibid.*, p. 64

« Comment élaborer une stratégie de recherche » rappelle l'importance de positionner son sujet par rapport aux travaux existants. Comme le souligne Jocelyn Létourneau, le chercheur se doit de pouvoir répondre « aux questions suivantes: Qu'ont fait les autres chercheurs dans ce domaine? Que puis-je faire ou que faut-il faire maintenant? Comment mon projet peut-il permettre l'avancement des connaissances ou d'un débat[40]? » Cette étape cruciale n'est pas nécessairement simple selon le sujet choisi. En effet, il n'existe pas d'outils conçus pour présenter le réseau d'ouvrages balisant un sujet précis. Lorsqu'on entre dans la longue durée, la contextualisation exhaustive de notre étude peut s'avérer difficile voire impossible. Aucune structure actuelle ne permet de valider l'exhaustivité d'une bibliographie. Dans l'absolu, une structure sémantique efficace et réfléchie peut soutenir une interface qui a comme objectif de constituer un réseau d'ouvrages selon un ensemble de critères énoncés. Cette interface est inconcevable dans la version 2.0 du Web.

Finalement, « Savoir communiquer sa pensée par écrit » rappelle que l'argumentaire soulevé et la documentation consultée ne sont pas les seuls facteurs qui valident la qualité d'un texte : « La clarté de l'argumentation développée, la logique du raisonnement tenu, la beauté de l'expression écrite et la capacité à soutenir l'intérêt du lecteur représentent quatre éléments qui ont une incidence déterminante sur la qualité finale d'un travail de recherche[41]. » Ce propos rejoint la proposition de Tim Sherratt sur l'importance de préserver la forme narrative pour présenter des faits historiques. La clarté est un objectif que l'historien doit maîtriser pour réussir à aplanir les différentes

[40] *Ibid.*, p. 195
[41] *Ibid.*, p. 217

couches historiques utilisées pour cerner son sujet. Cette démarche reste au cœur du travail de l'architecte de l'information historique puisque ce dernier doit s'assurer de la cohésion narrative des vocabulaires appliqués aux données. Dans ce deuxième paradigme concernant l'automatisation du processus historique, les règles ne sont plus grammaticales, mais ontologiques. Dans le prolongement de la méthodologie classique, l'historien aura pour nouveau mandat de diffuser ses données de recherche dans un format et des normes précises et non seulement dans une forme textuelle définitive.

Nouvelle littératie historique

Des données de recherche malléables comme sources pour s'éloigner de Google

Selon l'*Organisation de coopération et de développement économiques* (OCDE), la littératie est « l'aptitude à comprendre et à utiliser l'information écrite dans la vie courante, à la maison, au travail et dans la collectivité en vue d'atteindre des buts personnels et d'étendre ses connaissances et ses capacités[42]. » Dans l'environnement de l'architecte de l'information historique, il s'agit tout autant de la capacité à interagir avec les informations existantes que la diffusion de son propre contenu suivant les recommandations des différents langages.

Une nouvelle interface de saisie remplacera celle du simple traitement de texte. L'arrimage entre les données ne passera plus par une syntaxe, mais bien par des points d'entrées SPARQL en interaction entre eux. SPARQL est un outil utilisé afin

[42] Organisation de coopération et de développement économiques. *La littératie à l'ère de l'information: Rapport final de l'enquête internationale sur la littératie des adultes*, Paris, http://www.oecd.org/fr/edu/innovation-education/39438013.pdf, 2000, p. 10.

d'interroger des systèmes sémantiques qui sera présenté au prochain chapitre. Actuellement, la difficulté d'adaptation des nouveaux utilisateurs non techniques est qu'il n'existe pas de moyen d'explorer les données liées utilisant une version simplifiée d'un point d'entrée SPARQL. Le chercheur ou le catalogueur est alors confronté à une interface peu conviviale pour emmagasiner ou extraire des données.

À titre d'exemple, *Huma-Num* propose un service d'exposition de données nommé *Nakala*. Au même titre que *Calcul Québec*, *Huma-Num* est une très grande infrastructure de recherche (TGIR) qui facilite l'adoption du numérique, mais spécifiquement pour les sciences humaines. L'objectif principal de ce consortium est de coordonner la « production raisonnée et collective de corpus de sources[43]. » Proposant un lot de bonnes pratiques techniques à adopter, la TGIR modifie considérablement la littératie historique puisqu'il devient impératif de comprendre que le résultat qui s'affiche ou celui qu'on enrichit se constitue de données décentralisées reconstituées.

Huma-Num permet d'ouvrir « un accès persistant et interopérable[44] » aux données. D'un côté, la plate-forme attribue un identifiant pérenne *Handle* qui fournit un accès permanent et sécuritaire. *Handle*, système d'identifiants géré par la *Corporation for National Research Initiatives*, offre aux institutions la possibilité d'entreposer leurs données à l'extérieur de leur écosystème numérique[45]. On assure ainsi la pérennité même si des modifications sont apportées à notre base de données. De l'autre côté, par le

[43] Huma-Num, *À propos de Huma-Num*, http://www.huma-num.fr/la-tgir-en-bref, 24 mars 2015, consulté le 9 octobre 2015.
[44] Huma-Num, *Exposer ses données avec Nakala*, http://www.huma-num.fr/services-et-outils/exposer, 12 mai 2015, consulté le 9 octobre 2015.
[45] Corporation for National Research Initiatives, *Handle*, http://www.handle.net/, 25 août 2015, consulté le 9 octobre 2015.

biais des idées du Web sémantique et des métadonnées, *Nakala* permet de standardiser les données du chercheur pour que celles-ci puissent être réutilisées pour bâtir des applications. L'ensemble de ces fonctions sont accessibles par une interface simple et bien identifiée[46].

Nakala s'intègre dans l'outil de recherche *Isidore*[47]. Ce dernier est un outil de recherche en sciences humaines et sociales qui rassemble des données pour les chercheurs en s'appuyant sur les principes du Web sémantique[48]. Tentant de s'éloigner de la convergence que *Google* favorise, l'outil permet de moissonner des données scientifiques liées grâce à des normes internationales et en accès libre. Si on cherche « Samuel de Champlain » sur *Google*, on obtient une liste de sites web où cette chaîne de caractères se retrouve. Il est aussi possible de faire une recherche avancée qui s'intéresse aux métadonnées, mais son utilisation se limite à des champs ou des corpus précis. Sur *Isidore*, le processus est plus développé et dispose d'une interface intermédiaire qui présente les métadonnées de la source. Par exemple, « Samuel de Champlain » peut être recherché selon la métadonnée du type de source. On peut alors naviguer dans différentes bases de données, notamment *Gallica*, pour dénicher des documents en lien avec Champlain comme des plans dessinés à la main[49]. Cette nouvelle lecture rendue possible par les données liées, transforme les possibilités de recherche et d'agrégation des données. Il est donc fondamental pour l'historien de s'intéresser à

---

[46] Huma-Num, *Exposer ses données avec Nakala*, *loc. cit.*

[47] Huma-Num, *ISIDORE - Accès aux données et services numériques de SHS*, http://www.rechercheisidore.fr/, 2015, consulté le 9 octobre 2015.

[48] Huma-Num, *ISIDORE - À propos*, http://www.rechercheisidore.fr/apropos, 2015, consulté le 9 octobre 2015.

[49] Samuel de Champlain, *[Illustrations de Les Voyages de Champlain...] / [Non identifié] ; Samuel Champlain, aut. du texte*, s.l., http://gallica.bnf.fr/ark:/12148/btv1b2000019z via http://www.rechercheisidore.fr, 1613, consulté le 9 octobre 2015.

l'organisation de ces interfaces pour en profiter pleinement et ouvrir vers de nouvelles avenues de recherche.

Régulation par les pairs

L'analyse de textes, ou plus précisément la validation de ceux-ci, inclue l'exploration de comptes rendus de lecture. Cette méthode permet de savoir si les idées évoquées sont partagées par une communauté professionnelle importante. En prime, la révision par les pairs dans une revue scientifique exclut les textes ne répondant pas « aux exigences accrues du lectorat[50] ». Le Web sémantique ajoute une nouvelle dimension à ces validations, ce qui rend compte du caractère essentiel de l'évaluation par les pairs – ainsi que la reconnaissance scientifique qui l'accompagnerait alors –pour que la contribution à ces bases de données soit adoptée par les universitaires en sciences humaines.

En contribuant à un outil fédérateur, les historiens cumulent des faits et des opinions divergentes au même endroit. L'utilisateur est alors confronté à une contextualisation provenant de différentes sources, mais aussi aux théories appuyées par certains et réfutées par d'autres. Il devient donc plus simple de générer des statistiques concernant la validation par les pairs entourant différents faits.

Sans passer directement par les données liées, le groupe de recherche interdisciplinaire *Collaborative for Historical Information and Analysis* (CHIA) travaille

[50] Eva Charlebois, Louise Mallet et Julie Méthot, « L'ABC de la révision par les pairs », *Pharmactuel*, 2009, p. 42.

à créer un système fédérant les bases de données historiques[51]. Ce travail d'agrégation de l'information historique, jumelé à la mise en valeur des bases de données en histoire, est au cœur de leur système intitulé *World-Historical Dataverse*[52]. Ce moteur de recherche propose différentes facettes facilitant l'évaluation de la qualité des informations fournie par une base de données. Par exemple, faire une recherche par auteurs ou contributeurs de la base de données assure une certaine validité de l'information. De plus, le chercheur peut cibler précisément un cadre spatio-temporel (première date de la base de donnée, dernière date de la base de donnée, lieux géographiques couverts) ainsi que des sources qui ont permis la constitution de la base de données[53]. La méthodologie derrière ce système repose sur une étude intitulée *Data Hoover Project* (DHP) menée par Ruth Mostern et Marieka Arksey, qui recueillent des données sur la méthodologie des chercheurs lorsqu'ils explorent des jeux de données historiques. De cet échantillon, les deux chercheures en extraient des principes qu'elles présenteront dans un guide des bonnes pratiques prochainement[54].

Cependant, il est actuellement difficile de valider de l'information par l'analyse de consensus de chercheurs spécialistes. Le travail de validation actuel demande un investissement important des études et des sources afin de se rapprocher de l'exactitude d'un phénomène particulier. Si une masse critique d'historiens contribuent de manière coordonnée aux données liées, il sera possible de valider statistiquement différents

[51] Patrick Manning et David Ruvolo, *CHIA - Collaborative for Historical Information and Analysis*, http://www.chia.pitt.edu/index.php, 2016, consulté le 13 janvier 2016.
[52]Collaborative for Historical Information and Analysis, *World-Historical Dataverse*, https://dataverse.harvard.edu/dataverse/worldhistorical, 2015, consulté le 13 janvier 2016.
[53] *Ibid.*
[54]Ruth Mostern et Marieka Arksey, *The Data Hoover Project (DHP) and its aims*, http://www.chia.pitt.edu/datahoover.html, 2015, consulté le 13 janvier 2016.

consensus sans avoir à continuellement revenir à l'information présentée sous forme de texte. Évidemment, cette forme de portail demande des balises claires pour obtenir des résultats significatifs. Le théorème de Condorcet valide cette approche.

En 1785, Nicolas de Condorcet propose le théorème du jury dans son *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*[55]. Condorcet avance que « si la probabilité d'erreur individuelle est inférieure à ½, la probabilité d'erreur majoritaire décroît avec le nombre N de jurés et tend vers 0 quand N tend vers l'infini[56]. » Autrement dit, les chances d'exactitude augmentent avec le nombre d'individus qui se prononcent sur un sujet donné. Alors si on tend vers une infinité de personnes s'exprimant sur un sujet, le pourcentage d'erreur se rapproche de zéro. Évidemment ce théorème s'applique uniquement si les votants sont informés et possèdent tous le même objectif de réussite[57]. Il est donc nécessaire d'attribuer des comptes utilisateurs à des professionnels souhaitant collaborer intelligemment à ce grand chantier de contextualisation par le biais des données liées.

Le projet *The Historian's Macroscope: Big Digital History* de Shawn Graham, Ian Milligan et Scott Weingart, présenté précédemment, ouvre la porte à la validation collaborative, mais sans traiter de l'automatisation à grande échelle d'un tel système. C'est encore une fois le projet CHIA qui propose une solution qui préfigure un modèle qui faciliterait le passage d'une littératie historique classique vers une littératie

[55] Nicolas de Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendus à la pluralité des voix*, s.l., Paris, Imprimerie Royale, 1785, 514 p.
[56] Jean-François Laslier, « La norme majoritaire », *Revue économique*, 1999, vol. 50, n° 4, p. 671.
[57] *Ibid*., p. 671-672.

sémantique. Le système *Col*Fusion* (Collaborative Data Fusion) permet la fusion de différents jeux de données afin de faciliter leur réutilisation[58]. Une recherche de données exige souvent l'association de différentes bases de données afin de répondre à notre questionnement. *Col*Fusion* facilite ce travail en exposant un nombre important de jeux de données livrés par des chercheurs et en les fusionnant les uns aux autres selon la nature des données. Le résultat de notre recherche présenté sous forme de tableau peut donc être un amalgame entre deux ou plusieurs tableaux différents. Ces liaisons peuvent être faites automatiquement par reconnaissance de caractères ou manuellement par un chercheur[59]. Sans en être l'objectif principal, ce système pourrait facilement devenir un outil de croisement de jeux de données similaires afin de constituer un outil de validation de l'information. En regroupant toutes les bases de données traitant d'un même sujet, il deviendrait relativement simple de distinguer les incongruités et d'en faire des constats. Si une grande majorité des chercheurs en histoire utilisaient ce type de système, il serait possible de pouvoir tendre vers une plus grande validité des données historiques partagées par un ensemble de professionnels clairement identifié.

Cette validation pourrait être jumelée à une cote qui évaluerait la pertinence des données et des liens (dans le cas du Web sémantique) proposés par un auteur. Ce grade pourrait devenir un nouvel outil pour juger de la pertinence de projet de recherche qui s'articule autour de la création d'une base de données. Puisque ce type de travaux est peu financé, ne répondant pas aux exigences des organismes subventionnaires, une

[58] Collaborative for Historical Information and Analysis, *Col*Fusion - Your entry to the data world*, http://colfusion.exp.sis.pitt.edu/colfusion/, 2016, consulté le 13 janvier 2016.
[59] *Ibid.*

approche coopérative ouvre la voie à une nouvelle manière de démontrer la pertinence d'une base de données dans le monde de la recherche fondamentale.

Ce chapitre avait pour objectif de démontrer que le Web sémantique contribue au développement de la discipline historique en amenant de nouvelles possibilités d'analyses en résonance directe avec les idéaux de l'histoire moderne. Cet amalgame occasionne nécessairement des changements sous forme de trois paradigmes. L'historien devient un architecte de l'information historique, son travail de liaison s'automatise et les moyens de consultation affectent la littératie historique. Le Web sémantique permet une avancée considérable dans l'atteinte d'une large contextualisation de faits véridiques liés. La question est maintenant de savoir comment s'articulent le Web sémantique et les données liées.

CHAPITRE II : COMPRENDRE LE WEB SEMANTIQUE

Cette deuxième partie définit le Web sémantique dans un contexte historique. Après avoir présenté une série de concepts clés, le lecteur sera exposé à plusieurs projets phares qui allient Web sémantique et histoire. À travers ce parcours qui culminera sur une compréhension générale des possibilités que procurent les données liées, le chercheur sera confronté à différents problèmes qui serviront d'assises à l'argumentaire concernant la nécessité de l'implication historienne dans le processus de création d'une plate-forme sémantique pour l'histoire du Québec.

Web sémantique et données liées

À la recherche d'une définition simple pour les sciences historiques

L'idée du Web sémantique découle d'un texte de Tim Berners-Lee publié en 2001 intitulé *The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*[1]. Dans ce texte plutôt futuriste, l'auteur rappelle que le Web actuel fonctionne en silos. Par exemple, si nous souhaitons modifier tous nos mots de passe, nous devons aller sur chaque plate-forme où nous possédons un identifiant pour faire la modification. On peut faire le même constat avec la donnée que nous souhaitons modifier ou bonifier. La spécificité de chaque page web oblige un travail colossal pour créer des ponts entre les institutions gérant des bases de données. Puisqu'aujourd'hui les institutions culturelles subissent de plus en plus de

[1] Tim Berners-Lee, James Hendler et Ora Lassila, *loc. cit.*

pression pour fournir un accès intégré à leurs collections[2], le Web sémantique est la seule avenue qui permet cette agrégation.

Dans son article, Tim Berners-Lee définit le Web sémantique ainsi : « The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users[3]. » L'élément à retenir est l'importance de la structure des données. Il faut cesser de bâtir des bases de données structurées selon un contenu spécifique, mais bien en pensant à l'ensemble des possibilités de notre champ disciplinaire et même au-delà. Le projet est donc de créer une structure qui permettra aux ordinateurs de désambiguïser la donnée afin d'attribuer un sens à celle-ci; un Web sémantique.

Pour parvenir au Web sémantique, le moyen utilisé massivement est le Web de données (en anglais : *Linked data*). Pour qu'un ordinateur puisse désambiguïser une donnée, il faut lier celle-ci avec une grande quantité d'autres. La création de données liées permettra de mettre en place une structure évolutive puisque chaque individu devient maintenant un possible agent liant. On remarque donc qu'il y a une différence entre Web sémantique et Web de données (données liées). Le premier est un mouvement qui souhaite donner du sens au Web pour libérer les données des pages web et le second est l'outil permettant d'atteindre cet objectif. Tim Berners-Lee utilise à maintes reprises les termes « agent » ou « virtual agent » dans ses écrits sur le Web sémantique. L'agent

[2] Online Computer Library Center, *Library, Archive and Museum Collaboration*, http://www.oclc.org/research/activities/lamsurvey.html, 30 novembre 2011, consulté le 27 août 2015.
[3] Tim Berners-Lee, James Hendler et Ora Lassila, *loc. cit.*

peut être perçu de différentes façons, mais en général il s'agit d'une entité capable de naviguer au sein de cette structure pour en extraire un sens selon différentes directives sélectionnées par l'utilisateur. Dans l'imaginaire collectif, on imagine une représentation humaine comme une voix. Dans cette perspective, quoique limités dans leurs actions, les outils *Siri* et *Ok Google* sont deux exemples de cette agrégation de données.

Toutefois, ce concept est loin d'être concrétisé selon les dires de Ruben Verborg. Ce dernier, dans son article « Fostering intelligence by enabling it », rappelle que les données liées sont maintenant largement utilisées mais il manque un désir de développer un système performant, c'est-à-dire concis, homogène et ouvert aux normes internationales[4]. Puisque le Québec débutera la transformation de ses données en 2016-2017, il est impératif de s'intéresser à l'usage de ses éventuelles données liées.

Comment lier des données?

Les données sont présentement, dans la majorité des cas, compilées dans des tableaux. Les différentes entités et les différents champs forment une matrice à compléter. Cette forme simpliste fonctionne bien à l'interne. Par exemple, avec sa base de données relationnelle, le muséologue peut faire des requêtes simples ou complexes pour lister des objets selon un sujet d'exposition. Le problème avec cette forme relationnelle est qu'il est impossible d'associer des données qui possèdent des normes différentes. De plus, le tableau est une forme rigide qui ne permet pas une grande flexibilité de mise en relation des contenus. Il devient donc extrêmement difficile de

[4] Ruben Verborgh, *Fostering intelligence by enabling it*, http://ruben.verborgh.org/blog/2015/02/25/fostering-intelligence-by-enabling-it/, 25 février 2015, consulté le 27 août 2015.

répondre aux demandes imprévues ou plus complexes des utilisateurs[5]. La privatisation de ces systèmes empêche également le développement d'outil fédérateur à faible coût. Enfin, les bases de données intègrent à la fois des données qui peuvent être diffusées au public et d'autres qui ne peuvent être partagées pour des raisons de confidentialité ou de vie privée (information sur les donateurs ou information de gestion interne). Un système de données liées permet alors de ne diffuser que ce qui est d'intérêt public sous la forme la plus efficace.

C'est à partir de ce constat que le W3C développe le système des données ouvertes et liées cinq étoiles. Pour prétendre faire des données liées dans l'optique du Web sémantique, il faut qu'un système obtienne les cinq étoiles. Le gouvernement du Canada[6] propose une traduction de ces niveaux qui se déclinent ainsi :

Première étoile : fournir des données avec une licence ouverte. Selon *Statistique Canada*, une licence ouverte « contient peu de restrictions quant à la façon dont les données peuvent être utilisées et permet spécifiquement leur distribution ultérieure dans le cadre d'activités commerciales ainsi que non commerciales. Il n'y a aucun frais pour ce genre d'utilisation[7]. » Une licence ouverte peut toutefois imposer des restrictions. Par exemple, elle peut spécifier que la source ne peut pas être modifiée et que son utilisation doit se limiter à des activités non commerciales. Néanmoins, l'utilisation de celle-ci reste toujours gratuite. La question de licence ouverte ne sera que très peu abordée dans

[5] Réseau pancanadien du patrimoine documentaire*, loc. cit.*, p. 1.
[6] Gouvernement du Canada, *Cote de degré d'ouverture des données*, http://ouvert.canada.ca/fr/cote-degre-douvertrure-des-donnees, 25 avril 2013, consulté le 8 janvier 2016.
[7] Gouvernement du Canada et Statistique Canada, *Entente de licence ouverte de Statistique Canada - Foire aux questions (FAQ)*, http://www.statcan.gc.ca/fra/reference/licence-faq-fra#a1, 25 janvier 2013, consulté le 8 janvier 2016.

ce mémoire puisqu'il s'agit d'une étape préliminaire au projet de Web sémantique qui demande une connaissance juridique.

Deuxième étoile : présenter les données dans un format structuré. Cette étape est nécessaire pour qu'un programme puisse lire les données compilées. Les données non structurées (texte libre ou langage naturel) posent d'autres types de défis aux ordinateurs et ajouteraient d'autres couches de complexité à un système sémantique.

Troisième étoile : miser sur des formats ouverts. Par exemple, *Excel* étant un logiciel propriétaire produit par *Microsoft*, il faut le posséder (ou un autre logiciel disposant d'un module de conversion) pour lire les formats *.xls ou .xlsx*. Ainsi, un contenu doit être offert avec un format qui ne sera pas modifié au gré d'une entreprise pour en limiter l'accès et qui peut être lu par divers logiciels.

Quatrième étoile : proposer un *Uniform Ressource Identifier* (URI) pour chaque élément de notre base de données et l'inscrire dans un modèle *Ressource Description Framework* (RDF). À titre d'exemple, chaque donnée d'une cellule d'un chiffrier électronique (sauf pour les dates et les appellations) devrait pointer vers un URI, comme nous le verrons plus loin. Rappelons qu'un ordinateur ne comprend pas le sens des chaînes de caractères. Si deux personnes insèrent le terme « Jacques Cartier » dans leurs bases de données, il n'existe aucun lien entre ces deux chaînes sauf une équivalence de caractères. Encore plus problématique lorsqu'une personne est connue sous différents noms. C'est le cas, par exemple, de Marie Guyart dit Marie de l'Incarnation. Il est alors primordial de structurer l'information afin qu'on associe les différentes appellations de cette Ursuline à la même entité humaine. De cette manière, on peut trouver la

documentation souhaitée que l'on fasse la requête à partir de l'une ou l'autre des appellations. Pour créer un lien, il faut utiliser un URI, un identifiant unique, qui permet de référer au même élément sur le Web. Il peut s'agir d'un concept, d'une idée, d'une personne, d'un événement, d'un objet, d'un document, etc. Si l'on veut identifier « Jacques Cartier » dans notre réseau de données, on se doit d'utiliser un URI qui fait référence exclusivement à l'explorateur. Par exemple, il est possible d'utiliser l'URI proposé dans le *Library of Congress Subject Heading* (LCSH), catalogue qui identifie les différentes composantes des collections de l'institution. Dans ce système, l'explorateur Jacques Cartier est identifié par l'URI http://id.loc.gov/authorities/names/n50080987[8]. Il est aussi possible de créer ses propres URI et les héberger sur un serveur en assurant leur pérennité. De son côté, RDF propose une syntaxe simple et uniforme[9] qui repose sur des triplets. Cette forme permet rapidement et simplement d'expliciter le sens d'un concept, de créer des liens entre eux et s'applique à des contenus de toutes les disciplines[10]. Un triplet RDF prend toujours la forme *Sujet-Prédicat-Objet*. Pour reprendre l'exemple de Jacques Cartier, on pourrait proposer le triplet suivant : *Jacques Cartier-fonction-explorateur*. C'est l'accumulation de triplets autour d'un sujet qui donne du sens à celui-ci. Par exemple, le triplet précédent permet de s'assurer qu'il s'agit de Jacques Cartier l'explorateur et non du « député de la circonscription de Surrey à la Chambre d'assemblée du Bas-Canada de 1804 à 1809[11] »

---

[8] Library of Congress, *Cartier, Jacques, 1491-1557 - LC Linked Data Service*, http://id.loc.gov/authorities/names/n50080987.html, consulté le 27 août 2015.

[9] Ruben Verborgh, *Federated SPARQL queries in your browser*, http://ruben.verborgh.org/blog/2015/06/09/federated-sparql-queries-in-your-browser/, 9 juin 2015, consulté le 28 août 2015.

[10] *Ibid.*

[11] Ministère de la culture et des communications du Québec, *Cartier, Jacques*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=17323&type=pge#.Vd-sQK01uzx, 2013, consulté le 28 août 2015.

qui aurait comme triplet *Jacques Cartier-fonction-député*. Il est aussi possible de différencier les deux hommes par leurs années de naissance et de décès.

Cinquième étoile : utiliser les URI et RDF pour créer des ponts avec d'autres jeux de données. Comme il fut mentionné précédemment, il est plus simple de créer soi-même ses URI au lieu de pointer vers des bases de données externes. Cependant cette méthodologie recrée la forme silo que le Web sémantique tente de briser. L'utilisation de l'URI du LCSH pour identifier Jacques Cartier est la pratique à suivre pour obtenir cette cinquième et dernière étoile. Cet exemple s'inscrit dans une approche sémantique puisqu'on associe une base de données avec celle de la *Library of Congress*. Par la suite, quand une autre institution associe une de ses données avec le même URI, un lien est automatiquement créé entre les données des deux institutions par le biais de cet URI commun.

Les différents niveaux de données liées et diverses technologies associées

Le Web sémantique est une notion large qui peut être adaptée selon les moyens et le temps que possède chaque institution pour développer son projet. Il est possible de fournir sur son site web des fichiers en format *xml/rdf*. Ainsi une personne intéressée à utiliser ces données en RDF sera en mesure de les télécharger. Toutefois, cette étape simpliste ne permet pas une utilisation optimale des données. Pour qu'un programmeur puisse interroger votre système aisément, il est essentiel de déposer vos données dans un *TripleStore* et que ce dernier possède un point d'entrée SPARQL. Un *TripleStore* est un serveur qui permet d'emmagasiner des données RDF efficacement. En fait, un *TripleStore* est conçu pour gérer des triplets de manière efficace et ne peut pas servir à d'autres fins alors que les bases de données relationnelles gèrent différents types de

données selon la structure qui est utilisée. Sur le marché, il existe différents *TripleStores* et faire un choix peut rapidement devenir difficile.

Le joueur incontournable dans ce domaine est *Ontotext* qui développe différentes solutions pour les LAM. Un système qui gère la recherche collaborative, les téléchargements haute performance, les requêtes dans plusieurs répertoires à la fois, la réutilisation créative des ressources, la création de schémas de connaissances, l'agrégation des artéfacts pour les contextualiser, l'ouverture des données institutionnelles et la création des applications de recherche[12]. *Ontotext*, une solution payante, a contribué au développement des données liées du *British Museum*, du *Yale Center for British Art* et d'*Europeana*. Le marché regorge de multiples solutions en libre accès. En humanités numériques, on utilise souvent *Sesame* ou *Jena*. Le premier est développé par plusieurs partenaires, dont *Ontotext*[13], et le second fut lancé par un groupe de chercheurs de *HP Labs*[14]. En 2012, Thomas Francart, consultant en Web sémantique, affirme qu'il existe deux grandes écoles en ce qui concerne les *TripleStores*: adaptabilité de la plate-forme ou la convivialité de celle-ci. Pour sa part, il préfère *Sesame* puisque son fonctionnement, quoique moins performant que Jena, est plus intuitif[15]. Toutefois, le choix entre les deux demeure une question de goût. Évidemment, il existe plusieurs

[12] Ontotext, *Ontotext Semantic Solutions for Cultural Heritage*, http://ontotext.com/ semantic-solutions/galleries-libraries-archives-museums/, 2015, consulté le 28 août 2015.
[13] Sesame, *Contributors*, http://rdf4j.org/contributors.xhtml?view, 25 août 2015, consulté le 28 août 2015.
[14] Apache Jena. *What is Jena?*, http://jena.apache.org/about_jena/about.html, 2015, consulté le 28 août 2015.
[15] Thomas Francart, *RDF : Sesame, Jena, comparaison des fonctionnalités*, http://blog.sparna.fr/2012/05/08/rdf-sesame-jena-comparaison-des-fonctionnalites/, 8 mai 2012, consulté le 28 août 2015.

autres *TripleStores* sur le marché comme celui de *Virtuoso*[16] qui soutient *DBpedia* un vaste système de données liées qui sera présenté dans la prochaine partie.

Quant au point d'entrée SPARQL, il s'agit d'un moyen d'interagir avec la base de données. Il est donc possible de sélectionner des informations, poser des questions simples se répondant par vrai ou faux, créer des graphes de données, insérer, effacer et mettre à jour les données[17]. La syntaxe SPARQL étant tout aussi complexe que la syntaxe SQL pour des néophytes, des outils plus conviviaux seront nécessaires pour que ces systèmes soient plus facilement accessibles et interrogeables. Un exemple de la syntaxe SPARQL est présenté à l'Annexe I.

La majorité si ce n'est la totalité des institutions culturelles possèdent déjà des bases de données relationnelles lorsqu'elles souhaitent contribuer au nuage des données ouvertes et liées[18]. Ce processus de transformation nécessite la mise en place d'une stratégie, mais aussi le recours à différents outils. La liste qui suit n'est pas exhaustive, mais présente des outils libres et gratuits qui facilitent le travail d'un chercheur en sciences humaines qui souhaite produire des données ouvertes et liées.

La gestion de données dans un objectif de sémantisation doit nécessairement passer par une analyse globale et *OpenRefine* est l'outil tout désigné pour cette tâche. Développé par *Google*, mais maintenant disponible sous licence libre, *OpenRefine*

[16] OpenLink Software, *Virtuoso Universal Server*, http://virtuoso.openlinksw.com/, consulté le 13 avril 2016.
[17] World Wide Web Consortium, *SPARQL Query Language for RDF*, http://www.w3.org/TR/rdf-sparql-query/, 15 janvier 2008, consulté le 28 août 2015.
[18] Richard Cyganiak et Anja Jentzsch, *The Linking Open Data cloud diagram*, http://lod-cloud.net/, 2014, consulté le 31 octobre 2014.

permet d'effectuer trois grandes étapes obligatoires dans le processus de liaison. Premièrement, il permet d'explorer efficacement nos données. On peut regrouper nos données selon différentes facettes et ainsi repérer facilement les incongruités ou les faiblesses de celles-ci. Deuxièmement, il permet de nettoyer et de transformer nos données. Après l'étape de visualisation, il est inévitable de remarquer certaines erreurs dans la structure de nos données. Il est alors souhaitable d'uniformiser notre contenu et de créer des regroupements. Une des bonnes pratiques est de réduire à la plus petite unité informationnelle dans une cellule. Par exemple, il est préférable de présenter le concept « Jacques Cartier » en deux parties. Au lieu de regrouper les deux mots sous le champ « nom », il faut avoir les champs « prénom » et « nom ». Dans le même ordre d'idées, les champs « description », quoique souvent utiles pour synthétiser, ne doivent pas remplacer d'autres champs plus précis qui faciliteraient la sémantisation du contenu. Dernièrement, une quantité importante de champs vides peut causer différents problèmes d'interopérabilité.

Une fois le nettoyage et la réorganisation des données effectués, *OpenRefine* permet d'associer des données. À l'aide de différents logiciels de *Named-Entity Recognition* (NER), *OpenRefine* analyse des chaînes de caractères et identifie des URI susceptibles de les remplacer[19]. *AlchemyAPI*, *DBpedia Spotlight* et *Zemanta*, trois NER, ont été adaptés par le *Multimedia Lab* de l'*Université de Ghent* et *iMinds* en collaboration avec le groupe *MaSTIC* de l'*Université Libre de Bruxelles* afin d'être

[19] David Nadeau et Satoshi Sekine, « A survey of named entity recognition and classification », *Named Entities: Recognition, classification and use*, 2007, vol. 30, n° 1, (coll. « Lingvisticæ Investigationes »), p. 3.

compatibles avec *OpenRefine*[20]. L'efficacité de *DBpedia Spotlight* sera évaluée au chapitre trois à partir des données du RPCQ. Ces outils automatisent le processus de création de données liées pour éviter un travail d'association manuel qui serait impossible à réaliser sans disposer d'une grande équipe d'assistants de recherche.

La dernière étape consiste à associer nos données avec des structures existantes. *Karma Tool* développé par Craig Knoblock et Pedro Szekely de l'*Université de Southern California* permet de transformer des tableurs en réseau de liens suivant les principes du RDF. La prise en main du programme est rapide puisque l'interface est simple et compréhensible. Un des projets phares de l'équipe de recherche associé au programme concerne l'adaptation d'une collection d'œuvres d'art de la *Smithsonian Institution*[21].

Standards hiérarchisés

Standards de valeurs

Il existe des URI pour toutes les parties d'un triplet. Dans un premier temps, il sera question des URI des sujets et des objets. On réfère alors à des fichiers d'autorité. Puisque l'objectif principal du Web sémantique est de créer de la cohésion et du sens par des liaisons entre les données, l'architecte de l'information historique se doit de créer des ponts entre sa base de données et d'autres acteurs du nuage des données liées. Un des problèmes majeurs du Web sémantique est la multiplication des URI pour une même entité. Évidemment, il est plus simple de créer son propre URI pour Jacques Cartier que

[20] Ruben Verborgh, *Commentaire sur OpenRefine Named-Entity Recognition extension*, https://groups.google.com/forum/#!topic/openrefine/jeNxqeWo9Rg, 20 décembre 2012, consulté le 14 janvier 2016.
[21] Craig Knoblock et Pedro Szekely, *Karma: A Data Integration Tool*, http://usc-isi-i2.github.io/karma/, 2015, consulté le 6 octobre 2015.

de chercher ceux qui existent déjà. Toutefois, cette pratique laborieuse, tout en permettant d'atteindre les cinq étoiles, inscrit directement vos données dans le nuage partagé et peuvent répondre à des requêtes fédérées. Il existe plusieurs acteurs qu'un historien se doit de consulter pour associer ses données à des URI existants.

Au cœur du Web sémantique se trouve *DBpedia*. Celui-ci est une version des articles de *Wikipedia* sous forme de triplets[22]. Une large communauté travaille à l'automatisation de la transformation de l'information relationnelle vers des triplets. *DBpedia* est actuellement l'acteur central du Web sémantique selon la représentation de Richard Cyganiak et Anja Jentzsch[23], puisqu'il s'agit du plus large ensemble de données et de la base qui possède le plus de liens entrants. Rapidement, on peut trouver l'URI *DBpedia* de Jacques Cartier grâce à une simple recherche classique[24]. L'interface de visualisation de *DBpedia* n'a pas été améliorée depuis 2007[25], ce qui occasionne divers problèmes lors de la navigation. Par exemple, une image de Jacques Cartier sera intégrée à la page de ce dernier sous forme de lien et non en tant que fichier *jpg*. La formule en triplets se veut facile à lire pour l'ordinateur et pour l'humain, par contre le cumul de ceux-ci sur *DBpedia* nuit à la compréhension rapide de la plate-forme. Évidemment, *DBpedia* ne se veut pas une plate-forme grand public, mais son interface peu conviviale peut devenir un frein à la participation des historiens dans le processus de réflexion entourant les données liées. Toutefois, il existe des outils pour contrer cette

[22] DBpedia, *About | DBpedia*, http://dbpedia.org/about, 2014, consulté le 28 août 2015.
[23] Richard Cyganiak et Anja Jentzsch, *loc. cit.*
[24] DBpedia, *About: Jacques Cartier*, http://dbpedia.org/page/Jacques_Cartier, consulté le 28 août 2015.
[25] Denis Lukovnikov, Claus Stadler, Dimitris Kontokostas, Sebastian Hellmann et Jens Lehmann, « DBpedia Viewer - An Integrative Interface for DBpedia Leveraging the DBpedia Service Eco System », *Workshop on Linked Data on the Web*, Seoul, 2014, vol. 1184, http://ceur-ws.org/Vol-1184/ldow2014_paper_05.pdf, consulté le 28 août 2015.

problématique comme *Lodview*[26] qui crée une interface plus simple à comprendre pour les chercheurs en sciences historiques. Par contre, *DBpedia* étant un joueur central, il est essentiel de faire référence à ce dernier, mais il ne faut pas négliger d'autres joueurs plus près de la sensibilité historienne. C'est le cas de la *Library of Congress* dont il a été question précédemment. À titre de bibliothèque nationale des États-Unis, l'institution maintient à jour son catalogue de vocabulaires et d'autorités, comme le LCSH par exemple, depuis 1898[27].

Un autre joueur américain important est le *Getty Research Institute* qui propose des taxonomies sous forme d'URI. Une taxonomie est un vocabulaire hiérarchisé qui permet d'identifier un concept dans son réseau de liens parents-enfants. Parmi leurs outils, on retrouve premièrement l'*Art & Architecture Thesaurus* (AAT) qui permet d'identifier avec précision des concepts d'art et d'architecture. Grâce à ce vocabulaire hiérarchisé, un historien pourra trouver rapidement un URI de référence pour le terme « pilastres » (en anglais : *pilasters*) et une définition qui l'empêchera de le confondre avec « colonnes ». Deuxièmement, le *Thesaurus of Geographic Names* (TGN) permet d'obtenir un URI pour un lieu. En plus de fournir les coordonnées de celui-ci, le thésaurus présente les termes équivalents, qu'il s'agisse de traductions ou d'anciennes dénominations[28]. Troisièmement, on retrouve l'*Union List of Artist Names* (ULAN) qui établit un standard pour le nom des personnes associées au monde de l'art et de l'architecture. Finalement, la *Getty Research Institute* propose le *Cultural Objects Name*

[26]World Wide Web Consortium, *LodView*, http://www.w3.org/2001/sw/wiki/LodView, 22 décembre 2014, consulté le 28 août 2015.
[27] Library of Congress, *Library of Congress Subject Headings - LC Linked Data Service*, http://id.loc.gov/authorities/subjects.html, consulté le 28 août 2015.
[28] Getty Research Institute, *TGN: Frequently Asked Questions*, http://www.getty.edu/research/tools/vocabularies/tgn/faq.html, 1 juin 2015, consulté le 28 août 2015.

Authority (CONA). Ce vocabulaire permet d'identifier les œuvres d'art et les constructions architecturales.

Les standards pour les prédicats : les ontologies

Dans un triplet, la partie centrale, le prédicat, permet d'associer le sujet et l'objet. Tout comme les deux autres parties, le prédicat se doit d'être normalisé pour créer des ponts entre les différentes bases de données. Le domaine des sciences de l'information regorge de modèles d'identification. Par exemple, le *Dublin Core* est la méthode de standardisation la plus utilisée actuellement et sa création remonte à 1995 lors d'une séance d'atelier réunissant plus de 200 professionnels de disciplines diverses dans le but d'établir une liste des métadonnées permettant de structurer tous types de ressources (en particulier les livres et les archives)[29]. Cependant, Roger King affirme que la simplicité du modèle empêche *Dublin Core* de remplir sa fonction primaire : « One drawback of the Dublin Core is that it is very loosely defined. So, it often fails in its true purpose : to provide precisely-defined terms that all of us can use, and where we can be confident they will be uniformly interpreted[30]. » Les champs proposés se présentent ainsi sous forme de liste sans hiérarchie, mais cette dernière permet de circonscrire efficacement le domaine du patrimoine matériel. Les quinze champs principaux sont : titre, créateur, sujet, description, éditeur, contributeur, date, type de ressource, format, identifiant, source, langue, relation, couverture et droits[31]. L'ensemble de ces champs est maintenant disponible sous forme d'URI qui permet la validation et la précision des contenus.

---

[29] Marcia Lei Zeng, *Metadata*, New York, Neal-Schuman Publishers, 2008, p. 6.
[30] Roger King, *The Dublin Core and the Metadata Object Description Schema: a look at namespaces*, http://itknowledgeexchange.techtarget.com/semantic-web/the-dublin-core-and-the-metadata-object-description-schema-a-look-at-namespaces/, 2009, consulté le 9 novembre 2014.
[31] Dublin Core Metadata Initiative, *Dublin Core Metadata Element Set, Version 1.1*, http://dublincore.org/documents/dces/, 14 juin 2012, consulté le 28 août 2015.

Les modèles évoluent et migrent rapidement dans le monde des sciences de l'information. Actuellement le modèle *Functional Requirements for Bibliographic Records* (FRBR) et ses extensions prennent une place importante dans la migration des systèmes classiques vers des systèmes sémantiques[32], sans parler du modèle *Ressource : Description et Accès* (RDA). Cependant, ces classifications limitent les usages entourant le patrimoine immatériel ou l'histoire elle-même. Il existe aujourd'hui plusieurs autres modèles plus diversifiés qui permettent de mieux identifier certains éléments. Cette diversité permet de mieux représenter des domaines d'application comme l'histoire.

Lorsque notre vocabulaire est hiérarchisé et que les liens entre les concepts sont explicités, il est question d'une ontologie. Plus précisément, il s'agit de « l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances[33]. » Par exemple, si l'on s'intéresse aux relations de Jacques Cartier, il serait possible de placer sous sa propre personne Jacques Noël et Thomas Fromont[34]. Par contre, les relations entre Jacques Cartier et ces deux personnes sont différentes. Le premier est le neveu de l'explorateur tandis que le second est l'assistant de celui-ci sur la Grande Hermine. Une ontologie permet précisément d'effectuer la distinction entre des liens familiaux et des liens professionnels. Jacques Cartier serait associé à Jacques Noël par le triplet *Jacques Cartier – oncle de – Jacques Noël* et à Thomas Fromont par *Jacques Cartier – supérieur de – Thomas Fromont*. L'ontologie *Friend of a Friend*

[32] Ian Davis et Richard Newman, *Expression of Core FRBR Concepts in RDF*, http://vocab.org/frbr/core.html , 15 mai 2009, consulté le 28 août 2015.
[33] Wikipedia, *Ontologie (informatique)*, https://fr.wikipedia.org/wiki/Ontologie_ (informatique), consulté le 13 avril 2016.
[34] Université Laval et Université de Toronto, *Biographie – CARTIER, JACQUES (1491-1557)*, http://www.biographi.ca/fr/bio/cartier_jacques_1491_1557_1E.html, 2015, consulté le 28 août 2015.

(FoaF), développée par Dan Brickley et Libby Miller ainsi qu'une grande communauté de membres, permet de décrire des personnes et de créer un réseau entre celles-ci. La construction du réseau se réalise grâce au prédicat « connait » (en anglais : *knows*). Notons au passage que la notion « connait » est large et peut porter à confusion. Il faut lire la documentation qui accompagne l'ontologie pour éviter les faux liens. Par exemple, il est possible que Jacques Cartier connaisse le philosophe et écrivain Pétrarque de par ses écrits, mais le triplet *Jacques Cartier-knows-Pétrarque* est faux. Il est indiqué, dans la description de ce prédicat, que la relation doit être réciproque soit que Pétrarque connaisse aussi Jacques Cartier ce qui n'est évidemment pas le cas ici[35].

Lors du *Linked Open Data in Libraries Archives and Museums 2015* (LODLAM 2015) qui se tenait les 29 et 30 juin 2015 à Sydney en Australie, une discussion autour de la qualité des données liées a soulevé le manque d'intérêt des catalogueurs pour la documentation entourant les ontologies[36]. Une des solutions à ce problème est de gérer le type de données que le champ peut accueillir de manière à la fois englobante, mais précise. Par exemple, s'il souhaite compléter le champ « enfants » de Jacques Cartier, le catalogueur est sujet à une erreur classique. Puisque Jacques Cartier n'a pas eu d'enfants, on serait porté à inscrire « 0 » dans le champ « enfants ». Toutefois, cette pratique peut causer de graves problèmes de lecture des données. *Lodview*, présenté plus haut, permet de créer des arbres généalogiques automatiquement

[35] Dan Brickley et Libby Miller, *FOAF Vocabulary Specification*, http://xmlns.com/foaf/spec/#term_knows, 14 janvier 2014, consulté le 28 août 2015.
[36] Valentine Charles, *Data quality, validation, round-tripping*, Sydney, 2015, https://docs.google.com/document/d/16lcPBy1Cx4AKjLoTvlFtedDbr2fh0_NMr50PsnVN80c/edit?pli=1, consulté le 28 août 2015.

grâce aux données liées. Dans le cas illustré, si l'arbre de Jacques Cartier était créé, nous obtiendrions un enfant du nom de « 0 ».

En plus de connaître la méthodologie sous-jacente d'une ontologie, il faut aussi savoir laquelle est la plus adaptée à notre situation. Il existe en effet plusieurs ontologies qui se recoupent. FoaF est intégré par exemple dans BIO, un « vocabulary for describing biographical information about people, both living and dead[37]. » Bien que moins de liens pointent vers BIO, d'un angle historique il est plus performant que FoaF puisqu'on peut structurer la vie et les relations d'une personne sous forme d'événements. Par exemple, la relation entre Jacques Cartier et Catherine Des Granches[38] pourrait être représentée comme un intervalle entre le mariage et la mort de l'explorateur. Un tableau comparatif, à l'Annexe II, permet de constater le niveau de précision de BIO par rapport à FoaF, tout en exprimant les triplets nécessaires pour compiler ces informations sur Jacques Cartier.

Le concept d'événement est très prometteur pour le travail historique, mais il faut embrasser plus large que la biographie puisque l'historien est confronté à d'autres types de sources de différentes institutions comme les archives et les objets muséaux. C'est le domaine muséologique qui fournit un vocabulaire favorisant le plus l'interdisciplinarité des LAM. L'*International Council of Museums* (ICOM), par sa branche documentaliste nommée le *Comité international pour la documentation* (CIDOC) présente son

---

[37] Ian Davis et David Galbraith, *BIO: A vocabulary for biographical information*, http://vocab.org/bio/0.1/.html, 14 juin 2011, consulté le 28 août 2015.

[38] Louis-Marie Le Jeune, « Jacques Cartier » dans *Dictionnaire Général de biographie, histoire, littérature, agriculture, commerce, industrie et des arts, sciences, mours, coutumes, institutions politiques et religieuses du Canada*, Ottawa, Université d'Ottawa, 1931, vol. 1, p. 313-314, http://faculty.marianopolis.edu/c.belanger/QuebecHistory/encyclopedia/JacquesCartier-NaissancejeunesseetmariageHistoiredelaNouvelle-France.htm, consulté le 28 août 2015.

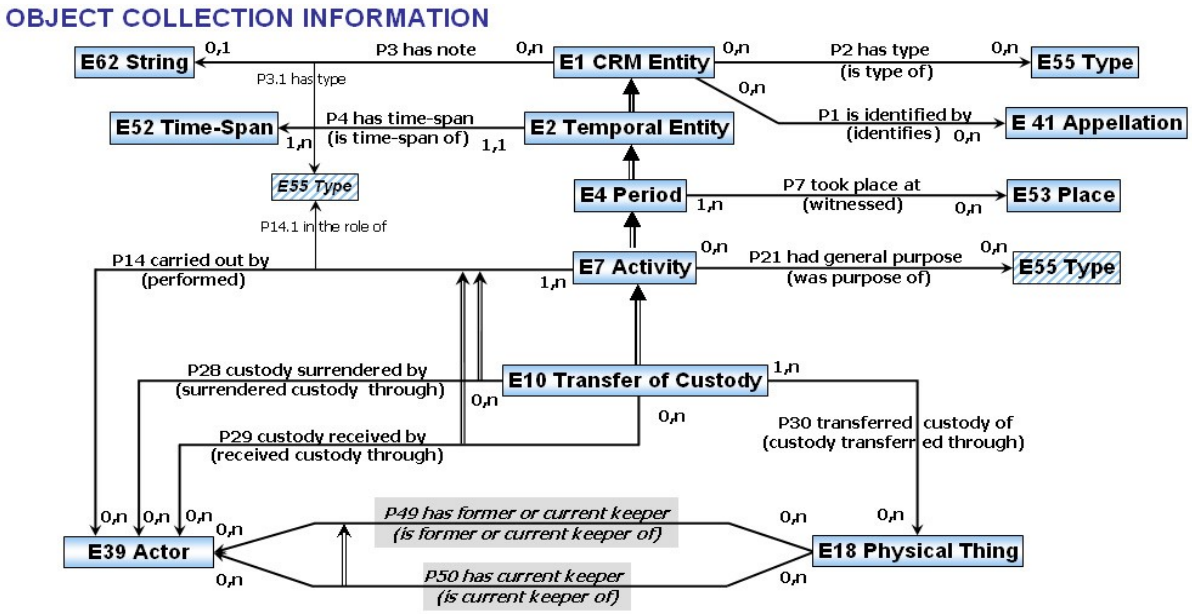
Conceptual Reference Model (CIDOC CRM). La définition de ce vocabulaire permet rapidement de constater son désir de réunir l'ensemble des institutions culturelles à caractère patrimonial :

> The CIDOC CRM is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. It is intended to be a common language for domain experts and implementers to formulate requirements for information systems and to serve as a guide for good practice of conceptual modelling. In this way, it can provide the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives.[39]

La structure s'appuie sur de courts codes débutant par les lettres « E » (pour entité) et « P » (pour propriété) suivies d'un ou plusieurs chiffres qui se rapportent à une courte définition. Rapidement, la conceptualisation d'un phénomène simple peut devenir complexe. Par exemple, la description d'un objet dans une collection muséale met en relation un nombre important d'entités suivant ce diagramme :

[39] International Council of Museums, *The CIDOC CRM*, http://cidoc-crm.org/, décembre 2014, consulté le 26 mars 2015.

Figure 2.1: Représentation hiérarchique d'un extrait du CIDOC CRM concernant des informations sur une collection d'objets



Source : International Council of Museums, *Object Collection Information*, http://www. cidoc-crm.org/ cidoc_graphical_representation_v_5_1/object_collection.html, consulté le 14 janvier 2016.

E18 Physical Thing (Objet physique) représente l'objet en question qui est mis en relation avec *E10 Transfert of Custody* (Transfert de garde) qui constitue une activité (*E7 Activity*) qui s'inscrit dans une période temporelle (*E4 Period* et *E2 Temporal Entity*). Cette période temporelle doit être balisée par une durée (*E52 Time-Span*) et cette dernière a eu lieu à un endroit précis (*E53 Place*). Précisons que les flèches à traits doubles représentent des relations de subsomption. C'est-à-dire que la classe où débute la flèche est la sous-classe de celle où arrive la flèche. Ce schéma est intéressant pour l'histoire puisqu'un processus (construction d'un immeuble, édition d'un livre, naissance d'une personne, catalogage d'une archive, etc.) est toujours associé à une date précise ou à un intervalle de temps. Un projet en cours de réalisation par Jean-Baptiste Pressac, analyste de bases de données au *Centre de recherche bretonne et celtique*, démontre toute la complexité de compréhension du modèle causée par l'héritage par les classes

inférieures des caractéristiques des classes supérieures[40]. Devant ses questions, Gauthier Poupeau, architecte de données à l'*Institut national de l'audiovisuel* et spécialiste du Web sémantique, lui conseille de retourner vers BIO pour faciliter son travail d'organisation puisque ce modèle répond aux besoins de Pressac[41]. Les historiens devraient participer à ce type de discussion afin de mieux baliser les modèles CIDOC CRM et BIO puisque ceux-ci s'appuient sur le concept d'événement pour structurer les données. La notion de temps, centrale dans la méthodologie historienne, est donc intrinsèquement liée au développement structurel de ces modèles. L'historien se doit de savoir quelle ontologie est la plus efficace selon les données à décrire.

L'implication historienne est d'autant plus importante étant donné que la gestion du temps est une problématique qui suscite actuellement des discussions entre les utilisateurs du modèle CIDOC CRM. À titre d'exemple, il est difficile de créer des passerelles sémantiques entre la définition du temps selon le CIDOC CRM et celle des ontologies fédératrices présentées dans la prochaine partie. La meilleure réponse de Georg Hohmann et Martin Scholz à cette problématique est de laisser la valeur attribuable au temps (heure, journée, année, etc) comme étant un type de données indéfini[42]. Ce manque de formalisation peut devenir inquiétant pour l'historien qui appuie la totalité de sa méthodologie sur la mise en relation d'une donnée avec son cadre temporel. Son implication dans la réflexion ontologique est donc plus que nécessaire.

---

[40] Jean-Baptiste Pressac, *Sémantiser une base de données relationnelle (1er épisode)*, http://bylg.hypotheses.org/96, 20 avril 2015, consulté le 28 août 2015.

[41] Gauthier Poupeau, *Commentaire sur l'article « Sémantiser une base de données relationnelle (1er épisode) » de Jean-Baptiste Pressac*, http://bylg.hypotheses.org/96, 23 avril 2015, consulté le 14 janvier 2016.

[42] Georg Hohmann et Martin Scholz, « Recommendation for the representation of the primitive value classes of the CRM as data types in RDF/OWL implementations », http://erlangen-crm.org/docs/crm-values-as-owl-datatypes.pdf, 24 février 2011, p. 1.

Les ontologies fondatrices

Pour qu'une ontologie soit pleinement compréhensible par les ordinateurs, il faut que sa structure soit exprimée en triplets. La force d'une structure RDF est d'utiliser des triplets autant pour l'ajout de données suivant une ontologie que pour expliquer les relations entre les entités et les prédicats. Pour effectuer cette seconde partie, il faut se tourner vers les premiers standards développés par le W3C dont le vocabulaire est présentement en anglais. Nous décrirons trois composantes essentielles du Web sémantique.

- RDFS : définition de la syntaxe de base et de son schéma
- OWL : définition des logiques de description afin de préciser la teneur des données
- SKOS : définition des relations hiérarchiques entre les données.

Le tout premier, développé à partir de 1999 et finalisé en 2004, se nomme *RDF Schema* (RDFS). Ce vocabulaire technique permet de développer certaines règles pour associer différents triplets entre eux suivant ainsi le fonctionnement hiérarchique du format *xml* duquel il est tiré. Seth van Hooland et Ruben Verborgh présentent, dans leur ouvrage *Linked Data for Libraries, Achives and Museums : How to Clean, Link and Publish Your Metadata*, différents exemples pour exprimer simplement les possibilités que permet le RDFS[43].

[43] Seth van Hooland et Ruben Verborg, *Linked data for libraries, archives and museums: how to clean, link and publish your metadata*, US-Ed., Chicago, Neal-Schuman, 2014, 254 p.

Avant de s'attaquer à un de leurs exemples d'utilisation de RDFS, rappelons deux principes techniques qui sous-tendent cette structure ontologique. Premièrement, la syntaxe utilisée permet d'abréger les références par l'utilisation d'un préfixe, équivalent d'une référence courte dans une note infra-paginale. L'URI pour exprimer le prédicat « créateur » de Dublin Core est http://purl.org/dc/elements/1.1/creator et celui de la date est http://purl.org/dc/elements/1.1/date. Pour éviter de recopier la partie commune de l'URI (http://purl.org/dc/elements/1.1/) à chaque utilisation, les préfixes (ou noms de domaines, en anglais: *namespace*) sont énumérés au début du fichier. Cette approche peut être considérée comme équivalente au style bibliographique APA, défini par l'*American Psychological Association*, lequel présente les références complètes en bibliographie et des références abrégées dans le texte[44]. Deuxièmement, les triplets eux-mêmes peuvent s'écrire de manière abrégée. La plus courante, soit le modèle *Turtle*, est simple à comprendre dès le premier contact[45]. À la manière du *ibid.* utilisé dans les notes infra-paginales ou des cinq tirets dans les bibliographies, les triplets qui décrivent le même sujet n'ont pas à répéter ce dernier au complet. Il est convenu de mettre une tabulation afin de faciliter la lecture et l'alignement ainsi produit des prédicats indique qu'ils sont associés au sujet précédent. Ces deux principes permettent de lire de simples fichiers RDF comme celui présenté par Hooland et Verborgh :

[44] American Psychological Association, *Publication manual of the American Psychological Association*, 6e ed., Washington, DC, American Psychological Association, 2010, 272 p.
[45] Seth van Hooland et Ruben Verborg, *op. cit.*, p. 47.

Figure 2.2: Exemple de la syntaxe des triplets RDF

```
@prefix ex: < http://example.org/ontology# >.
@prefix rdf: < http://www.w3.org/1999/02/22-rdfsyntax-ns# >.
@prefix rdfs: < http://www.w3.org/2000/01/rdf-schema# >.

ex:hasWritten rdf:type rdf:Property;
              rdfs:domain ex:Person;
              rdfs:range ex:LiteraryWork;
              rdfs:subPropertyOf ex:hasAuthored.

:HermanMelville ex:hasWritten :MobyDick.

:HermanMelville a ex:Person.
:MobyDick a ex:LiteraryWork.
:HermanMelville ex:hasAuthored :MobyDick.
```

Source: Seth van Hooland et Ruben Verborgh, *Linked data for libraries, archives and museums: how to clean, link and publish your metadata*, US-Ed., Chicago, Neal-Schuman, 2014, p.126-127.

La partie du haut permet à l'ordinateur d'identifier les vocabulaires qui seront utilisés dans le fichier – lui permettant de s'y référer pour valider la structure du document – et de nommer les préfixes (ex, RDF et RDFS). Cet exemple renvoie d'abord à une ontologie fictive qui se nomme *example*, qui permettrait ici de décrire des œuvres littéraires et les personnages qui y sont associés. L'ontologie *example* est pour sa part construite en respectant la syntaxe des méta-ontologies RDF et RDFS. Les triplets permettent de définir l'URI « hasWritten » (en français: *a écrit*), crucial dans la description d'une œuvre. Le premier triplet utilise le prédicat « rdf:type » afin de définir « hasWritten » comme étant lui-même un prédicat dans cette syntaxe. L'élément « rdfs:domain » définit les classes possibles pour les sujets de ce prédicat, soit des personnes. Le « rdfs:range » limite pour sa part les valeurs possibles de la partie objet du triplet à des œuvres littéraires. Le dernier des quatre triplets souligne que « hasWritten » est une sous-propriété de « hasAuthored » (en français: *est l'auteur de*). À partir de cette structure minimaliste, un ordinateur peut générer des triplets automatiquement à partir d'une seule déclaration. À partir du triplet *:HermanMelville ex:hasWritten :MobyDick*,

l'ordinateur saura qu'Herman Melville est une personne, que Moby Dick est une œuvre littéraire et que Herman Melville en est l'auteur. Les triplets suivants découlent alors du premier. La création d'une ontologie et l'expression de concepts sous forme de triplets est une opération intellectuelle qui intéressera probablement fort peu d'historiens. Il est toutefois important de comprendre comment ces structures sont élaborées afin de s'assurer qu'elles correspondent aux besoins des historiens.

Le langage de représentation des connaissances nommé *Web Ontology Language* (OWL) recommandé par le W3C en 2004, élargit les possibilités d'association entre les triplets. OWL présente des caractéristiques d'égalité, de restrictions, de versions et d'annotations[46]. Parmi la liste de nouveaux prédicats, *owl:sameAs* est une fonction très utilisée puisqu'elle permet de spécifier que deux URI différents représentent la même entité.

La plus récente ontologie de base est devenue une recommandation du W3C en 2009, soit le *Simple Knowledge Organization System* (SKOS). Cette structure permet de construire facilement des systèmes de classification qui demandent la création de listes hiérarchiques. Les ontologies thématiques, comme les vocabulaires du *Getty Research Institute*, décrivent bien des situations particulières alors que SKOS vise plutôt à associer des termes de manière plus générale grâce à des prédicats comme *skos:broader* qui permet d'affirmer que le sujet est dans une catégorie plus englobante que l'objet auquel

[46] World Wide Web Consortium, *OWL Web Ontology Language Overview*, http://www.w3.org/TR/owl-features/, 10 février 2004, consulté le 28 août 2015.

il est associé[47]. Au final, les ontologies se chevauchent et se complètent et il existe une infinité de combinaisons possibles. Il est donc important de définir les besoins du projet avant de choisir le bouquet d'ontologies pertinent pour exprimer un corpus de données. Les études de cas permettront de mieux comprendre l'importance des choix ontologiques et les possibilités d'applications que permet le Web sémantique.

Analyse d'études de cas

CLAROS et le* British Museum *: Cas pratiques de CIDOC CRM

Encore aujourd'hui, peu de systèmes peuvent se vanter d'utiliser CIDOC CRM comme ontologie de classification. Parmi les rares cas, on compte CLAROS et le *British Museum*. CLAROS est une fédération internationale de recherche interdisciplinaire qui utilise les dernières technologies pour rendre l'art accessible. Dirigé par l'*Oxford e-Research Centre* (OeRC), ce projet débute en 2000, soit au commencement du développement des technologies entourant le Web sémantique. Le projet semble moins actif, leur dernier article de blogue datant du 23 mai 2011[48], mais cette équipe de recherche a rendu accessible une grande partie de leur démarche technique qui permet de comprendre rapidement le fonctionnement de leur portail.

Le contenu de CLAROS étant décentralisé, le travail du centre et de ses partenaires devait être judicieusement réparti et en symbiose. Pour qu'un contenu apparaisse sur CLAROS, l'organisation partenaire doit respecter trois règles. Premièrement, chaque élément doit être associé à un URL unique, ce qui permet à

[47] World Wide Web Consortium, *Introduction to SKOS*, https://www.w3.org/2004/02/skos/intro, février 2004, consulté le 26 janvier 2016.
[48] Alexander Dutton, *Constraining the CLAROS SPARQL endpoint*, https://clarosdata.wordpress.com/, 23 mai 2011, consulté le 28 août 2015.

CLAROS de pointer directement vers le site web du partenaire. Deuxièmement, les institutions déterminent les différentes licences à accorder selon les contenus et associent ces derniers avec les termes de CIDOC CRM. Finalement, la base de données doit permettre d'extraire des déclarations RDF en format *xml*[49]. La structure de données utilisée est perçue comme un travail en progression et donc adaptable à différentes situations selon les réalités des partenaires.

Malgré ses qualités, CLAROS souffre d'un problème récurrent dans le monde des données liées, voire contraire à la philosophie du Web sémantique, soit l'insistance à contrôler l'ensemble à l'interne, incluant le rendu des données. La plate-forme n'accueille que des partenaires ciblés, empêchant ainsi le plein potentiel d'agrégation que permettent les données liées.

La description du Web sémantique sous-entend la transformation du Web en une énorme base de données où tout serait lié. Cette idée novatrice repose sur l'utilisation de standards internationaux comme ceux de la *Getty Research Institute*. Tout en s'appuyant sur les principes du Web sémantique, le premier objectif de CLAROS s'inscrit plutôt dans le Web traditionnel, soit de créer un portail qui permet de réunir au même endroit plusieurs collections pour en dégager la cohérence. CLAROS a souhaité ouvrir un peu son système en utilisant le standard *Geonames*[50] afin de géolocaliser les collections CLAROS sur *Google Maps*. S'imposant en soi comme un standard, CLAROS permet

[49]CLAROS, *Principles of CLAROS data extraction*, http://www.clarosnet.org/XDB/ASP/claroshome/technicalData.html, consulté le 14 janvier 2016.
[50] Christophe Boutreux, *GeoNames*, http://www.geonames.org, consulté le 26 janvier 2016.

alors à son corpus d'être interrogé par de simples programmes conçus pour arpenter le nuage de données liées.

Ce problème d'ouverture au sens large se retrouve aussi dans le projet de données liées du *British Museum*. Cette institution a simplement transposé ses données classiques vers des données RDF en souhaitant une harmonisation future :

> It provides access to the same collection records available through the Museum's web presented Collection Online, but in a computer readable format. The use of the W3C open data standard, RDF, allows the Museum's collection data to join and relate to a growing body of linked data published by other organizations around the world interested in promoting accessibility and collaboration[51].

La fiche de l'œuvre *Hoa Hakananai'a* représente bien le modèle sémantique du musée[52]. Les URI des objets pointent pratiquement tous vers l'ontologie propre au musée, la *British Museum Ontology* (BMO). Par exemple, l'œuvre présentée ci-haut a été réalisée sur l'Île de Pâques (en anglais: *Easter Island*)[53], pourtant le musée utilise son propre URI pour identifier cet endroit au lieu de prendre un vocabulaire de référence comme *Geonames* qui en propose déjà un[54]. Même constat pour les techniques de fabrication puisque l'AAT propose un URI pour le terme « incrustation » (en anglais: *inlay*)[55] qui n'est pas repris par le *British Museum*[56]. Le choix d'utiliser des URI « internes » au lieu de référer à une ontologie existante découle de plusieurs facteurs liés

[51] British Museum, *British Museum Semantic Web Collection Online*, http://collection.britishmuseum.org, consulté le 30 août 2015.
[52] British Museum, *Hoa Hakananai'a*, http://collection.britishmuseum.org/id/object/EOC3130, 2012, consulté le 30 août 2015.
[53] British Museum, *Easter Island*, http://collection.britishmuseum.org/id/place/x69553, 2012, consulté le 30 août 2015.
[54]Geonames, *Easter Island-URI*, http://www.geonames.org/maps/google_-27.117_-109.367.html, consulté le 30 août 2015.
[55] Getty Research Institute, *Inlay (process)*, http://vocab.getty.edu/aat/300053850, 27 décembre 2013, consulté le 30 août 2015.
[56] British Museum, *Inlaid*, http://collection.britishmuseum.org/id/thesauri/x12176, 2012, consulté le 30 août 2015.

à la culture institutionnelle et des rapports entre les institutions. Malheureusement, cette mentalité empêche le plein développement du Web sémantique et la mise en valeur de son potentiel associatif. L'utilisation d'*owl:sameAs* est une solution à ce problème, mais une meilleure cohésion entre les institutions augmenterait davantage le potentiel associatif en évitant des pertes de sens.

Malgré le manque de connexions avec d'autres institutions, le modèle du *British Museum* est extrêmement intéressant pour la recherche et pour des partenariats futurs puisqu'il utilise une version sémantique de CIDOC CRM nommée *Erlangen CRM/OWL* développée par Berhard Schiemann, Martin Oischinger et Günther Görz de l'*Université Friedrich-Alexander*[57]. Cette interprétation du CIDOC CRM vise à refléter le contenu de la source le plus possible[58], un principe épousé par Manfred Thaller dans sa critique des bases de données commerciales[59]. Cette ontologie a aussi l'avantage de ne pas pointer vers un document descriptif en texte libre, mais plutôt vers un format interopérable sous forme de triplets.

Sémanticpédia *et* Europeana *: portails collaboratifs avec communautés francophones*

Le Web sémantique est un outil fort prometteur pour le traitement multilingue. Les exemples précédents ont des contenus essentiellement anglophones, mais un système de traduction automatique peut facilement être implanté grâce aux triplets RDF, un des avantages incontestables des structures de données atomisées. Une des

[57] Bernhard Schiemann, Martin Oischinger, Günther Görz, Georg Hohmann, Judith Merges, Mark Fichtner et Martin Scholz, *Erlangen CRM OWL*, http://erlangen-crm.org/, 2013, consulté le 30 août 2015.
[58] *Ibid.*
[59] Susan Schreibman, Raymond George Siemens et John Unsworth, *A companion to digital humanities*, Malden, MA, Blackwell Pub (coll. « Blackwell companions to literature and culture »), 2004, p. 60.

applications qui montre la force de ce procédé est le système d'information géographique (SIG) *Maphub* développé par le *Cornell Information Science* avec le support de l'*Université de l'Illinois* et l'*Université de Vienne*[60]. Ce logiciel permet d'annoter des cartes historiques à partir du modèle RDF pour lier ses données avec *DBpedia* et ainsi offrir un moteur de recherche multilingue. Par exemple, si une note signale sur une carte la présence de Montréal, le logiciel associera cette entrée avec l'URI *DBpedia* de Montréal : http://dbpedia.org/ressource/Montreal. Ainsi, grâce au prédicat *owl:sameAs*, *Maphub* pourra rendre cette annotation accessible à un usager grec qui indiquerait Μόντρεαλ pour identifier Montréal[61].

La francophonie ne peut évidemment pas se contenter d'attendre que les contenus anglais soient traduits en français. Le développement de contenus francophones au sein des données liées repose surtout sur les efforts menés en France avec le projet *Sémanticpédia*. Il s'agit d'« une plateforme de collaboration entre le *Ministère de la culture et de la communication* (MCC), *Inria* et *Wikimedia France* pour réaliser des programmes de recherche et de développement appliqués à des corpus ou des projets collaboratifs culturels, utilisant des données extraites des projets de *Wikimedia*[62]. » Ce partenariat repose sur le projet *DBpédia en français* qui libère, sous forme de données liées, le contenu des pages francophones de *Wikipedia*. Dans un effort d'internationalisation de *DBpedia*[63], et suivant son modèle de présentation, l'*Inria* propose le préfixe http://fr.dbpedia.org/ressource/ pour obtenir du contenu

[60] MapHub, *Historic Map Annotation Portal*, http://maphub.github.io/, consulté le 30 août 2015.
[61] DBpedia, *About: Montréal*, http://dbpedia.org/page/Montreal, consulté le 30 août 2015
[62] Ministère de la Culture et des Communications de France, Inria et Wikimedia France, *Sémanticpédia*, http://www.semanticpedia.org/, consulté le 30 août 2015.
[63] DBpedia, *Internationalization*, http://wiki.dbpedia.org/Internationalization/, 2014, consulté le 30 août 2015.

francophone[64]. En plus de ce projet, *Sémanticpédia*, grâce à un programme de sémantisation piloté par le MCC français, développe *Muséosphère* et *JocondeLab*.

Muséosphère a pour objectif d'identifier un musée susceptible de répondre aux intérêts d'un usager en s'appuyant sur les données de *Wikipédia* et celles de la base *Muséofile* détenue par le ministère. Développé dans le cadre d'un cours en Web sémantique par Mara Dumitru en 2011, le projet a été ouvert au mode collaboratif. Comme il s'appuie sur *DBpédia en français*, si ce dernier est incomplet, l'outil de recherche le sera tout autant. Les applications construites par-dessus des immenses jeux de données permettent de constater les lacunes de ces derniers. Une notice d'un musée moins précise sur le site de *Muséosphère* signifie que la boîte d'informations (en anglais : *infobox*) de *Wikipedia* n'est pas complète. Cependant, dès qu'une personne ajoute des champs descriptifs à ce musée sur *Wikipedia*, *Muséosphère* se mettra à jour pratiquement automatiquement[65].

Le Québec ne semble pas présent sur cette plate-forme, bien que la province possède des œuvres artistiques importantes. Les œuvres de Gustave Courbet répertoriées sur le réseau *Info-Muse* de la *Société des musées québécois* (SMQ) qui se trouvent au sein de la collection du *Musée des Beaux-Arts de Montréal* (MBAM)[66] n'apparaissent

[64] Par contre, étant une entité distincte de *DBpedia*, il est impossible, pour l'instant, d'interroger la plate-forme à partir d'un NER comme *DBpedia Spotlight*. Le point d'entrée SPARQL permet bien de faire des requêtes sur l'ensemble des triplets, mais lorsqu'arrive le moment d'aligner nos données avec *DBpédia en français*, la tâche demande un temps considérable.

[65] Mara Dumitru, *Museosphere*, http://museosphere.net/apropos?language=fr, 2013, consulté le 30 août 2015.

[66] Société des musées québécois, *Info-Muse Recherche Gustave Courbet*, http://infomuse.smq.qc.ca/Infomuse/f_MasterLayout.cgi?la=f&db=1&style=99&realm=2&es=1&rs=1&who_i=WHOO&who_t=Gustave+Courbet&who_o=and&sort=NO_SORT, consulté le 30 août 2015.

pas sur *Muséophère*[67]. Une recherche rapide permet de constater que l'article *Wikipedia* sur Gustave Courbet[68] ne fait pas, non plus, mention des œuvres conservées au MBAM et qu'il n'existe donc pas de triplets permettant à l'application d'inclure ce musée comme destination possible. Les institutions culturelles québécoises devront donc veiller à être présentes sur les grandes bases de données du monde pour que le contenu québécois apparaisse sur le Web sémantique.

Le second projet, *JocondeLab*, « permet d'accéder à près de 300 000 notices décrivant des œuvres des musées en de [sic] France en 14 langues[69]. » Cette expérimentation a pour objectif de « démontrer les possibilités du web sémantique ou "Web 3.0" et de la mise en relation de données culturelles "liées", tant en terme de multilinguisme, que d'ergonomie et d'interactivité[70]. » Le portail unit deux types de sources : premièrement, le catalogue des collections des musées de France, *Joconde*[71] et deuxièmement, *DBpédia en français* qui permet de décrire divers éléments structurant les éléments du catalogue. Trois types de recherches sont possibles : par lieux sur une carte du monde, par chronologie sur une ligne du temps ou par thèmes sous forme de liste. Le volume considérable du corpus permet d'obtenir quelques résultats reliés au

[67]Mara Dumitru, *Gustave Courbet*, Museosphere, http://museosphere.net/inspire/fr?artist=Gustave+Courbet&mouvement=&localisation=, 2013, consulté le 30 août 2015.
[68] Wikipedia, *Gustave Courbet*, https://fr.wikipedia.org/wiki/Gustave_Courbet, consulté le 30 août 2015.
[69]Ministère de la Culture et de la Communication de France. *Sémanticpédia : construire le web de données culturelles - Langue française et langues de France*, http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/Web-semantique-web-de-donnees-liage-de-donnees/Semanticpedia-construire-le-web-de-donnees-culturelles, 23 octobre 2014, consulté le 30 août 2015.
[70] Ministère de la Culture et des Communications de France, *JocondeLab - À propos*, http://jocondelab.iri-research.org/jocondelab/about/, consulté le 30 août 2015.
[71] Ministère de la Culture et des Communications de France, *Joconde, portail des collections des musées de France*, http://www.culture.gouv.fr/documentation/joconde/fr/pres.htm, consulté le 31 août 2015.

Québec mais tous concernent Montréal[72]. Cette application lie donc les objets des champs *Artiste(s)*, *Domaine(s)*, *Sujet représenté*, *Datation* et *Provenance* pour augmenter la richesse informationnelle multilingue grâce à *DBpédia en français*. Ce projet plus abouti que *Muséosphère* au niveau de l'interface de visualisation démontre réellement la puissance d'interconnexion du Web sémantique.

L'ensemble des projets présentés précédemment s'inscrit dans un idéal de convergence de l'information pour en faciliter la consultation. Le Web sémantique étant encore une notion récente, il est difficile d'envisager le potentiel d'association qu'il permet d'atteindre. Un des projets d'envergure les plus prometteurs est *Europeana* déjà présenté sommairement. L'ensemble des données d'*Europeana* est sous licence *CCO Public Domain Dedication*, soit une licence ouverte[73]. La question des droits d'auteurs ne se pose donc pas lorsqu'on souhaite développer une application utilisant leurs données liées. Effectivement, depuis octobre 2012, une grande proportion des données disponibles sur *Europeana* a été transformée en données liées et accessibles via l'adresse www.data.europeana.eu[74]. Le développement de ce portail d'envergure a débuté en février 2012 avec quelques fournisseurs de données afin d'élaborer un modèle de données efficace. Aujourd'hui, on compte plus de 20 millions de sources, tous types confondus[75]. En plus de fournir les fichiers brutes de données liées, il est possible

[72]Ministère de la Culture et des Communications de France, *JocondeLab - Montréal*, http://jocondelab.iri-research.org/jocondelab/map/#http%3A%2F%2Ffr.dbpedia.org%2Fresource%2FMontr%25C3%25A9al, consulté le 31 août 2015.
[73] Bernhard Hashofer et Antoine Isaac, *Europeana Linked Open Data*, http://labs.europeana.eu/api/linked-open-data-introduction, consulté le 31 août 2015.
[74] *Ibid.*
[75] *Ibid.*

d'interroger celles-ci par un point d'entrée SPARQL. Un exemple d'interrogation de ce point d'entrée SPARQL est présenté à l'Annexe I.

Europeana développe aussi sa propre ontologie, soit l'*Europeana Data Model* (EDM). Ce modèle permet donc à *Europeana* de s'inscrire dans le nuage des données liées où la plate-forme se classe dans le noyau central[76]. Les trois besoins principaux que ce modèle gère sont la capacité de distinguer la différence entre un élément réel et sa reconstitution numérique, la possibilité de différencier un élément réel des métadonnées le décrivant et la faculté d'ingérer ces différentes entrées pour un même élément qui seront constituées inévitablement de propos contradictoires[77]. Dans une optique d'économie de temps et d'argent, l'organisme se tourne vers différents vocabulaires existants comme *Dublin Core* et SKOS présentés précédemment. Du même coup, on souhaite faciliter l'adoption du modèle, ce qui ne semble pas être gagné d'avance. En effet, selon l'outil LOV, qui permet de comparer différents vocabulaires[78], un seul lien pointe vers l'EDM soit *Simple Service Status Ontology* (SSSO), une ontologie de gestion d'événements[79]. Ce résultat concerne seulement les ontologies qui réutilisent l'EDM dans leurs propres structure, mais n'identifie pas quelles institutions, hors de l'organisation, utilisent ce modèle. Constat normal puisqu'il s'agit d'une ontologie conçue, en premier lieu, pour répondre à des besoins organisationnels. L'Annexe III propose un complément technique sur le vocabulaire de l'EDM.

---

[76] Richard Cyganiak et Anja Jentzsch, *loc. cit.*

[77] Bernhard Haslhofer et Antoine Isaac, *Data structure Europeana*, http://labs.europeana.eu/api/linked-open-data-data-structure, consulté le 31 août 2015.

[78] Open Knowledge Foundation, *Linked Open Vocabularies (LOV)*, http://lov.okfn.org/dataset/lov/, 28 août 2015, consulté le 31 août 2015.

[79] Antoine Isaac, *Europeana Data Model vocabulary (edm)*, http://lov.okfn.org/dataset/lov/vocabs/edm, 2013, consulté le 31 août 2015.

Au-delà des tranchées: projet pilote canadien sur le potentiel du Web sémantique

Hébergé sur le portail *Canadiana* pour en faciliter une large consultation[80], le projet *Au-delà des tranchées: Un projet des Données ouvertes liées*, a été développé en 2012 par le RPCPD. Il s'agit d'entreprendre « une "démonstration de faisabilité" pour présenter un sous-ensemble de la richesse du réseau des ressources numériques en utilisant les "données ouvertes liées" et le Web sémantique[81]. » Le groupe avait déjà noté le rôle important que les données liées jouent au sein des bibliothèques, des archives et des musées. L'équipe du projet offre trois justifications pour cette initiative. Premièrement, il s'agit d'un moyen de fournir l'accès aux données qui pourront être réutilisées et réorganisées pour répondre aux besoins imprévisibles des chercheurs. Deuxièmement, le terme « Web sémantique » prend tout son « sens » lorsqu'on crée des connexions entre une vaste communauté d'institutions et qu'on construit ainsi une expérience cohérente et riche pour le public. Troisièmement, on libère les données des institutions et on favorise ainsi la mise en réseau d'un plus grand nombre de connaissances qui pourront rehausser la valeur informationnelle de nos données[82].

Le projet regroupe cinq institutions partenaires qui ont chacune libéré une partie de leurs données : BAnQ (chansons sur la guerre), *Université McGill* (affiches sur la guerre), *Université de l'Alberta* (archives d'articles de journaux, cartes postales et dossiers du temps de la guerre), *Université de Calgary* (archives de portraits de soldats

[80]Canadiana, *Au-delà des tranchées : Un projet des Données ouvertes liées*, http://www.canadiana.ca/rpcpd-dol, 15 juillet 2012, consulté le 31 août 2015.
[81] Réseau pancanadien du patrimoine documentaire*, loc. cit.*
[82] *Ibid*., p. 8.

du Corps expéditionnaire canadien (CED) et de documents de la Première Guerre mondiale) et *Université de la Saskatchewan* (documents d'archives du projet *Saskatchewan War Experience*)[83]. *Bibliothèque et Archives Canada* (BAC) est venu compléter la documentation et le *Musée canadien pour les droits de la personne* a proposé de publier les données sous la licence *Open Data Commons Public Domain Dedication and License* (ODC-PDDL). Bien que le projet utilise ses propres URI pour les sujets et les objets, un effort supplémentaire a été fourni pour utiliser des vocabulaires contrôlés reconnus comme *Geonames* et le *Thesaurus for Graphic Materials* (TGM) de la *Getty Research Institute*[84]. Le modèle ontologique du projet *Au-delà des tranchées* ainsi que les recommandations qui en découlent se retrouvent à l'Annexe IV.

Le RPCPD s'est grandement inspiré de projets australiens pour l'élaboration de sa plate-forme puisque l'équipe ne souhaitait pas fournir un simple outil de recherche fédéré. Le premier est la plate-forme *the real face of white australia*[85] développée dans le cadre du projet *Invisible Australians*. On emprunte ici le visuel de l'application qui invite à la recherche par les visages des Australiens qui étaient enregistrés et limités dans leurs actions en vertu de la *White Australia Policy*[86]. En effet, l'interface principale présente uniquement des photos d'archives de ces Australiens. Alors, la première invitation à explorer les contenus associés débute autour du désir d'en apprendre d'avantage sur un visage. Cette approche, pouvant être considérée comme

[83] *Ibid.,* p. 2.
[84] *Ibid.*, p. 7.
[85] Archives nationales de l'Australie, *The real face of white australia :: experimental browser*, http://invisibleaustralians.org/faces/, consulté le 31 août 2015.
[86] *Ibid.*

contemplative, permet d'explorer les fiches produites dans le cadre de cette politique et conservées par les *Archives nationales d'Australie* sans passer par un moteur de recherche traditionnel. On camoufle complètement la structure derrière un visuel simple et invitant.

Le second projet qui a servi de modèle est un partenariat entre le *Museum of Australian Democracy* et, encore une fois, les *Archives nationales d'Australie* autour de la collection Mildenhall[87]. Cette collection met en valeur les premiers développements de la capitale, Canberra, grâce à 7 600 photographies prises entre 1920 et 1940. Elle est enrichie par l'ajout de commentaires textuels ainsi qu'une géolocalisation des clichés. Ce travail s'effectue de manière collaborative puisque les deux institutions encouragent le public à contribuer à ce vaste chantier. À la fin du projet, les données recueillies ont été transférées, sous la licence *Creative Commons*, sur data.gov.au[88], un dépôt de données ouvertes qui possède son équivalent au Québec: donnees.gouv.qc.ca[89].

Troisièmement, le projet s'appuie sur l'article *Every story has a beginning* de Tim Sherratt[90], un historien numérique australien présenté au chapitre précédent. Il est un des rares historiens qui a réfléchi à l'apport de la discipline au développement du Web sémantique.

---

[87] Museum of Australian Democracy et Archives nationales de l'Australie, *About · Mildenhall's Canberra*, http://mildenhall.moadoph.gov.au/about, 2013, consulté le 31 août 2015.
[88] Gouvernement de l'Australie, *Data.gov.au*, http://data.gov.au/, consulté le 31 août 2015.
[89] Gouvernement du Québec, *Données ouvertes*, http://donnees.gouv.qc.ca/?node=/accueil, 2015, consulté le 31 août 2015.
[90] Tim Sherratt, *Every story has a beginning*, *loc. cit.*

L'objectif du deuxième chapitre était de familiariser l'historien avec un nombre de concepts clés qui permettent de saisir les possibilités qu'apporte le Web sémantique. On a défini ce dernier comme étant une conceptualisation du Web orienté vers l'uniformisation des données pour favoriser une meilleure contextualisation et une meilleure agrégation des informations historiques. Le chercheur qui maîtrise ces concepts pourra plus facilement dialoguer avec des informaticiens pour la mise en place de ce type de données. Les différents projets qui lient histoire et Web sémantique, synthétisés suivant les cinq étoiles à l'Annexe V, permettent de relever plusieurs problèmes que l'historien doit analyser et tenter de corriger pour que les données liées puissent répondre aux besoins de la discipline. Pour se faire l'historien doit s'initier aux nouveaux modes de présentation de ses données afin de répondre aux normes ontologiques présentées dans ce chapitre. De plus, il doit devenir un acteur central dans la réflexion entourant les notions de temps et d'événement que l'on retrouve dans les modèles CIDOC CRM et BIO par exemples. Évidemment, l'historien devra, en plus de comprendre la théorie, être capable d'appliquer ces concepts de manière pratique.

CHAPITRE III : DE LA THEORIE A LA PRATIQUE, LE CAS DU REPERTOIRE DU PATRIMOINE CULTUREL DU QUEBEC

Dans le cadre du PCNQ, le RPCQ sera modifié pour favoriser son ouverture et sa réutilisabilité dans l'objectif de promouvoir la cohésion des données québécoises dans le réseau francophone. Ce grand chantier nécessite l'intervention historienne pour assurer que la transformation fasse écho à la réalité historique et que le résultat améliore réellement les possibilités de recherche en histoire, question soulevée dans le premier chapitre. Quel rôle doit jouer l'historien dans la mise en place et le maintien d'une structure de données liées? Quelles sont les limites des compétences historiennes dans le domaine du Web sémantique? Ces deux questions guident la rédaction de ce dernier chapitre qui met en évidence les avantages d'inclure un historien dans le processus de création d'une plate-forme sémantique.

Le répertoire dans sa forme actuelle

Le paysage des bases de données patrimoniales au Québec

Le Québec, tout comme le Canada, ne possède pas de plate-forme fédératrice de contenus culturels qui pourrait se comparer à *Europeana*. L'information est alors dispersée dans un nombre important de bases de données parfois accessibles par un site web institutionnel. Il est difficile de tracer un portrait clair de l'ensemble des bases de données patrimoniales puisqu'il n'existe aucun outil qui les identifie tous. Notre connaissance du paysage des bases de données est alors limitée par trois facteurs intimement liés. Premièrement, on consulte plus souvent les bases de données

accessibles en ligne[1]. La consultation d'une base de données locale demande un investissement de temps et d'argent pour la repérer et pour évaluer sa pertinence. Si la thématique de celle-ci ne correspond pas à notre sujet, peut-être passerons-nous à côté de données capitales. Deuxièmement, les mises à jour constantes d'un site web augmentent sa visibilité. Il devient donc difficile de trouver ceux qui ne sont plus actualisés, mais qui, dans le cadre de la discipline historique peuvent toujours servir de matière première. À ce titre, il est plus facile de trouver un travail écrit qu'un jeu de données, les méthodes de repérage des ouvrages étant mieux balisées que celles des bases de données. Troisièmement, les bases de données institutionnelles sont plus accessibles que les bases de données de chercheurs. Encore une fois, les ressources humaines et financières d'un projet de recherche qui a produit des données sont souvent limitées et empêchent de maintenir une visibilité du contenu. Il faut donc connaître directement le chercheur pour avoir accès à ses données. Le portrait actuel des données est donc majoritairement composé d'informations provenant des bibliothèques, des centres d'archives et des musées. Il y a ainsi une création de silos puisque le contenu n'est pas organisé par la teneur des données, mais bien par le type de supports (livres, archives, objets).

BAnQ joue un rôle central dans le portrait des données patrimoniales. Deux catalogues facilitent le repérage découlant de la fusion des *Archives nationales* avec la *Bibliothèque nationale* : IRIS pour les livres et PISTARD pour les archives. Cette distinction en deux catalogues montre la division des contenus au sein même de

[1] Ron Stewart, Vivek Narendra et Axel Schmetzke, « Accessibility and usability of online library databases », Library Hi Tech, http://www.emeraldinsight.com/doi/abs/10.1108/07378830510605205, 2005, vol. 23, no 2, p. 265286.

l'organisme. Il n'y a aucun moyen de faire des recherches fédérées dans les deux systèmes en s'appuyant sur des métadonnées communes. Tandis que BAnQ et les centres d'archives régionaux conservent et diffusent le patrimoine documentaire québécois, les musées préservent les objets. La SMQ regroupe quelques 300 institutions muséales et propose un outil de recherche fédéré du nom d'*Info-Muse*. Ce dernier a été créé en 1991 en collaboration avec le *Réseau canadien d'information sur le patrimoine* (RCIP) et favorise l'échange d'information[2]. Les outils technologiques de cette période ne permettaient pas une exploitation dynamique comme le permettent les données liées aujourd'hui. Malgré cet esprit de fédération, on focalise exclusivement sur la donnée muséale, ce qui résulte majoritairement en des corpus d'objets.

La question des personnes et du patrimoine immatériel permet d'ouvrir vers d'autres bases de données. Le *Dictionnaire biographique du Canada* (DBC) permet justement de traiter de la question des personnes en proposant des articles biographiques. Ce dictionnaire regroupe un nombre considérable de personnages historiques associés par leurs activités au Québec, soit 5 448[3] dont 1 964[4] sont nés au Québec, en date du 26 novembre 2015. Cet outil ne permet toutefois pas d'associer ces personnages avec des objets ou des archives conservés dans des institutions québécoises ou canadiennes. Le même constat peut être établi du côté de la politique québécoise avec l'outil de recherche de l'*Assemblée nationale du Québec*. Bien qu'il soit possible d'associer des

[2] Société des musées québécois, *Info-muse*, http://infomuse.smq.qc.ca/basisbwdocs/infm/Info/f_HumanitiesInfoHead.html, 1999, consulté le 15 décembre 2015.
[3]Université Laval et Université de Toronto, *Résultats de la recherche – Région d'activités*, http://www.biographi.ca/fr/resultats.php?partial=0&stemmed=1&count=20&l_ft_2=and&l_ft_3=and&cp=126+125+124+123+122, 2015, consulté le 26 novembre 2015.
[4]Université Laval et Université de Toronto, *Résultats de la recherche – Région de naissance*, http://www.biographi.ca/fr/resultats.php?partial=0&stemmed=1&count=20&l_ft_2=and&l_ft_3=and&bp=122+123+124+125+126, 2015, consulté le 26 novembre 2015.

députés avec des projets de loi ou des événements particuliers, comme des conférences de presse, les possibilités se limitent encore une fois au contenu interne de la base de données[5].

La notion de lieu, quant à elle, peut être mise à profit grâce aux développements des SIG et la base de données de la *Commission de la toponymie du Québec*. Cet outil permet de repérer des noms actuels de lieux[6]. Malheureusement, il n'existe pas d'outils dynamiques en ligne qui permettent d'identifier le nom d'un lieu à une période donnée. Par exemple, si on s'intéresse à un événement qui se déroulait à l'emplacement géographique de Québec avant 1608, il faudrait utiliser le toponyme « Stadaconé » au lieu de celui de « Québec ». Une méconnaissance de ce changement toponymique peut empêcher un chercheur d'obtenir toutes les données existantes sur son sujet de recherche. Le Web sémantique permettrait de pallier ce type d'erreur de saisie en liant ces différentes appellations avec leurs années d'apparition et de disparition et en définissant des équivalences.

Ce repérage sommaire des bases de données sur le patrimoine québécois rend compte qu'il existe peu de systèmes conçus pour tisser des liens entre différents contenus qui, ensemble, facilitent la compréhension du passé. Le seul outil qui regroupe différentes composantes de notre patrimoine québécois est le RPCQ.

---

[5] Assemblée nationale du Québec, *Recherche avancée*, http://www.assnat.qc.ca/fr/recherche/recherche-avancee.html, 10 octobre 2012, consulté le 15 décembre 2015.

[6] Commission de la toponymie du Québec, *Banque de noms de lieux du Québec*, http://www.toponymie.gouv.qc.ca/ct/accueil.aspx, 2012, consulté le 15 décembre 2015.

Le contenu et le contenant du RPCQ

Le RPCQ est avant tout une plate-forme de diffusion pour les données conservées par le MCC. Le répertoire est la vitrine publique du système PIMIQ qui conserve les informations « sur les biens culturels reconnus et classés[7] ». Le contenu du répertoire est vaste puisqu'articulé autour de la notion de patrimoine culturel qui, depuis les années 1980, est un terme grandement inclusif[8]. À ce titre, la *Loi sur le patrimoine culturel*, adoptée le 19 octobre 2011, stipule qu'il « est constitué de personnages historiques décédés, de lieux et d'événements historiques, de documents, d'immeubles, d'objets et de sites patrimoniaux, de paysages culturels patrimoniaux et de patrimoine immatériel[9]. » D'un point de vue légal, un bien n'est patrimonial que si celui-ci est classé et protégé par une loi[10]. Cependant, il est tout à fait possible de simplement considérer une composante patrimoniale comme étant un élément témoin de l'histoire qui devient un ancrage dans le présent qui cristallise la mémoire collective.

Le RPCQ inclut dans sa base de données autant les biens protégés et classés que des composantes inventoriées par ses partenaires. Ce vaste mandat nécessite la collaboration de différents acteurs dont le *Gouvernement du Québec*, les communautés autochtones, les municipalités et d'autres partenaires[11]. On y traite autant de patrimoine matériel qu'immatériel. Ainsi, on catalogue le patrimoine mobilier et immobilier, mais

---

[7] Gouvernement du Québec, *PIMIQ - Thésaurus de l'activité gouvernementale*, http://www.thesaurus.gouv.qc.ca/tag/terme.do?id=CBC291, 2015, consulté le 16 décembre 2015.
[8] Pierre Nora, « L'explosion du patrimoine », *Revue de l'Institut national du patrimoine*, 2006, no 2, p. 7.
[9] Gouvernement du Québec, *Loi sur le patrimoine culturel*, http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/P_9_002/P9_002.html, 19 octobre 2012, consulté le 26 janvier 2015.
[10] Pierre Nora, *loc. cit.*, p. 6.
[11] Ministère de la Culture et des Communications du Québec, *À propos du Répertoire du patrimoine culturel du Québec*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/redirection.do?go=about, 2013, consulté le 16 décembre 2015.

aussi les événements, les groupes ainsi que les personnes. Une place est aussi accordée au folklore et aux traditions.

Face à la diversité des contenus, le RPCQ propose un modèle de fiche classique où les champs sont modifiés selon l'information présentée. Le titre de la fiche donne le nom de l'élément enregistré. Un lot de champs permet de l'identifier. Une section présente différents médias qui permettent une visualisation de la pièce patrimoniale, que ce soit une image, une vidéo, une carte géographique, etc. Par ailleurs, on retrouve également une description assez détaillée ainsi que le statut patrimonial de l'élément. Cependant, la section la plus utile à des fins de données liées s'intitule « Éléments associés ». Cette catégorie permet de lier des fiches entre elles. Par exemple, un bâtiment peut être associé à un site patrimonial, un architecte, un propriétaire, un locataire, un événement, etc. Derrière ces liens établis de façon manuelle, on réalise que ce sont les principes du Web sémantique qui sous-tendent cette pratique, bien qu'ils ne soient pas utilisés pour le moment. Évidemment, la procédure vise à effectuer des croisements entre les données du répertoire, mais il serait possible d'utiliser ce système pour pointer vers d'autres plates-formes.

Volume, représentativité et corpus de données

Les fiches du RPCQ sont classées selon différentes listes de vocabulaires utilisées dans le répertoire. Ces vocabulaires ont été créés avant que le Web sémantique ne s'impose comme solution à la mise en relation des données et ils présentent ainsi certains problèmes. À titre d'exemple le champ « Occupation », courant pour toute étude des personnes, comporte trois limites à son utilisation sémantique. La première lacune

est le manque de définition des différents termes se trouvant dans le vocabulaire. Quelle est la différence entre un guerrier et un militaire? Pourquoi l'historien est-il classé dans la catégorie des sciences et non pas dans culture? Pourquoi avoir créé une catégorie « Travail » pour y insérer des concepts comme « Employé de bureau » ou même « Esclave »? Le second problème, qui découle du premier, est le manque d'uniformité, ce qui complexifie la compréhension du modèle. Le dernier obstacle est que ce vocabulaire – conçu spécifiquement pour le RPCQ – est infrastructurel et n'est donc pas exposé à l'usager. Ce dernier peut prétendre comprendre le système de recherche sans réaliser que les vocabulaires sous-jacents peuvent biaiser sa recherche. Dans un idéal de fédération, il faut plutôt utiliser des modèles généraux conçus pour associer des schémas spécifiques répondant à des corpus diversifiés. Dans le cas des occupations, la *Classification nationale des professions* (CNP) de 2011 de Statistique Canada[12] permettrait l'uniformisation d'une partie des professions du Canada. Son analyse et sa mise en relation avec d'autres modèles seront présentées dans la dernière section du présent chapitre.

Malgré ces quelques problématiques, le RPCQ comporte un nombre imposant de fiches très variées, ce qui en fait un exemple idéal pour montrer le potentiel d'une structure en données liées. En effet, en date du 17 novembre 2015 on retrouve 106 011 fiches dans le RPCQ. Ce total se divise en cinq ensembles : 55 351 fiches de biens mobiliers, 34 729 fiches de biens immobiliers, 14 831 fiches d'événements, de groupes et de personnes, 31 fiches de biens immatériels et 1 009 plaques commémoratives. On

[12] Gouvernement du Canada et Statistiques Canada, *Types des professions*, http://www.statcan.gc.ca/fra/concepts/profession, 21 novembre 2011, consulté le 16 décembre 2015.

remarque alors que 60 fiches ne semblent pas associées à aucune de ces catégories[13]. Du grand total, 38 % représentent des biens protégés et valorisés par une loi. Dans les biens mobiliers, les œuvres d'art et les biens ethno-historiques en représentent 58 %. Ainsi, on constate que le patrimoine matériel prend une place considérable dans le RPCQ.

Dans le cadre de ce mémoire, le corpus choisit concernera les 11 883 personnes fichées dans le RPCQ. Ces fiches-personnes découlent de l'objectif de représenter le patrimoine immatériel dans le répertoire, en identifiant les personnes à partir des fiches concernant les biens. Il s'agit donc du raisonnement inverse du modèle CIDOC CRM qui sera étudié dans le cadre de cet exercice pratique. En effet, les fiches-personnes du RPCQ découlent des biens tandis que CIDOC CRM utilise les événements et les acteurs de ceux-ci pour tisser les liens entre les sources. Un autre point de départ aurait pu être les événements classés dans le répertoire, mais le nombre actuel de 71 fiches ne permettait pas de valider convenablement la méthodologie et les compétences nécessaires pour créer des données liées.

Les 11 883 personnes seront traitées automatiquement, mais un échantillon de 1 000 (8,42 %) servira à valider ce processus algorithmique. La liste a été sélectionnée en ordre croissant des identifiants uniques octroyés par le RPCQ. Ayant un lien uniquement avec les fiches des biens associés, cette sélection permet d'avoir un échantillon aléatoire représentatif.

[13] Ministère de la Culture et des Communications du Québec, *Recherche - Répertoire du patrimoine culturel du Québec*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/rechercheProtege.do?methode=afficher, 2013, consulté le 16 décembre 2015.

Association avec DBpedia

Méthodologie et outils de liaison

La première étape lorsqu'on souhaite intégrer le nuage des données liées est d'attribuer un URI à chacune des données. Évidemment, afin de briser le modèle en silo, il est préférable de lier les données à des URI déjà existants, si c'est possible. Dans le cas présent, les personnes du RPCQ seront associées directement avec celles de *DBpedia*. Il existe d'autres vocabulaires, mais *DBpedia* est le joueur le plus important du nuage des données liées et possède un NER du nom de *DBpedia Spotlight* qui automatise le repérage d'URI[14]. Le LCSH aurait probablement été un partenaire disciplinaire plus intéressant puisque le contenu qui s'y trouve est mieux organisé, mais sans la possibilité d'utiliser un NER, le travail devient rapidement laborieux, voire irréalisable. Un autre avantage est que *DBpedia* s'appuie sur *Wikipedia* pour créer des URI. Sachant que BAnQ propose, chaque premier mardi du mois depuis février 2014, des ateliers pour enseigner les rudiments de la saisie de données sur cette encyclopédie libre, peut-être qu'une résonance de cet effort de compilation de contenus francophones sera perceptible sur *DBpedia*[15].

Une demande a été faite au ministère pour obtenir la liste des personnes fichées au RPCQ, dans un format tabulaire, puisqu'il est impossible de créer automatiquement un tableau à partir de leur site web. Ensuite, le logiciel *OpenRefine* a été utilisé pour traiter les données. Après avoir fusionné les noms et les prénoms des personnes au sein

[14] DBpedia, *DBpedia Spotlight*, https://dbpedia-spotlight.github.io/demo/, consulté le 16 décembre 2015.
[15] Bibliothèque et Archives nationales du Québec, Mardi c'est Wiki, https://fr.wikipedia.org/w/index.php?title=Wikip%C3%A9dia:BAnQ/Mardi_c%27est_Wiki&oldid=120982301, 2 décembre 2015, consulté le 16 décembre 2015.

de la même cellule pour faciliter le travail du NER, ce dernier fut lancé à plusieurs reprises afin de paramétrer le logiciel le plus efficacement possible[16]. En effet, l'algorithme utilisé pour repérer les noms permet une plus grande ou plus faible sensibilité aux caractères afin de maximiser sa performance. Cette automatisation fut vérifiée grâce à l'échantillon de 1 000 personnes présenté plus tôt.

Résultats bruts et second nettoyage

Après une analyse textuelle d'environ deux heures, *DBpedia Spotlight* a déniché 707 correspondances sur les 11 883 noms proposés, ce qui correspond à 5,95 % du corpus. Plusieurs erreurs sont survenues lors de cette automatisation. En effet, le logiciel associait souvent un seul nom du nom complet de la personne. Autrement dit, lorsqu'un nom se compose de plusieurs mots, le logiciel a tendance à les décortiquer de manière erronée. Par exemple, il associera Matthew Henry Cochrane à un URI référant à Matthew Henry et un autre URI pour les personnes possédant le nom de famille Cochrane. Dans la grande majorité des cas, ce type d'association donnait de mauvais liens. Les prénoms dont seulement les initiales sont présentes dans la fiche créent aussi des liaisons souvent erronées. Des lettres peuvent servir pour un acronyme ou faire référence à plusieurs personnes. C'est le cas de G.P. McWhaw lié avec l'URI des initiales G.P. qui correspondent à une série télévisée australienne. Un autre problème qui revient souvent est l'association avec l'URI « De La » qui ne correspond à aucune entrée sur *DBpedia*. Par exemple, Henri De La Blanchère sera associé avec cet URI fantôme. Les accents ne causent pas de problèmes, mais aucun nom possédant des apostrophes n'a pu être identifié sur *DBpedia*.

[16] À titre indicatif, voici les paramètres retenus: *Support* de 30 et *Confidence* de 0.5

Ces erreurs étaient évidentes et faciles à éliminer de la liste de départ. Le travail le plus complexe concerne la vérification des noms trouvés par le logiciel *DBpedia Spotlight*. Pour valider que le nom associé correspond à la bonne personne, chaque nom a été transformé en URI *DBpedia*. Autrement dit, il suffit d'ajouter http://dbpedia.org/resource/ devant le nom et de remplacer tous les espaces par des barres de soulignement. Cette manipulation se fait facilement avec *OpenRefine* ou *Excel*. Il faut comprendre que la grande majorité des noms trouvés, à l'exception des personnalités connues du grand public, sont des noms souvent génériques. C'est le cas de John Scott qui possède deux entrées au RPCQ. Évidemment, ce nombre limité de fiches permet rapidement de distinguer les deux personnes grâce à leurs dates de naissance et de décès. Toutefois, sur *DBpedia*, il y a une liste de 54 personnes nommées John Scott[17]. La classification de *DBpedia*, malgré son manque d'uniformisation, permet, dans ce cas-ci, de repérer rapidement qu'un des John Scott est un politicien canadien ce qui semble correspondre à l'une de nos données. Après une analyse rapide des triplets de ce nouvel URI, par exemples les dates de naissance et de décès, on peut confirmer qu'il s'agit bien de la même personne. Dans un cas comme celui-ci, l'URI trouvé par *DBpedia Spotlight* est bon, mais pas suffisamment précis. Il est donc préférable de remplacer http://dbpedia.org/resource/John_Scott par http://dbpedia.org/resource/John_Scott_(Canadian_politician). Les URI sur *DBpedia* peuvent être extrêmement différents selon le choix d'organisation d'une entité. Dans le cas précédent, il fallait ajouter la fonction de la personne afin de retenir la bonne fiche. Si on devait lier un autre John Scott, peut-être aurions-nous eu besoin d'inscrire ses années de naissance et de décès ou encore seulement la mention décès avec une date. Parfois, il peut s'agir de

[17] DBpedia, *About: John Scott*, http://dbpedia.org/page/John_Scott, consulté le 16 décembre 2015.

la combinaison de ces différents champs. Par exemple, il y a un John Scott identifié en tant que joueur de criquet né en 1841 ayant donc comme URI http://dbpedia.org/resource/John_Scott_(cricketer,_born_1841)[18].

Après ce nettoyage et cette spécification d'URI, le corpus de bonnes liaisons entre *DBpedia* et le RPCQ passe de 5,95 % à 2,96 % soit 350 entrées sur les 707 trouvées par *DBpedia Spotlight* originellement. Évidemment, il existe plusieurs autres fichiers d'autorités qu'il faudrait analyser de la même façon. Cependant, dans le cas de *DBpedia*, il y a un faible pourcentage d'associations potentielles avec le RPCQ et donc un grand nombre d'URI à créer pour positionner le patrimoine québécois sur l'échiquier des données liées. Avant de conclure trop rapidement, il est nécessaire de vérifier manuellement ce faible rendement du programme.

Rendement de **DBpedia Spotlight** *et approximation finale*

Reprenant manuellement les milles premières personnes analysées par *DBpedia Spotlight*, on remarque un écart important entre les liens trouvés par le logiciel et le nombre véritable. La méthodologie fut la même que pour l'étape précédente à l'exception que les 1 000 noms ont été cherchés sur *DBpedia* afin de voir si une réponse s'affichait ou non. Sur les 1 000, le logiciel avait associé avec justesse 47 personnes. Ce nombre exclut les liens qui associent des homonymes n'étant pas la bonne personne. On obtient donc un pourcentage légèrement différent que pour l'échantillon total soit 4,70 % au lieu de 2,96 %.

[18] DBpedia, *About: John Scott (cricketer, born 1841)*, http://dbpedia.org/page/John_Scott_(cricketer,_born_1841), consulté le 16 décembre 2015.

La distinction majeure se trouve plutôt du côté du nombre de bonnes associations que le logiciel aurait dû répertorier. En effet, 290 entrées sur 1 000 concordent avec *DBpedia*, soit un total de 29 %. On peut alors estimer un total de 3 446 personnes sur les 11 883 du RPCQ qui pourraient avoir un URI au sein de *DBpedia*. En contrepartie, 61 % des fiches du RPCQ nécessiteraient la création d'URI par le ministère. On constate alors qu'une grande partie de l'histoire du Québec, du moins à ce qui a trait aux personnages, ne peut être trouvée par des requêtes sémantiques et donc qu'il est impossible d'enrichir nos connaissances par d'éventuelles liaisons externes avec *DBpedia*.

Une autre question est celle du type de personne que l'on retrouve sur *DBpedia*. La majeure partie des personnes répertoriées provient des sphères politique, religieuse et des arts. De cette dernière, les architectes, les chanteurs et les écrivains prennent une place considérable, reflet de l'approche sous-jacente à la construction du répertoire. L'histoire locale et surtout l'histoire des femmes sont peu représentées, autant sur le RPCQ que sur *DBpedia*. Plus précisément, sur l'échantillon de 1 000 personnes du RPCQ, seulement 32 femmes sont présentes, soit 3,2 % et *DBpedia* en trouve 5 sur 290, soit 1,72 %. Les biens culturels étant souvent associés à des hommes, cela favorise la présence de ceux-ci au détriment des femmes dans la représentation du passé. Par ailleurs, les données liées étant une nouvelle méthodologie pour construire l'histoire, ses premières étapes risquent de reprendre les œillères qui ont caractérisé la construction de l'histoire traditionnelle.

Ce travail d'arrimage exige un investissement de temps considérable surtout si on intègre d'autres fichiers d'autorité que *DBpedia*, comme le LCSH par exemple. Est-ce

que ces liaisons éventuelles peuvent favoriser une meilleure contextualisation de la donnée et devenir un gain de temps et d'argent pour les institutions culturelles et les chercheurs? La réponse semble évidente à la lumière de ce court exercice.

Premièrement, le croisement des sources fait ressortir les données différentes d'une source à l'autre et donc potentiellement erronées. Les dates de naissance et de décès sont des exemples parlant pour l'échantillon sélectionné. La date de décès de Robert Montgomery Martin, qui est fiché dans les deux systèmes, est le 6 septembre 1868. Cependant sa date de naissance est approximative dans les deux outils. Le RPCQ attribue l'année 1803[19] et *DBpedia* l'année 1801[20]. L'absence de triplets permettant de valider la provenance de chacune de ces informations empêche d'arriver à une date de naissance définitive. Pour John Sparrow David Thompson, 4e Premier ministre du Canada, la fiche du RPCQ lui attribue le 10 novembre 1844 comme date de naissance[21] tandis que *DBpedia* stipule qu'il s'agit de la même date, mais en 1845[22]. Rapidement, on peut penser qu'il s'agit d'une simple erreur de saisie et une validation auprès du DBC semble donner raison à *DBpedia*[23]. La liaison ou la juxtaposition de données force l'historien à confronter directement des éléments contradictoires en contextualisant la donnée au sein de différentes institutions.

---

[19] Ministère de la Culture et des Communications du Québec, *Martin, Robert Montgomery*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=15111&type=pge#.VnHK_-I1uzw, 2013, consulté le 16 décembre 2015.

[20] DBpedia, *About: Robert Montgomery Martin*, http://dbpedia.org/page/Robert_Montgomery_Martin, consulté le 16 décembre 2015.

[21] Ministère de la Culture et des Communications du Québec, *Thompson, John Sparrow David*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=19600&type=pge#.VnHQJuI1uzy, 2013, consulté le 16 décembre 2015.

[22] DBpedia, *About: John Sparrow David Thompson*, http://dbpedia.org/page/John_Sparrow_David_Thompson, consulté le 16 décembre 2015.

[23] Université Laval et Université de Toronto, *Biography – THOMPSON, Sir JOHN SPARROW DAVID*, http://www.biographi.ca/en/bio/thompson_john_sparrow_david_12E.html, 2015, consulté le 16 décembre 2015.

Deuxièmement, rappelons que les fiches-personnes du RPCQ découlent d'une démarche d'inscrire au sein du répertoire, le patrimoine immatériel. Ce dernier a donc été extrait des fiches des biens matériels afin de les exposer comme éléments du patrimoine à part entière. Cependant cette mise de l'avant n'a pas toujours été accompagnée d'une recherche approfondie permettant de comprendre le rôle de l'individu dans l'histoire du Québec. La majorité des fichiers d'autorité de *DBpedia* peut contribuer à l'enrichissement des fiches ou, du moins, à leur validation.

L'exemple de Roméo Leblanc permet de démontrer concrètement le potentiel d'enrichissement. Sa fiche personnelle sur le site du RPCQ est pratiquement vide. Seules trois informations permettent d'identifier le personnage. La première mentionne qu'il est une personne, la seconde qu'il est un homme et la troisième l'associe à une plaque commémorative. Cette plaque de l'*Édifice-Louis-S.-St-Laurent* mentionne : « Restauré par *Travaux Publics Canada*, l'Honorable Roméo Leblanc, Ministre, inauguré par l'Honorable Pierre Bussières, Ministre du Revenu National le 11 juin 1984[24]. » L'information que propose le RPCQ se limite à ces champs. En créant une liaison avec *DBpedia*, le répertoire pourrait obtenir un grand nombre d'informations sans devoir effectuer une recherche supplémentaire. En plus d'une excellente description qui retrace le parcours du 25e gouverneur du Canada, on associe à Roméo Leblanc des dates, des lieux et des personnes. Par exemple, on apprend qu'il est né à Memramcook, au Nouveau-Brunswick, et que Fernand Robichaud fut son successeur en tant que député de la circonscription de Westmorland-Kent. Cette dernière liaison n'est pas aussi explicite

[24]Ministère de la Culture et des Communications du Québec, *Plaque de l'Édifice-Louis-S.-St-Laurent*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=111400&type=bien#.VnHVX-I1uzx, 2013, consulté le 16 décembre 2015.

puisque l'ontologie de *DBpedia* mentionne uniquement que Fernand Robichaud a succédé à Roméo Leblanc sans préciser le poste[25]. Dans ce cas-ci, il a fallu passer par la page *Wikipedia* de la circonscription pour confirmer que Fernand Robichaud a été député de Westmorland-Kent et non gouverneur général[26]. Malgré les lacunes ontologiques de *DBpedia*, le potentiel associatif est prometteur pour l'enrichissement du contenu du RPCQ. Il s'agit maintenant d'inclure ces données dans un modèle ontologique efficient qui facilite la cohésion et la contextualisation du patrimoine québécois.

Ontologie(s), plate-forme fédératrice et limite des compétences historiennes

Vocabulaires transdisciplinaires, l'exemple du CNP

Le MCC doit miser sur différents aspects pour développer un système de données liées durable. Sans parler des besoins d'une politique sur l'interopérabilité des données et des formations en humanités numériques à soutenir et instaurer, il faut repenser notre approche sectorielle face à la donnée. À ce sujet Bernhard Haslhofer et Antoine Isaac soulignent:

> A vast number of Europe's cultural heritage objects are digitised by a wide range of data providers from the library, museum, archive and audio-visual sectors, and they all use different metadata standards. This data needs to appear in a meaningful way in a crosscultural, multilingual context such as Europeana. Numerous cultural heritage resources such as thesauri exist worldwide and have the potential to add valuable content at low cost when re-used. Duplication of effort, however, needs to be avoided. The Linked Open Data environment lacks authoritative data from the cultural heritage community to contribute to the development of new knowledge[27].

[25]DBpedia, *About: Roméo LeBlanc*, http://dbpedia.org/page/Rom%C3%A9o_LeBlanc, consulté le 16 décembre 2015.
[26]Wikipedia, *Westmorland—Kent (circonscription fédérale)*, https://fr.wikipedia.org/wiki/Westmorland%E2%80%94Kent_(circonscription_f%C3%A9d%C3%A9rale, consulté le 16 décembre 2015.
[27]Bernhard Haslhofer et Antoine Isaac, « The Europeana Data Model for Cultural Heritage », *Europeana*, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Factsheet.pdf, p. 1.

En d'autres mots, il faut que le RPCQ s'intègre à une structure plus large au lieu de se percevoir comme un ensemble uniquement fermé mais cohérent. En effet, l'ouverture ne nécessite pas la disparition du système de classification traditionnelle. L'association avec d'autres bases de données permet d'ajouter du contenu complémentaire sans avoir à effectuer la recherche tout en augmentant la visibilité des données via un nouveau vecteur d'informations, le Web sémantique.

Les professions répertoriées dans le RPCQ devront, par exemple, s'appuyer sur un modèle plus général comme le CNP. Dans un système de données liées complet et national, on pourrait directement lier les données de *Statistiques Canada* avec les données du RPCQ afin de contextualiser une profession. À partir d'une fiche d'un marchand, il serait possible de calculer automatiquement le pourcentage de personnes actives dans le secteur commercial et dans les autres secteurs représentées dans le RPCQ pour sa ville afin d'évaluer sa représentativité. Cependant, la comparaison des catégories générales annonce le travail d'arrimage complexe qu'il faudra faire entre les deux vocabulaires. Le RPCQ possède un corpus circonscrit lui permettant d'avoir une plus large catégorisation de départ (16 entrées[28]) que celle du CNP (10 entrées[29]). Un autre problème majeur du CNP est qu'il s'intéresse aux professions actuelles. Il serait donc impossible d'attribuer la profession d'explorateur à Jacques Cartier en se basant strictement sur le modèle du CNP. Par ailleurs, le vocabulaire du RPCQ propose

[28] Ministère de la Culture et des Communications du Québec, *Recherche - Répertoire du patrimoine culturel du Québec*, *op. cit.*

[29] Gouvernement du Canada et Statistique Canada, *Classification nationale des professions (CNP) 2011*, http://www23.statcan.gc.ca/imdb/p3VD_f.pl?Function=getVDPage1&db=imdb&dis=2&adm=8&TVD=122372, 6 janvier 2012, consulté le 17 décembre 2015.

l'occupation de « Navigateur / pilote » qui pourrait aussi être attribué au personnage malgré que ce choix n'a pas été fait pour la fiche de l'explorateur. En extrapolant le modèle du CNP, Jacques Cartier, à titre de capitaine du bateau, serait différencié d'un matelot de pont par exemple. Le matelot serait classé dans « Matelots de ponts et matelots de salle des machines du transport par voies navigables[30] » (identifiant: 7532) et Jacques Cartier hypothétiquement dans « Opérateurs/opératrices de bateau à moteur, de bac à câble et personnel assimilé[31] » (identifiant: 7533). La notion d'embarcation motorisée exclut nécessairement la flotte de Jacques Cartier. Puisque les thésaurus sont extensibles, on pourrait croire que l'ajout d'une catégorie « Opérateurs/opératrices de bateau à voiles » permettrait de résoudre le problème. Toutefois, ce type d'approche est susceptible d'exposer, au même niveau de correspondance, des éléments anachroniques. À ce titre, la meilleure intervention semble être d'aligner les classes principales des vocabulaires d'occupations constitués selon des cadres spatio-temporels. Par exemple, au Royaume-Uni, le *Primary, Secondary, Tertiary System* propose une nomenclature des professions basée sur l'analyse d'un projet de recherche qui se concentre sur la période 1379-1911[32]. La première division s'effectue selon le secteur d'activités (primaire, secondaire et tertiaire). Plus localement, l'étude sur la structure professionnelle de Montréal en 1825 de Jean-Paul Bernard, Paul-André Linteau et Jean-Claude Robert

[30]Gouvernement du Canada et Statistique Canada, *CNP 2011 - 7532 - Matelots de pont et matelots de salle des machines du transport par voies navigables*, http://www23.statcan.gc.ca/imdb/p3VD_f.pl?Function=getVD&TVD=122372&CVD=122376&CPV=7532&CST=01012011&CLV=4&MLV=4, 6 janvier 2012, consulté le 17 décembre 2015.
[31]Gouvernement du Canada et Statistique Canada, *CNP 2011 - 7533 - Opérateurs/opératrices de bateau à moteur, de bac à câble et personnel assimilé*, http://www23.statcan.gc.ca/imdb/p3VD_f.pl?Function=getVD&TVD=122372&CVD=122376&CPV=7533&CST=01012011&CLV=4&MLV=4, 6 janvier 2012, consulté le 17 décembre 2015.
[32]Edward Anthony Wrigley, « The PST system of classifying occupations », http://www.campop.geog.cam.ac.uk/research/projects/occupations/britain19c/papers/paper1.pdf, 2010, p. 13-17.

identifie la nomenclature la plus représentative selon la structure économique de l'époque grandement orientée autour du commerce[33]. Afin de structurer les professions dans le temps, il faut nécessairement débuter avec un cadre spatial qui délimitera les types de professions en vigueur pour ensuite subdiviser ses espaces en périodes temporelles afin de modéliser l'évolution du travail selon les époques. Les classes supérieures à aligner définiront alors un cadre spatio-temporel et les principes fondamentaux du système économique en place.

Repenser la structure par l'événement

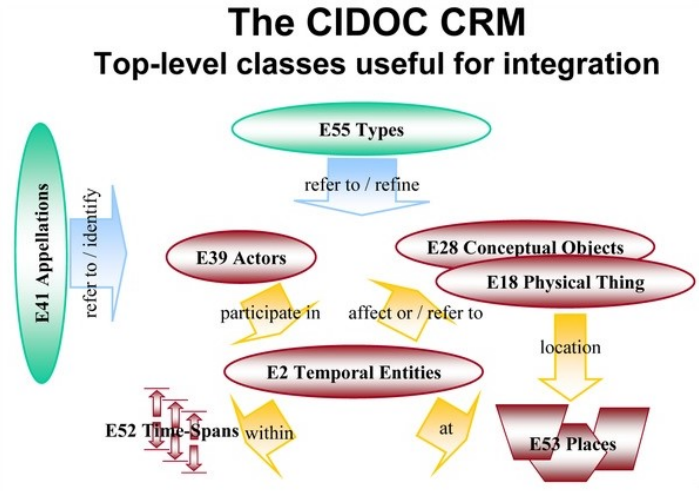
Pour le modèle général d'association, CIDOC CRM semble être le schéma de données le plus susceptible de répondre aux objectifs de fédération en sciences historiques. Les fiches-personnes du RPCQ permettent la compréhension du modèle et de ses capacités tout en affichant le chaînon manquant des données patrimoniales québécoises, soit la notion d'événement. Sur les 106 011 fiches du répertoire, seulement 71 sont des événements soit 0.07 % de celles-ci.

La fiche de la bataille des Plaines d'Abraham préfigure très bien le pouvoir de l'événement pour la contextualisation d'une partie du patrimoine québécois. Ce lieu de mémoire permet de lier 29 autres fiches du RPCQ et trois documents d'archives de BAnQ[34] tout en illustrant parfaitement les grandes classes du modèle CIDOC CRM tel que présentées par Stephen Stead :

---

[33]Jean-Paul Bernard, Paul-André Linteau et Jean-Claude Robert, « La structure professionnelle de Montréal en 1825 », *Revue d'histoire de l'Amérique française*, décembre 1976, vol. 30, n° 3, p. 388-391.
[34]Ministère de la Culture et des Communications du Québec, *Bataille des Plaines d'Abraham*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=25651&type=pge#.VnLjfuI1uzx, 2013, consulté le 17 décembre 2015.

Figure 3.1: Les classes générales du modèle CIDOC CRM



Source: Stephen Stead, *The CIDOC CRM, a Standard for the Integration of Cultural Information*, Capsule Web, CIDOC CRM, http://cidoc-crm.org/cidoc_tutorial/index.html, 2008, consulté le 14 janvier 2016.

La bataille des Plaines d'Abraham est donc une appellation d'une entité temporelle qui possède une durée. C'est à partir de cet événement qu'il est possible d'associer des acteurs qui y ont participé comme Louis-Joseph de Montcalm et James Wolfe[35]. Ces deux noms sont par ailleurs des appellations qui font références à deux entités humaines. Ensuite, cet événement entre en résonance avec des objets qui rappellent la bataille, comme un lot de dix plaques commémoratives ainsi que cinq livres. La notion de concept n'est pas utilisée directement dans la fiche du RPCQ, mais elle pourrait s'insérer parfaitement dans le schéma. Par exemple, le concept de conquête est intimement lié à cet événement qui engrangea cette idée de défaite des Français. La bataille s'est déroulée sur les plaines, ce qui complète le diagramme avec la notion d'endroit qui permet d'obtenir le cadre spatio-temporel. L'ensemble de ces classes peut être catégorisé par la notion de type. C'est par cette dernière que l'on peut intégrer les occupations structurées en SKOS selon le CNP et les autres vocabulaires professionnels.

[35] *Ibid.*

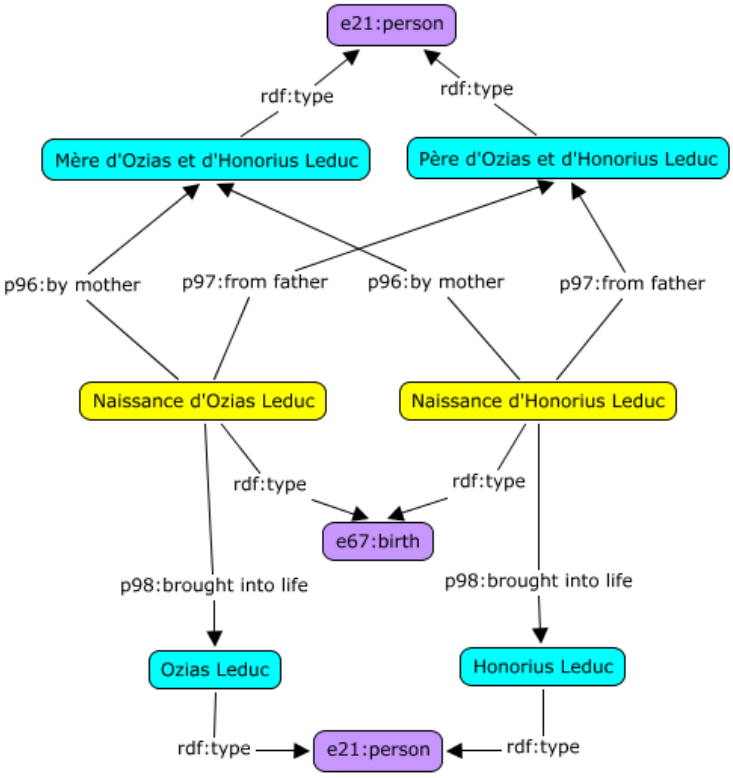
James Wolfe a participé à la bataille des Plaines d'Abraham en tant que commandant militaire de l'armée britannique, qui est un type d'occupation.

Le CIDOC CRM précisera les liens entre deux données. Les liaisons actuelles entre les différentes fiches du RPCQ ne permettent pas toujours de comprendre la raison de l'association. Par exemple, il est impossible de connaître la nature de l'implication de James Wolfe dans la bataille des Plaines d'Abraham. Ce rôle peut uniquement être connu dans le texte descriptif qui accompagne la fiche du personnage ou de l'événement.

Certains liens seront plus explicites grâce à l'accumulation des événements. Prenons en exemple les liens de parenté. Le peintre Ozias Leduc est lié à Honorius Leduc, mais aucune mention dans les éléments associés ne précise la nature de ce lien[36]. Il faut lire le descriptif pour réaliser qu'il s'agit de son frère. Leur événement de naissance respectif permettrait de trouver son frère par une requête relativement simple sans avoir à connaître le nom de ses parents.

[36] Ministère de la Culture et des Communications du Québec, *Leduc, Ozias*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=7700&type=pge#.VnLmR-I1uzw, 2013, consulté le 17 décembre 2015.

Figure 3.2: Modélisation d'une partie de la famille d'Ozias Leduc avec CIDOC CRM



On demande au système de trouver qui a donné naissance à Ozias ensuite on lui demande de trouver toutes les personnes qui sont nées lors de naissances impliquant la mère. Afin d'éviter qu'on trouve des demi-frères et demi-sœurs, il est possible d'ajouter le père dans la requête. On réalise encore que toute la recherche implique la notion d'événement. La naissance d'Ozias Leduc n'est plus un champ dans une fiche, mais un nœud majeur de liaisons. La création de données événementielles demande de repenser le modèle institutionnel omniprésent, mais dynamise grandement le potentiel associatif interinstitutionnel.

Plate-forme fédératrice: Rôle de l'historien et les limites de la discipline

Le RPCQ devra passer à travers diverses étapes pour mettre en place une plate-forme fédératrice en données liées. L'historien aura des responsabilités dans cette démarche ainsi que dans la pérennité du projet comme évoqué au premier chapitre. Comment diviser les rôles disciplinaires et professionnels pour développer un projet phare?

L'URI est au cœur du Web sémantique. Il est l'outil permettant l'insertion et la liaison de données au sein du nuage. Le MCC doit faire la promotion des URI en proposant des partenaires fiables et de qualité comme le LCSH par exemple. En plus, il doit encourager le développement d'URI québécois pour que le Québec devienne la figure d'autorité pour ses données culturelles. Rappelons que *DBpedia* possède une banque de données considérable sur les personnages historiques du Québec, mais qu'il est difficile de connaître la provenance de l'information. Si les institutions culturelles ne rendent pas accessible leurs données sous forme d'URI, le chercheur ou l'institution souhaitant utiliser le Web sémantique pour une recherche sera limité à la grande histoire et non aux différentes particularités de celle-ci.

Par la suite, il faut se questionner sur la méthode de fédération, où se distinguent deux méthodologies. La première consiste à lancer des projets de moindre ampleur en utilisant des ontologies reconnues et ainsi favoriser l'émergence de projets communs. Si deux institutions utilisent la même ontologie, un projet de recherche pourrait faire des requêtes fédérées facilement. Cependant, l'accumulation de différentes ontologies

risque, sur le long terme, de rendre les requêtes de plus en plus complexes. Le requérant aura besoin de connaître précisément le fonctionnement des différents systèmes fédérés pour en dégager des constats pertinents. De plus, un alignement des ontologies ne garantit pas une préservation du sens de chaque concept normalisé. Même si techniquement il est possible de multiplier les ontologies et les URI à l'infini et de créer des correspondances, la complexité de l'histoire nécessite un alignement réfléchi des classes généralistes. Si on limite les ontologies d'un projet patrimonial, on assure une correspondance exacte entre les classes et les propriétés. La deuxième méthodologie de fédération consiste donc à limiter les projets individuels au profit d'une plate-forme commune. Cette approche est à privilégier puisqu'elle limite la multiplication des ontologies. Au lieu de voir un chercheur et une institution comme des auteurs, ceux-ci prennent le rôle de participant à un ensemble plus vaste. Le système est alors conçu pour répondre aux besoins des différents partenaires et facilite les recherches fédérées.

À titre d'exemple d'accumulation, Matthew Lincoln offre un court article sur le site *The Programming Historian* pour expliquer comment faire une requête SPARQL sur un ensemble de données liées et termine sur cette question de chevauchement des ontologies[37]. Par exemple, *Europeana* ne détient aucune donnée de localisation, on doit plutôt interroger *DBpedia* pour ce type de question. Même si le point d'entrée SPARQL permet de faire des recherches décentralisées, le requérant doit connaître la structure de chaque modèle pour identifier les bons préfixes ainsi que les bons URI. Bref, en misant

[37] Matthew Lincoln, « Using SPARQL to access Linked Open Data », *Programming Historian*, http://programminghistorian.org/lessons/graph-databases-and-SPARQL, 24 novembre 2015, consulté le 17 janvier 2016.

sur une fédération par le bas, l'assemblage de données peut rapidement devenir difficile puisque les ramifications se multiplieront au fil des nouveaux projets en données liées.

Le RPCQ peut alors devenir l'instigateur d'un portail de données liées avec sa propre méthodologie et ses ontologies. Rappelons qu'en plus d'une réflexion théorique, un tel projet nécessitera un investissement massif. Il faut financer la création des URI, l'implantation de bases de données, le développement d'interfaces, les différentes mises à jour et la formation académique. À l'instar d'*Europeana*, le RPCQ ferait office de bibliothèque culturelle numérique, au sens large, où les partenaires viendraient contribuer. Par ailleurs, étant détenu par le MCC, le RPCQ fait figure d'autorité auprès des grandes institutions d'état. C'est aussi le MCC, en collaboration avec BAnQ, qui chapeaute la réflexion entourant le Web sémantique entamée par un comité d'experts mis sur pied dans le cadre du PCNQ. Les professionnels en sciences historiques doivent donc réfléchir à un langage transversal qui permettra de réunir différents concepts définis différemment selon les disciplines. CIDOC CRM est un point de départ incontournable, mais son adaptation aux modèles culturels et institutionnels québécois est obligatoire.

Cette réflexion conceptuelle ne nécessite pas de compétences techniques et est primordiale avant d'impliquer des programmeurs ou des informaticiens. Ces derniers sont toutefois nécessaires pour la mise en place de la plate-forme avec un point d'entrée SPARQL et un *TripleStore* pour emmagasiner les triplets. Évidemment, il est possible de simplement produire un fichier RDF qui peut être inséré dans divers programmes, mais l'objectif ici est d'avoir une interface simple d'utilisation pour démocratiser la

création de données liées. Il faut transformer la consultation SPARQL en interface classique sous forme de liste de champs. Cette dernière étape est sans doute la plus importante puisqu'elle déterminera l'acceptation ou non de l'outil auprès des acteurs du milieu. Bien que le programmeur possède les compétences pour réaliser une interface conviviale, il ne peut identifier seul les besoins spécifiques des sciences historiques. Les chercheurs en humanités numériques facilitent ce transfert de connaissances, ayant un pied dans chaque discipline.

La dernière étape est la pérennité de cet outil. Le numérique est avantageux pour l'accessibilité à l'information et le Web sémantique en est une démonstration éloquente, mais, en contrepartie, extrêmement éphémère[38]. Comment assurer la pérennité des supports et des contenus sous forme de données liées? Pour le support, il faut que l'informaticien sépare clairement les URI des interfaces de consultation. À ce titre, il peut utiliser des *Persistent URL* (*PURL*) qui permettent de pérenniser des identifiants sur des sites web basiques et qui ne sont pas influencés par les changements rapides du Web[39]. L'information est compilée sur des pages web simples en attribuant un chemin vers le fichier du serveur correspondant. Dans le domaine culturel purl.org, soutenu et développé par l'*Online Computer Library Center* (OCLC), est un choix incontournable puisque le modèle a été réfléchi par des bibliothécaires professionnels dans l'objectif premier de créer une structure facilitant les recherches décentralisées[40].

---

[38] Isabelle Boydens, « La conservation numérique des données de gestion », *Document numérique*, 1er juin 2004, vol. 8, no 2, p. 13.
[39] Online Computer Library Center, *PURL*, https://purl.org/docs/index.html, consulté le 17 décembre 2015.
[40] *Ibid.*

L'essentiel ici demeure d'assurer la pérennité des données liées culturelles québécoises en faisant en sorte que les institutions et les chercheurs collaborent à l'outil sur une base régulière. Cette multiplication des auteurs demande une implication particulière des historiens et des chercheurs en sciences historiques afin d'assurer la qualité des données. L'historien est l'un des seuls professionnels apte à maintenir une cohésion et une justesse des données et des liens entre celles-ci. De plus, Jonathan Rochkind propose une mise en garde face à l'idée que le Web sémantique offre une solution magique aux problèmes de fédération[41]. Pour que les données liées soient efficaces, il faut une réflexion en amont centrée sur les besoins de nos utilisateurs :

> « Linked data » can't be your goal. You are using linked data to accomplish something to add value to your patrons. We must understand what our patrons are doing, and how to intervene to improve their lives. We must figure out what services and systems we need to do that. Some work to that end, even incomplete and undeveloped if still serious and engaged, comes *before* figuring out what data we need to create those services. To the extent it's about data, make sure your data modeling work and choices are about creating the data we need to serve our users, not just fitting into the linked data model[42].

Les données liées sont un outil et non un objectif en soi. Dans l'idée d'amener plus de cohésion dans le domaine des sciences historiques, le Web sémantique semble une avenue fort prometteuse. La réussite ou l'échec d'un tel projet n'en tient qu'à l'équipe de professionnels qui le pilotera. La connaissance historique est interdisciplinaire et l'outil qui permettra sa découverte se doit de l'être aussi.

Ce dernier chapitre met en lumière certaines difficultés d'application des données liées dans le contexte culturel québécois. En démontrant que le RPCQ est une base de

---

[41] Jonathan Rochkind, *Linked Data Caution*, https://bibwild.wordpress.com/2015/11/23/linked-data-caution/, 23 novembre 2015, consulté le 17 décembre 2015.
[42] *Ibid.*

données suffisamment diversifiée pour porter un projet de plate-forme fédératrice et collaborative, une expérimentation sur les personnages du répertoire montre les avantages des données liées et les compétences à posséder pour manœuvrer un tel projet. L'association avec des URI déjà existants peut se faire par des NER. Cependant, *DBpedia Spotlight* nécessite une rétroaction humaine pour vérifier les correspondances. Au cœur du processus de création de la plate-forme se trouvent les professionnels en sciences historiques qui doivent miser sur des vocabulaires et des ontologies interdisciplinaires répondant aux besoins des utilisateurs. Les données liées historiques québécoises montrent seulement les grands axes de l'histoire québécoise sans entrer dans les particularités auxquelles les chercheurs actuels s'intéressent.

CONCLUSION

Le Web sémantique comme nouvelle approche en histoire

L'historien, en tant qu'artisan de l'archive et transmetteur du passé au présent, doit inévitablement s'adapter aux nouvelles technologies afin d'éviter de se voir imposer des outils inadaptés à sa discipline. Le Web sémantique en fait certainement partie. Les données liées qui sous-tendent ce Web 3.0 demandent à l'historien une interdisciplinarité qui l'amènera vers les sciences de l'information. L'historien qui choisit de s'y investir se rapprochera du rôle d'architecte de l'information historique. En contribuant au Web sémantique, l'historien augmentera l'interopérabilité et la cohésion des données historiques afin d'offrir une nouvelle option de recherche fédérée.

Le Web sémantique nécessite un lot de connaissances externes aux domaines historiques, à commencer par la classification suivant des normes internationales. Il faut connaître les fichiers d'autorité qui permettront de lier nos données avec des partenaires compétents dans un domaine qu'on ne peut investir pleinement. Le Web sémantique mise sur le concept de décentralisation de l'information présentée dans une interface fédérée. L'URI, cet identifiant unique attribuable à n'importe quel type de contenu disponible sur le Web, désambiguïse les homonymes et les données décontextualisées. Le modèle RDF, sous forme de triplets, permet l'association d'URI afin de faciliter les mises en relations.

Cette conceptualisation du Web et ses premières applications remontent au début des années 2000. Cependant, le Québec, comparativement à la France, l'Australie et les

États-Unis, possède un net retard dans l'assimilation et l'utilisation de cette nouvelle méthode. Contrairement à ces pays, aucune structure nationale ou provinciale n'a vu le jour afin de donner aux chercheurs et aux institutions un point d'ancrage pour démarrer des projets sémantiques. *Europeana*, qui fut maintes fois utilisé à titre d'exemple dans ce mémoire, est un modèle dont le Québec et le Canada devraient grandement s'inspirer.

Pour augmenter le taux de cohésion des données, il fut proposé de miser sur une fédération par le haut, c'est-à-dire qu'une institution instaure un modèle interdisciplinaire auquel les institutions et les chercheurs adhèrent. Si l'on souhaite qu'une masse critique d'historiens investissent le Web sémantique, il faut que les institutions patrimoniales majeures instaurent des standards et proposent des URI. À ce titre, le RPCQ semble un point de départ intéressant, de par sa diversité de contenus. Chapeauté par le MCC et BAnQ, deux organisations qui font figures d'autorité en culture québécoise, le répertoire est la plate-forme qui peut prétendre à un financement suffisant pour déployer une stratégie en données liées.

Cette stratégie débute avec l'utilisation d'URI. À ce sujet, il faut éviter de multiplier la création d'URI déjà existants. Si un acteur du Web sémantique, prenons en exemple le LCSH, a déjà créé un identifiant pour une entité qu'on souhaite décrire, il est souhaitable d'utiliser son URI plutôt que d'en ajouter un nouveau qui ne fait qu'augmenter les points d'entrées vers une même information. Par contre, la création d'identifiants uniques est nécessaire lorsque le contenu n'existe pas dans le nuage sémantique. Le Québec doit gérer les URI qui concernent son patrimoine culturel

puisqu'il est l'autorité qui possède les sources capables de documenter et de valider chaque triplet.

Ensuite, le ministère et BAnQ devront réfléchir à la structure des données. Les ontologies qu'on affecte à la partie centrale du triplet (le prédicat) inscrivent chaque triplet dans une plus large modélisation. Dans le cadre de ce mémoire, le CIDOC CRM est le schéma le plus susceptible de répondre aux besoins des disciplines historiques. Cette structure oblige à repenser le modèle organisationnel actuel pour y inclure un plus grand nombre d'événements comme agrégateurs du patrimoine. Rappelons, en exemple, que seulement 0,06 % des fiches du RPCQ concernent des événements. Néanmoins CIDOC CRM facilite l'interopérabilité des institutions culturelles en offrant un langage transversal où chaque discipline peut inscrire son vocabulaire.

C'est par cet environnement que les données liées permettent aux ordinateurs de mieux traiter l'information qu'ils partagent entre eux. On place la donnée au cœur de l'architecture numérique au lieu de transiger par l'intermédiaire de sites web qui font office de silos. Au lieu de demander à l'historien de connaître parfaitement chaque système de données, le Web sémantique lui propose de constituer sa propre interface selon ses intérêts de recherche. Cependant, l'utilisation des données liées n'est pas synonyme d'un système utile et performant. L'historien et les professionnels en sciences historiques permettront la création de vocabulaires pertinents tout en assurant la pérennité des systèmes en établissant une veille et une méthodologie de validation des liens. Il faut nécessairement un investissement intellectuel et financier afin de

développer une expertise dans la compréhension et l'utilisation de cette nouvelle méthodologie.

Pour ce faire, l'historien n'a pas à devenir un programmeur comme le prédisait Emmanuel Leroy Ladurie, il doit simplement être en mesure de contribuer au nuage des données liées par le biais d'une interface simple d'utilisation ainsi que de repérer les données susceptibles de l'intéresser. En d'autres mots, il s'agit d'automatiser la méthodologie historienne qui depuis les balbutiements de la discipline associe des faits pour en dégager des récits et des constats. À ce titre, il faut suivre le projet de *Système modulaire de gestion de l'information historique* (SyMoGIH) développé actuellement par le *Laboratoire de recherche historique Rhône-Alpes* et qui ouvre la voie à des vocabulaires spécifiquement conçus pour la discipline historique[1]. Cette approche favorisera grandement l'interopérabilité des données de l'historien. Au lieu d'un travail à l'échelle humaine et individuelle, les données liées proposent de présenter les récits historiques sous la forme d'une toile de liens qui peut être explorée et analysée de différents angles. On évite alors l'aplanissement des faits et la réduction contextuelle qu'oblige le récit textuel linéaire.

Le Web sémantique permettra-t-il l'accomplissement des idéaux de l'histoire totale? Sans pouvoir affirmer que le Web serait en mesure de présenter l'histoire dans sa totalité, les données liées ouvrent certainement à nouveau la porte aux études en histoire de longue durée. Si l'historien accepte de participer à un projet collectif au lieu d'être

[1]Laboratoire de recherche historique Rhône-Alpes, *Le projet symogih.org : un système modulaire de gestion de l'information historique*, http://www.symogih.org/, 2015, consulté le 15 avril 2016.

l'auteur d'une parcelle isolée de l'histoire, l'ensemble pourrait se rapprocher d'une compréhension totale de l'histoire ou, du moins, d'une totalité que peuvent appréhender l'ensemble des utilisateurs des données liées. Ensuite, l'ordinateur pourrait en morceler et livrer des parties selon les vocabulaires et les ontologies choisies par un chercheur.

Une fusion des modèles de classification institutionnel et académique

Les données liées et leur objectif de fédération de l'information révèlent un écart majeur entre la classification institutionnelle et celle employée en recherche. En institution, on organise selon le support ou la provenance, soit une classification à partir des métadonnées. En recherche, on organise à partir de la donnée, ce sont les contenus qu'on doit associer entre eux. Cette différence est corolaire à la division du patrimoine en différentes disciplines et institutions.

Le Web sémantique permet de constater cette dichotomie et son utilisation démarrera très certainement une discussion sur l'arrimage entre le monde institutionnel et académique. Derrière les lignes de ce mémoire se cache la nécessité de briser les barrières disciplinaires, mais aussi celles professionnelles. Il faut repenser l'expérience utilisateur pour que l'historien n'ait plus à supporter le rôle de fédérateur de l'histoire. Un nouvel environnement doit venir se juxtaposer ou remplacer le modèle actuel pour que les bibliothécaires professionnels, les archivistes, les muséologues, les archéologues, les historiens de l'art et les historiens travaillent en symbiose. Il faut éviter l'isolement justifié par une accessibilité aux données culturelles. Rendre accessible ces corpus à un chercheur n'est pas un exemple concret d'interdisciplinarité.

Pour y parvenir, le domaine culturel peut s'inspirer du *Cadre commun d'interopérabilité* du *Gouvernement du Québec*[2]. Bien qu'il s'agisse uniquement d'une recommandation pour les organismes gouvernementaux québécois, ce document démontre l'importance de l'interopérabilité afin d'accroître le rendement des systèmes d'information. Évidemment, les auteurs proposent une série de mesures pour favoriser la mise en place de ces interconnexions, dont RDF fait partie. L'arrimage entre l'institutionnel et l'académique doit passer par trois approches qui synthétisent parfaitement les propos évoqués dans ce mémoire : un point de vue stratégique pour établir une vision précise des besoins et des objectifs, un point de vue conceptuel pour définir les standards nécessaires à la conception et, finalement, un point de vue opérationnel pour la mise en œuvre des protocoles et la réalisation des systèmes[3]. Cette stratégie de normalisation et de fédération est plus que nécessaire pour assurer la cohérence des données patrimoniales au sein, par exemple, du RPCQ. Une approche interdisciplinaire et interprofessionnelle permettra d'étudier plus aisément le patrimoine québécois et d'en dégager de nouvelles problématiques. D'un côté, il serait plus facile de traiter des questions de genres en patrimoine, d'intégration des communautés ethniques dans un modèle français du patrimoine, d'enjeux de mémoire en résonance avec des débats historiographiques, etc. De l'autre côté, une telle approche permettrait d'étudier les relations entre le patrimoine québécois et les patrimoines mondiaux puisque ceux-ci seraient structurés suivant des normes internationales évitant ainsi les barrières

[2] Talel Kokobi, *Cadre commun d'interopérabilité du Gouvernement du Québec: Normaliser, S'aligner, Performer*, http://www.tresor.gouv.qc.ca/fileadmin/PDF/ressources_informationnelles/cadre_commun_interoperabilite.pdf , 2014, p. 70.
[3] *Ibid.,* p.18-19.

culturelles et linguistiques. Malgré que le statisticien George Box rappelle qu'aucun modèle n'est parfait, il faut se souvenir que parfois certains sont fort utiles[4].

[4] George Box et Norman Richard Draper, *Empirical model-building and response surfaces*, New York, Wiley (coll. « Wiley series in probability and mathematical statistics »), 1987, p. 424.

ANNEXE I

Fonctionnement du point d'entrée SPARQL d'*Europeana*

En s'appuyant sur les exemples de Matthew Lincoln[1], l'interrogation de ce point d'entrée permet de comprendre le potentiel de validation d'un tel outil. Par exemple, je souhaite connaître les auteurs, répertoriés sur *Europeana*, qui ont écrit un livre en langue française. À titre de réponse, j'aimerais connaître, en plus du nom des auteurs, le titre de leur ouvrage ainsi que l'institution qui confirme cette information. Pour ce faire je devrai effectuer la requête suivante :

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT ?livre ?id ?auteur
WHERE {
?livre dc:creator ?auteur .
?livre dc:identifier ?id .
?livre dc:language "fr" .
}
LIMIT 10
```

Les deux premières entrées concernent les URI qui seront utilisés dans la présente requête. FoaF permettra de répertorier des personnes lors de la recherche, malgré qu'aucun prédicat de ce dernier ne soit spécifié de prime abord. Ensuite, la fonction SELECT permet de désigner les variables qui seront nécessaires à l'obtention de nos réponses. On peut écrire les variables de notre choix, simplement utiliser les mêmes lors de la rédaction de la requête. On cherche alors les livres, les identifiants (id), qui correspondent dans *Europeana* aux institutions, et les auteurs. Ensuite une série de triplets permet d'identifier ce qu'est un livre tout en précisant les informations souhaitées. Les deux premiers servent à spécifier nos deux secondes variables et le

[1] Matthew Lincoln, *loc. cit.*

dernier à préciser la langue française (sur *Europeana*: fr). La limite (en anglais: *limit*) permet simplement d'obtenir un nombre maximal de requêtes puisque, parfois, le volume peut devenir très élevé et le résultat tardera à s'afficher.

La réponse se présente alors sous forme de tableau avec les trois variables spécifiées. Dans le cas présent, on obtient 10 résultats (notre limite), mais on remarque que seulement quatre sont différents. On retrouve des doublons puisque certaines informations proviennent de deux institutions différentes. Il s'agit d'un excellent moyen de validation de l'information puisque deux institutions fournissent exactement le même résultat.

ANNEXE II

Tableau 2.1: Comparatif entre les ontologies FoaF et BIO

Faits historiques	Ontologies	Sujet du triplet	Prédicat du triplet	Objet du triplet
Jacques Cartier est l'oncle de Jacques Noël	FoaF	Jacques Cartier	knows	Jacques Noël
	BIO	Jacques Cartier	child	Parents de Jacques Cartier
		Jacques Noël	child	Jehanne Cartier
		Jehanne Cartier	child	Parents de Jacques Cartier
Jacques Cartier est le supérieur de Thomas Fromont	FoaF	Jacques Cartier	knows	Thomas Fromont
	BIO	La traversée de l'Atlantique	employer	Jacques Cartier
		La traversée de l'Atlantique	agent	Thomas Fromont
Jacques Cartier fut marié avec Catherine Des Granches de leur mariage à la mort de l'explorateur	FoaF	Jacques Cartier	knows	Catherine Des Granches
	BIO	Jacques Cartier	participant	Relation conjugale de Jacques Cartier et Catherine Des Granches
		Catherine Des Granches	participant	Relation conjugale de Jacques Cartier et Catherine Des Granches
		Relation conjugale de Jacques Cartier et Catherine Des Granches	interval	Période de vie conjuale de Jacques Cartier et Catherine Des Granches
		Période de vie conjuale de Jacques Cartier et Catherine Des Granches	initiating event	Mariage de Jacques Cartier et Catherine Des Granches
		Mariage de Jacques Cartier et Catherine Des Granches	date	1519
		Période de vie conjuale de Jacques Cartier et Catherine Des Granches	concluding event	Décès de Jacques Cartier
		Décès de Jacques Cartier	date	1557

Sources: Ian Davis et David Galbraith, *BIO: A vocabulary for biographical information*, http://vocab.org/bio/0.1/.html, 14 juin 2011, consulté le 28 août 2015 et Dan Brickley et Libby Miller, *FOAF Vocabulary Specification*, http://xmlns.com/foaf/spec/ #term_knows, 14 janvier 2014, consulté le 28 août 2015.

ANNEXE III

Exploration détaillé du EDM

Tout comme l'ontologie CIDOC CRM, l'EDM s'intéresse à tous types d'institutions culturelles et adapte son modèle en conséquence, bien que sa gestion de l'évolution dans le temps soit inexistante et que les types de relations entre les personnes soient limitées à un concept mal défini.

Tableau 2.2: Propriétés d'un agent dans l'EDM

Properties for the edm:Agent	
+ skos:prefLabel	foaf:name
skos:altLabel	rdaGr2:biographicalInformation
skos:note	rdaGr2:dateOfBirth
dc:date	rdaGr2:dateOfDeath
dc:identifier	rdaGr2:dateOfEstablishment
dcterms:hasPart	rdaGr2:dateOfTermination
dcterms:isPartOf	rdaGr2:gender
edm:begin	rdaGr2:placeOfBirth
edm:end	rdaGr2:placeOfDeath
edm:hasMet	rdaGr2:professionOrOccupation
edm:isRelatedTo	owl:sameAs

Source: Robina Clayphan, Valentine Charles, et Antoine Isaac, *Europeana Data Model - Mapping Guidelines v.2.2*, s.l., 2014, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.2.pdf, p. 9.

Dans cet exemple, tiré du *Europeana Data Model – Mapping Guidelines v2.2*, on constate l'impossibilité de créer des liens de parenté clairs entre des personnes[1] comme peut aisément le faire CIDOC CRM. L'utilisation excessive de *edm :hasMet* (en français: a rencontré) est aussi problématique puisque le prédicat est défini de cette façon : « The identifier of an agent, a place, a time period or any other identifiable entity

[1] Robina Clayphan, Valentine Charles, et Antoine Isaac, *Europeana Data Model - Mapping Guidelines v.2.2*, s.l., 2014, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.2.pdf, p. 9.

that the CHO (Cultural Heritage Object) may have “met” in its life[2]. » La présence des guillemets est somme toute particulière puisqu’elles semblent inviter à créer des liens questionnables. Le schéma présenté en page 13 d’*Europeana Data Model Primer* en est un exemple éloquent[3].

Figure 2.3: Liaison du « Portrait de Mona Lisa » sur *Europeana*

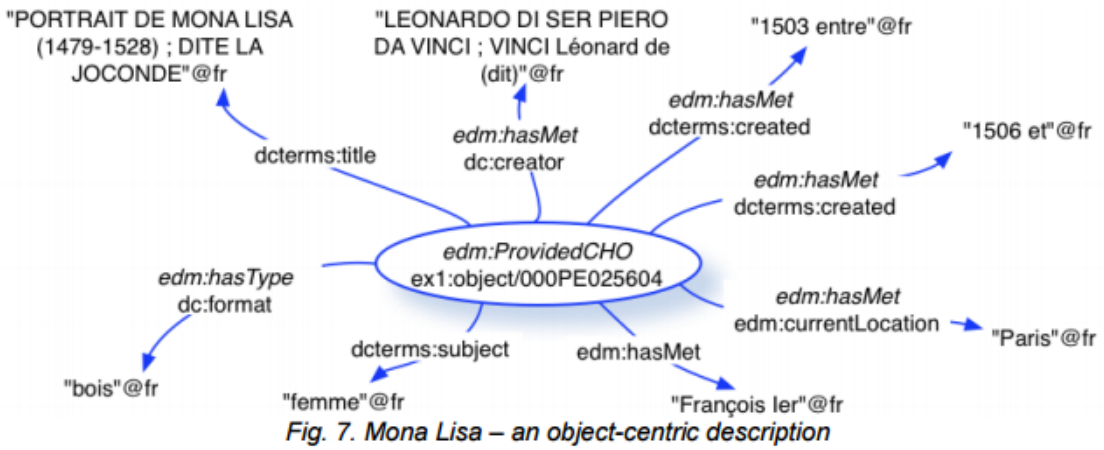


Fig. 7. Mona Lisa – an object-centric description

Source: Antoine Isaac, *Europeana Data Model Primer*, s.l., 2013, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf, p.13.

On définit ici les métadonnées entourant l’œuvre de Léonard de Vinci « Portrait de Mona Lisa ».

On utilise à cinq reprises le concept d’*edm: hasMet*. La première question serait de se demander si une œuvre peut réellement rencontrer quelqu’un ou quelque chose. Dans l’exemple, on peut « lire » que l’œuvre a rencontré son créateur Léonard de Vinci ainsi que François 1er. Les deux autres cas sont encore plus particuliers. On mentionne que l’œuvre a rencontré la ville de Paris ainsi que la période de 1503 à 1506. Ce dernier exemple est d’autant plus faux puisque le « Portrait de Mona Lisa » est toujours à Paris

[2]Robina Clayphan, Valentine Charles, et Antoine Isaac, *op. cit.*, p.20.
[3]Antoine Isaac, *Europeana Data Model Primer*, s.l., 2013, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf, p.13.

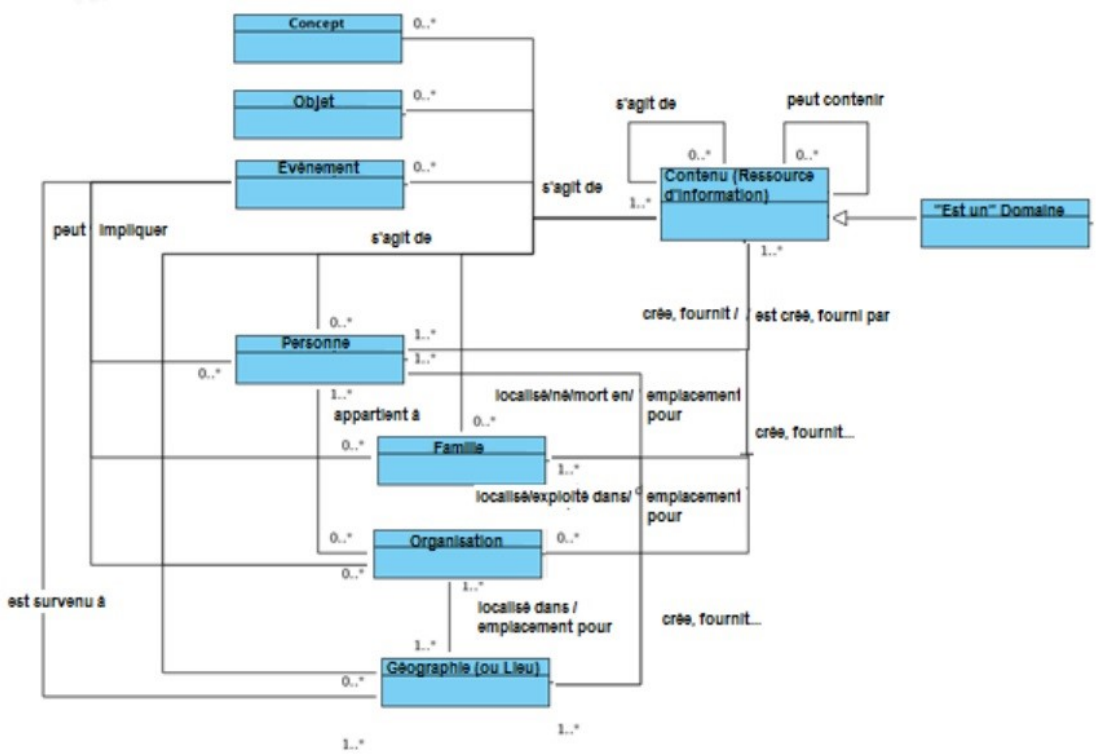
et continue donc de « rencontrer » Paris. Ce type d'erreur est normal lorsqu'on tente de généraliser notre méthode de classification pour correspondre aux besoins de multiples institutions et disciplines. Cependant, il faut constamment réfléchir à son amélioration pour obtenir un rendu conceptuel reflétant la réalité. Malgré tout, l'Europe possède un outil favorisant la cohésion et la mise en valeur de son patrimoine, ce qui n'est pas le cas du Canada et du Québec.

ANNEXE IV

Modèle ontologique et recommandations du projet *Au-delà des tranchées*

Un modèle ontologique a été spécialement développé pour ce projet spécifique qui s'appuie sur une quantité impressionnante d'autres ontologies comme DC et FoaF[1]. Puisqu'il s'agit d'un projet spécifique avec des balises connues, le modèle ontologique utilisé est relativement simple et se résume schématiquement ainsi :

Figure 2.4: Schéma conceptuel de l'EDM



Source: Réseau pancanadien du patrimoine documentaire, *« Démonstration de faisabilité » de la Visualisation des Données ouvertes liées (LOD)*, Canada, 2012, http://www.canadiana.ca/sites/pub.canadiana.ca /files/PCDHN%20Proof-of-concept_Final-Report-FRA.pdf, consulté le 27 août 2015, p.6.

Cette illustration montre un avantage important d'une ontologie, soit la possibilité de retrouver un concept à différents niveaux interreliés. Si on organise ce schéma suivant une hiérarchie, on retrouve au sommet la notion de *Contenu* (Ressource d'information)

[1] Pour la liste complète : Réseau pancanadien du patrimoine documentaire*, loc. cit.,* p.6.

qui englobe tous les autres encadrés (*Concept*, *Objet*, *Événement*, *Personne*, *Géographie (ou Lieu)*, etc.). Cependant, on vient créer des ponts entre les données pour augmenter les possibilités de consultations selon les visées du chercheur. Par exemple, même si un événement et un lieu sont tous les deux des contenus, il existe une correspondance entre ces deux entités soit qu'un événement est survenu dans un lieu précis. Toutefois, un lieu n'est pas seulement un endroit où survient un événement. Les multiples facettes du lieu s'exposent par les liens qu'il possède avec d'autres entités comme les personnes, les familles et les organisations. De plus, les liens avec ces autres entités ne sont pas du même ordre. Par exemple, une personne peut être localisée, née ou décédée dans un lieu et l'utilisation d'un terme générique peut porter à confusion.

La démonstration de faisabilité se termine sur des leçons et des recommandations du RPCPD qui affirment le rôle des données liées dans le domaine des sciences historiques et, par le fait même, confirment que l'historien doit collaborer à la mise en place de cette structure. « RDF et LOD (Linked Open Data) sont une approche élégante pour l'exploration des ressources intégratrices à travers différents domaines, institutions et services[2]. » Les données liées facilitent l'intégration des données du RPCPD dans des récits et des expositions virtuelles d'organisations externes. Cette intégration peut se faire consciemment ou inconsciemment. Cette seconde forme se produit lorsqu'une institution utilise les mêmes vocabulaires que le RPCPD. Une économie est aussi faite au niveau de la transmission des données puisque le processus est automatique et ne nécessite plus de transiger par une personne. Dans un même ordre d'idées, les données liées ne nécessitent pas une programmation immense et se concrétisent alors rapidement.

[2] *Ibid.*, p. 15.

Par exemple, dans le cas du projet *Au-delà des tranchées*, seulement trois mois furent nécessaires pour créer un logiciel de transformation des données et pour effectuer la migration de trois ensembles de celles-ci[3]. De plus, la réutilisation par l'utilisateur peut se faire selon ses propres objectifs. Autrement dit, l'institution ne fournit plus le chemin informationnel que doit suivre l'utilisateur. Le projet a démontré la richesse des ontologies existantes. Il est plus facile aujourd'hui d'inscrire nos données dans le nuage des données liées puisque des modèles existent déjà. De plus, la possibilité de jumeler différentes ontologies a évité au RPCPD de développer son propre ensemble d'éléments spécifique[4]. Malheureusement, aucune suite du projet ne semble être envisagée pour le moment malgré une liste de recommandations[5] publiées par le RPCPD.

[3] *Ibid.*, p.17.
[4] *Ibid.*, p.15.
[5] Pour la liste complète : *Ibid.*, p. 19.

ANNEXE V

Tableau 2.3: Comparatif des projets en données liées culturelles selon la charte des cinq étoiles

Projets	Première étoile	Deuxième étoile	Troisième étoile	Quatrième étoile	Cinquième étoile	Commentaires généraux sur la 5e étoile
CLAROS	☑	☑	☑	☑	☑	Limitée: Les liens pointent uniquement vers *Geonames*
British Museum	☑	☑	☑	☑	☒	Aucun URI des sujets ou des objets ne pointent ailleur que sur la collection même du musée
Maphub	☑	☑	☑	☑	☑	Limitée: Les liens pointent uniquement vers *DBpedia*
Muséosphère	☑	☑	☑	☑	☑	Limitée: Les liens pointent uniquement vers *DBpedia*
JocondeLab	☒	☑	☑	☑	☑	Limitée: Les liens pointent uniquement vers *DBpedia*
Europeana	☑	☑	☑	☑	☑	Modèle prometteur puisque des ponts sont créés avec, par exemple, le AAT
Au-delà des tranchées	☑	☑	☑	☑	☑	Modèle prometteur puisque des ponts sont créés avec, par exemple, le LCSH

BIBLIOGRAPHIE

I. Sources

AMERICAN PSYCHOLOGICAL ASSOCIATION. *Publication manual of the American Psychological Association*, 6e ed., Washington, DC, American Psychological Association, 2010, 272 p.

APACHE JENA. *What is Jena?*, http://jena.apache.org/about_jena/about.html, 2015, consulté le 28 août 2015.

AUSTRALIE, ARCHIVES NATIONALES D'. *The real face of white australia : experimental browser*, http://invisibleaustralians.org/faces/, consulté le 31 août 2015.

AUSTRALIE, GOUVERNEMENT DE L'. *Data.gov.au*, http://data.gov.au/, consulté le 31 août 2015.

BIBLIOTHEQUE ET ARCHIVES NATIONALES DU QUEBEC. *Mardi c'est Wiki*, https://fr.wikipedia.org/w/index.php?title=Wikip%C3%A9dia:BAnQ/Mardi_c%27est_Wiki&oldid=120982301, 2 décembre 2015, consulté le 16 décembre 2015.

BOUTREUX, Christophe. *GeoNames*, http://www.geonames.org, consulté le 26 janvier 2016.

BRICKLEY, Dan et MILLER, Libby. *FOAF Vocabulary Specification*, http://xmlns.com/foaf/spec/#term_knows, 14 janvier 2014, consulté le 28 août 2015.

BRITISH MUSEUM. *British Museum Semantic Web Collection Online*, http://collection.britishmuseum.org, consulté le 30 août 2015.

BRITISH MUSEUM. *Easter Island*, http://collection.britishmuseum.org/id/place/x69553, 2012, consulté le 30 août 2015.

BRITISH MUSEUM. *Hoa Hakananai'a*, http://collection.britishmuseum.org/id/object/EOC3130, 2012, consulté le 30 août 2015.

BRITISH MUSEUM. *Inlaid*, http://collection.britishmuseum.org/id/thesauri/x12176, 2012, consulté le 30 août 2015.

CALCUL QUEBEC. *À propos de Calcul Québec*, http://www.calculquebec.ca/fr/a-propos-de-cq/calcul-quebec, 2015, consulté le 8 octobre 2015.

CANADA, GOUVERNEMENT DU. *Cote de degré d'ouverture des données*, http://ouvert.canada.ca/fr/cote-degre-douvertrure-des-donnees, 25 avril 2013, consulté le 8 janvier 2016.

CANADA, GOUVERNEMENT DU et STATISTIQUE CANADA. *Classification nationale des professions (CNP) 2011*, http://www23.statcan.gc.ca/imdb/p3VD_f.pl?Function=getVDPage1&db=imdb&dis=2&adm=8&TVD=122372, 6 janvier 2012, consulté le 17 décembre 2015.

CANADA, GOUVERNEMENT DU et STATISTIQUE CANADA. *CNP 2011 - 7532 - Matelots de pont et matelots de salle des machines du transport par voies navigables*, http://www23.statcan.gc.ca/imdb/p3VD_f.pl?Function=getVD&TVD=122372&CVD=122376&CPV=7532&CST=01012011&CLV=4&MLV=4, 6 janvier 2012, consulté le 17 décembre 2015.

CANADA, GOUVERNEMENT DU et STATISTIQUE CANADA. *CNP 2011 - 7533 - Opérateurs/opératrices de bateau à moteur, de bac à câble et personnel assimilé*, http://www23.statcan.gc.ca/imdb/p3VD_f.pl?Function=getVD&TVD=122372&CVD=122376&CPV=7533&CST=01012011&CLV=4&MLV=4, 6 janvier 2012, consulté le 17 décembre 2015.

CANADA, GOUVERNEMENT DU et STATISTIQUE CANADA. *Entente de licence ouverte de Statistique Canada - Foire aux questions (FAQ)*, http://www.statcan.gc.ca/fra/reference/licence-faq-fra#a1, 25 janvier 2013, consulté le 8 janvier 2016.

CANADA, GOUVERNEMENT DU et STATISTIQUES CANADA. *Types des professions*, http://www.statcan.gc.ca/fra/concepts/profession, 21 novembre 2011, consulté le 16 décembre 2015.

CHAMPLAIN, Samuel de. *[Illustrations de Les Voyages de Champlain. ..] / [Non identifié] ; Samuel Champlain, aut. du texte*, s.l., http://gallica.bnf.fr/ark:/12148/btv1b2000019z via http://www.rechercheisidore.fr, 1613, consulté le 9 octobre 2015.

CLAROS. *Principles of CLAROS data extraction*, http://www.clarosnet.org/XDB/ASP/claroshome/technicalData.html, consulté le 14 janvier 2016.

COLLABORATIVE FOR HISTORICAL INFORMATION AND ANALYSIS. *Col*Fusion - Your entry to the data world*, http://colfusion.exp.sis.pitt.edu/colfusion/, 2016, consulté le 13 janvier 2016.

COLLABORATIVE FOR HISTORICAL INFORMATION AND ANALYSIS. *World-Historical Dataverse*, https://dataverse.harvard.edu/dataverse/worldhistorical, 2015, consulté le 13 janvier 2016.

CORPORATION FOR NATIONAL RESEARCH INITIATIVES. *Handle*, http://www.handle.net/, 25 août 2015, consulté le 9 octobre 2015.

DAVIS, Ian et GALBRAITH, David. *BIO: A vocabulary for biographical information*, http://vocab.org/bio/0.1/.html, 14 juin 2011, consulté le 28 août 2015.

DBPEDIA. *About | DBpedia*, http://dbpedia.org/about, 2014, consulté le 28 août 2015.

DBPEDIA. *About: Jacques Cartier*, http://dbpedia.org/page/Jacques_Cartier, consulté le 28 août 2015.

DBPEDIA. *About: John Scott*, http://dbpedia.org/page/John_Scott, consulté le 16 décembre 2015.

DBPEDIA. *About: John Sparrow David Thompson*, http://dbpedia.org/page/John_Sparrow_David_Thompson, consulté le 16 décembre 2015.

DBPEDIA. *About: Robert Montgomery Martin*, http://dbpedia.org/page/Robert_Montgomery_Martin, consulté le 16 décembre 2015.

DBPEDIA. *About: Roméo LeBlanc*, http://dbpedia.org/page/Rom%C3%A9o_LeBlanc, consulté le 16 décembre 2015.

DBPEDIA. *About: Montréal*, http://dbpedia.org/page/Montreal, consulté le 30 août 2015.

DBPEDIA. *Internationalization*, http://wiki.dbpedia.org/Internationalization/, 2014, consulté le 30 août 2015.

DBPEDIA. *DBpedia Spotlight*, https://dbpedia-spotlight.github.io/demo/, consulté le 16 décembre 2015.

DUBLIN CORE METADATA INITIATIVE. *Dublin Core Metadata Element Set, Version 1.1*, http://dublincore.org/documents/dces/, 14 juin 2012, consulté le 28 août 2015.

DUMITRU, Mara. *Andy Warhol*, Museosphere, http://museosphere.net/inspire/fr?artist=Andy+Warhol&mouvement=&localisation= , 2013, consulté le 30 août 2015.

DUMITRU, Mara. *Gustave Courbet*, Museosphere, http://museosphere.net/inspire/fr?artist=Gustave+Courbet&mouvement=&localisation= , 2013, consulté le 30 août 2015.

DUMITRU, Mara. *Museosphere*, http://museosphere.net/apropos?language=fr, 2013, consulté le 30 août 2015.

EUROPEANA. *Europeana - About us*, http://www.europeana.eu/portal/aboutus.html, consulté le 27 août 2015.

FRANCE, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DE, INRIA et WIKIMEDIA FRANCE. *Sémanticpédia*, http://www.semanticpedia.org/, consulté le 30 août 2015.

FRANCE, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DE. *JocondeLab - À propos*, http://jocondelab.iri-research.org/jocondelab/about/, consulté le 30 août 2015.

FRANCE, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DE. *JocondeLab - Montréal*, http://jocondelab.iri-research.org/jocondelab/map/#http%3A%2F%2Ffr.dbpedia.org%2Fresource%2FMontr%25C3%25A9al, consulté le 31 août 2015.

FRANCE, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DE. *Joconde, portail des collections des musées de France*, http://www.culture.gouv.fr/documentation/joconde/fr/pres.htm, consulté le 31 août 2015.

FRANCE, MINISTERE DE LA CULTURE ET DE LA COMMUNICATION DE. *Sémanticpédia : construire le web de données culturelles - Langue française et langues de France*, http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/Web-semantique-web-de-donnees-liage-de-donnees/Semanticpedia-construire-le-web-de-donnees-culturelles, 23 octobre 2014, consulté le 30 août 2015.

GETTY RESEARCH INSTITUTE. *Inlay (process)*, http://vocab.getty.edu/aat/300053850, 27 décembre 2013, consulté le 30 août 2015.

GETTY RESEARCH INSTITUTE. *TGN: Frequently Asked Questions*, http://www.getty.edu/research/tools/vocabularies/tgn/faq.html, 1 juin 2015, consulté le 28 août 2015.

GEONAMES. *Easter Island-URI*, http://www.geonames.org/maps/google_-27.117_-109.367.html, consulté le 30 août 2015.

GULDI, Jo et JOHNSON-ROBERSON, Chris. *Paper Machines | Visualize Your Zotero Collections*, http://papermachines.org/, 2015, consulté le 8 octobre 2015.

HASLHOFER, Bernhard et ISAAC, Antoine. *Data structure Europeana*, http://labs.europeana.eu/api/linked-open-data-data-structure, consulté le 31 août 2015.

HASLHOFER, Bernhard et ISAAC, Antoine. *Europeana Linked Open Data*, http://labs.europeana.eu/api/linked-open-data-introduction, consulté le 31 août 2015.

HUMA-NUM. *À propos de Huma-Num*, http://www.huma-num.fr/la-tgir-en-bref, 24 mars 2015, consulté le 9 octobre 2015.

HUMA-NUM. *Exposer ses données avec Nakala*, http://www.huma-num.fr/services-et-outils/exposer, 12 mai 2015, consulté le 9 octobre 2015.

HUMA-NUM. *ISIDORE - Accès aux données et services numériques de SHS*, http://www.rechercheisidore.fr/, 2015, consulté le 9 octobre 2015.

HUMA-NUM. *ISIDORE - À propos*, http://www.rechercheisidore.fr/apropos, 2015, consulté le 9 octobre 2015.

INTERNATIONAL COUNCIL OF MUSEUMS. *Object Collection Information*, http://www.cidoc-crm.org/cidoc_graphical_representation_v_5_1/object_collection.html, consulté le 14 janvier 2016.

INTERNATIONAL COUNCIL OF MUSEUMS. *The CIDOC CRM*, http://cidoc-crm.org/, décembre 2014, consulté le 26 mars 2015.

ISAAC, Antoine. *Europeana Data Model Primer*, s.l., 2013, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf, 35 p.

ISAAC, Antoine. *Europeana Data Model vocabulary (edm)*, http://lov.okfn.org/dataset/lov/vocabs/edm, 2013, consulté le 31 août 2015.

JACQUES, Isabelle. *Statistiques de fréquentation du Répertoire du patrimoine culturel du Québec*, Québec, Ministère de la Culture et des Communications du Québec, 2015.

KNOBLOCK, Craig et SZEKELY, Pedro. *Karma: A Data Integration Tool*, http://usc-isi-i2.github.io/karma/, 2015, consulté le 6 octobre 2015.

LABORATOIRE DE RECHERCHE HISTORIQUE RHONE-ALPES. *Le projet symogih.org : un système modulaire de gestion de l'information historique*, http://www.symogih.org/, 2015, consulté le 15 avril 2016.

LIBRARY OF CONGRESS. *Cartier, Jacques, 1491-1557 - LC Linked Data Service*, http://id.loc.gov/authorities/names/n50080987.html, consulté le 27 août 2015.

LIBRARY OF CONGRESS. *Library of Congress Subject Headings - LC Linked Data Service*, http://id.loc.gov/authorities/subjects.html, consulté le 28 août 2015.

MANNING, Patrick et RUVOLO, David. *CHIA - Collaborative for Historical Information and Analysis*, http://www.chia.pitt.edu/index.php, 2016, consulté le 13 janvier 2016.

MAPHUB. *Historic Map Annotation Portal*, http://maphub.github.io/, consulté le 30 août 2015.

MOSTERN, Ruth et ARKSEY, Marieka. *The Data Hoover Project (DHP) and its aims*, http://www.chia.pitt.edu/datahoover.html, 2015, consulté le 13 janvier 2016.

MUSEUM OF AUSTRALIAN DEMOCRACY et ARCHIVES NATIONALES DE L'AUSTRALIE. *About · Mildenhall's Canberra*, http://mildenhall.moadoph.gov.au/about, 2013, consulté le 31 août 2015.

ONLINE COMPUTER LIBRARY CENTER. *Library, Archive and Museum Collaboration*, http://www.oclc.org/research/activities/lamsurvey.html, 30 novembre 2011, consulté le 27 août 2015.

ONLINE COMPUTER LIBRARY CENTER. *PURL*, https://purl.org/docs/index.html, consulté le 17 décembre 2015.

ONTOTEXT. *Ontotext Semantic Solutions for Cultural Heritage*, http://ontotext.com/semantic-solutions/galleries-libraries-archives-museums/, 2015, consulté le 28 août 2015.

OPEN KNOWLEDGE FOUNDATION. *Linked Open Vocabularies (LOV)*, http://lov.okfn.org/dataset/lov/, 28 août 2015, consulté le 31 août 2015.

OPENLINK SOFTWARE. *Virtuoso Universal Server*, http://virtuoso.openlinksw.com/, consulté le 13 avril 2016.

POUPEAU, Gauthier. *Commentaire sur l'article « Sémantiser une base de données relationnelle (1er épisode) » de Jean-Baptiste Pressac*, http://bylg.hypotheses.org/96, 23 avril 2015, consulté le 14 janvier 2016.

QUEBEC, ASSEMBLEE NATIONALE DU. *Recherche avancée*, http://www.assnat.qc.ca/fr/recherche/recherche-avancee.html, 10 octobre 2012, consulté le 15 décembre 2015.

QUEBEC, COMMISSION DE TOPONYMIE DU. *Banque de noms de lieux du Québec*, http://www.toponymie.gouv.qc.ca/ct/accueil.aspx, 2012, consulté le 15 décembre 2015.

QUEBEC, GOUVERNEMENT DU. *Données ouvertes*, http://donnees.gouv.qc.ca/?node=/accueil, 2015, consulté le 31 août 2015.

QUEBEC, GOUVERNEMENT DU. *Loi sur le patrimoine culturel*, http://www2.publicationsduquebec.gouv.qc.ca/dynamicSearch/telecharge.php?type=2&file=/P_9_002/P9_002.html, 19 octobre 2012, consulté le 26 janvier 2015.

QUEBEC, GOUVERNEMENT DU. *PIMIQ - Thésaurus de l'activité gouvernementale*, http://www.thesaurus.gouv.qc.ca/tag/terme.do?id=CBC291, 2015, consulté le 16 décembre 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *06 – Aider le réseau de la culture à s'approprier les technologies du Web sémantique afin de maximiser la présence des données culturelles québécoises dans le Web : Plan culturel numérique*, http://culturenumerique.mcc.gouv.qc.ca/aider-le-reseau-de-la-culture-a-sapproprier-les-technologies-du-web-semantique-afin-de-maximiser-la-presence-des-donnees-culturelles-quebecoises-dans-le-web-banq/ , 2015, consulté le 26 mars 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *À propos du Répertoire du patrimoine culturel du Québec*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/redirection.do?go=about, 2013, consulté le 16 décembre 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *À propos : Plan culturel numérique*, http://culturenumerique.mcc.gouv.qc.ca/a-propos/, 2015, consulté le 26 mars 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *Bataille des Plaines d'Abraham*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=25651&type=pge#.VnLjfuI1uzx, 2013, consulté le 17 décembre 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *Cartier, Jacques*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=17323&type=pge#.Vd-sQK01uzx , 2013, consulté le 28 août 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *Leduc, Ozias*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=7700&type=pge#.VnLmR-I1uzw, 2013, consulté le 17 décembre 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *Martin, Robert Montgomery*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=15111&type=pge#.VnHK_-I1uzw, 2013, consulté le 16 décembre 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *Plaque de l'Édifice-Louis-S.-St-Laurent*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=111400&type=bien#.VnHVX-I1uzx, 2013, consulté le 16 décembre 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *Thompson, John Sparrow David*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=19600&type=pge#.VnHQJuI1uzy , 2013, consulté le 16 décembre 2015.

QUEBEC, MINISTERE DE LA CULTURE ET DES COMMUNICATIONS DU. *Recherche - Répertoire du patrimoine culturel du Québec*, http://www.patrimoine-culturel.gouv.qc.ca/rpcq/rechercheProtege.do?methode=afficher, 2013, consulté le 16 décembre 2015.

SESAME. *Contributors*, http://rdf4j.org/contributors.xhtml?view, 25 août 2015, consulté le 28 août 2015.

SHERRATT, Tim. *Inigo Jones - The weather prophet*, http://lodbookdev.herokuapp.com/#!/text/1/, consulté le 14 janvier 2016.

SOCIETE DES MUSEES QUEBECOIS. *Info-muse*, http://infomuse.smq.qc.ca/basisbwdocs/infm/Info/f_HumanitiesInfoHead.html, 1999, consulté le 15 décembre 2015.

SOCIETE DES MUSEES QUEBECOIS. *Info-Muse Recherche Gustave Courbet*, http://infomuse.smq.qc.ca/Infomuse/f_MasterLayout.cgi?la=f&db=1&style=99&realm=2&es=1&rs=1&who_i=WHOO&who_t=Gustave+Courbet&who_o=and&sort=NO_SORT, consulté le 30 août 2015.

STEAD, Stephen. *The CIDOC CRM, a Standard for the Integration of Cultural Information*, Capsule Web, CIDOC CRM, http://cidoc-crm.org/cidoc_tutorial/index.html, 2008, consulté le 14 janvier 2016.

UNIVERSITE LAVAL et UNIVERSITE DE TORONTO. *Biographie – CARTIER, JACQUES (1491-1557)*, http://www.biographi.ca/fr/bio/cartier_jacques_1491_1557_1E.html, 2015, consulté le 28 août 2015.

UNIVERSITE LAVAL et UNIVERSITE DE TORONTO. *Biography – THOMPSON, Sir JOHN SPARROW DAVID*, http://www.biographi.ca/en/bio/thompson_john_sparrow_david_12E.html , 2015, consulté le 16 décembre 2015.

UNIVERSITE LAVAL et UNIVERSITE DE TORONTO. *Résultats de la recherche – Région de naissance*, http://www.biographi.ca/fr/resultats.php?partial=0&stemmed=1&count=20&l_ft_2=and&l_ft_3=and&bp=122+123+124+125+126, 2015, consulté le 26 novembre 2015.

UNIVERSITE LAVAL et UNIVERSITE DE TORONTO. *Résultats de la recherche – Région d'activités*, http://www.biographi.ca/fr/resultats.php?partial=0&stemmed=1&count=20&l_ft_2=and&l_ft_3=and&cp=126+125+124+123+122, 2015, consulté le 26 novembre 2015.

VERBORGH, Ruben. *Commentaire sur OpenRefine Named-Entity Recognition extension*, https://groups.google.com/forum/#!topic/openrefine/jeNxqeWo9Rg, 20 décembre 2012, consulté le 14 janvier 2016.

WILLIAMS, Evan. *Medium*, https://medium.com/, août 2012, consulté le 7 octobre 2015.

WORLD WIDE WEB CONSORTIUM. *LodView*, http://www.w3.org/2001/sw/wiki/LodView, 22 décembre 2014, consulté le 28 août 2015.

WORLD WIDE WEB CONSORTIUM. *OWL Web Ontology Language Overview*, http://www.w3.org/TR/owl-features/, 10 février 2004, consulté le 28 août 2015.

WORLD WIDE WEB CONSORTIUM. *Introduction to SKOS*, https://www.w3.org/2004/02/skos/intro, février 2004, consulté le 26 janvier 2016.

WORLD WIDE WEB CONSORTIUM. *SPARQL Query Language for RDF*, http://www.w3.org/TR/rdf-sparql-query/, 15 janvier 2008, consulté le 28 août 2015.

II. Ouvrages généraux

LE JEUNE, Louis-Marie. « Jacques Cartier » dans *Dictionnaire Général de biographie, histoire, littérature, agriculture, commerce, industrie et des arts, sciences, mours, coutumes, institutions politiques et religieuses du Canada*, Ottawa, Université d'Ottawa, 1931, vol.1, p. 313-314, http://faculty.marianopolis.edu/c.belanger/QuebecHistory/encyclopedia/JacquesCartier-NaissancejeunesseetmariageHistoiredelaNouvelle-France.htm, consulté le 28 août 2015.

WIKIPEDIA. *Gustave Courbet*, https://fr.wikipedia.org/wiki/Gustave_Courbet, consulté le 30 août 2015.

WIKIPEDIA. *Ontologie (informatique)*, https://fr.wikipedia.org/wiki/Ontologie_(informatique), consulté le 13 avril 2016.

WIKIPEDIA. *Westmorland—Kent (circonscription fédérale)*, https://fr.wikipedia.org/wiki/Westmorland%E2%80%94Kent_(circonscription_f%C3%A9d%C3%A9rale, consulté le 16 décembre 2015.

III. Études

AYERS, Edward L. *History in Hypertext*, http://www.vcdh.virginia.edu/Ayers.OAH.html, 1999, consulté le 27 février 2014.

BERNARD, Jean-Paul, LINTEAU, Paul-André et ROBERT, Jean-Claude. « La structure professionnelle de Montréal en 1825 », *Revue d'histoire de l'Amérique française*, décembre 1976, vol. 30, n° 3, p. 383-414.

BERNERS-LEE, Tim, HENDLER, James et LASSILA, Ora. « The Semantic Web », *Scientific American*, 17 mai 2001,http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html, consulté le 27 août 2015.

BLAIS, Mireille et MARTINEAU, Stéphane. « L'analyse inductive générale: description d'une démarche visant à donner un sens à des données brutes », *Recherches qualitatives*, 2006, vol. 26, n° 2, p. 1-18.

BLOCH, Marc. *Apologie pour l'histoire ou métier d'historien*, Paris, Colin, 2011, 159 p.

BOX, George et DRAPER, Norman Richard. *Empirical model-building and response surfaces*, New York, Wiley (coll. « Wiley series in probability and mathematical statistics »), 1987, 669 p.

BOYDENS, Isabelle. « La conservation numérique des données de gestion », *Document numérique*, 1er juin 2004, Vol. 8, n° 2, p. 13-22.

BRAUDEL, Fernand. *Écrits sur l'histoire*, Flammarion, Paris, 1969, 314 p.

BURNARD, Lou. « The Historian and the Database » dans Evan Mawdsley (ed.), *History and computing III: historians, computers, and data: applications in research and teaching*, Manchester, Manchester University Press : Distributed exclusively in the United States and Canada by St. Martin's Press, 1990, p. 3-7.

CANADIANA. *Au-delà des tranchées : Un projet des Données ouvertes liées*, http://www.canadiana.ca/rpcpd-dol, 15 juillet 2012, consulté le 31 août 2015.

CHARLEBOIS, Eva, MALLET, Louise et METHOT, Julie. « L'ABC de la révision par les pairs », *Pharmactuel*, 2009, p. 42-52.

CHARLES, Valentine. *Data quality, validation, round-tripping*, Sydney, 2015, https://docs.google.com/document/d/16lcPBy1Cx4AKjLoTvlFtedDbr2fh0_NMr50PsnVN80c/edit?pli=1, consulté le 28 août 2015.

CHITIKA. *The Value of Google Result Positioning*, http://chitika.com/google-positioning-value, 7 juin 2013, consulté le 14 janvier 2016.

CLAIRMONT, Nicholas. *What Is The Difference Between « Facts » and « Opinions »?*, http://bigthink.com/the-proverbial-skeptic/what-is-the-difference-between-facts-and-opinions, consulté le 6 octobre 2015.

CLAYPHAN, Robina, CHARLES, Valentine et ISAAC, Antoine. *Europeana Data Model - Mapping Guidelines v.2.2*, s.l., 2014, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.2.pdf, 46 p.

COHEN, Daniel J. « From Babel to Knowledge: Data Mining Large Digital Collections », mars 2006, http://www.dlib.org/dlib/march06/cohen/03cohen.html, vol. 12, n° 3.

CONDORCET, Nicolas de. *Essai sur l'application de l'analyse à la probabilité des décisions rendus à la pluralité des voix*, s.l., Paris, Imprimerie Royale, 1785, 514 p.

CYGANIAK, Richard et JENTZSCH, Anja. *The Linking Open Data cloud diagram*, http://lod-cloud.net/, 2014, consulté le 31 octobre 2014.

DACOS, Marin. *Manifeste des Digital humanities*, http://tcp.hypotheses.org/318, 26 mars 2011, consulté le 27 février 2014.

DAVIS, Ian et NEWMAN, Richard. *Expression of Core FRBR Concepts in RDF*, http://vocab.org/frbr/core.html , 15 mai 2009, consulté le 28 août 2015.

DRUCKER, Johanna. « Humanities Approaches to Graphical Display ». *DHQ:Digital Humanities Quarterly*, volume 5, no. 1, 2011, p. 1-23.

DUTTON, Alexander. *Constraining the CLAROS SPARQL endpoint*, https://clarosdata.wordpress.com/, 23 mai 2011, consulté le 28 août 2015.

FOGEL, Robert et ENGERMAN, Stanley. *Time on the Cross: The Economics of American Negro Slavery*, Boston, Little Brown, 1974, 336 p.

FRANCART, Thomas. *RDF : Sesame, Jena, comparaison des fonctionnalités*, http://blog.sparna.fr/2012/05/08/rdf-sesame-jena-comparaison-des-fonctionnalites/, 8 mai 2012, consulté le 28 août 2015.

FURET, François. « De l'histoire-récit à l'histoire-problème », *Diogène*, 1 janvier 1975, n° 89, p. 113-130.

GRAHAM, Shawn, MILLIGAN, Ian et WEINGART, Scott. *Exploring Big Historical Data: The Historian's Macroscope*, http://www.themacroscope.org/2.0/, 2015, consulté le 7 octobre 2015.

GRAHAM, Shawn. *Reflecting on our process*, http://www.themacroscope.org/?page_id=303, 11 septembre 2013, consulté le 7 octobre 2015.

GRENZINGER, Yannick. *Qu'est-ce que l'expérience utilisateur ? | Ergonomie, Expérience Utilisateur, Design Thinking*, http://ux-fr.com/experience-utilisateur-definition/, consulté le 7 octobre 2015.

GULDI, Jo et ARMITAGE, David. *The History Manifesto*, Cambridge University Press, Cambridge, 2014, 175 p.

HABERT, Benoit, SALAÜN, Jean-Michel et MAGUE, Jean-Philippe. « Architecte de l'Information: Un métier », *ADBS*, 2012, vol. 49, n° 1, (coll. « Documentaliste - Sciences de l'Information »), p. 4-5.

HASLHOFER, Bernhard et ISAAC, Antoine. « The Europeana Data Model for Cultural Heritage », *Europeana*, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Factsheet.pdf, p. 1-2.

HOHMANN, Georg et SCHOLZ, Martin. « Recommendation for the representation of the primitive value classes of the CRM as data types in RDF/OWL implementations », http://erlangen-crm.org/docs/crm-values-as-owl-datatypes.pdf, 24 février 2011, p. 1.

HOOLAND, Seth van et VERBORGH, Ruben. *Linked data for libraries, archives and museums: how to clean, link and publish your metadata*, US-Ed., Chicago, Neal-Schuman, 2014, 254 p.

KOENING, Glenn. *FileMaker Early History*, http://www.dancing-data.com/filemakerhist.html, 2004, consulté le 30 octobre 2014.

KOKOBI, Talel. *Cadre commun d'interopérabilité du Gouvernement du Québec: Normaliser, S'aligner, Performer*, http://www.tresor.gouv.qc.ca/fileadmin/PDF/ressources_informationnelles/cadre_commun_interoperabilite.pdf , 2014, p. 70.

LASLIER, Jean-François. « La norme majoritaire », *Revue économique*, 1999, vol. 50, n° 4, p. 669-698.

LE DEUFF, Olivier. *Quelles littératies et formations pour les humanités digitales ?*, Humanités numériques 2015, http://hn2015.org/programme/, Montréal, 12 août 2015.

LE GOFF, Jacques. *Faut-il vraiment découper l'histoire en tranches?*, Paris, Seuil, 2014, 207 p.

LETOURNEAU, Jocelyn. *Le coffre à outils du chercheur débutant: guide d'initiation au travail intellectuel*, Montréal, Boréal, 2006, 259 p.

LINCOLN, Matthew. « Using SPARQL to access Linked Open Data », *Programming Historian*, http://programminghistorian.org/lessons/graph-databases-and-SPARQL, 24 novembre 2015, consulté le 17 janvier 2016.

LUKOVNIKOV, Denis, STADLER, Claus, KONTOKOSTAS, Dimitris, HELLMANN, Sebastian et LEHMANN, Jens. « DBpedia Viewer - An Integrative Interface for DBpedia Leveraging the DBpedia Service Eco System », *Workshop on Linked Data on the Web*, Seoul, 2014, vol.1184, http://ceur-ws.org/Vol-1184/ldow2014_paper_05.pdf, consulté le 28 août 2015.

MARIAN, Jakub. *A curiosity about the F-word in Google Ngram Viewer*, https://jakubmarian.com/a-curiosity-about-the-f-word-in-google-ngram-viewer/, consulté le 14 janvier 2016.

NADEAU, David et SEKINE, Satoshi. « A survey of named entity recognition and classification », *Named Entities: Recognition, classification and use*, 2007, vol. 30, n° 1, (coll. « Lingvisticæ Investigationes »), p. 3-26.

NORA, Pierre. « L'explosion du patrimoine », *Revue de l'Institut national du patrimoine*, 2006, n° 2, p. 6-11.

ORGANISATION DE COOPERATION ET DE DEVELOPPEMENT ECONOMIQUES. *La littératie à l'ère de l'information: Rapport final de l'enquête internationale sur la littératie des adultes*, Paris, http://www.oecd.org/fr/edu/innovation-education/39438013.pdf, 2000, 211 p.

PETIT, François-Xavier. *Qu'est-ce qu'être historien?*, http://www.histoire-pour-tous.fr/education/179-metiers-histoire/3416-quest-ce-quetre-historien-.html, 21 décembre 2010, consulté le 6 octobre 2015.

PRESSAC, Jean-Baptiste. *Sémantiser une base de données relationnelle (1er épisode)*, http://bylg.hypotheses.org/96, 20 avril 2015, consulté le 28 août 2015.

RESEAU PANCANADIEN DU PATRIMOINE DOCUMENTAIRE. *« Démonstration de faisabilité » de la Visualisation des Données ouvertes liées (LOD)*, Canada, 2012, http://www.canadiana.ca/sites/pub.canadiana.ca/files/PCDHN%20Proof-of-concept_Final-Report-FRA.pdf, consulté le 27 août 2015, 83 p.

ROCHKIND, Jonathan. *Linked Data Caution*, https://bibwild.wordpress.com/2015/11/23/linked-data-caution/, 23 novembre 2015, consulté le 17 décembre 2015.

RUIZ, Émilien. *Les historiens seront-ils finalement programmeurs ?*, http://www.boiteaoutils.info/2011/09/les-historiens-seront-ils-finalement/, 22 septembre 2011, consulté le 8 octobre 2015.

SALAÜN, Jean-Michel. *Référentiel de compétences en Architecture de l'information*, http://archinfo01.hypotheses.org/453, 7 octobre 2013, consulté le 7 octobre 2015.

SCHIEMANN, Bernhard, OISCHINGER, Martin, GÖRZ, Günther, HOHMANN, Georg, MERGES, Judith, FICHTNER, Mark et SCHOLZ, Martin. *Erlangen CRM OWL*, http://erlangen-crm.org/, 2013, consulté le 30 août 2015.

SCHREIBMAN, Susan, SIEMENS, Raymond George et UNSWORTH, John. *A companion to digital humanities*, Malden, MA, Blackwell Pub (coll. « Blackwell companions to literature and culture »), 2004, 611 p.

SHERRATT, Tim. *Every story has a beginning*, http://discontents.com.au/every-story-has-a-beginning/, 4 octobre 2011, consulté le 31 août 2015.

SHERRATT, Tim. *Stories for machines, data for humans*, http://discontents.com.au/stories-for-machines-data-for-humans/, 10 avril 2015, consulté le 7 octobre 2015.

SONNTAG, Emmanuelle. *L'écoute dans les humanités numériques*, Humanités numériques 2015, http://hn2015.org/ programme/, Montréal, 13 août 2015.

SONNTAG, Emmanuelle. *L'écoute dans les humanités numériques : je suis écoute*, https://medium.com/@lvrdg/df17f12968b5, 13 août 2015, consulté le 7 octobre 2015.

THALLER, Manfred. « Controversies around the Digital Humanities: An Agenda », *Kontroversen um die Digitalen Geisteswissenschaften: Ein Arbeitsplan.*, septembre 2012, vol. 37, n° 3, p. 7-23.

VERBORGH, Ruben. *Federated SPARQL queries in your browser*, http://ruben.verborgh.org/blog/2015/06/09/federated-sparql-queries-in-your-browser/, 9 juin 2015, consulté le 28 août 2015.

VERBORGH, Ruben. *Fostering intelligence by enabling it*, http://ruben.verborgh.org/blog/2015/02/25/fostering-intelligence-by-enabling-it/, 25 février 2015, consulté le 27 août 2015.

VEYNE, Paul. *Comment on écrit l'histoire*, Éditions du Seuil, Paris, 1971, 342 p.

WOOLLARD, Matthew et DENLEY, Peter. *A Tutorial for Kleio*, St. Katharinen, Max-Planck-Institut für Geschichte, http://www.hki.unikoeln.de/kleio/old.website/tutorial/intro.htm, 1993, consulté le 1er avril 2014.

WRIGLEY, Edward Anthony. « The PST system of classifying occupations », http://www.campop.geog.cam.ac.uk/research/projects/occupations/britain19c/papers/paper1.pdf, 2010, 24 p.

ZENG, Marcia Lei. *Metadata*, New York, Neal-Schuman Publishers, 2008, 365 p.