ÉTUDE

Archivistique, histoire et Web sémantique : une approche interdisciplinaire basée sur l'événementiel

PHILIPPE MICHON
Consultant en informatique appliquée à l'histoire

INTRODUCTION : MISE EN SILOS NUMÉRIQUES

L'appropriation du Web 2.0

Difficile de concevoir une recherche scientifique en 2017 sans l'utilisation du Web. Cette technologie de diffusion et de partage des connaissances fait intrinsèquement partie des méthodologies archivistiques et historiennes. Plus précisément, c'est le développement du Web 2.0 et de ses interfaces conviviales qui a favorisé la création de sites Web ainsi que des bases de données, par de multiples individus et organisations, sans la nécessité d'acquérir des compétences techniques poussées. D'un côté, il est facile d'instaurer une vitrine des collections institutionnelles et des données de recherche. De l'autre, l'utilisateur est en mesure de consulter facilement ces différents systèmes grâce à des outils de requêtes. La question reste à savoir si cette appropriation massive

du Web 2.0 dans la pratique engendre une distanciation des disciplines archivistique et historique en camouflant à l'utilisateur néophyte l'organisation informationnelle derrière ces systèmes de recherche. Est-ce que cette organisation pourrait aussi causer des biais majeurs quant à l'analyse des données ?

La dérive de la spécialisation

L'organisation des connaissances par discipline n'est pas attribuable à l'apparition du Web, mais ce dernier, en simplifiant les moyens d'accès à l'information, arrive à dissocier le contenant de la forme. À titre d'exemple et fort heureusement, il n'est plus nécessaire de comprendre une norme de classification archivistique pour effectuer des recherches dans ce domaine. Cependant, une structure organisée dans une ou plusieurs tables, telle qu'une base de données relationnelle, oriente nécessairement la construction du contenu qu'il faut placer dans des cellules en relation avec des lignes et des colonnes. L'exemple suivant montre la difficulté d'ajouter la notion d'oncle dans un système qui n'en fait point mention sans devoir ajouter un nouveau tableau ou une nouvelle colonne qui dédoubleraient l'information :

Père	Mère	Enfant	Frère de la mère	Frère du père
Pierre	Sylvie	Damien	Gérald	Victor
Jacques	Annette	Flavie	Olivier	Aucun

Tableau 1 : Exemple de table avec la notion d'oncle

La recherche par facettes peut sembler libre, mais elle est un dérivé de cette construction tabloïde. Ensuite arrive la connaissance des modèles de description ou d'indexation utilisés en archivistique, mais peu connus du monde historien. Pensons par exemple au *Dublin Core* (DC) ou aux *Règles pour la description des documents d'archives* (RDDA) qui demandent une étude plus approfondie. L'historien, de son côté, a plutôt tendance à normaliser de manière organique, c'est-à-dire qu'il partira du contenu et construira une nomenclature utile pour sa question de recherche. Nous avons donc deux disciplines qui organisent leurs contenus sur des bases méthodologiques très peu communes, ce qui accentue la difficulté

de mettre en place des interfaces de recherches interdisciplinaires. Sans prétendre qu'on se dirige vers un monde où les vocabulaires disciplinaires seront entièrement déconnectés les uns par rapport aux autres comme le dépeint George Orwell (2005), une prise de conscience est plus que nécessaire afin de réfléchir à une véritable mise en commun des savoirs de manière à mieux intégrer des données extérieures à notre champ disciplinaire. L'utilisation du Web ne peut se concevoir comme une simple translation du monde analogique vers un monde numérique, mais bien comme une opportunité de réformer notre rapport à la source.

Nouveau Web, nouvelle méthodologie

L'archivistique a clairement amorcé un virage numérique majeur tandis que l'histoire tarde à investir pleinement ce média. Le besoin de formation est décrié par plusieurs professionnels comme le montre l'enquête dirigée par Émilien Ruiz (2015). L'historien a souvent joué un rôle passif dans le domaine des technologies numériques. À l'exception des tentatives d'histoire quantitative des années 1960 et 1970 (Floud 1979), ainsi que les travaux de Manfred Thaller sur un logiciel adapté aux historiens, qui suivront dans les années 1990 (Thaller, 1995), il n'y a jamais eu une implication massive de l'historien dans la réflexion entourant les outils numériques. La tendance était plutôt d'utiliser les logiciels une fois ceux-ci pleinement développés et qui, évidemment, n'étaient pas conçus selon les préoccupations historiennes de temps et d'espace, à titre d'exemple (Thaller, 2012). Par ailleurs, encore aujourd'hui, les projets en histoire numérique à grand déploiement sont relativement rares. Souffrant d'un manque de compétences techniques, l'historien est souvent contraint à travailler avec certaines technologies seulement après leur démocratisation. C'est le cas avec les systèmes de gestion de contenu (SGC), les systèmes d'information géographique (SIG) et la modélisation 3D. Actuellement, une des grandes mouvances numériques concerne les données liées et le Web sémantique. Ces derniers ont pour objectif de décloisonner la donnée du document. En effet, Tim Berners-Lee constatait au début des années 2000 que l'information disponible sur le Web était encapsulée dans des sites Web et des documents ne permettant pas une pleine interopérabilité des contenus. Devant ce constat, celui-ci propose de donner un sens à la donnée en l'associant avec d'autres. Ainsi, la donnée se désambiguïse et son traitement automatisé est plus

efficient. On passe alors d'un Web de documents à un Web de données (Bibliothèque nationale de France, 2015). Cependant, il n'y a pas uniquement des gens enthousiastes devant ce modèle. Notons, par exemple, la critique de Florian Cramer qui affirme que le Web sémantique est irréalisable, même non souhaitable, puisqu'il s'appuie sur la mise en place d'un vocabulaire unique pour couvrir tous les aspects du monde (Cramer, 2007). Cette inquiétude est valable, mais se tempère avec la notion d'alignement qui semble aujourd'hui être la clé de voûte de l'interopérabilité du Web sémantique. En d'autres mots, il suffit d'aligner différents vocabulaires spécialisés afin de pouvoir unir des méthodologies et des contenus. Le Web sémantique est actuellement la seule avenue prometteuse pour décloisonner les données. Malgré cette conception du Web qui se développe depuis le début des années 2000 (BernersLee, 2001), il n'existe à ce jour aucun outil simple d'utilisation permettant de transformer des données relationnelles vers des données liées bien organisées. En effet, la majorité des bases de données culturelles et historiques utilisent *My Structured Query Language (MySQL)*, un système de gestion de base de données relationnelle très répandu. Il est donc nécessaire de développer des outils de conversion de données conviviaux afin que l'historien et l'archiviste puissent adapter leurs données et leurs métadonnées aux exigences du Web sémantique, sans devoir passer par des informaticiens. Cependant, en amont de cette nécessité, l'archiviste et l'historien doivent réfléchir à un vocabulaire qui faciliterait les partages informationnels entre les deux disciplines. Comme nous le verrons dans ce court article, le Web sémantique s'articule autour de vocabulaires qui ne nécessitent pas une connaissance technique pointue. Après avoir vu certains concepts fondateurs du Web sémantique, nous explorerons l'ontologie *Conceptual Reference Model* développée par le *Comité international pour la documentation* (CIDOC CRM) de l'*International Council of Museums* (ICOM). CIDOC CRM semble un excellent point de départ pour construire des ponts entre les disciplines qui étudient les sources patrimoniales. Cette exploration d'un schéma de représentation de données soulèvera la question des compétences de chaque discipline et la nécessité de les arrimer.

1. LE WEB SÉMANTIQUE : PRISE DE CONSCIENCE DE L'ABSENCE D'UN LANGAGE FÉDÉRATEUR

1.1. Le fonctionnement du Web sémantique

Le Web sémantique s'appuie sur deux concepts. Le premier est l'identification précise de tout élément compilé sur le Web. Cette étape se concrétise par l'utilisation d'un *Uniform Resource Identifier* (URI), un identifiant unique que l'on exprime avec le protocole *http* et qui permet de localiser précisément une ressource. (W3C 2001) Malgré son extrême ressemblance avec un *Uniform Resource Locator* (URL) qui pointe uniquement vers des pages Web, l'URI désigne n'importe quel type de ressource : une personne, un concept, un événement, un objet, un document d'archives, etc. Le deuxième concept est celui de liaison entre les données pour attribuer du sens à celles-ci, ce qui signifie qu'un ordinateur peut interpréter une donnée en analysant les données qui lui sont associées. Cette configuration présentée sous la forme de triplets *Sujet-Prédicat-Objet*, se nomme *Resource Description Framework* (RDF). (W3C 2014a) C'est donc en incluant des URI dans une structure RDF que l'on inscrit nos données dans un vaste nuage de données liées. Il existe ensuite différents niveaux de diffusion de ce contenu. Il est possible de rendre simplement disponibles ces fichiers RDF sur n'importe quelle plateforme Web classique, mais la manière la plus dynamique de partager ces données est de les publier sur un serveur dédié aux données liées. Ce dernier, que l'on nomme un *TripleStore*, est souvent associé à un point d'entrée *SPARQL Protocol and RDF Query Language* (SPARQL). Celui-ci permet d'effectuer différentes actions sur la base de données liées comme d'ajouter, de modifier et de supprimer du contenu ainsi que d'effectuer différentes requêtes sur les données. (W3C 2008) C'est par ces interrogations de multiples points d'entrées SPARQL que l'on constitue des applications de mise en commun des données. Un tutoriel d'utilisation de ce langage de requêtes orienté vers la pratique historienne est disponible sur le site *The Programming Historian* (Lincoln, 2015). Évidemment, pour parvenir à des requêtes fédérées, il faut aussi intégrer au modèle RDF des normes connues et utilisées par les professionnels du domaine visé, dans le cas présent, le patrimoine documentaire culturel.

1.2. Mise en place de figures d'autorités

La création d'un URI est une manipulation simpliste comparée à la recherche d'un URI existant. Un archiviste ou un historien peut très bien instaurer un nouvel URI pour représenter une personne mentionnée dans un document d'archives. Par exemple, si un document a été signé par un dénommé Jacques Cartier, il serait tout à fait possible de créer un URI pour le désigner. Cependant, s'il s'agit bel et bien de l'explorateur, il serait plus pertinent de faire appel à un URI existant dans un fichier d'autorité. L'explorateur Jacques Cartier se retrouve sur *DBpedia*, une version en triplets de *Wikipedia* actuellement un incontournable du nuage des données liées (Cyganiak et Jentzsch, 2014). *DBpedia* est l'ensemble de données le plus vaste et auquel on réfère le plus souvent. Cependant, dans un esprit de réutilisation scientifique, les triplets fournis par *DBpedia* permettent uniquement d'identifier, et non pas d'ouvrir vers, une analyse en profondeur du personnage. L'utilité première de *DBpedia* est donc d'identifier correctement Jacques Cartier, car la division d'une fiche de classification en triplets nécessite une définition précise de chaque élément. « Jacques Cartier » peut nous paraître simple à identifier lorsqu'on retrouve le terme sur la même page qu'une photographie ou un lieu de naissance. Mais lorsque ce terme est décontextualisé dans le nuage des données liées, la tâche peut rapidement se compliquer (Verborgh, 2015). L'archiviste et l'historien sont grandement sensibilisés à ce besoin de désambiguïsation, puisqu'il est nécessaire d'identifier avec exactitude le contenu d'une source. Il n'existe pas un seul et unique Jacques Cartier, mais bien plusieurs dizaines. Une courte recherche au sein du *Répertoire du patrimoine culturel du Québec* (RPCQ) confirme l'existence de deux « Jacques Cartier » totalement différents. Un étant évidemment l'explorateur et l'autre député de la circonscription de Surray de 1804 à 1809 (Répertoire du patrimoine culturel du Québec, 2016). Sachant que seulement environ 29 % des personnages fichés au RPCQ possèdent un URI *DBpedia*, on réalise qu'il faudra aussi mettre en place des URI au Québec afin de mieux baliser notre histoire (Michon, 2016). Autant l'archiviste que l'historien ne pourront utiliser pleinement les possibilités de fédération des données liées si les institutions ne créent pas leurs propres URI pour leurs collections uniques. Sans lieux communs auxquels se référer, l'histoire du Québec ne pourra prendre sa juste place dans cette nouvelle conception du Web.

1.3. Plateforme fédératrice et participative

En plus de la création massive d'URI, il faut un projet mobilisateur afin de confirmer aux autorités l'importance d'investir le Web sémantique et de constituer une méthodologie interdisciplinaire efficace. Il existe deux manières de concevoir un projet en données liées. Une première vision s'exprime de bas en haut et s'appuie sur l'éducation efficiente des institutions afin de produire des données interopérables basées sur des vocabulaires phares afin de favoriser une mise en commun naturelle (Kamath Sowmya, 2013). Si tout le monde applique les principes du Web sémantique, il va inévitablement se créer des liens qui mèneront à de nouvelles connaissances. Certes, cette méthodologie fonctionne, mais la juxtaposition et la multiplication des modèles et des vocabulaires peuvent rapidement devenir intraitables pour un néophyte. Pour cette raison, une préférence est portée à la conception du haut vers le bas, soit qu'une institution ou un groupe d'experts instaure une plateforme en y incluant déjà des normes et une marche à suivre. Ni le Québec ni le Canada ne possèdent d'infrastructure nationale qui encourage le versement et la création de liens entre les données, comme le propose par exemple la bibliothèque numérique paneuropéenne *Europeana*. Il faut aussi miser sur une plateforme collaborative afin de tisser des liens entre la méthodologie archivistique et historienne. La schématisation classique actuelle est de voir l'archiviste comme un diffuseur de contenus organisés et l'historien comme étant celui qui s'approprie ces données afin de répondre à une problématique. Cette approche unidirectionnelle empêche le dialogue entre les deux disciplines à un point tel que l'historien a adapté sa méthodologie aux méthodes de classification archivistiques sans pour autant en connaître les subtilités. Si l'on souhaite ouvrir les frontières disciplinaires, il faut penser à une structure participative. D'un côté, on a l'archiviste qui organise les archives de manière à les repérer facilement dans une abondance de contenus et de l'autre, l'historien qui augmente la portée du document d'archives en l'associant directement avec son contexte historique. Mais comment arriver à unir deux disciplines qui ont des objectifs communs mais dont les méthodologies sont différentes ?

2. CIDOC CRM : L'ÉVÉNEMENT COMME CLÉ DE VOÛTE

2.1. Schémas de représentation de données et ontologies

La réponse à cette question se trouve dans la dernière composante du Web sémantique, soit l'utilisation des modèles ontologiques afin de porter un regard particulier sur les données. Il existe une panoplie de définitions d'une ontologie, mais Bruno Bachimont en exprime l'essentiel lorsqu'il dit :

> On peut caractériser une ontologie comme une structuration des concepts d'un domaine. Ces concepts sont rassemblés pour fournir les briques élémentaires et exprimer les connaissances dont on dispose dans ce domaine. (Bachimont, 2006)

Dans le contexte du Web sémantique, il s'agit de réfléchir à une organisation de liens entre des concepts d'une discipline ou d'un champ donné. En effet, le langage extensible *RDF Schema* (RDFS) permet de constituer des ontologies suivant la forme en triplet (W3C, 2014b). Dès lors, on peut constituer des classes, des propriétés, des domaines et des portées qui sont tous des composantes essentielles d'un triplet. Par la suite, d'autres ontologies peuvent être modélisées suivant ces principes fondateurs. Sans parler d'une ontologie, *Dublin Core* a été modélisé suivant la nomenclature RDF (Nilsson, Powell, Johnston et Naeve , 2008). Il existe des ontologies concernant un lot impressionnant de thématiques telles que les personnes (BIO et *Friend of a Friend*), le commerce en ligne *(GoodRelations)*, la musique (*Music Ontology*), etc. (Semantic Web, 2012) Il existe aussi des ontologies institutionnelles construites pour répondre aux besoins spécifiques d'une collection. C'est le cas du *Europeana Data Model* (EDM) créé pour répondre aux besoins spécifiques de la plateforme (Europeana, 2014). Existe-t-il une ontologie pour le patrimoine culturel ? Il n'existe toujours pas d'alternative ontologique idéale actuellement mais la norme CIDOC CRM développée par l'ICOM semble une avenue de réflexion incontournable.

2.2. Fonctionnement du modèle CIDOC CRM

Le modèle CIDOC CRM provient de la branche muséologique, mais est proposé comme une ontologie permettant la fédération des institutions

patrimoniales et culturelles (International Council of Museums, 2014a). Pour y arriver, la structure se bâtit autour de la notion d'entité temporelle (un événement) afin de reconstruire et de lier des faits. La démonstration suivante découle du tutoriel présenté par Stephen Stead, disponible en ligne :

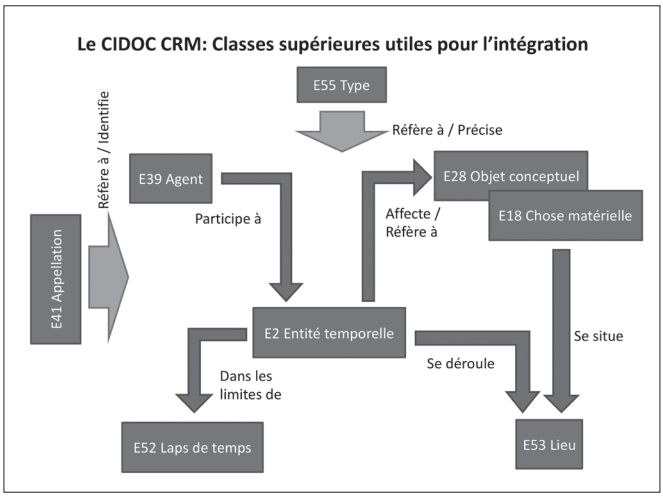

Figure 1: Le CIDOC CRM. Classes supérieures utiles pour l'intégration. Source : Traduction de l'auteur d'après le schéma de Stead, 2008.

Il y a donc, au centre du modèle, un événement auquel se greffent différents autres objets. En premier lieu, on retrouve des agents qui sont des entités humaines qui participent à l'événement. Cependant, il n'y a pas nécessairement toujours une implication humaine pour qu'une action se produise, pensons au jaunissement du papier par exemple. Ensuite, des concepts et des objets physiques sont affectés ou font référence à l'événement. Si le papier jauni, ce dernier subit une modification et sera donc modifié. Certains événements ont un impact sur la compréhension d'un concept. En effet, un concept existe dans la mémoire de ceux qui le comprennent. Un cours sur les méthodes de préservation du papier peut modifier votre compréhension du concept de conservation. Arrive alors le lieu. Chaque événement, objet et concept peut se retrouver dans un lieu. La tenue d'un colloque dans une université, le document d'archives dans un centre d'archives et le concept de conservation dans le cerveau des étudiants en archivistique sont tous des associations entre des entités CIDOC CRM et un lieu. On remarque d'ailleurs que le lieu peut être mobile comme c'est le cas des cerveaux des étudiants ou d'un texte

écrit qui se retrouve sur une feuille de papier. Peu importe où se situe la feuille, le texte y sera toujours. En dernier lieu, on associe l'événement avec une durée. La durée étant toujours imprécise (il est impossible de savoir à la milliseconde près la durée d'un événement), il est possible de la baliser avec deux bornes externes ou une borne interne. Les bornes externes permettent de signaler que l'événement s'est déroulé entre ces deux moments. La borne interne permet de signaler qu'un événement était en cours à un moment précis. Il existe aussi deux grandes catégories qui englobent le modèle : les appellations qui permettent de nommer toutes les entités CRM. Il existe donc une distinction très claire entre une entité vivante et son nom. « Jacques Cartier » est une norme sociale qui identifie différentes personnes, dont l'explorateur bien connu. Cette distinction permet de s'intéresser à la transmission des noms par la toponymie. Pourquoi avons-nous un pont nommé « Jacques Cartier » qui relie Montréal et sa rive sud ? L'autre catégorie est le typage. En effet, il est possible de typer toutes les classes de CIDOC CRM afin de préciser la teneur du contenu. Par exemple, si l'objectif de la modélisation est d'informer l'utilisateur que Jacques Cartier est un explorateur, il faudra créer un type d'occupation associé au personnage puisque CIDOC CRM ne permet pas d'aller à ce niveau de profondeur descriptive. Le type qu'on attribue à une classe peut provenir par exemple d'un thésaurus et peut même posséder son propre URI afin de le définir précisément dans un plus grand réseau de concepts.

2.3. Ajouter du contenu et combattre la perte de sens

Les contraintes techniques de l'historien face à une ontologie sont moins grandes que celles de l'archiviste. Puisque l'historien n'adhère pas déjà à un ensemble défini de vocabulaires de description, il peut plus facilement en adopter ou en créer un selon ses besoins. Le défi est plutôt du côté de l'assimilation des connaissances informatiques. Si par exemple CIDOC CRM répond à l'entièreté de sa méthodologie, il ne lui reste qu'à comprendre et à appliquer le schéma. L'archiviste a un travail plus imposant de par son ancrage initial dans des modèles descriptifs qui permettent la transmission des connaissances dans sa sphère d'activités. Malgré qu'il existe des outils qui facilitent la mise en commun de différents schémas avec CIDOC CRM comme l'*Encoded Archival Description* (EAD) (International Council of Museums, 2014b), il n'existe encore aucune mise en commun entre les RDDA et le CIDOC

CRM. Si une institution archivistique utilise les RDDA pour la description de ses données, le travail d'association se complexifie grandement. Cette difficulté est augmentée par la nécessité d'inclure des événements dans le schéma de classification afin de créer des liens entre les différentes composantes du patrimoine. Autant l'archiviste que l'historien ne cataloguent que très peu l'événementiel au premier niveau, c'est-à-dire que l'événement se présente souvent comme une valeur du champ d'un enregistrement. En histoire, on retrouve souvent des lignes du temps, mais sinon l'événement est tellement ancré dans la pratique que l'on oublie de le ficher. À titre d'exemple, sur les 106 011 fiches du RPCQ, seulement 71 sont des événements, soit 0,07 % de celles-ci (Michon, 2016). Si CIDOC CRM est une option intéressante pour la fédération des données, il faut miser sur une restructuration du contenu orientée vers l'événement. La communication entre archivistes et historiens s'accentuera si les deux disciplines contribuent à la définition d'événements selon leurs compétences respectives. Avant cette restructuration, il faut aussi s'interroger sur la perte de sens que peut laisser entendre le modèle CIDOC CRM. Ce dernier ne peut être le seul schéma de métadonnées utilisé dans un projet de données liées puisqu'il ne permettrait pas une analyse précise des événements et des relations entre ceux-ci. Pour y parvenir, la solution semble être le typage du modèle CIDOC CRM de manière extrêmement rigoureuse. Il faut élaborer une stratégie qui amalgame un lot de vocabulaires en résonance avec CIDOC CRM afin de proposer une méthodologie orientée autour de la source et de son contexte, qui faciliterait le travail des professionnels en sciences historiques. Une perte de sens est inévitable lorsque l'on souhaite associer des composantes de diverses disciplines. Par contre, l'utilisation d'une ontologie RDF n'oblige pas une institution ou un chercheur à abandonner ses autres méthodes. Il s'agit au contraire d'un nouvel outil qui s'utilise conjointement avec les systèmes en place. Par exemple, l'archiviste peut utiliser les RDDA afin d'être en résonance avec les professionnels de son domaine tout en utilisant CIDOC CRM afin d'ouvrir ces données à un public non spécialisé en archivistique. Reste qu'il ne faut pas oublier que le Web sémantique et les vocabulaires rattachés ne sont pas les uniques éléments qui contribuent aux échanges interdisciplinaires. Le développement des compétences face aux technologies du numérique est nécessaire à la mise en place d'une structure informatique adaptée aux besoins des professionnels.

3. COMPÉTENCES TRANSVERSALES : ARCHITECTURE DE L'INFORMATION HISTORIQUE

3.1. Compétences de l'historien

La méthodologie de l'historien préfigure le fonctionnement du Web sémantique. Ce dernier devient un moyen concret pour l'historien d'ajouter un processus argumentatif à son analyse qui se concrétise actuellement, dans la vaste majorité des cas, en un texte écrit. Affirmer que le texte écrit est une formule inadéquate en histoire serait une grave erreur, puisque celui-ci permet de fixer le récit et la conservation d'une expressivité plus que nécessaire pour la compréhension historique (Sherratt, 2015). Cependant, il aplanit le propos historique, le simplifie, afin de respecter des règles d'écriture et d'édition. La donnée de recherche se retrouve encastrée dans un bloc texte qu'il faut lire et décortiquer (revenir à la donnée) afin d'aller plus loin. Un gain de temps est à prévoir si l'historien parvient à modéliser numériquement sa méthodologie et à la diffuser sous forme malléable sur le Web. L'approche la plus pertinente qui allie données liées et texte historique est le travail de l'historien australien Tim Sherratt qui propose un texte intitulé *Inigo Jones: The Weather Prophet*. Son projet consiste à inclure des balises RDF au sein même du texte écrit. Une application peut alors utiliser les URI associés au texte pour développer des visualisations de celui-ci. Par exemple, il est possible de constituer un réseau familial sous forme d'arbre généalogique (Sherratt, 2015). Le texte peut alors être réinterprété selon différentes facettes. De plus, un terme peut être approfondi par le lecteur sans que l'auteur ait besoin de le définir. Il y a donc association de contenus qui permet une explosion de la linéarité de l'histoire et le lecteur peut alors moduler sa lecture selon ses connaissances et ses intérêts. Les données liées ouvrent la voie à une nouvelle forme d'histoire, définie par Edward L. Ayers :

> *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.* (Ayers, 1999)

Cependant, il faut bien comprendre que les compétences en informatique de Tim Sherratt dépassent considérablement celles de

l'historien moyen. Trois axes de compétences présentés dans l'ouvrage collectif *Le coffre à outils du chercheur débutant : Guide d'initiation au travail intellectuel*, dirigé par Jocelyn Létourneau, encouragent l'intégration du Web sémantique à la pratique historienne, (Michon, 2016) soit la documentation à l'ère numérique, l'élaboration d'une stratégie de recherche et la communication de sa pensée par écrit (Létourneau, 2006). Le premier axe rappelle l'importance d'une exploration exhaustive de notre champ de recherche à l'aide des catalogues numériques. Comme il fut mentionné ci-dessus, la facilité de manipulation de ces catalogues amène l'historien à évacuer de sa réflexion le fonctionnement même de l'outil de recherche. Avec le Web sémantique, l'historien se doit de connaître les ontologies pour effectuer des requêtes SPARQL. Il y a donc une union forte entre le contenant et le contenu, ce qui permet d'éviter les biais qu'occasionne le traitement de masse de l'information. Le deuxième axe est celui de l'élaboration d'une stratégie de recherche afin de positionner son travail dans l'historiographie. Avec un cadre spatio-temporel restreint, la création d'une bibliographie complète est somme toute réalisable. Cependant, si on agrandit ce cadre afin d'entrer dans la longue durée, concept qui refait surface depuis quelques années en histoire (Guidi et Armitage, 2015), la tâche s'avère beaucoup plus complexe, voire impossible. Les contraintes linguistiques et de temps sont deux facteurs de cette difficulté de définition d'une stratégie de recherche complète. À ce titre, l'exhaustivité d'une bibliographie est un critère difficilement atteignable à l'échelle humaine. Il est tout à fait envisageable qu'une plateforme sémantique participative à grand déploiement puisse constituer des bibliographies thématiques à l'échelle mondiale. Finalement, le dernier axe rejoint les propos de Tim Sherratt qui rappelle l'importance de préserver la forme narrative, élément essentiel de l'explication historique. Le travail de l'historien est simplement de fournir en supplément de son texte écrit des données normalisées qui faciliteront la réutilisation de son contenu. On migre alors vers des compétences actuellement détenues par des archivistes.

3.2. Compétences de l'archiviste

Le rôle de l'archiviste au sein d'une organisation est avant tout d'organiser, de conserver et d'assurer l'accessibilité des archives. La structure des données est donc un élément central de la réflexion archivistique à l'ère du numérique. En milieux culturels, il expose ses

données aux différents publics, dont les historiens. Selon l'Association des archivistes du Québec (AAQ), l'une des tâches de l'archiviste est d'élaborer et de développer des outils archivistiques (2016). À ce titre, l'explicitation du contenu est un enjeu majeur afin de constituer un environnement sémantique pertinent pour les sciences historiques. Suivant les règles du Web sémantique, l'archiviste sera amené à positionner un document d'archives dans son contexte, ce qu'il fait déjà, mais aussi à le mettre en relation avec un ensemble d'autres éléments du patrimoine. Il sera alors confronté à différents niveaux de normalisation. Les bibliothèques et les musées ont deux façons nettement différentes de regarder la donnée. Le bibliothécaire a l'avantage d'avoir un grand nombre de référents externes lorsqu'il ajoute un élément à sa base de données, ce qui n'est pas le cas du muséologue qui doit cataloguer un objet unique ou possédant un contexte particulier. Dans un projet de données culturelles liées, l'archiviste devra élaborer un programme de gestion des sources pluridisciplinaires, notamment en instaurant une ontologie propre au projet, qui s'inspirera grandement de CIDOC CRM. Sa connaissance accrue des bases de données et de leur fonctionnement fait de lui un joueur essentiel dans la mise en place des outils collaboratifs. Évidemment, son expertise devra entrer en résonnance avec les autres disciplines des sciences historiques, mais aussi avec les programmeurs. Les sciences de l'information ont un accès privilégié aux professionnels en informatique que l'historien n'a pas. Cette connaissance des vocabulaires, des bases de données et des langages de requêtes facilite l'interaction et favorise le développement d'une plateforme réfléchie pour les disciplines historiques.

3.3. Favoriser les rapprochements

Afin de bâtir une méthodologie commune entre l'archivistique et l'histoire orientée autour des concepts du Web sémantique, l'architecture de l'information offre une définition qui se fait l'écho aux deux disciplines :

> 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. (Salaün, 2013)

Pour construire une plateforme sémantique commune aux sciences historiques, l'archiviste et l'historien devront miser sur cinq compétences

mitoyennes présentées par Jean-Michel Salaün, professeur en sciences de l'information et de la communication de l'École normale supérieure (ENS) de Lyon. Premièrement, la gestion dynamique des projets facilitera la segmentation du projet en étapes et permettra d'établir les mécanismes de collaboration et ce, dès le départ. Deuxièmement, la coopération et le dialogue avec les métiers connexes comme les autres disciplines historiques et les partenaires techniques permettront la mise en application d'un processus de création interdisciplinaire concerté. Troisièmement, il faut réfléchir à l'expérience utilisateur telle que définie par 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. » (Grenzinger, 2013) Dans le domaine des sciences historiques, il faut penser aux chercheurs, mais aussi à la diffusion et à l'appropriation du contenu par le grand public. Quatrièmement, l'historien et l'archiviste doivent structurer l'information selon différents vocabulaires adaptés à leurs problématiques. Cette étape nécessitera de la part de l'historien une transposition d'une méthodologie organique, soit construire selon le sujet de recherche, vers une méthodologie structurée pour favoriser les croisements interdisciplinaires. La dernière compétence consiste à savoir constituer des prototypes. Dans un environnement en données liées, les chercheurs ne peuvent plus concevoir la transmission du savoir uniquement au travers de projets complétés. Intégrant une structure en continuelle expansion, chaque projet devient un ajout ou un complément à d'autres données déjà compilées. C'est cet effort d'intégration à des corpus interdisciplinaires plus vastes qui permettra la création d'une plateforme sémantique en sciences historiques.

CONCLUSION : PROJETS INTÉGRATEURS ET INTERDISCIPLINAIRES

Le Web sémantique comme outil d'intégration

Le Web 2.0 s'apparente à l'interdisciplinarité, définie comme mode de coopération (Payette, 2001) Le Web tel qu'on le connaît est une vitrine extraordinaire pour les projets de recherche. La coopération se fait via des sites Web qui permettent de partager des points de vue, de partager des expertises, de transmettre de l'information, de se consulter et de travailler en équipe. Cependant, comme il fut mentionné précédemment,

il s'agit d'une juxtaposition de contenus qui demande une agrégation par l'humain. Le Web sémantique, comme il fut mentionné à plusieurs reprises dans cet article, permet une interdisciplinarité basée sur l'intégration des contenus (Payette, 2001). Il y a donc une couche généraliste qui facilite la migration des données. Dans le cas des données liées, il s'agit des vocabulaires contrôlés et des ontologies qui font office de tronc commun. Pour qu'une intégration fonctionne, il faut que le modèle d'ensemble soit réfléchi afin que chaque participant futur puisse intégrer de façon harmonisée le modèle fédérateur. Ce procédé permettra de décloisonner les données disciplinaires au sein d'une structure qui facilite l'acquisition d'une information originellement distincte d'une méthodologie particulière. En d'autres mots, un historien pourra assimiler avec rigueur une donnée normalisée en RDDA, car cette dernière peut être présentée selon le modèle CIDOC CRM, plus facilement compréhensible pour un utilisateur externe.

Projets pilotes interdisciplinaires

Le processus théorique présenté ici prétend que le Web sémantique est l'outil par excellence afin d'harmoniser les contenus archivistiques et historiques. L'application pratique et l'acquisition des compétences peuvent rapidement nuire au développement des données culturelles québécoises liées. S'appuyer sur des études de cas devient un essentiel pour éviter différentes dérives méthodologiques. Un des projets en contexte canadien est celui du Réseau pancanadien du patrimoine documentaire (RPCPD) intitulé *Au-delà des tranchées* qui s'intéressait à la mise en commun des archives associées aux soldats canadiens ayant participé à la Première Guerre mondiale (Réseau pancanadien du patrimoine documentaire, 2012). Étant une démonstration de faisabilité, le projet se justifie en trois points. Premièrement, le Web sémantique permet d'ouvrir l'accès aux données qui seront malléables et est donc en mesure de répondre aux besoins littéralement imprévisibles des chercheurs. Deuxièmement, les données liées favorisent la création d'une expérience cohérente et riche pour l'utilisateur. Troisièmement, on encourage la mise en réseau des connaissances ce qui, avec le temps, rehaussera la valeur informationnelle d'un jeu de données. L'utilisation d'un sujet historique afin de rallier cinq institutions patrimoniales canadiennes, comme le propose le projet du RPCPD (Réseau pancanadien du patrimoine documentaire, 2012), est l'approche à retenir pour une expérimentation efficace, puisque l'histoire nécessite la mise en commun de diverses sources. Ensuite, suivant

le processus réalisé par le RPCPD, il faut regrouper les métadonnées et les organiser selon un modèle ontologique réfléchi qui s'appuie sur des normes existantes. La création d'URI sera aussi nécessaire, mais il faut miser sur une réutilisation de ceux existants afin de positionner le projet dans le nuage des données liées.

Plan culturel numérique du Québec : une occasion à saisir

Le désir d'une plus grande collaboration entre l'archiviste et l'historien était présent bien avant l'apparition du Web sémantique. Au sein des institutions, les deux disciplines se côtoient et échangent fréquemment. Il n'est pas rare de retrouver un duo archiviste-historien dans un centre d'histoire local par exemple. Le Web sémantique ouvre simplement la porte à une normalisation de cette méthodologie interdisciplinaire à différents niveaux, que ce soit local, provincial, fédéral ou international. La difficulté reste de mobiliser les chercheurs et les professionnels des milieux à l'enjeu du Web sémantique pour les données patrimoniales québécoises. C'est le défi que se donne le ministère de la Culture et des Communications du Québec (MCC) avec le *Plan culturel numérique du Québec* (PCNQ) qui propose la mesure 06 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* (2014). Avec un comité d'experts qui s'est réuni pour la première fois en mars 2016, il faut espérer que des recommandations concrètes et novatrices ressortiront de ce processus de réflexion. À la lumière de ce court article, le comité se doit d'analyser trois composantes fondamentales d'un projet en données patrimoniales liées, soient la démocratisation des concepts techniques, la normalisation des contenus et l'intégration des disciplines.

Philippe Michon

BIBLIOGRAPHIE

ASSOCIATION DES ARCHIVISTES DU QUÉBEC. (2016). Fonctions et tâches de l'archiviste. Repéré à https://archivistes.qc.ca/devenir-archiviste/

AYERS, E. L. (1999). History in hypertext. Repéré le 27 février 2014 à http://www.vcdh.virginia.edu/Ayers.OAH.html

BACHIMONT, B. (2006). Qu'est-ce qu'une ontologie ? Repéré le 1 avril 2016 à http://www.technolangue.net/imprimer.php3?id_article=280

BERNERS-LEE, T., Hendler, J. et Lassila, O. (2001). The semantic Web. Repéré le 1 avril 2016 à http://www.scientificamerican.com/article/the-semantic-web/

BIBLIOTHÈQUE NATIONALE DE FRANCE. (2015). Web sémantique, web de données : définitions. Repéré le 28 avril 2016 à http://www.bnf.fr/fr/professionnels/anx_web_donnees/a.web_donnees_definitions.html

CRAMER, F. (2007). Critique of the « Semantic Web ». Repéré le 28 avril 2016 à http://www.nettime.org/Lists-Archives/nettime-l-0712/msg00043.html

CYGANIAK, R. et JENTZSCH, A. (2014). The Linking Open Data cloud diagram. Repéré le 31 octobre 2014 à http://lod-cloud.net/

EUROPEANA. (2014). Definition of the Europeana Data Model v5.2.6. Repéré le 1 avril 2016 à http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM%20Definition%20v5.2.6_01032015.pdf (Lien non fonctionnel).

FLOUD, R. (1979). Introduction to quantitative methods for historians. Londres: Methuen

GRENZINGER, Y. (2013). Qu'est-ce que l'expérience utilisateur ? Ergonomie, Expérience Utilisateur, Design Thinking. Repéré le 1 avril 2016 à http://ux-fr.com/experience-utilisateur-definition/

GULDI, J. et ARMITAGE, D. (2014). *The History Manifesto* (Cambridge University Press). Cambridge Repéré le 1 avril 2016 à http://historymanifesto.cambridge.org/read

INTERNATIONAL COUNCIL OF MUSEUMS. (2014a). The CIDOC CRM. Repéré le 1 avril 2016 à http://www.cidoc-crm.org/

INTERNATIONAL COUNCIL OF MUSEUMS. (2014b). The CIDOC CRM mappings, specializations and data examples. Repéré le 1 avril 2016 à http://cidoc-crm.org/crm_mappings.html (Lien non fonctionnel).

LÉTOURNEAU, J. (2006). *Le coffre à outils du chercheur débutant : guide d'initiation au travail intellectuel*. Montréal : Boréal.

LINCOLN, M. (2015). Using SPARQL to access linked open data. *Programming Historian*. Repéré le 1 avril 2016, à l'adresse http://programminghistorian.org/lessons/graph-databases-and-SPARQL.html

MICHON, P. (2016). *Vers une nouvelle architecture de l'information historique : L'impact du Web sémantique sur l'organisation du Répertoire du patrimoine culturel du Québec* (Mémoire). Université de Sherbrooke, Sherbrooke.

MINISTÈRE DE LA CULTURE ET DES COMMUNICATIONS DU QUÉBEC. (2014). 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. Repéré le 10 février 2016 à 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/

NILSSON, M., POWELL, A., JOHNSTON, P. et NAEVE, A. (2008). Expressing Dublin Core metadata using the Resource Description Framework (RDF). Repéré le 1 avril 2016 à http://dublincore.org/documents/dc-rdf/

ORWELL, G. (2005). *1984* (Gallimard). Paris : Gallimard.

PAYETTE, M. (2001). Interdisciplinarité : clarification des concepts. *Interactions*, *5*(1), 1936.

RÉPERTOIRE DU PATRIMOINE CULTUREL DU QUEBEC. (2013). Cartier, Jacques. Repéré le 14 novembre 2015 à http://www.patrimoine-culturel.gouv.qc.ca/rpcq/detail.do?methode=consulter&id=17323&type=pge#.%20VkaXKGs1uzw

RÉSEAU PANCANADIEN DU PATRIMOINE DOCUMENTAIRE. (2012). *« Démonstration de faisabilité » de la Visualisation des Données ouvertes liées (LOD) - Au-delà des tranchées* (p. 83). Repéré le 1 avril 2016 à http://www.canadiana.ca/sites/pub.canadiana.ca/files/PCDHN%20Proof-of-concept_Final-Report-FRA.pdf (Lien non fonctionnel).

RUIZ, É. (2015). Les historien-nes et le numérique : usages et besoins de formation. Repéré le 31mars 2016 à http://www.boiteaoutils.info/2015/03/historiens-numerique/

SALAÜN, J.-M. (2013). Référentiel de compétences en Architecture de l'information [Billet]. Repéré le 1 avril 2016 à http://archinfo01.hypotheses.org/453

SEMANTIC WEB. (2012). Ontology. Repéré le 1 avril 2016 à http://semanticweb.org/wiki/Ontology

SHERRATT, T. (2011). Every story has a beginning. Repéré le 31 août 2015 à http://discontents.com.au/every-story-has-a-beginning/

SHERRATT, T. (2015). Inigo Jones - The weather prophet. Repéré le 14 janvier 2016 à http://lodbookdev.herokuapp.com/#!/text/1/

SOWMYA, S. K. (2013). A bottom-up approach towards achieving semantic web services. Dans *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (p. 13171322). http://doi.org/10.1109/ICACCI.2013.6637368

STEAD, S. (2008). The CIDOC CRM, a standard for the integration of cultural information. Repéré le 1 avril 2016 à http://old.cidoc-crm.org/cidoc_tutorial/index.html

THALLER, M. (2012). Controversies around the digital humanities: an agenda. *Historical Social Research / Historische Sozialforschung*, *37*(3 (141)), 723.

THALLER, M. (1995). KLEIO Introduction. Repéré le 1 avril 2016 à http://www.hki.uni-koeln.de/kleio/old.website/tutorial/intro.htm#mark0 (Lien non fonctionnel).

VERBORGH, R. (2015). Turtles all the way down. Repéré le 14 novembre 2015 à http://ruben.verborgh.org/blog/2015/10/06/turtles-all-the-way-down/

W3C. (2001). URIs, URLs, and URNs: clarifications and recommendations 1.0. Repéré le 1 avril 2016 à https://www.w3.org/TR/uri-clarification/

W3C. (2008). SPARQL Query Language for RDF. Repéré le 1 avril 2016 https://www.w3.org/TR/rdf-sparql-query/

W3C. (2014a). RDF - semantic Web standards. Repéré le 1 avril 2016 à https://www.w3.org/RDF/

W3C. (2014b). RDF Schema 1.1. Repéré le 1 avril 2016 à https://www.w3.org/TR/rdf-schema/