LES ENJEUX DU WEB SEMANTIQUE

ET DES DONNEES OUVERTES LIEES

POUR LES HUMANITES NUMERIQUES

Mémoire de Master d'Alice DEFOURS

Sous la direction de Mme Sabine LOUDCHER

Illustration de couverture : Graphe des évènements ayant eu lieu à Toronto, des organisations domiciliées dans cette ville et des périodiques en faisant mention, constitué à partir des données ouvertes liées du projet Lesbian and Gay Liberation in Canada (LGLC) piloté par Constance Crompton et Michelle Schwartz, 2019.

Université Lumière Lyon 2

Institut de la communication (ICOM)

Master Humanités Numériques

Les enjeux du web sémantique et des données ouvertes liées pour les Humanités Numériques

Présenté par Alice Defours,
Etudiante en deuxième année du Master Humanités Numériques

Sous la direction de Mme Sabine Loudcher,
Enseignante-chercheure, professeure des universités de Lyon 2, Laboratoire ERIC, ICOM

Mémoire soutenu le 10 Septembre 2019

Année Universitaire 2018 - 2019

Sommaire

Liste des abréviations

ASK : *Association for Social Knowledge*

CSV : *Coma-Separated Values*

CWRC : *Canadian Writing Research Collaboratory*

DCMI : *Dublin Core Metadata Initiative*

DDC : *Digital Dinah Craik*

DNS : *Domain Name System*

DOL : Données Ouvertes Liées

GLF : *Gay Liberation Front*

HN : Humanités Numériques

HTML : *Hypertext Markup Langage*

HTTP : Protocole de transfert hypertexte (*Hypertext Transfer Protocol*)

LINCS : *Linked Infrastructure for Networked Cultural Scholarship*

LGLC : *Lesbian and Gay Liberation in Canada*

LOD : *Linked Open Data*

OWL : *Web Ontology Language*

RDF : *Resource Description Framework*

RDFS : *Resource Description Framework Schema*

SHS : Sciences Humaines et Sociales

URI : *Uniform Resource Identifiers*

URL : *Uniform Resource Locator*

VIAF : *Virtual International Authority File*

XHTML : *Extensible Hypertext Markup Langua*

XPATH : *XML Patch Language*

XSLT : *eXtensible Stylesheet Language*

WWW : *World Wide Web*

W3C : *World Wide Web Consortium*

Remerciements

Toute ma plus sincère gratitude va à Sabine Loudcher, enseignante chercheure à l'Université Lyon 2, pour sa détermination précieuse qui porte le projet Franco-canadien depuis le premier jour, pour son implication dans la réussite de cet échange et pour son soutien incroyable depuis le début de cette aventure.

Un énorme merci à Constance Crompton, PI CRC en Humanités Numérique à l'Université d'Ottawa, pour sa bienveillance unique dès notre rencontre, pour sa confiance fabuleuse en mes compétences et pour toutes les merveilleuses opportunités professionnelles qu'elle m'offre, thank you !

Je souhaite aussi remercier Elisa Beshero-Bondar (Université de Pittsburg) et David J. Birnbaum (Université de Pittsburg) pour leur détermination et pour les heures qu'ils ont passées, après chaque journée de cours à DHSI, avec chacun de leurs élèves pour nous aider à améliorer nos codes personnels.

J'adresse enfin ma reconnaissance la plus totale à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail mais aussi à l'équipe du Humanities Data Lab. pour sa bonne humeur indéfectible, à mes relecteurs qui ont abattu un travail colossal, ainsi qu'aux membres du jury dont les conseils auront une grande valeur.

INTRODUCTION

En Février 2009, Tim Berners-Lee s'exprimait en ces mots : « J'aimerais maintenant que l'on imagine, non pas deux jeux de données connectés ensemble, […] mais un monde où tous auraient partagé des données sur le web pour que virtuellement, tout ce dont vous ayez rêvé soit sur le web ; j'appelle cela les Données Connectées. »[1]. A ce moment-là, le World Wide Web qu'il avait conçu fêtait alors ses vingt ans. Aujourd'hui, après dix années de travail supplémentaires, le potentiel jusqu'alors inexploité des données est devenu la clé de voute d'un système qui n'a pas encore fini d'évoluer.

Comment le web des données est-il né ? Projetons-nous quelques années en arrière à une époque où les chercheurs utilisaient des technologies différentes. En 1950, on travaillait déjà sur des projets de recherche novateurs mais chaque équipe avait son matériel et son organisation technique qui lui étaient propres[2]. Chaque projet était sauvegardé sur des supports de stockage individuels dans leur format natif, ces derniers n'étant pas toujours identiques d'une institution à l'autre. Lorsque l'on voulait étudier ces projets et visualiser leurs données, il fallait à chaque fois se connecter sur des machines différentes, ou être muni de / des CD-ROMs, et être en possession des logiciels adaptés. Si la diversité de ces projets était déjà louable, ils ne pouvaient pas communiquer entre eux : la pluralité des méthodes de stockage, des logiciels et des formats ne rendait pas ces documents accessibles au plus grand nombre. C'est alors qu'on a eu l'idée de les intégrer à un système de documentation virtuel qui permettrait d'y accéder depuis n'importe quel poste dans le monde. Le partage de documents sur Internet a notamment été rendu possible par la création d'un langage de communication entre les clients et les serveurs unifiés : le protocole de transfert hypertexte. Chaque document pouvait être lié à un autre, c'est-à-dire que chaque projet pouvait avoir un endroit dédié où publier ses recherches et connecter ses documents à d'autres grâce aux liens

[1] Tim Berners-Lee s'est exprimé en ces mots lors des conférences de TED2009 ayant eu lieues à Long Beach, en Californie, en Février 2009 : « I want us now to think about not just two pieces of data being connected […] but of a world where everybody has put data on the web and so virtually everything you could imagine is on the web, and I'm calling that Linked Data. ».

[2] HITZLER P., KROTZSCH M., RUDOLPH S., *Foundations of Semantic Web Technologies*, New York, 2009 : Aux Etats-Unis, plusieurs sites avec des infrastructures informatiques et déjà des projets requerraient un échange de données entre eux.

hypertextes. N'importe quel lien placé dans une page web pointait vers une autre et par extension vers d'autres documents. Ce système a donc inauguré les bases du web que l'on connait aujourd'hui.

Si la connaissance est ainsi devenue davantage accessible, cet outil de partage n'en restait pas moins qu'une bibliothèque de stockage qui ne permettait pas le dialogue entre les contenus. Chaque document est précieux par les informations qu'il contient. Isolées les unes des autres, à cause du format initial du web, ces informations ou données ne peuvent pas communiquer entre elles. Les chercheurs ont vu dans ce problème un potentiel inexploité : c'est ainsi que sont nées les premières problématiques orbitant autour des données. Le concept au cœur des données liées est la relation qu'elles ont entre elles. Prenons un exemple simple : si deux données identiques existent dans deux documents différents, elles créent alors un lien entre ces deux documents. Ce phénomène instaure une relation à double-sens telle que les deux documents liés partagent du contenu similaire. Ils peuvent donc s'enrichir mutuellement en nouvelles données qui viennent compléter le contexte de l'un et de l'autre. Ce type de relation pourrait être visualisé grâce à un graphe qui rendrait compte de la multitude et de la complexité du réseau de données.

Nous comprenons donc maintenant comment s'enchevêtrent documents et données. Evidemment, il existe des règles à respecter pour publier des données sur le web et faire en sorte qu'elles soient liées et ouvertes. Nous pourrons traiter ce sujet en détail plus tard dans cette discussion. Cependant, en guise d'introduction il faut retenir que les protocoles de transfert hypertexte ne sont plus seulement utilisés pour décrire des documents : ils décrivent dorénavant le contenu de ces documents (noms de personnes, dates, lieux, etc.). La donnée, qui est finalement une ressource, est donc localisée sur le web grâce à cette adresse que l'on connait aujourd'hui communément sous l'appellation *Uniform Resource Identifier* (URI)[3]. Enfin, insistons sur le fait qu'une donnée n'est pas brute ni isolée de contexte : elle a des relations, concept central des données « liées ». Leur accessibilité est aussi au cœur de notre raisonnement, d'où le terme données « ouvertes ». Les connaissances stockées sur des bases qui ne sont pas partagées ont un potentiel inexploité. Les rassembler et promouvoir leur

[3] ALLEMANG D., HENDLER J., *Semantic Web for the working ontologist effective modeling in RDFS and OWL*, San Francisco, 2011 : nous recommandons d'étudier le chapitre dédié aux URI.

partage déverrouille l'accès vers de nouvelles découvertes : c'est l'opportunité de se poser de nouvelles questions et d'établir des relations entre différents contextes qui participent à la construction d'un pont entre plusieurs disciplines.

Motivé par une initiative du W3C, *Linking Open Data*, le projet *DBpedia* a promu l'étude de sources issues de sites comme Wikipédia[4]. C'est l'un des premiers projets qui a vu le jour dans ce climat d'intérêt croissant pour le web des données. L'interface de requêtage SPARQL ainsi développée permet, en effet, à l'utilisateur de questionner les bases de données ouvertes du site[5]. Les jeux de données sont en partie composés des informations brutes que l'on trouve notamment dans les encarts des pages web du site. Ce projet est un exemple parmi d'autres qui illustre les aboutissements techniques que permettent les données ouvertes et liées[6]. Les méthodes de traitement de ce type de données varient mais ont dans tous les cas de grandes conséquences sur les projets en Sciences Humaines et Sociales[7]. D'ailleurs, les SHS qui gèrent leurs données à travers le prisme des méthodes de traitement que nous venons de présenter deviennent des Humanités Numériques.

Si le but de cet exercice n'est pas de donner une définition précise de ce que sont les Humanités Numériques, nous ne pouvons toutefois pas omettre d'en mentionner le terme[8]. Interdisciplinaire par essence, ce domaine au sens large est né de l'alliance entre l'Histoire, les Sciences Humaines et les Sciences de l'Informatique. Au même titre que les données ouvertes liées créent des relations entre les disciplines via le croisement des connaissances, les Sciences de l'Informatique, en offrant de nouveaux outils aux SHS, soutiennent la création de ces ponts interdisciplinaires. Les chercheurs ont pris conscience de la valeur de cette interdisciplinarité et de ses bénéfices : les méthodes de travail nouvelles sont plus efficaces et offrent davantage d'opportunités d'interprétation. Cela donne naissance à des projets novateurs, au sein du

[4] DBPEDIA [en ligne, consulté le 14/05/2019, disponible à https://wiki.dbpedia.org/services-resources/interlinking].

[5] DUCHARNE B., Learning SPARQL, Sebastopol, 2013.

[6] Dans la sphère de recherche anglophone, les chercheurs utilisent le terme original *Linked Open Data*.

[8] Voir, CROMPTON C., LANE R., SIEMENS R., *Doing Digital Humanities*, Londres, 2016 pour une définition en détail des Humanités Numériques.

domaine de l'édition par exemple, avec la création de livres augmentés[9] ou la publication de corpus enrichis[10] en métadonnées.

Il est aussi impossible de mentionner les métadonnées, ces données qui décrivent d'autres données, sans introduire le web sémantique. Une communauté participative est ce qui constitue la base même du *World Wide Web* : tout le monde peut dire ce qu'il veut sur n'importe quel sujet. En conséquence, plus cette communauté s'agrandit, plus le web croît. Ce dernier souffre alors de ce que les ontologistes appellent en anglais *data wilderness*[11]. Traduisible en français par « données indomptables », ce terme désigne le résultat engendré par le grand nombre de ressources non-triées qui existent sur le web.

Quelle solution a-t-on trouvé à ce problème ? Trier les ressources revient à les ranger et c'est ce que le web sémantique fait : il fournit une nouvelle infrastructure servant à ordonner le web au niveau des données au moyen d'un modèle appelé *Resource Description Framework* (RDF). Le RDF sert donc à décrire les données et métadonnées présentes sur le web. Un peu plus tôt nous avons mentionné à quel point il était important que les données soient liées ensemble pour qu'elles puissent créer des ponts interdisciplinaires. Elles pointent l'une vers l'autre en utilisant des références globales : il s'agit des URI qui sont finalement une des technologies du web sémantique.

Le web sémantique et ses technologies sont des outils largement utilisés aujourd'hui dans les projets en Humanités Numériques aujourd'hui qui doivent traiter des corpus de documents textuels. C'est le cas pour les projets portés par le laboratoire *Humanities and Data Lab*. de l'Université d'Ottawa (Canada) dont les données ont alimenté ce mémoire et qui seront présentées plus tard.

[9] Lettre d'information de l'Université de Lille 3 [en ligne, consulté le 24/06/2019, disponible à https://www.univ-lille3.fr/actualites/?actu=15011].

[10] Dossiers du projet Bouvard & Pécuchet [en ligne, consulté le 17/07/2019, disponible à http://www.dossiers-flaubert.fr/projet-originel].

[11] ALLEMANG, op.cit.

Pour l'heure, cette introduction contextualise le développement de notre exercice. Un projet incapable de dialoguer avec d'autres projets manque de puissance scientifique. Nous nous proposons, dans cette réflexion, d'explorer de quelle manière nous pourrions au mieux connecter les projets, leurs corpus et leurs données pour augmenter leur potentiel d'utilisation, d'interprétation et leur pérennité. Nous allons donc étudier comment les techniques de sémantisation du web sont utilisées pour servir au mieux la communauté scientifique en Humanités Numériques. Inévitablement, nous rencontrerons les conséquences qu'ont ces méthodes de travail sur les Sciences Humaines et Sociales[12].

Pour répondre à ces questions, nous étudierons d'abord l'état de la recherche sur la question des données ouvertes liées et du web sémantique. Nous verrons comment la transition d'un web de documents à un web de données s'est faite plus en détails et présenterons des clés favorisant la compréhension du contexte dans lequel nous travaillons. Notre matériel de recherche étant essentiel composé de données textuelles, nous en ferons la présentation puis expliquerons de quelle manière elles ont été analysées et dans quel but. Ces dernières remarques feront office de transition vers la fin du développement. Nous analyserons à plus grande échelle les conséquences de l'utilisation du numérique en Sciences Humaines et Sociales par le biais d'autres travaux en cours issus du même champ de recherches.

Nous souhaitons également prévenir le lecteur que ces travaux de recherche sont largement alimentés par des sources anglophones et qu'il risque, occasionnellement, de rencontrer des termes anglais, comme c'est déjà le cas dans cette introduction, qu'il ne serait pas élégant de traduire en français.

[12] Etude qui a des relations connexes aux sciences sociales notamment en Histoire et en Ethnologie.

PARTIE 1 : Petite Histoire des Données Ouvertes Liées et présentation du matériel de recherche

1. Historiographie

Avant les données ouvertes liées et le web sémantique, chaque page web était un document qui parfois pointait vers un autre document, une autre page, grâce à des liens hypertextes. La recherche en sciences de l'informatique a fait évoluer le web jusqu'à ce qu'il soit maintenant possible de créer des séries de données interconnectées et interopérables. Quelle est la trame de fond de cette histoire ? Quelles limites ou besoins ont-ils déclenché des mutations structurelles ?

A. A propos de la naissance du web

Nous ne pouvons contextualiser cette étude sans mentionner quelques jalons ayant marqué la naissance du web. Cette courte histoire ne saurait se montrer exhaustive en faits mais fournit au lecteur un repère chronologique essentiel à la compréhension de l'évolution des sciences de l'informatique fondée par Alan Turing. Notre Histoire raccourcie des données ouvertes liées commence avec la création des premiers réseaux de communication distribués. C'est dans un climat de conflits qu'ARPANET est né en 1969[13] et il est le premier réseau opérationnel de commutateurs par paquets. En d'autres mots, c'est un réseau par lequel transitent de larges volumes de données découpés, que l'on appelle des paquets, qui ont notamment donné son nom à cette technologie en anglais *packet-switching*[14]. Conçu pour améliorer les tactiques militaires du gouvernement américain, il a fini par glisser vers la sphère publique lorsque l'on s'est rendu compte qu'il apportait une solution à des problèmes de télécommunication. ARPANET fonctionne aux antipodes des réseaux centralisés précédant sa naissance. En décentralisant la globalité du réseau de communication entre plusieurs

[13] LUKASIK S., "*Why the Arpanet Was Built*" in *IEEE Annals of the History of Computing*, 1058-6180/11, 2011, pp. 4-20.

[14] L'idée de diviser l'information en paquets est attribuée à Paul Baran dans les années 1960.

machines, les ingénieurs ont réalisé qu'il était plus résistant à des attaques : si une machine était endommagée, un chemin coupé, l'information était redirigée et pouvait quand-même arriver à destination par d'autres biais[15]. L'information voyage dans le réseau sous forme de paquets, la disparition d'une machine affecte seulement une route potentielle de livraison du paquet. Dans le cas d'un réseau centralisé, la disparition d'une machine cause l'extinction entière de celui-ci. Les chercheurs tenaient donc un moyen d'améliorer substantiellement les télécommunications au sens large. C'est ainsi que la vocation d'ARPANET avec sa technologie et ses protocoles s'est lentement transformée jusqu'à servir de base à l'une des plus grandes innovations du XXème siècle : Internet.

En 1989[16], un informaticien britannique, Tim Berners-Lee, posait une autre pierre à l'édifice en proposant un outil standardisant l'échange de documents entre différents systèmes informatiques indépendants : le protocole de transfert hypertexte (HTTP). Il a mis au point, au cours des dix années suivantes, les piliers fondateurs du World Wide Web dont un langage pour construire des pages web (*Hypertext Markup Langage*, HTML) et un serveur destiné à l'héberger. L'ensemble de ces innovations permet alors l'échange, à grande échelle, d'informations tout autour du monde. Lorsqu'en 1991, le premier serveur web, en dehors de l'Europe, est mis en marche, Berners-Lee s'exprime en ces mots : « Le projet World Wide Web vise à encourager le partage et la création de liens vers des informations où qu'elles soient. Ce projet a d'abord été créé pour permettre à des physiciens d'échanger des données et de la documentation. Nous sommes impatients d'élargir les perspectives qu'offre le web à d'autres disciplines »[17]. Il venait finalement de créer un web de documents.

[15] LUKASIK, op.cit. : ARPANET a d'abord été conçu pour les forces armées en prévision d'un conflit international pour assurer que les intallations restent opérationnelles en cas de dégradation de certaines infrastructures.

[16] HITZLER, op. cit.

[17] "The *World Wide Web* (WWW) project aims to allow links to be made to any information anywhere […] The WWW project was started to allow high energy physicists to share data, news, and documentation. We are very interested in spreading the web to other areas and having gateway servers for other data. Collaborators welcome!", Tim Berner-Lee in HITZLER, op. cit.

B. Transition d'un web de documents à un web de données

Il est maintenant légitime de se demander quel facteur a été l'élément déclencheur de la prise de conscience du potentiel des données et dans quelles circonstances cette transition s'est opérée. La réponse à ces questions se situe sans doute dans le format même du web initial. La plupart des pages étaient alors seulement conçues pour être lisibles par l'œil humain et les machines ne pouvaient pas comprendre la nature du contenu partagé. C'est ce que les chercheurs ont parfois appelé le web syntactique où la mise en page du contenu est assurée par l'ordinateur et où l'interprétation des informations est déléguée aux utilisateurs[18]. Pourquoi est-il donc arrivé à bout de souffle ?

Nous avons introduit un peu plus tôt l'idée qu'un effet de réseau soit à l'origine de l'agrandissement de la communauté du web. Les chercheurs, ou le grand public en l'occurrence, ont eu accès à un outil leur permettant de partager n'importe quelle information librement avec le World Wide Web. Karin Breitman souligne d'ailleurs que la taille des résultats de recherche devenait de plus en plus difficile à interpréter par les opérateurs humains et la documentation pertinente avait tendance à se perdre dans les méandres de ce web. Etant donné que les informations numériques abondent et que le nombre des utilisateurs croît sans précédent, la gestion de l'ensemble des données partagées devient un défi. Breitman utilise aussi le terme *information overload,* synonyme du *data wilderness* d'Allemang,[19] pour nommer ce phénomène et le considère comme la plus grande menace du web. Elle nuance son propos en insistant sur le fait que les pages web n'ont aucun moyen de référencer ou d'identifier le contenu et les sujets auxquels elles réfèrent[20]. C'est un peu comme si l'on entreposait à un endroit une quantité folle de ressources sans les trier et sans même savoir ce dont elles parlent : le lecteur perd alors un temps considérable à effectuer sa

[18] BREITMAN K., CASANOVA M., TRUSZKOWSKI W., *Semantic Web: Concepts, Technologies and Applications*, Londres, 2007 : A l'image de certains de ses contemporains, que nous avons ou citerons bientôt dans ce développement, l'auteur souligne que le format du web des années 2000 n'était pas fait pour supporter l'échange de données mais pour afficher, de manière embellie, des informations interprétées par l'opérateur qui effectuait la recherche.

[19] ALLEMANG, op. cit., chapitre 1.

[20] L'une des solutions qui a été trouvée à ce problème est la méthode du *tagging* : on donne des *tags* à du contenu, des mots-clés pour identifier les notions-clés dont il est fait mention afin de pouvoir localiser l'information : c'est un peu la première carte d'identité des données sur le web.

recherche. L'auteur termine finalement son raisonnement sur le web de documents en reformulant les trois principaux problèmes existants. Premièrement, vu que les recherches contiennent un nombre conséquent d'entrées, le résultat peut manquer de précision. Deuxièmement, ces recherches sont sensibles au vocabulaire employé. Les utilisateurs peuvent employer des termes synonymes de ceux présents dans les pages web : dans ce cas, la recherche peut échouer. Enfin, un même site peut être cité plusieurs fois parmi les résultats de recherche puisque l'outil n'a aucun moyen de savoir ce que les pages web contiennent et qu'il restitue simplement les entrées susceptibles de correspondre à la requête de l'utilisateur[21].

Nous sentons poindre que la sémantique entre en jeu dans l'efficience du web. En fonctionnant avec cette structure-ci, le contenu sémantique (un ensemble de données pertinentes) n'est accessible qu'aux humains. La machine, elle, ne supporte que les rôles de stockage, d'affichage ou de pointage (liens hypertextes d'un document renvoyant à un autre). Plus qu'une nécessité c'est donc un besoin d'efficacité qui a poussé la communauté à chercher une transition vers un web répondant davantage à ses exigences.

C. Classifier le web

Le web sémantique est né alors que les ingénieurs tentaient d'améliorer les performances de la connaissance partagée sur le web en imaginant des outils capables de la parcourir et de la trier. Armer les machines d'applications leur permettant de comprendre les données a été le concept fer de lance d'une décennie de recherches. Dans son format initial, le web contient des informations précieuses mais il n'existe aucune garantie qu'elles seront ordonnées ou parcourables facilement étant donné leur abondance et la vitesse à laquelle la communauté l'abreuve. Comment peut-on construire un web davantage consistant et intelligent ? C'est l'une des questions que se posent Allemang et Jim Hendler. Dans leurs travaux, ils expliquent que les pages web présentent simplement de l'information au lieu de décrire les entités qu'elles contiennent. Il en résulte en un web « idiot » et déconnecté des

[21] BREITMAN, op. cit.

réalités scientifiques auxquelles il prétend[22]. De leur point de vue, il existe un problème qui a aussi été abordé par Breitman : les doublons. Sur le web de documents, chacun est libre de désigner une entité avec son propre vocabulaire. Plusieurs ressources peuvent donc mentionner la même entité sans s'en rendre compte. Nous avons affaire au même problème résolu par Berners-Lee concernant la communication entre projets. Si deux projets ne peuvent pas communiquer entre eux, ils manquent peut-être une occasion de partager des données et d'étayer le contenu l'un de l'autre. Ici, c'est la même chose. Si deux ressources concernent la même entité mais ne la référencent pas de la même manière, aucune connexion entre les deux ne peut s'établir. Il s'agit à nouveau d'une question de sémantique auquelle la nouvelle structure du web apportera une réponse grâce à la standardisation des formats de données. De la même manière que le web traditionnel, les technologies du web sémantique vont être mises au point pour proposer un modèle standardisé de référencement des données. Ce dernier sera un format digeste pour les machines. Explorons maintenant comment ce système a été élaboré.

La structure du web en elle-même est déjà clairement spécifiée par des protocoles et des langages qui permettent l'échange de documents au-delà des frontières logicielles (HTTP, HTML). Les données sont quant à elles désordonnées et indépendamment publiées ; elles ont besoin du même type de schémas sur les bases desquels le web fonctionne. Pour désigner des objets d'études, les biologistes ont été parmi les premiers à vouloir utiliser des catalogues de termes bien précis, lesquels ont été adoptés et partagés par la communauté de scientifiques, afin que tout le monde comprenne les recherches des uns et des autres[23]. Quels principes ont donc utilisés les ingénieurs des sciences de l'informatique pour que des données deviennent interopérables ? Cette classification devait faciliter l'accès aux données en proposant un référencement standardisé. Il s'agissait donc de réviser la façon de structurer le contenu sur le web. Exprimer des données qui ne sont ni plus ni moins que des connaissances est donc une tâche de modélisation. La manière dont cette modélisation est effectuée dépend aussi des usages que l'on souhaite en faire, qui dans ce cas, sont de faciliter l'interopérabilité et l'accessibilité. Les chercheurs en arrivent à se poser trois questions que le web sémantique

[22] Autrement dit en anglais "*Dumb web".*

[23] BREITMAN, op. cit.

doit résoudre afin de fonctionner : Comment construire les modèles ? Comment capturer la connaissance pour qu'elle soit interprétable par les machines ? Comment échanger au mieux toutes ces informations ?

Nous assistons alors à un véritable effort pour construire des machines capables de raisonner et tirer des conclusions significatives du contenu qu'elles lisent. Construire des modèles est une tâche descriptive qui doit faciliter la manipulation de ces entités lues. Un modèle n'est finalement qu'une description de termes, dont on se sert pour faciliter la compréhension de ceux-ci par les ordinateurs. Créer un modèle revient à concevoir un langage de communication universellement utilisable et compréhensible soit, en d'autres termes, « une ontologie ». En Sciences de l'Informatique, une ontologie est bien un catalogue utilisé pour décrire un domaine de connaissances particulier grâce à des termes spécifiques [24]. La signification des termes décrits est alors compréhensible par les machines. Nous aurons l'occasion de travailler avec des ontologies dans cette étude et nous réservons la possibilité d'étayer cette explication en utilisant notre cadre de recherche personnel plus tard dans ce développement.

Classer des données dans des catégories uniques n'est pas la seule solution d'organisation que les chercheurs ont explorée. Ils ont aussi tenté une approche permettant de décrire des entités en fonction de ce qu'elles évoquent : c'est ce qu'on a appelé des facettes[25]. Ces dernières sont un moyen de définir un objet selon une combinaison de critères différents plutôt que par un label unique. S'il existe un effort derrière cette méthode qui ne stéréotype pas les entités, elle est aussi d'une grande complexité à mettre en place. Dans tous les cas, le développement de modèle requiert l'utilisation de nouveaux langages que nous aborderons dans la présentation de notre méthode de travail.

Le web sémantique vise donc à améliorer la logique computationnelle pour faciliter l'échange de ressources entre utilisateurs via des machines capables de les distribuer et de les

[24] BREITMAN, op. cit.

[25] BRUNETTI J., GIL R., GARCIA R., ''*Facets and pivoting for flexible and usable linked data exploration*'' in UNGER C., CIMIANO P., LOPEZ V. (eds.), *et al*., *Proceedings of Interacting with Linked Data*, Heraklion, 2012.

lier. Il a été pensé comme une sorte d'extension du *World Wide Web* qui, en se superposant à lui, entraine des modifications internes de la façon dont structurer les données.

Il resterait encore énormément de concepts à aborder mais cette étude se concentre sur les données ouvertes liées et les technologies du web sémantique grâce auxquelles elles existent. Les ordinateurs deviennent alors capables de parcourir le contenu web et de le traiter plus intelligemment grâce aux modèles et aux langages de description. Ils constituent l'un de nos sujets d'études principaux et que nous aborderons bientôt dans ce développement dont le *Resource Description Framework* (RDF) et le *Web Ontology Language* (OWL).

2. Les données ouvertes liées

Le web de données amène avec lui son lot d'innovations et de nouvelles problématiques. La meilleure gestion des ressources web, rendue possible grâce aux mutations structurelles, s'accompagne de cette envie de rendre la connaissance accessible et connectée au monde scientifique qui l'entoure. En 2010, Tim Berners-Lee démontre dans un amphithéâtre rempli à la Gov Expo 2.0, que comprendre les données ouvertes liées est aussi facile que de lire un sac de chips[26]. Le paquet est le document et les informations imprimées dessus, que les hommes ne sont pas toujours capables de comprendre, sont des données que les machines peuvent gérer. Il insiste toutefois sur les conditions à remplir pour qu'une donnée soit considérée comme ouverte et liée : c'est ce que nous allons explorer maintenant.

D. Définition des DOL

Le principe est le suivant : si tous les jeux de données sur le web sont publiés afin qu'ils soient tous accessibles dans le même format, il serait possible de les réutiliser à volonté. Rendre les données interopérables est la tâche du web sémantique mais cela ne peut se faire sans que trois conditions se rencontrent. Premièrement les données doivent utiliser un format typique des données ouvertes liées, c'est-à-dire que ce dernier doit structurer les ressources

[26] BERNERS-LEE T., Open, Linked Data for a Global Community, Gov 2.0 Expo 2010 [vidéo en ligne, consultée le 17/07/2019, disponible à http://bit.ly/2jRSmf1].

en utilisant des standards connus par les machines qui pourront alors les reconnaître et les traiter en évitant des problèmes de compatibilité. Deuxièmement, une entité doit être référencée de la même façon pour tout le monde. Les données à propos de personnes, de lieux ou d'évènements, par exemple, devraient toujours être référencées de la même manière dans tous les cas où elles sont citées. Troisièmement, les données doivent être publiées de manière ouverte : n'importe qui doit être capable de les utiliser sans frais et sans que la compatibilité logicielle ne soit un souci.

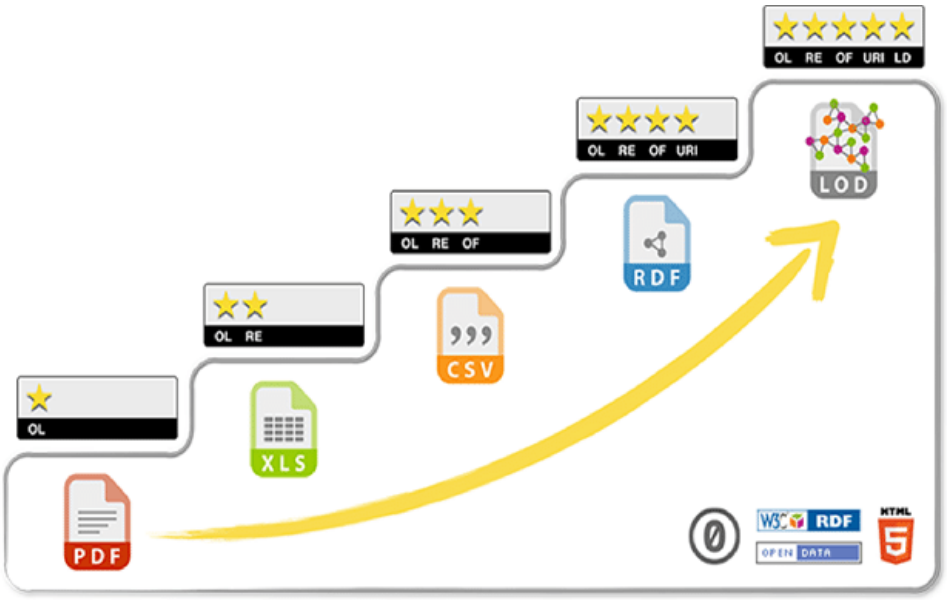

Figure 1 : Les 5stars Linked Open Data de Tim Berners-Lee (CC0 1.0)

Définissant l'essence basale du système, ces trois nécessités, qui sont des critères d'évaluation de la qualité des DOL, ont été schématisées sous forme d'étoiles, par Berners-Lee[27]. Prenons l'exemple d'un chercheur qui met en ligne, sous n'importe quel format (PDF par exemple), son jeu de données sur le web : il obtient une étoile. En faisant un effort, il pourrait aussi publier ce contenu en l'organisant à l'aide d'une feuille de calcul, au format XLS. Ses données deviennent donc manipulables et méritent deux étoiles. Toutefois, ce format a besoin d'un logiciel comme Microsoft® Excel pour obtenir l'accès à toutes les fonctionnalités de calcul. La suite Office®, étant un logiciel payant, l'accessibilité aux données est moyenne. En publiant son jeu au format *Coma-Separated Values* (CSV), le chercheur offrirait la possibilité

[27] Projet 5-star Open Data Plan de Tim Berners-Lee [en ligne, consulté le 16/05/2019, disponible à http://5stardata.info/en/].

à n'importe quel utilisateur de le consulter et manipuler le contenu. En effet, CSV est le format le plus standard d'enregistrement des données tabulaires où la séparation n'est plus opérée par une cellule mais par une virgule : les données remplissent suffisamment de critères pour obtenir leur troisième étoile. Finalement, les deux dernières étapes utilisent des technologies qui ouvrent et lient les données. Pour obtenir une quatrième étoile, les entités contenues dans un jeu doivent être identifiés par des URI afin que d'autres personnes puissent y faire référence et être sûres de parler de la même chose. Enfin, la cinquième étoile s'acquiert en liant ce travail à d'autres travaux qui possèderaient des références similaires et mentionneraient les mêmes entités. Notre chercheur peut se féliciter d'avoir créé des données ouvertes liées !

Le moment est venu de définir plus précisément le fonctionnement de certaines normes exigées par les DOL. Nous allons voir à quoi correspondent concrètement des entités, pour introduire l'une des structures fondamentales des DOL : les triplets. Ces derniers nous conduiront à nous intéresser à la technologie des URI puis aux mécanismes des ontologies qui décrivent nos ressources sur le web.

E. Identification des entités

Assez simplement une entité peut désigner un large éventail d'éléments. Elles sont aussi bien des noms de personnes, de lieux, d'organisations que des informations numériques comme des dates, des indications de métrologie, etc. Dans le cadre de la publication de données ouvertes et liées, une entité est finalement un attribut auquel correspond une valeur. Dans le cas où l'attribut est une personne, la valeur peut-être un nombre. L'idée principale derrière cette notion est qu'une chose, peu importe sa nature, sera toujours unique et désignée par une valeur qui la rendra unique. La capitale de l'Angleterre sera toujours Londres peu importe le nombre d'études qu'elle alimente, et Londres aura toujours le même identifiant sur le web de telle sorte que s'il existait plusieurs villes du même nom, nous soyons sûrs de bien parler de la capitale. Nous avons déjà largement abordé ce sujet précédemment mais il est nécessaire de souligner l'importance de l'unicité des entités puisqu'il s'agit du cœur même des DOL.

Désambiguïser les entités est une tâche à laquelle s'attellent des fichiers spécifiques. Ils constituent une librairie d'identifiants uniques d'entités. Ces fichiers, appelés « autorités », sont mis à jour régulièrement et entretenus quotidiennement par des sources fiables. L'un de ces fichiers, largement connu et utilisé dans le monde, est le Fichier d'Autorité International Virtuel ou *Virtual International Authority File* (VIAF)[28]. VIAF est un projet qui a pour objectif de promouvoir l'accessibilité aux fichiers d'autorité des bibliothèques nationales et de créer des liens entre eux. C'est l'un des plus grands catalogues assemblant, à ce jour, des noms de personnes et des noms géographiques[29]. En plus de permettre l'identification des entités sans ambiguïté, ces notices sont également établies pour normaliser les noms propres. En effet, une entité peut avoir une ou plusieurs orthographes pour de multiples raisons, qu'elles soient historiques, linguistiques, politiques ou culturelles. Finalement, étant donné que les données et les fichiers sont liés ensemble, il est parfois possible de trouver des éléments de contexte pour une entité. Si l'on effectue une recherche sur Tim Berners-Lee dans VIAF, on constate que le fichier propose plusieurs possibilités d'écriture du nom propre, tout en précisant la forme officielle retenue[30]. En poursuivant la recherche, nous avons accès, non seulement à tout le contenu des fichiers d'autorité où Berners-Lee est cité à travers le monde, mais aussi à d'autres éléments liés comme ses travaux. VIAF attribue donc soixante-neuf travaux à Berners-Lee dont certains ont été co-écrits. L'outil propose aussi un planisphère où l'on peut voir les pays ayant publié ces travaux : voilà un exemple de la force des données ouvertes liées.

Bien que ces fichiers soit d'une aide précieuse, il en existe plusieurs et c'est alors à l'utilisateur de décider s'il souhaite travailler avec les notices les plus utilisées ou d'autres qui peuvent parfois correspondre davantage à ses besoins. Parfois, il est possible que les entités n'existent pas dans les autorités. C'est notamment le cas lorsqu'une étude traite d'un sujet inédit qui nécessite la création de nouvelles entités. La plus simple des solutions reste de créer ses propres identifiants. Cela dit, le créateur prend le risque que ses entités existent peut-être

[28] Virtual International Authority File (VIAF) [en ligne, consulté le 14/05/2019, disponible à https://viaf.org/].

[29] GeoNames Authority Files [en ligne, consulté le 14/05/2019, disponible à http://www.geonames.org]. Voir http://www.geonames.org/2643743/london.html

[30] VIAF, op. cit. : https://viaf.org/viaf/85312226/#Berners-Lee,_Tim

dans un fichier d'autorité moins populaire ou bien que ses identifiants ne soient pas réutilisés car ils manqueraient de visibilité.

Identifier les données sur le web suit les mêmes principes que localiser une page grâce à une URL[31]. Ces identifiants sont finalement des pointeurs qui décrivent une entité sur le web et peuvent conduire à une page où de plus amples informations, sur l'entité, sont fournies (des notices d'autorité par exemple). Les *Uniform Resource Identifiers* (URI) sont finalement la carte d'identité des données sur le web. C'est un moyen très fiable et surtout unique de représenter une entité de sorte qu'elle soit utilisable par n'importe quel utilisateur. Reprenons notre exemple d'entité nominale pour décomposer la structure d'une URI. Sur VIAF, Tim Berners-Lee est une personne, un attribut, qui a pour valeur 85312226 : son URI est donc http://viaf.org/viaf/85312226/. La façon d'écrire des URI et des URL est similaire mais leur sens est complètement différent. Alors que les URL sont l'adresse d'une page ou d'un fichier sur le web, les URI sont seulement des identifiants. Elles n'aboutissent pas toujours à une page car elles ne sont pas conçues pour cela (les URL le sont). Finalement, toutes les URL sont aussi des URI (identification d'une page sur le web) mais la réciproque est incorrecte. En d'autres mots, une URI peut identifier n'importe quel objet, tandis qu'une URL n'indique que sa localisation sur le web. Jonathan Blaney utilise l'exemple d'un livre pour éclaircir cette notion : chaque ouvrage possède un ISBN mais ce numéro ne dit pas au lecteur où il peut trouver une copie du livre[32].

Pourquoi une URI ressemble donc-t-elle autant à une URL ? La norme d'écriture des URI requiert qu'elles suivent la syntaxe standard mise en place pour le *World Wide Web* d'où les similarités visuelles qu'elles ont parfois avec les URL[33]. En effet, c'est en partie parce qu'elles utilisent des noms de domaines enregistrés dans le *Domain Name System* (DNS) qu'on peut garantir leur unicité. Les noms de domaines décrivent parfois quelle organisation ou quel auteur est responsable de la description de la ressource : c'est une indication supplémentaire

[32] BLANEY J., *Introduction to the Principles of Linked Open Data*, s.l., 2017 [consulté le 20/07/2019, disponible à https://programminghistorian.org/en/lessons/intro-to-linked-data].

[33] ALLEMANG, op. cit. : L'auteur explique très longuement la composition des URI et leurs différences avec les URL.

sur la fiabilité de l'identification. Une URI entière, comme celle de Berners-Lee, ne comporte aucune ambiguïté quant à la provenance de l'entité. Nous savons que l'URI identifie l'informaticien britannique et que cette information est fiable car le fichier d'autorité VIAF fait foi.

F. Le *Resource Description Framework* (RDF) et ses triplets

Avant tout, il semble important de présenter le *Resource Description Framework* (RDF)[34]. Cette ressource est un modèle de données décrivant la façon dont les données doivent être structurées sur le web : c'est le langage clé de voûte du web sémantique. Ce modèle décrit notamment le fonctionnement des triplets et fournit aux utilisateurs un mode d'emploi du web sémantique. L'étape suivant l'attribution d'un identificateur unique à des entités est l'étude de leurs relations mutuelles : l'identification rend les données accessibles et encode les relations qu'elles entretiennent les lie (données ouvertes et liées). Ces relations sont exprimées en utilisant ce qu'on appelle un triplet RDF.

Reprenons l'exemple de notre informaticien britannique favori. Un triplet est composé de trois éléments : un sujet, un prédicat et un objet. Fondamentalement, le prédicat détermine la relation existant entre le sujet et l'objet du triplet, le sujet étant la ressource décrite et l'objet sa valeur. Dans notre exemple, un triplet ressemblerait à ce qu'illustre la Figure 2 ci-dessous. Cela dit, la syntaxe des triplets varie selon la méthode d'encodage utilisée : ils peuvent s'écrire en *Turtle* mais le RDF reste la plus employée[35]. Sur un jeu de données tabulaires, un triplet serait constitué du nom de la ligne (sujet), du nom de la colonne (prédicat) et de la valeur contenue dans la cellule (objet).

[34] POWERS S., *Practical RDF*, Sebastopol, 2003 : le RDF est au web sémantique ce que le HTML est au web traditionnel d'une certaine manière. Il apporte des solutions technologiques aux problèmes résolus par le web sémantique et permet l'échange d'informations.

[35] Pour en apprendre davantage sur la validation des schémas RDF voir : BONEVA I., GAYO J., KONTOKOSTAS D. et al., "*Validating RDF Data*" in *Synthesis Lectures on the Semantic Web: Theory and Technology 7*, 2017.

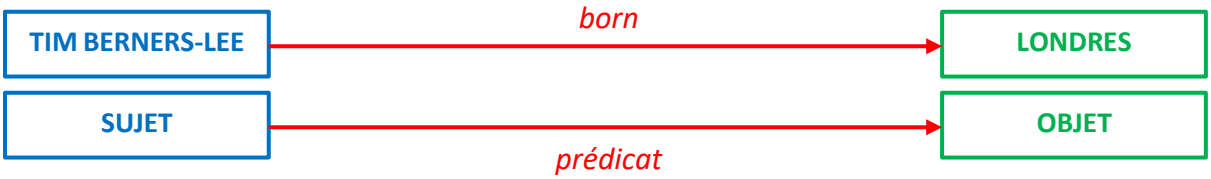

Figure 2 : Exemple d'un triplet (CdA)

En résumé, les données ouvertes liées doivent être accessibles. Pour cela elles ont besoin d'identificateurs uniques décrivant des entités qui entretiennent des relations entre elles représentées par des triplets.

G. Les ontologies essentielles du web sémantique

Ce système devient intéressant lorsque plusieurs triplets font référence à la même entité. C'est d'ailleurs ce à quoi aspirent les données ouvertes liées : faire en sorte que les données communiquent entre elles et se complètent pour enrichir les connaissances[36]. Lorsque l'on crée un nombre conséquent de triplets dont les entités et relations s'entrecroisent, un réseau d'échange d'informations apparaît. Il peut modifier la vision que l'on a de ses données : c'est exactement ce que veulent les DOL. La Figure 3 illustre, par exemple, le réseau de relations existant entre plusieurs personnes qui entretiennent des

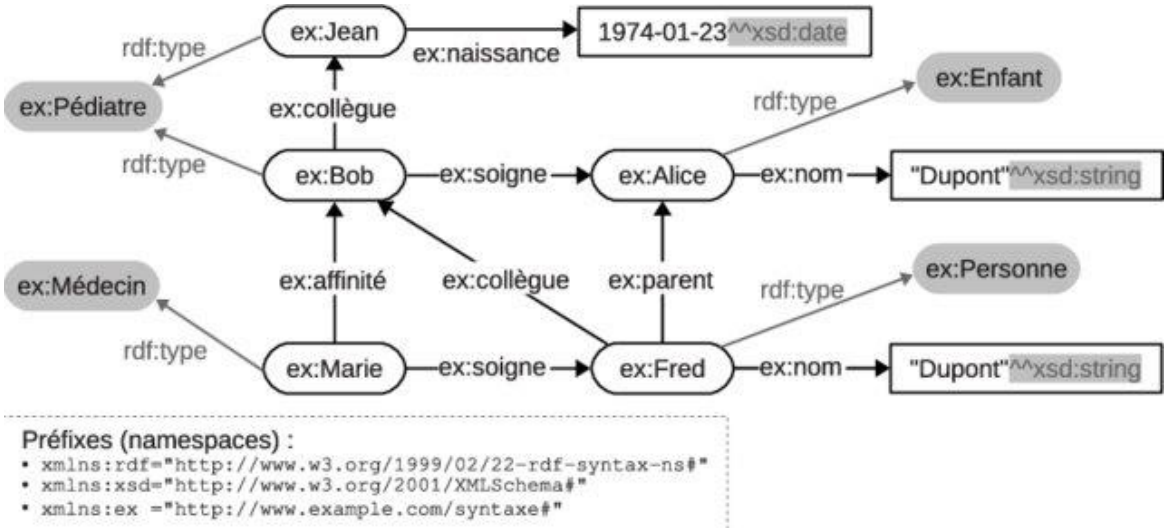

Figure 3 : Document RDF typé selon l'ontologie RDFS en figure 3 et les types XSD (Maillot et al., 2014)

[36] ALLEMANG, op. cit., adresse une partie de son analyse aux données tabulaire, à leur fonctionnement et à leur richesse.

rapports professionnels et leurs affinités. Les préfixes désignent l'ontologie où les chercheurs sont allés piocher des classes pour caractériser ces relations.

Plus tôt, nous avions expliqué que les modèles de représentation des données étaient essentiels à la restructuration du web de documents en web de données. Ces modèles et le vocabulaire associé sont donc des ontologies qui servent à décrire les relations entre les sujets et les objets des triplets. Elles sont essentielles au bon fonctionnement du système : sans elles une machine ne serait pas capable de reconnaître le contenu d'un triplet.

Utiliser une ontologie pour décrire des ressources est un premier pas, reste à faire en sorte que la machine comprenne ce qu'est une ontologie et comment elle fonctionne. C'est pourquoi on a inventé le *Web Ontology Language* (OWL) qui est un langage RDF, développé par le W3C, définissant ce que sont des classes et des propriétés[37]. Il a été construit en s'inspirant du *Resource Description Framework Schema* (RDFS), un langage également conçu par le W3C qui avait aussi pour mission de créer le schéma de base de ce que devait être et comment devait fonctionner une ontologie. Les rouages des ontologies ont finalement été assemblés par le travail conjoint de RDFS puis d'OWL[38].

Il existe pléthore d'ontologies thématiques[39] : plus nous en concevrons, plus nous serons en mesure de décrire le monde qui nous entoure et plus les machines seront capables de traiter les données intelligemment. Nous mentionnerons plus tard dans l'explication de notre projet de recherche quelles ontologies ont été utiles. Pour l'heure nous souhaitons brièvement présenter l'une des ontologies les plus populaires, celle née sous l'impulsion de la *Dublin Core Metadata Initiative* (DCMI). Les bibliothèques et les musées ont joué un rôle fondamental très tôt dans le développement du web sémantique et des ontologies. Traditionnellement disponibles sur le web, les catalogues de ces bibliothèques proposent un

[37] BLACE R., FISHER M., HEBELER J., et al., *Semantic Web Programming, Indianapolis*, 2009 : les auteurs consacrent une grande partie de leur travail à décrire de manière très détaillée toutes les technologies qui fonctionnent sous le web sémantique dont le OWL et ses profils dérivés (OWL-Lite, OWL DL et OWL FULL).

[38] SEGARAN T., EVANS C., TAYLOR J., *Programming the Semantic Web*, Sebastopol, 2009 : Les auteurs explorent comment OWL est devenu l'aboustissement d'un long travail de recherches et comment il s'impose comme le standard officiel du W3c pour les défitions des schémas du web sémantique.

[39] ARP R., SMITH B., SPEAR A., *Building Ontologies with Basic Formal Ontology*, Londres, 2015 : Décrit BFO.

panel d'outils permettant de parcourir leurs collections. Ces collections sont enregistrées dans leurs bases de données où un livre est finalement une entité. Lorsqu'un utilisateur recherche un livre, il peut avoir recours à une série de mots-clés qui selon lui définissent le mieux le contenu dont il a besoin. Une fois effectuée, la recherche lui livre un certain nombre de résultats plus ou moins pertinents selon le type de mots-clés qu'il a utilisés. Ces mots-clés sont finalement des tags appliqués aux livres afin de décrire leur contenu : c'est ce qu'on appelle des métadonnées, littéralement des « données à propos d'autres données ».

Le travail d'indexation des documents est largement facilité par l'existence des métadonnées ainsi que la mise en évidence des relations entre eux. Des métadonnées communes à plusieurs entités forment un réseau de relations. L'ontologie *Dublin Core* propose un modèle de description des métadonnées applicable à des documents textuels mis en ligne sur le web : elle offre une liste de quinze classes destinées à décrire les métadonnées des livres. Autrement dit, elle fournit aux éditeurs une nomenclature pour la description des métadonnées à propos des éléments constitutifs d'un livre tels que : le titre, l'auteur, l'éditeur, le sujet du livre mais aussi la langue originale, son format, etc.

3. Conclusion de la première partie

Lors de ce voyage dans les méandres du web, nous avons exploré les évènements qui ont provoqué sa formation mais aussi ceux qui l'on construit et amélioré au fil des années. Le web sémantique agit en modifiant la structure web traditionnel et autorise de nouvelles avancées scientifiques dans le traitement des données. Les différentes technologies innovantes développées par la communauté des ingénieurs en sciences de l'informatique offrent de nouvelles perspectives d'interprétation des données. Grâce aux outils qui permettent de dépasser les frontières imposées par les logiciels, les chercheurs peuvent dorénavant confronter leurs données avec d'autres ressources provenant de projets différents et élargir leurs perspectives de recherche. Ces liens érigent éventuellement des ponts entre les disciplines : fait nouveau et précieux pour l'avancée de la recherche.

Dans un contexte professionnel, nous avons conduit des recherches sur les possibilités offertes par les outils du web sémantique. En tant qu'assistante de recherches, nos

expérimentations concernent la traduction de fichiers XML/TEI en triplets RDF. Les projets sur lesquels nous avons travaillé s'inscrivent, en effet, dans le climat actuel de la recherche qui évolue, non sans difficultés, vers un monde où un grand nombre de données seraient accessibles sur le web en *open access*, toutes interconnectées et interopérables.

PARTIE 2 : Méthodologie et traitement des données

Maintenant que nous détenons les clés de compréhension nécessaires au traitement des données du web sémantique, il est temps d'entrer dans le vif du sujet technique. Quelles méthodes avons-nous mises en œuvre pour analyser les documents et leurs données ? Dans les grandes lignes, la transformation du fichier XML/TEI en triplets RDF passe par une conversion du premier grâce à l'*eXtensible Stylesheet Language* (XSLT). Nous tenterons donc d'éclaircir, pas à pas, le fonctionnement de cette transformation.

4. Projet de Recherche

Le contenu scientifique alimentant ce mémoire provient d'un projet de recherche et représente l'ensemble du travail réalisé en tant qu'assistante de recherche à l'Université d'Ottawa. Constance Crompton, *Principal Investigator* (PI) et superviseure de ce projet, a donné son accord pour que nous puissions à la fois présenter les données étudiées et les conclusions auxquelles nous avons abouti.

H. *Lesbian and Gay Liberation in Canada* (LGLC)[40] : le projet

Figure 4 : LGLC Logo (CC 3.0)

Apparu à New York dans les années 1970, le premier élan de prise de conscience des inégalités de genre en Amérique du Nord a été baptisé *Gay Liberation Front* (GLF). Ce mouvement s'est construit à l'opposé des politiques assimilationnistes conduites dans les années 1980 : les militants ont rejeté la propagande des familles nucléaires hétérosexuelles et se sont battus pour la libération des mœurs et l'égalité de tous les genres. Les libérationnistes ont voulu se détacher radicalement de toutes les normes culturelles standard. Ils ont été les instigateurs d'une nouvelle culture visuelle, haute en couleurs, ont œuvré pour faire améliorer l'ouverture d'esprit de la communauté et ont même réussi à influencer des

[40] *Lesbian and Gay Liberation in Canada* (LGLC) [en ligne, consulté le 19/03/2019, disponible à https://lglc.ca/].

réformes politiques. Prises dans ce tourbillon de changement, les femmes ont aussi pris part à cette lutte contre les inégalités et l'on a vu, aux côtés des mouvements gays, des mouvements féministes se créer.

Le projet LGLC s'abreuve des progrès de ces mouvements au fil des années, de leurs réussites et défaites politiques et de leur diversité. Les évènements canadiens[41] ont été passés au crible par Don McLeod dans deux ouvrages. Le premier[42] est une chronologie annotée des douze premières années de contestation du mouvement GLF au Canada, dont le premier jalon marque la fondation de l'*Association for Social Knowledge* (ASK) à Vancouver[43]. Son second livre couvre les évènements survenus à partir de 1976 et jusqu'en 1981 où les premières études sur le Syndrome d'Immunodéficience Acquise (SIDA) ont commencé à être publiées. Avec leurs équipes de recherche, Constance Crompton (Université d'Ottawa) et Michelle Schwartz (Université de Ryerson) ont converti les deux livres en une base de données qui permet aux utilisateurs de parcourir toute l'histoire du mouvement de libération gay au Canada[44]. Le projet LGLC a donc pour objectif de proposer une ressource digitale interactive pour l'étude de l'Histoire qui concerne les lesbiennes, les gay, les bisexuels et les transgenres (LGBT) au Canada. Le texte numérisé est continuellement mis à jour et enrichi de nouvelles données à partir des trente-quatre mille enregistrements de personnes, des lieux, des organisations et des articles de presse déjà recensés par McLeod. Cet enrichissement se traduit, par exemple, par l'ajout de données biographiques pour chaque personne présente dans la base de données.

Le projet agit en récipiendaire d'une Histoire souvent peu documentée ou mal interprétée. A long terme, LGLC souhaite substantiellement augmenter la quantité

[41] MCLEOD D., *Lesbian and gay liberation in Canada: a selected annotated chronology, 1964-1975*, Toronto, 1996 puis MCLEOD D., *Lesbian and gay liberation in Canada: a selected annotated chronology*, 1976–1981, Toronto, 2016 : catalogues des manifestations, des actions politiques, des réformes légales et de diverses actions menées contre les lobbys, de tous les évènements liés au mouvement GLF au Canada.

[42] MCLEOD, op. cit.

[43] ASK est la première organisation canadienne soutenant les mouvements de libération gay au Canada.

[44] Base de données LGLC accessible ici : https://lglc.ca/search

d'informations accessibles à propos de ces groupes persécutés pour leurs idées, afin de la transmettre aux générations futures.

I. *Lesbian and Gay Liberation in Canada* : les données

Les données avec lesquelles nous avons travaillé sont donc toutes les entités décrites par McLeod dans ses deux volumes de chronologie : des évènements, des lieux, des organisations, des manifestations et des périodiques. Les entités sont encodées dans cinq fichiers différents correspondant au quatre types étudiés pour des raisons logiques : un évènement ne peut pas être classifié de la même manière qu'une organisation et la description d'un lieu sera forcément différente de celle d'un périodique.

Nous avons donc pu travailler directement sur ces fichiers qui sont disponibles en suivant ces cibles[45] :

- demonstrations.xml : liste complète des manifestations conduites entre 1965 et 1981,
- organizations.xml : index des organisations impliquées dans les évènements,
- periodicals.xml : bibliographie de la production médiatique générée par le mouvement,
- places.xml : répertoire de tous les endroits hébergeant des actions du mouvement,
- lglc.xml : fichier le plus complet, c'est un condensé chronologique des quatre précédents, davantage détaillé, où une entrée correspond à un évènement regroupant tous les acteurs impliqués dans celui-ci.

Après avoir parcouru ces fichiers, nous pouvons nous demander pourquoi l'équipe du projet a souhaité encoder ses documents en TEI et les enregistrer au format XML. En réalité LGLC souhaite utiliser les technologies offertes par le web sémantique pour traiter ses données, afin de faciliter leur accessibilité et surtout de rendre ces données ouvertes et liées. De cette manière, LGLC peut profiter des bénéfices du web sémantique et faire de ses entités des triplets identifiés par des URI ouvrant la porte à des connexions potentielles avec d'autres projets. Nous sommes donc partis de ces fichiers textuels notés entre 3 et 4 étoiles, pour en

[45] Tous les liens fonctionnent et ont été testés sur différentes machines pour la dernière fois le 30/08/2019. Si un problème technique devait se présenter lors du téléchargement, merci d'envoyer un email à alice.defours@gmail.com pour obtenir une solution technique rapide.

faire des données ouvertes liées véritables, méritant les 5 étoiles. Cette transformation, dans le cas de documents textuels, passe par la sérialisation XML/RDF. Dans la suite du développement, nous nous appliquerons à expliquer le fonctionnement des technologies du web sémantique appliquées à ce projet et de quelle manière elles lui sont utiles.

La qualité des données n'a pas, ici, été un problème. Hormis quelques erreurs d'encodage du document TEI, le contenu n'a pas besoin d'être remis en question puisqu'il provient de travaux publiés et légitimes[46]. Personnellement, la question du stockage des données ne s'est pas posée puisque nous avons simplement travaillé à la conversion des fichiers du format XML/TEI vers le format XML/RDF.

Le dernier point important à aborder vis-à-vis de ces données est leur confidentialité. Les données personnelles sont délicates à gérer lorsqu'elles concernent des individus puisque l'utilisation que l'on en fait pourrait affecter soit ladite personne, soit ses proches. Le 14 Mai 2019, le Dr Marie-Hélène Roy-Gagnon a donné une conférence à propos des outils statistiques et bio informatiques pour l'infrastructure multisectorielle i-BALSAC visant à créer une cartographie de la population franco-canadienne destinée à la recherche en sciences biologiques et sociales. Ce projet repose en partie sur l'intégration de données généalogiques, génétiques et cartographiques à un outil interactif qui pourra être utilisé pour « identifier des variations génétiques associées aux maladies complexes et la mise en place de stratégies de traitement, de dépistage et de prévention »[47]. Ce projet crée donc des connexions entre trois types de données (génétiques, généalogiques et cartographiques) : objectif que l'on ne peut que souhaiter car il permettra de se poser de nouvelles questions et connectera ces trois ensembles de façon inédite. Ici les données sont d'une sensibilité extrême (surtout celles génétiques et généalogiques) et l'équipe fait face à un véritable défi de gestion, explorant plus avant les possibilités de protéger les individus concernés par ces données.

[46] MCLEOD, op. cit.

[47] Dr Marie-Hélène ROY-GAGNON (Faculté de médecine de l'Université d'Ottawa) lors d'une conférence donnée dans le cadre du *Digital Humanities Institute : Technologies East 2019* (DHSITE2019), le 14 Mai 2019.

La gestion de ces données est moins compliquée pour le projet LGLC. En effet, la plupart des informations émanent de sources publiques (articles de journaux, etc.) et ne nécessitent pas d'autorisation éthique particulière pour être publiées. Cela dit, ces données sont mises en ligne d'une façon totalement inédite, même si elles sont issues du domaine public. En dehors de la confidentialité, les chercheurs se doivent de raconter fidèlement l'Histoire des mouvements de libération LGBT, sans la déformer[48]. Cela dit, l'équipe se protège elle-même ainsi que sa communauté en proposant un formulaire de rétractation disponible sur le site web, au cas où l'une des parties mentionnées dans les données souhaiterait faire retirer les informations mises en ligne qui la concernent.

J. Projets connexes : présentation et données

Lors de notre investigation, nous avons également eu recours à un autre jeu de données sur lequel nous avons testé d'intéressants mécanismes qui méritent d'être exposés dans ce mémoire. Ce jeu n'est pas la propriété du laboratoire de l'Université d'Ottawa mais de l'un de ses partenaires de recherche qui approuve leur partage dans le respect des conditions fixée par les *Creative Commons*[49] : il s'agit des données du *Digital Dinah Craik* (DDC) appartenant à une équipe de recherche de l'Université de Calgary, dirigée par Karen Bourrier (*Assistant Professor*, Université de Calagary)[50].

Dinah Mulock Craik, une écrivaine britannique, a rédigé plus de mille lettres manuscrites pendant sa longue carrière de quarante ans. Le projet *Digital Dinah Craik*, à l'image du *Lesbian and Gay Lberation in Canada*, s'applique à conserver la trace de ses correspondances en numérisant toutes les lettres, en analysant leur contenu, en l'enrichissant et souhaite aussi en faire une ressource de données ouvertes et liées. L'équipe a un but noble. En décomposant ses relations avec ses correspondants, ils tentent de comprendre comment

[48] CROMPTON C., SCHWARTZ M., "*Lesbian and Gay Liberation in Canada: Representing the Dyke Dynamo*" in Digital Studies/Le champ numériques, Vol. 5, n°1, 2016. Dans cet ouvrage, les auteures soulignent l'importance de raconter une histoire qui respecte le plus fidèlement la vie des communautés LGBT.

[49] Les *Creative Commons* protègent leur contenu licencié et partagé sur le web : il peut être copié, distribué et réutilisé dans le respect des lois sur le droit d'auteur.

[50] *Digital Dinah Craik* [en ligne, consulté le 19/03/2019, disponible à http://www.digitaldinahcraikproject.org/].

l'auteure a construit un réseau professionnel qui a contribué à son succès pendant si longtemps. A long terme, ils souhaitent brosser un portrait des écrivaines du XIXème siècle.

Cette prosopographie a été transposée au sein d'un fichier en TEI : le *Craik Site Index*. Celui-ci nous a servi de cobaye lors de nos tests de conversion d'un fichier XML/TEI vers des triplets RDF notamment car il contient des éléments intéressants à transformer : les relations entre les personnes. Nous reviendrons bientôt sur cette activité qui s'est révélée utile et a amélioré notre compréhension des langages de description. Nous pouvons dire avec confiance que ces données sont fiables puisqu'elles émanent d'archives officielles et du *Oxford Dictionary of National Biography*.

5. Passage du format TEI au format XML/RDF

Pour expliquer notre méthode de travail, commençons par aborder la conversion en elle-même : comment passer du format TEI au format RDF. Nos explications veilleront à rester abordables pour que notre propos soit à la fois compréhensible par des néophytes et intéressant pour les spécialistes.

K. L'eXtensible Stylesheet Language (XSLT)

Le XSLT est un langage servant à transformer des fichiers nativement en XML[51]. Le format cible peut être très varié : nous pouvons convertir un fichier XML en du *Hypertext Markup Language* (HTML), du *Extensible Hypertext Markup Language* (XHTML) ou du texte brut par exemple[52]. Cette opération nécessite d'abord la création d'un document vierge où l'on écrira des instructions en XSLT. Pendant notre étude, nous avons utilisé le logiciel *Oxygen® XML Editor* pour accomplir la conversion. Cet environnement de travail est dédié à la manipulation des formats XML. Il possède toutes les fonctionnalités utiles à nos transformations, y compris un outil de construction intégré en XSLT.

[51] FITZGERALD M., *Learning XSLT: A Hands-On Introduction to XSLT and XPath*, Beijing, 2003.

[52] TIDWELL D., *XSLT : 2nd Edition*, Sebastopol, 2009.

La première étape à accomplir avant toute chose est d'associer à un document TEI, la feuille de style en XSLT qui servira d'opérateur de transformation. Dès lors que l'utilisateur lancera une transformation, le document TEI transitera par la feuille de style à laquelle il est associé. Cette dernière exécutera les instructions en XSLT qui auront été écrites à l'intérieur et le résultat de cette opération sera un fichier de sortie dans un format décidé par l'utilisateur. Dans notre cas, le fichier de sortie portera l'extension .rdf puisque ce sont des triplets dans un format interopérable que nous souhaitons obtenir.

La structure d'une feuille de style est toujours la même peu importe le format cible et commence toujours par l'élément **<xsl:stylesheet>** que l'on ferme en pied de page par la balise suivante : **</xsl:stylesheet>**. Une feuille de style en XSLT fonctionne de manière hiérarchique, certains éléments ne peuvent exister que s'ils sont contenus dans un autre élément. C'est le cas ici : **<xsl:stylesheet>** dit à la machine qu'elle va devoir traiter une feuille de style. Celle-ci pourra contenir des instructions, matérialisées par d'autres éléments. Afin que le système comprenne la linguistique utilisée. Les éléments doivent s'accompagner de la description de leur nom de domaine ou *namespaces*. Sans cela, la machine sera incapable de localiser le lexique du langage sur le web et ne pourra pas le comprendre. L'en-tête d'une feuille de style agit comme son professeur : elle lui apprend tout ce dont elle doit savoir pour poursuivre son chemin dans le fichier. Les noms de domaines sont définis dans des attributs apposés aux éléments. Pour une feuille standard, la description prendrait cette forme :

<xsl:stylesheet xmlns:xsl=http://www.w3.org/1999/XSL/Transform>
</xsl:stylesheet>

Ici, nous voyons que l'attribut **@xmlns** décrit le préfixe « xsl » comme étant une ressource localisable à l'URL qui suit. L'acronyme « xmlns » signifie tout simplement *eXtensible Markup Language Namespace* soit « nom de domaine XML »[53]. Notre processeur de documents est maintenant capable de comprendre que nous allons lui demander d'effectuer une transformation utilisant des mécanismes exprimés en XSLT. L'en-tête est donc un endroit pratique où l'on définit la totalité des ressources linguistiques utilisées dans la transformation.

[53] Rappelons-nous que le XSLT est un langage de transformation XML.

Pour nous, qui allons aussi utiliser des ontologies afin de créer nos triplets, c'est l'endroit idéal pour les déclarer afin que la machine apprenne le langage et soit capable de le réutiliser dans les instructions qui suivront.

Le deuxième élément obligatoire que doit contenir une feuille de style est **<xsl:template>**. Celui-ci a toujours un attribut **@match** égal à un nœud. Le premier *template* doit toujours s'appliquer à la racine du document en TEI exprimée par le caractère « / ». Nous aborderons plus tard l'ordre d'application des modèles mais pour l'heure contentons-nous de dire que le premier modèle doit obligatoirement correspondre à la racine « / ». Les *template* suivants pourront eux s'appliquer à d'autres nœuds du document TEI[54].

L'instruction **<xsl:template>** commande au processeur d'appliquer un modèle (''*template*'') à du contenu. Nous avons ici deux termes à expliciter : modèle et nœud. En XSLT, une *template rule* (« règle de modèle ») consiste à appliquer une ou plusieurs instructions à du contenu que l'on appellera contexte. Celui-ci n'est ni plus ni moins que les données textuelles du fichier TEI que l'on va aller attraper grâce à nos nœuds. Qu'est-ce qu'un nœud ? Un document XML pourrait être dessiné comme un arbre où chaque nœud serait l'endroit exact où deux branches se séparent. Les nœuds peuvent donc être des éléments, des attributs ou du texte pour restreindre la liste aux plus importants. Récupérer ces nœuds pour les intégrer aux feuilles de style requiert l'utilisation d'un outil connexe au XSLT, le *XML Path Language* (XPATH).

XPATH est un langage employé pour naviguer dans un fichier XML et en récupérer des portions[55]. Il contient même des fonctions intégrées capables d'exécuter des requêtes en XQUERY (booléens, etc.). A l'image de la plupart des technologies du web sémantique, XPATH

[54] La *Text Encoding Initiative* définit les critères d'encodage des documents textuels sur le web. Notre sujet de mémoire portant sur les DOL, nous avons choisi de ne pas s'attarder sur la TEI. Pour plus d'informations, voir : https://teibyexample.org/

[55] Connaissances acquises lors de la participation à l'enseignement « XPATH for processing XML and managing project » dispensé par Elisa BESHERO-BONDAR et David J. BIRNBAUM lors du programme annuel de formation en Humanités Numériques organisé par l'Université de Victoria (Vancouver, BC) : le *Digital Humanities Summer Institute* (DHSI). Le contenu du cours est librement accessible en ligne en suivant ce lien.

comprend et fonctionne par relations. On dit qu'un nœud peut avoir des enfants, des parents, des ancêtres et des descendants. Prenons cet exemple de *snippet* :

```
<div type="poem">
        <head>Invictus</head>
            <lg>
                <l>Out of the night that covers me, </l>
                <l>Black as the Pit from pole to pole,</l>
                <l>I thank whatever gods may be</l>
                <l>For my unconquerable soul. </l>
            </lg>
</div>
```

Dans cet extrait, nous avons encodé la strophe d'un poème[56] en suivant les standards d'encodage TEI d'un document textuel. Si l'on réutilise notre vocabulaire XPATH, ** ** est l'élément racine « / », **<head>** est un nœud tout comme **<lg>** et **<l>**. Ce qui détermine la relation entre ces éléments est le contexte[57]. Dans le cas où notre contexte serait **<head>**, son parent serait ** **, **<lg>** serait son frère, et **<l>** seraient des enfants de **<lg>**. Avec une notation en XPATH, naviguer en partant du contexte **<head>** jusqu'à la cible **<l>[1]**[58] reviendrait à écrire ceci : **parent::div/lg[1]/l[1]**. Cette expression XPATH illustre typiquement le genre de formulation que nous allons utiliser dans notre transformation en XSLT pour sélectionner des portions de TEI sur lesquelles viendront **@match** des **<xsl:template>**.

<xsl:template match=''/''> applique donc un modèle à l'ensemble des nœuds contenus dans la racine du document, c'est-à-dire qu'il attrape la totalité de celui-ci. Dans notre exemple, le *template* s'applique à ** ** et attrape donc tout ce que l'on peut trouver à l'intérieur (**<head>, <lg>** et **<l>**). Il est ensuite d'usage d'utiliser **l'élément <xsl:apply-templates select=" "/>,** avant de fermer le premier **</xsl:template>,** afin d'indiquer au processeur de continuer à travailler. Sans la mention de cette instruction, il se contentera d'appliquer un modèle vide à la totalité du contenu en TEI. L'attribut **@select**, comme sa désignation l'indique, dirigera la machine vers le contexte sujet à la transformation. En

[56] HENLEY W., *Invictus, Book of Verses*, Londres, 1888.

[57] Le contexte est défini par une instruction en XPATH qui cible un nœud. Ce nœud devient alors un contexte.

[58] Premier élément **<l>**. Ici nous en avons quatre qui seront respectivement notés **<l>[1], <l>[2], <l>[3] et <l>[4].**

d'autres mots, cet attribut doit contenir un nœud, visant ce contexte, qui sera l'objet de la transformation en XSLT.

Pour transformer le contexte sélectionné, nous allons avoir besoin d'écrire un second **<xsl:template>** qui lui correspondra (**@match**). Si l'on souhaitait sélectionner les vers du poème *Invictus* comme contexte, il faudrait donc écrire :

<xsl:template match="l">

Les instructions contenues à l'intérieur du modèle s'appliqueront donc à tous les éléments **<l>** existant dans le document. Nous aurons bientôt l'occasion d'étudier en détails la syntaxe de ces instructions. Toutefois, dans un souci constant de vouloir éclaircir notre propos, voici un exemple complet de ce à quoi ressemblerait une feuille de style pour transformer les vers de poème en texte brut :

Feuille de style

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
                xmlns:xs="http://www.w3.org/2001/XMLSchema"
                exclude-result-prefixes="xs"
                version="2.0">

        <xsl:template match="/">
            <xsl:apply-templates select="l"/>
        </xsl:template>

        <xsl:template match="l">
            <xsl:value-of select="."/>
        </xsl:template>
</xsl:stylesheet>
```

Sortie

Out of the night that covers me,
Black as the Pit from pole to pole,
I thank whatever gods may be
For my unconquerable soul.

Nous venons d'aborder les principes d'une conversion d'un fichier du format TEI au format RDF. Dans le cadre de notre étude, reste à en évoquer l'intérêt très spécifique qui a servi notre projet.

L. Etude de cas 1 : Conversion des données LGLC

Avant de poursuivre la lecture de ce développement, nous invitons le lecteur à se munir de ce fichier et de faire des allers-retours entre notre explication et la feuille de style pour mieux comprendre les mécanismes de transformation. Notre plus grande tâche pour faire avancer le projet LGLC consistait à transformer les informations des fichiers XML/TEI, susmentionnés, en triplets RDF, à l'aide de la feuille de style présentée ci-dessus. Nous avons essentiellement travaillé avec le fichier lglc.xml car c'est celui qui rassemble la totalité des informations pertinentes à l'étude et à la création de triplets. Ce fichier est immense. Voici la preuve avec cet aperçu de son arbre (accessible ici en taille réelle) :

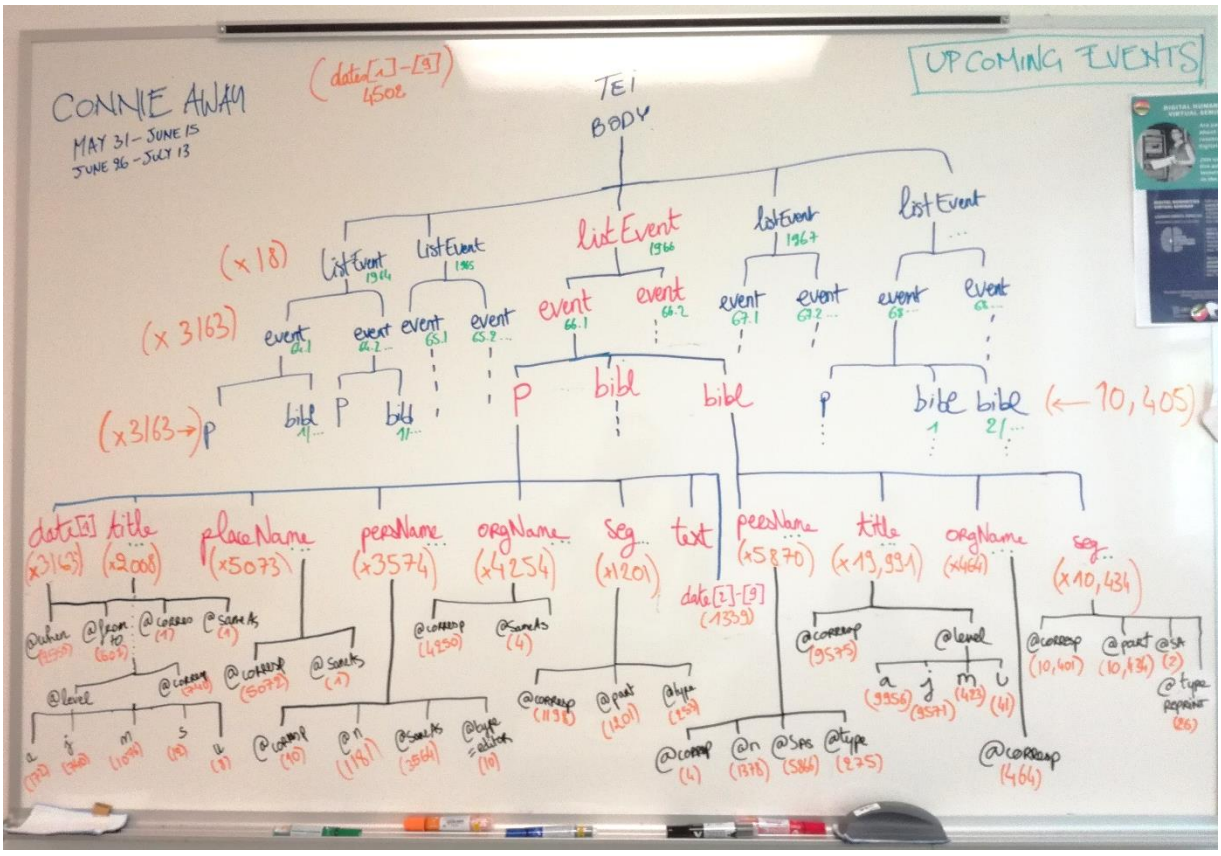

Figure 5 : Arbre représentant la structure du document XML/TEI lglc.xml (CdA)

Nous avons dû faire des choix éditoriaux pour sélectionner les entités que nous souhaitions transformer en triplets. Il a été décidé que chaque évènement de la liste serait un sujet (la ressource décrite par le triplet). En observant bien l'encodage des éléments dans lglc.xml, nous sommes capables d'anticiper la localisation des différentes ressources formant nos triplets. En effet, si les éléments **<event>** sont nos sujets, les informations qu'ils contiennent (dates, noms de lieux, d'organisation) en sont les objets et la relation entre les deux sera déterminée par l'ontologie utilisée.

Notre feuille de style peut être expliquée comme suit. Au tout début, nous retrouvons la déclaration de tous les préfixes qui seront utilisés dans la transformation : ils désignent des ressources qui sont toutes des ontologies que nous utiliserons pour former nos triplets et que nous présenterons un peu plus tard. Comme l'exige le fonctionnement d'une feuille de style, nous devons ensuite appliquer un modèle à la racine de lglc.xml qui va sélectionner la totalité de son contenu dans l'attente d'instructions. C'est ici que notre chemin dans le *Resource Description Framework* débute. Dans ce premier modèle, nous créons un élément **<rdf:RDF>**. Celui-ci sera l'en-tête de notre fichier de sortie et contiendra tous nos triplets puisque nous spécifions clairement au modèle d'appliquer toutes les instructions qui vont suivre à cet endroit à l'aide d'un **<xsl:apply-templates select="descendant::listEvent"/>**. En effet, l'expression XPATH précédente peut se traduire par « applique les modèles suivants à tous les descendants des éléments **<listEvent>** que tu rencontreras et à rien d'autre puis attends que je te donne d'autres instructions ».

Une fois ces dispositions liminaires prises, nous pouvons commencer la création des triplets. A l'aide d'un autre modèle, nous allons récupérer le sujet de notre triplet **<event>**. Celui-ci est caractérisé par un identificateur au sein du projet LGLC : son URI. L'instruction suivante crée donc un triplet liant le sujet à son identification :

```
<xsl:template match="event">
    <schema:Event rdf:about="{concat('http://lglc.ca/lglc.rdf#',@xml:id)}">
    ...
    </schema:Event>
</xsl:template>
```

Décomposons cette instruction :

1) Le processeur applique un modèle au contenu de l'élément **<event>**,
2) Nous lui demandons ensuite de créer un triplet utilisant la sérialisation XML/RDF. Pour ce faire, nous écrivons <schema:Event...> qui exprime, dans cette sérialisation, un sujet.
3) La relation entre le sujet et son objet se décrit en utilisant des ontologies. Dans ce cas-ci, c'est l'ontologie RDF et sa classe « about » que nous utilisons. Nous devons donc convoquer le préfixe « *rdf* :… » la machine comprend alors que l'attribut **@about** appartient à ce langage et que son utilisation est autorisée à cet endroit, dans cette syntaxe.
4) La valeur de l'attribut constitue ensuite l'objet de notre triplet. Dans cet exemple, nous souhaitons identifier chaque évènement. L'équipe s'est chargée, dans le fichier XML/TEI, de donner à chaque entité un **@xml:id** unique qui est son numéro d'identité. Pour former l'URI, nous avons donc besoin d'écrire du texte brut. Notre instruction se charge ensuite de concaténer ce texte avec la valeur de l'**@xml:id** propre à chaque **<event>** sur lequel le modèle se déroule.
5) Voici le résultat, un triplet exprimé en XML/RDF :
   <schema:Event rdf:about="http://lglc.ca/lglc.rdf#n64.1">

Cette balise n'est pas fermée. En effet, notre triplet contiendra finalement un lot de triplets au sujet principal commun : **<event>**. Pour faire en sorte que la transformation créée des triplets pour tous les éléments pertinents contenus dans **<event>**, il faut lui signifier d'appliquer tous les modèles suivants à l'intérieur du modèle actuel, celui qui **@match** sur **<event>** :

```
<xsl:template match="event">
        <schema:Event rdf:about="{concat('http://lglc.ca/lglc.rdf#',@xml:id)}">
                <xsl:apply-templates select="p" mode="main"/>
                <xsl:apply-templates select="bibl/seg[@corresp]" mode="citationSeg"/>
                        <xsl:apply-templates select="p" mode="markup"/>
        </schema:Event>
</xsl:template>
```

L'utilisation des **<xsl:apply-templates>** force alors le processeur à se diriger vers le prochain contexte et à suivre les instructions qui s'y trouvent, jusqu'à ce qu'il atteigne un point où il n'aura plus d'instructions à exécuter et s'arrêtera. En déroulant ainsi notre fichier de transformation, la machine va pouvoir créer un répertoire de triplets à propos des évènements.

Notre expérimentation permet l'utilisation de fonctions avancées pour interroger les données. Ces interrogations fonctionnent comme des algorithmes : si un attribut possède une valeur que nous désirons récupérer, le système déclenche une suite d'opérations destinées à le faire. Sinon, il ne fait rien. Ce genre de tour de magie est très utile lorsque certains sujets ont des objets que d'autres sujets n'ont pas : le programme traite automatiquement toutes les occurrences présentes et ignore les cas où elles sont absentes.

Nous avons utilisé ces fonctions à plusieurs reprises dans notre transformation de façon à créer une bibliographie, sous forme de triplets, pour les évènements lorsqu'ils avaient suscité l'intérêt des auteurs et alimenté des articles de presse, de périodiques, etc. Dans le fichier TEI, ces ressources, lorsqu'elles existent, sont encodées à l'aide de la balise **<bibl>**. Nous voulions faire en sorte que, lorsque la transformation convertit chaque lot de triplets avec un sujet commun, elles convertissent également la bibliographie associée. Pour réussir, sans perturber le déroulement des modèles créant les triplets des **<event>**, nous avons dû appliquer des modèles séparément. Cela se joue au niveau du deuxième et du troisième modèle.

```
<xsl:template match="listEvent">
        <xsl:apply-templates select="event"/>
        <xsl:apply-templates select="descendant::bibl"/>
</xsl:template>
```

Lorsque nous appliquons le deuxième modèle sur « listEvent », nous dirigeons le programme vers deux autres modèles qui vont exécuter deux commandes différentes mais qui placeront les informations dans le fichier cible au même endroit, si elles concernent le même sujet. Le premier **<xsl:apply-templates>** exécute donc les instructions relatives à la création des triplets, ce que nous avons expliqué ci-dessus. Le deuxième **<xsl:apply-**

templates> va aller chercher le modèle suivant et appliquer les instructions qu'il contient. La particularité de ce contexte est qu'on doit dire au programme de regarder parmi les descendants de **<listEvent>** pour trouver tous les **<bibl>** qui constitueront notre contexte. Ceci nous amène au modèle suivant :

```
<xsl:template match="bibl">
        <xsl:variable name="segCorresp">
                <xsl:for-each select="seg">
                        <xsl:value-of select="@corresp"/>
                </xsl:for-each>
        </xsl:variable>
        <bf:Instance rdf:about="{concat('http://lglc.ca/lglc.rdf', $segCorresp)}">
                <xsl:apply-templates select="title" mode="titleBF"/>
                <xsl:apply-templates select="persName" mode="persNameBF"/>
        </bf:Instance>
</xsl:template>
```

En français, ces instructions signifient :

1) Fabrique une variable qui aura pour nom « *segCorresp* ». Pour chaque élément **<seg>** que tu trouves, sélectionne son attribut **@corresp** et stocke-le dans ma variable.
2) Crée ensuite un triplet à partir de l'ontologie BIBFRAME[59], l'URI de notre segment, qui est réellement la ressource bibliographique portant un **@xml:id** forgé par LGLC.
3) Sache que ce triplet a pour sujet une ressource bibliographique ayant un titre et un auteur. Applique donc, au sein de ce triplet principal, le modèle « *title* » dont le mode est « *titleBF* » et ensuite le modèle « *persName* » dont le mode est « *persNameBF* ».

Pour chaque **<event>** qui possèdera des ressources textuelles publiées et encodées par **<bibl>** et **<seg>**, le programme concevra une bibliographie qui ressemblera à ceci :

```
<bf:Instance rdf:about="http://lglc.ca/lglc.rdf#n71.1.1">
        <bf:title>The Male Homosexual in Literature: A Bibliography</bf:title>
        <bf:Contributor rdf:resource="http://lglc.ca/people.rdf#IYOU"/>
</bf:Instance>
```

[59] Pour une étude sur l'importance des métadonnées, BIBFRAME et le rôle des librairies dans l'avancée du Web sémantique, voir : HARDESTY J., "*Transitioning from XML to RDF considerations for effective moves towards Linked data and the Semantic Web*" in *Information Technology and Libraries* 35, Avril 2016, pp. 51-64 et BECKER B., "*Web of Data: A Primer of Linked Data and the Semantic Web*" in *Behavioral & Social Sciences Librarian 35*, 2016, pp. 42-45.

Cette bibliographie se trouve, dans le résultat, à la suite de tous les triplets sur les évènements. Les listes d'évènements **<listEvent>** étant classées par ordre chronologique, nos triplets ne seront pas « recrachés » dans le désordre par le programme. Ils apparaîtront, ainsi que la bibliographie qu'ils ont alimentée, par groupes datés de la même année. C'est pour cela que nous verrons des jeux de triplets alterner entre une liste de **<schema:Event>** et une liste de **<bf:Instances>**.

En fonctionnant de cette manière, la bibliographie est désolidarisée du contexte de l'évènement auquel elle renvoie (le modèle d'applique sur **<listEvent>**). Nous avons quand même voulu que nos jeux de triplets formés sur le modèle **<event>** contiennent aussi un renvoi à ces références. Jetons un œil à la structure du modèle **<event>** à nouveau :

```
<xsl:template match="event">
        <schema:Event rdf:about="{concat('http://lglc.ca/lglc.rdf#',@xml:id)}">
                <xsl:apply-templates select="p" mode="main"/>
                <xsl:apply-templates select="bibl/seg[@corresp]" mode="citationSeg"/>
                <xsl:apply-templates select="p" mode="markup"/>
        </schema:Event>
</xsl:template>
```

Après avoir créé le triplet principal formé sur l'URI de l'évènement, nous demandons au programme d'appliquer trois modèles caractérisés par des modes. Brièvement, les modes servent à désambiguïser l'encodage si l'on fait appel aux mêmes nœuds plusieurs fois, de sorte que le programme n'écrase pas son travail au fur et à mesure de sa progression. Prenons l'exemple du deuxième sous-modèle : il s'applique aux segments qui ont un attribut **@corresp**.

```
<xsl:template match="seg[@corresp]" mode="citationSeg">
    <bf:citationUri rdf:resource="{concat('http://lglc.ca/lglc.rdf', @corresp)}"/>
</xsl:template>
```

Grâce à celui-ci, nous pouvons intégrer au sein des jeux de triplets, les références bibliographiques comme montré ci-dessus. Le processeur va, en effet, attraper la valeur de l'attribut du segment et forge une URI à partir de celui-ci. L'ensemble décrit une association

sujet : référence bibliographique, prédicat : ressource et objet URI. Nous avons lié notre bibliographie à l'évènement auquel elle correspond !

Le dernier sous-modèle d'**<event>** s'applique à tous les modèles restants qui n'ont pas déjà été exécutés grâce à l'élément **<xsl:apply-templates/>** qui ne cible pas de nœud particulier en raison de l'absence d'un attribut **@select**. Les dates, organisations et toutes les données pertinentes à partir desquelles nous avons défini des instructions seront donc convertis.

M. Etude de cas 2 : Conversion des données DDC

La conversion des données du projet LGLC est l'ouverture d'un nouveau chapitre le connectant un peu plus au web sémantique. Ce travail est le résultat d'une expérimentation longue et riche en rebondissements qui passe par la mise au point de tentatives d'encodage que nous avons appliquées à des données différentes. C'est l'index du projet *Digial Dinah Craik* qui a servi de cobaye à nos essais. Afin de conserver un propos clair, nous n'allons pas détailler la totalité des expériences que nous avons menées. Nous souhaitons simplement en présenter une ayant abouti à des résultats très intéressants. Les fichiers que nous avons utilisés sont disponibles ci-dessous :

- CraikSiteIndex_2017-03-23.xml : document TEI contenant les données,
- CraikSiteIndex_Stylesheet.xsl : feuille de transformation en XSLT[60].

L'index *Craik* se présente sous la forme d'un catalogue de toutes les entités citées dans les lettres de l'auteure victorienne. Nous avons décidé de n'utiliser que la prosopographie de la division n°4, « *Historical People* »[61], pour tester notre encodage avant de l'appliquer aux documents de LGLC. Cette division liste des personnes les unes à la suite des autres : chacune est identifiée par un **@xml:id** et enveloppe une série de données la concernant (titres, noms, date de naissance, de décès, biographie, etc.). La plupart des protagonistes de cette liste ont

[60] Si la feuille de style est le résultat de notre travail personnel, l'index en XML appartient et est maintenu par le projet DDC qui accepte avec bienveillance de partager ses données dans le respect des droits d'auteurs et dans un contexte de recherche académique.

[61] Que nous abrègerons DIV4.

également leurs nationalités et leurs professions documentées. Ces éléments se sont révélés retords à transformer mais le résultat, convaincant, est une belle performance d'encodage dont voici le détail.

Chaque tag **<nationality>** ou **<occupation>** décrit une caractéristique à propos de la personne qu'il concerne. Les auteurs du document TEI ont décidé que l'information contenue dans ces balises serait exprimée sous la forme de texte brut. Pour les transformer en données ouvertes liées, ces nationalités et professions, qui sont des entités, doivent posséder une URI. Le premier travail a donc été de créer une liste de toutes les professions ainsi que de toutes les nationalités mentionnées dans la DIV4, à l'aide de requêtes en XPATH prenant, par exemple, cette forme : **distinct-values(//div[4]//nationality)**. Cette requête nous renvoie une liste de toutes les informations contenues dans les balises **<nationality>** en omettant les doublons : au total, nous devrons donc transformer treize nationalités différentes[62] et trente-six professions[63].

Dans un deuxième temps, nous sommes allés chercher les URI de ces professions sur Wikidata[64]. Ressource extrêmement pertinente, Wikidata est un outil de référence lorsque l'on travaille avec des données ouvertes liées. Nous en avons d'ailleurs fait usage quotidiennement. Munis de tous nos outils, nous n'avions donc plus qu'à exécuter la transformation.

La conversion des nationalités et des occupations fait appel à une nouveauté récente du XSLT : les cartes (« *maps* »). Conçues comme des tables de stockage, les *maps* sont un outil de construction de séquences, c'est-à-dire qu'elles permettent d'associer deux éléments ensemble : une clé (« *key* ») et une valeur. La clé définit une donnée qui, lorsqu'on fera appel

[62] La requête renvoie quatorze résultats : en effet, l'encodage du document n'est pas uniforme. Parfois, l'absence d'information sur la nationalité ou l'occupation est encodée à l'aide d'une **<nationality/>** ou rien du tout. La requête n'est pas fausse, elle compte également ces balises auto-fermantes. Lorsqu'une information était manquante, il aurait sans doute fallu ne rien mettre plutôt que d'écrire une balise vide : ce type de coquilles est le témoin des différentes mains qui ont écrit le document et qui n'avaient pas les mêmes pratiques d'encodage.

[63] Il existe davantage d'occupations différentes dans le document en TEI : l'objectif était de tester des méthodes d'encodage sans perdre trop de temps à trouver les URI de rôles compliqués.

[64] *Wikidata* est la base de toutes les données de *Wikipédia*.

à elle, retournera sa valeur, contenue dans la seconde expression **@select**. Voici, ci-dessous, la *map* construite pour les nationalités :

```
<xsl:variable name="nationalitiesPrefix"
    select="'http://www.wikidata.org/wiki/Special:EntityData/Q'"/>
<xsl:variable name="nationalities" as="map(xs:string, xs:string)">
        <xsl:map>
                <xsl:map-entry key="'English'" select="'842438'"/>
                <xsl:map-entry key="'British'" select="'842438'"/>
                <xsl:map-entry key="'Irish'" select="'170826'"/>
                …
        </xsl:map>
</xsl:variable>
```

La première instruction établit une variable qui contient le préfixe de nos URI jusqu'au dernier caractère commun entre toutes : la lettre « Q ». La valeur qui terminera une URI sera l'une de celles contenues dans la *map* sélectionnée par le programme lorsqu'il trouvera une nationalité correspondant à une clé. Ce *snippet* peut exister indépendamment du reste du code sans le corrompre : une variable (dans laquelle la *map* est stockée) ne s'active que lorsque l'on fait appel à elle. Dans le cadre du DDC, nous avons été obligés d'utiliser ce système pour transformer les données car le contenu des balises XML était du texte brut[65].

Comment utiliser cette *map* ? Le fichier de transformation de l'index est construit comme celle de la transformation LGLC. Les modèles se succèdent en respectant les règles imposées par le langage. Cette transformation est similaire à celle du projet LGLC car elle a permis de poser les bases du schéma que nous avons ensuite perfectionné dans LGLC. Voici le résultat de la conversion qui utilise les *maps* établies pour les nationalités :

```
<xsl:template match="nationality">
        <xsl:choose>
                <xsl:when test="map:contains($nationalities, current())">
```

[65] Il est aussi possible d'utiliser une API de réconciliation des données qui va aller attraper les données de Wikidata, les confronter à celles du fichier TEI et appliquer une URI cible à celles-ci. Ceci est possible grâce au logiciel OpenRefine, par exemple. Toutefois, nous n'avons exploré cette piste à l'heure où nous avons développé cette solution de transformation car nous n'avions pas connaissance de cette autre solution. Tout compte fait, ce n'est pas un problème car nous avons aussi pu chercher une solution par nous-même et approfondir considérablement nos connaissances du XSLT.

```
                <schema:nationality
                rdf:resource="{concat($nationalitiesPrefix, $nationalities(current()),
                '.rdf')}"/>
            </xsl:when>
            <xsl:otherwise>
                <xsl:comment>No URI for nationality
                <xsl:value-of select="current()"/>
                </xsl:comment>
            </xsl:otherwise>
        </xsl:choose>
</xsl:template>
```

Pour qu'une *map*, contenue dans une valeur fonctionne, il faut l'appeler. Nous créons donc un modèle qui cible spécifiquement la nationalité des personnes de la DIV4. Etant donné que certaines personnes n'ont parfois pas de nationalité ni de profession documentée, nous utilisons un **<xsl:choose>** qui choisira l'action à effectuer en fonction du cas par cas. Lorsqu'une personne n'a pas de données renseignées, le modèle a pour instruction de créer un commentaire XML qui servira seulement à indiquer l'absence d'information. A l'inverse, lorsqu'une information présente déclenchera le processus, le modèle effectue d'abord un test. Si la *map* contient n'importe quelle clé également présente dans son contexte actuel **(<nationality>**), il déclenche l'instruction enveloppée à l'intérieur. Celle-ci exécute une concaténation entre plusieurs éléments. Elle attrape d'abord le préfixe de notre URI auquel elle vient apposer la clé de la nationalité qu'elle a trouvée dans la *map*.

Cependant, rappelons-nous que la clé d'une *map* sélectionne une valeur qui lui est associée. L'instruction va donc chercher le numéro d'identification *Wikidata* de la nationalité. Enfin, nous lui enjoignons d'ajouter « .rdf » à la fin de l'ensemble car c'est le format des données ouvertes liées sur Internet. Nous nous sommes surtout appuyés sur l'exemple des nationalités dans cette explication mais elle s'applique au modèle conçu pour transformer les professions via leur propre *map*. Voici le résultat de l'exécution du modèle, un triplet RDF complet :

```
<schema:nationality
rdf:resource="http://www.wikidata.org/wiki/Special:EntityData/Q842438.rdf"/>
```

Le deuxième exemple d'encodage testé sur DDH, qui nous a permis d'aboutir à une transformation LGLC efficace, concerne les relations entre les personnes. Il s'agit ici d'une manipulation plus compliquée que tout ce que l'on a pu aborder dans ce mémoire. C'est pourquoi nous tenterons simplement d'expliquer son objectif et non la totalité de l'encodage afin de ne pas dérouter les lecteurs moins spécialistes de la discipline. Cependant, un lecteur spécialiste pourra consulter la feuille de style s'il souhaite découvrir notre encodage en détail.

A la toute fin de la DIV4 de l'index, les équipes ont décidé d'aller plus loin dans leur étude des protagonistes en encodant les relations qu'ils entretiennent entre eux dans une **<listRelation>**. Chaque **<relation>** est caractérisée par un nom et un attribut contenant les **@xml:id** des personnes concernées par la relation. Une relation exprimant un lien marital ressemblera deux époux, parfois des divorcés, par exemple. Tout le défi de cette transformation réside dans cette gymnastique à effectuer entre la liste de **<relation>** et les personnes de la DIV4 via l'unique élément connectant les deux listes : l'**@xml:id.** Tout se joue ensuite dans la feuille de style sur le modèle lié à **<person>**.

Réaliser cette gymnastique requiert l'emploi de la fonction **<xsl:for-each>** qui va nous couper du contexte du modèle. C'est pourquoi nous devons d'abord stocker la valeur de l'**@xml:id** du contexte dans une variable. Les opérations suivantes vont utiliser deux **<xsl:key>** au sein de la fonction **<xsl:for-each>**. La modèle va donc être capable de comparer l'**@xml:id** de chaque **<person>** de DIV4 aux attributs **@mutual** des **<relation>** grâce à une **<xsl:key>** que nous avons programmée pour exclure le **#**[66]. Après cette comparaison, si le modèle trouve un ID commun entre les personnes et les relations, il crée un élément XML configuré pour exprimer le type de la relation entretenue, c'est la première partie de notre triplet. La deuxième partie de ce triplet concatène les premiers caractères d'une URI (ici, c'est un *dummy*) avec le contexte de notre modèle, soit ici la valeur de **@mutual** pêchée par le programme. Cette valeur correspond aux personnes entretenant une relation avec le sujet à partir duquel nous sommes en train de créer un jeu de triplets. Le résultat de cette voltige apparaît alors, au sein du triplet d'une personne, sous cette forme :

[66] Initialement, nous ne pouvions pas utiliser les valeurs des attributs de **@mutual** car les **@xml:id** des personnes ne contiennent pas de #, il faut donc s'en débarrasser avec ***tokenize***.

```
<schema:Person rdf:about="http://example.org/tei#HallSamuelCarter">
        […]
        <cwrc:hasGender rdf:resource="http://sparql.cwrc.ca/testing/cwrc#man"/>
        <schema:additionalName>Hall</schema:additionalName>
        <schema:givenName>Samuel</schema:givenName>
        <schema:givenName>Carter</schema:givenName>
        […]
        <schema:spouse rdf:resource="http://example.org/tei#HallAnnaMaria"/>
</schema:Person>
```

Lorsqu'il a traité « Samuel Carter Hall », le programme a exécuté les modèles permettant d'obtenir des informations à propos de lui comme son nom, son genre, etc. Il a également comparé son **@xml:id** au contenu de l'attribut **@mutual** de **<relation>** et s'est rendu compte qu'il était marié à Anna Maria. Il a donc récupéré l'ID d'Anna Maria du contexte de **<relation>** et créé une URI à partir de lui qu'il a retourné dans le *snippet* de triplets RDF à propos de Samuel.

6. Conclusion de la partie 2

Réaliser la transformation en XSLT d'un document TEI est un exercice très intéressant qui permet de manipuler des fonctions et travailler avec des contextes variés pour sculpter nos données dans un format RDF. L'efficacité de cet outil n'est plus à démontrer. Les astuces d'encodage présentées, dans la dernière sous-partie à propos de l'index *Craik,* ont été très bénéfiques à l'élaboration d'une conversion complète des données du projet LGLC.

PARTIE 3 : Analyse des résultats

Cet exposé approche de sa conclusion. Avant d'aborder des pistes de réflexion destinées à aller plus loin, il convient d'analyser une partie des résultats de recherche afin d'expliquer pourquoi certains choix ont été faits.

7. Ontologies et choix d'encodage

La sérialisation XML/RDF fonctionne par la superposition d'éléments XML nichés les uns dans les autres. Toutes les entités doivent donc être expliquées par des classifications qui permettent à l'ordinateur de comprendre la donnée. Ces classes s'expriment à l'aide d'ontologies comme *Dublin Core.* Nous avions d'ailleurs souligné qu'il en existait une très grande variété répondant à différents besoins d'encodage. Dans le cas de cette étude, les deux principales ontologies avec lesquelles nous avons travaillé sont *Schema* et CWRC.

N. *Schema* : une ontologie très vaste

Comme toutes les ontologies, *Schema* propose une collection de termes pour encoder du contenu destiné à être publié sur le web, de sorte qu'il soit compréhensible par les machines[67]. Très vaste, elle permet de décrire un grand nombre d'entités. Elle pourrait être considérée comme une ontologie généraliste puisqu'elle n'est pas spécialisée dans la description d'un domaine précis de concepts comme la *Music Ontology*[68]. Par exemple, l'item le moins contraignant de l'ontologie est *thing* : il peut littéralement s'appliquer à n'importe quelle entité. Chaque classe nécessite ensuite une description plus précise à l'aide de propriétés. Pour *thing*, celles-ci sont *name* (une chose a un nom), *description* (une chose peut être décrite), *url* (elle peut être localisée par une adresse sur le web) et *image* (elle peut être

[67] Ontologie de *Schema* [en ligne, consulté le 7/04/2019, disponible à https://schema.org/docs /full.html].

[68] *The Music Ontology* fournit un modèle de structuration des données liées au domaine musical. Elle propose l'utilisation de classes comme « compositions », « musicArtist » et autre vocabulaire bien spécifique au monde de la musique qui est absent des ontologies plus généralistes comme *Schema*.

picturale). Le résultat de la conversion dans sa totalité est disponible en suivant ce lien, au format RDF. Regardons de plus près cet exemple de *snippet* en triplets RDF :

```
<schema:Event rdf:about="http://lglc.ca/lglc.rdf#n71.104">
        <bf:citationUri rdf:resource="http://lglc.ca/lglc.rdf#n71.104.1"/>
        <schema:startDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date"/>
        <schema:endDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date"/>
        <schema:Location rdf:resource="http://lglc.ca/places.rdf#TOR"/>
        <schema:contributor rdf:resource="http://lglc.ca/organizations.rdf#CBC"/>
</schema:Event>
```

Schema fourni au projet LGLC plusieurs classes pertinentes. Décrivant le sujet de chaque triplet, **<schema:Event>** indique à la machine qu'elle est en train de traiter un évènement qui a eu lieu à une certaine date dans un certain lieu. Cette expression admet un lot de propriétés qui décrivent l'évènement. Le tag **<schema:Event>** admet donc l'utilisation de la propriété **<schema:startDate>** pour indiquer la date à laquelle un évènement a eu lieu, la propriété **<schema:Location>** pour désigner le lieu de l'évènement, etc.

Individuellement, chaque ligne de ce *snippet* est un triplet. Un **<schema:Location>** est un lieu qui possède un objet : son identification sous la forme d'une URI. Cette relation s'exprime à l'aide de l'ontologie RDF qui admet que les contributeurs sont des types de ressources pouvant avoir une URI. Sur le *snippet* ci-dessus, le premier triplet **<schema:Event>** se termine à la fin de l'énumération de tous les autres triplets. Cela signifie que tous les triplets contenus dans un **<schema:Event>** le concernent directement lui et seulement lui. Un autre évènement pourra avoir comme localisation « Toronto » mais il aura eu lieu à une date différente en des circonstances différentes.

O. Analyse d'une configuration de *Schema* au service du projet LGLC

Ce parti pris d'encodage est un moyen de conserver, au sein d'un même noyau de triplets, toutes les données qui concernent un seul et même évènement. La transformation en XSLT va attraper dans le fichier TEI les trois mille cent-soixante-trois évènements qu'il contient et convertit les milliers de données qui se rapportent à chacun d'entre eux en triplets

RDF. La création d'une structure telle que celle-ci pour organiser les triplets a été en partie motivée par la quantité de données à transformer.

Le projet LGLC en est à ses débuts de l'exploration des possibilités offertes par les Données Ouvertes Liées appliquées à son contenu. Ces étapes liminaires au développement d'un schéma de transformation des données définitif est une expédition à tâtons des technologies du web sémantique pertinentes dans ce cadre-ci. L'ensemble de ce mémoire est à considérer comme le premier jalon de cette conversion des données de LGLC en *open linked data*.

Convertir les données de cette manière est une façon de tester des solutions d'encodage pour atteindre deux objectifs. Premièrement, vu la grande quantité des données disponibles, il s'agissait de trouver des solutions viables et efficaces de transformation en triplets RDF. Deuxièmement, le résultat de la transformation mettait en exergue les erreurs d'encodage par l'absence de résultats. Etant donné la quantité de données et le nombre d'éditeurs ayant altéré le fichier TEI, mettre au point une transformation permettait de filtrer facilement les erreurs. Nous avons, par exemple, rencontré le cas où un élément avait deux attributs différents pour exprimer le même objet[69]. Le nombre différent d'entités en entrée et en sortie nous a permis de localiser les sources d'erreurs et de prendre un parti d'encodage[70].

Ces arguments justifient les raisons pour lesquelles nous avons mis au point une transformation lisible par un opérateur humain en sortie.

[69] **@corresp** et **@sameAs** par exemple alors qu'un seul suffirait.

[70] Ne pouvant pas modifier le fichier TEI, nous avons décidé de ne pas traiter les cas où le nombre d'erreurs ne dépassait pas 20.

P. A propos du *Canadian Writing Research Collaboratory* (CWRC)

En plus d'être une infrastructure pilote du projet *Canadian Writing Research Collaboratory* (CWRC), le projet LGLC souhaite, à long terme utiliser leur ontologie pour traiter ses données prosopographiques[71].

Pour le moment, nous avons pu tester CWRC sur le document XML/TEI que nous a partagé le projet *Digital Dinah Craik* (DDC). Le résultat de la conversion du document TEI en RDF est disponible en suivant de lien. Cette ontologie est intéressante pour deux raisons. La première est qu'elle a un grain de précision bien plus spécifique que Schema pour décrire certaines relations ou entités. En effet, CWRC admet une description plus fine du genre en proposant non-seulement des propriétés comme « homme » et « femme » mais il offre aussi la possibilité de décrire des personnes revendiquant un genre LGBT : une aubaine pour le projet LGLC. La deuxième est qu'elle impose une ligne de conduite au projet LGLC qui respecte scrupuleusement le format RDF pour l'encodage des données. En théorie, une DOL prend la forme d'un triplet au format RDF. Dans certains cas, nous avons dû transformer des chaînes de caractère de texte en prétendant que cela était autorisé. Des Données Ouvertes Liées de bonne qualité n'ont pas de texte brut, seulement des liens, c'est là tout leur intérêt. Si Schema autorise l'adjonction de texte brut dans sa configuration, ce n'est pas le cas de CWRC qui est bien plus strict à ce sujet et n'autorise que l'utilisation d'URI. Si le web sémantique a un rêve c'est bien celui de proposer des ontologies permettant d'absolument tout décrire jusqu'à ce que toutes les données se connectent ensemble.

Q. Bénéfices pour les Sciences Humaines et Sociales

Après avoir encodé l'ensemble de ces données en triplets RDF, il est nécessaire de disposer d'un environnement où les stocker. La France héberge déjà des projets ayant aboutit à la construction de *triplestores*. Conçu spécifiquement pour le stockage de données RDF, le

[71] Ontologie du *Canadian Writing Research Collaboratory* (CWRC) [en ligne, consulté depuis mars 2019, disponible à http://sparql.cwrc.ca/ontologies/cwrc.html].

triplestore français le plus célèbre qui nous vient à l'esprit est celui de Persée[72]. Ces bases de données RDF ne sont pas inaccessibles : un utilisateur peu naviguer à l'intérieur à partir d'un SPARQL Endpoint, un navigateur de recherches RDF, d'une certaine manière. Parcourir les collections de triplets requiert néanmoins de posséder quelques compétences de bases dans la manipulation du langage SPARQL. Celui-ci sert à formuler des requêtes, dans un format spécifique, afin de pouvoir sillonner les collections. En définitif, un *triplestore* est l'aboutissement du travail de préparation et de conversion qui rend les DOL accessibles.

Au Canada, si LGLC travaille en étroite relation avec CWRC, c'est finalement pour fournir du contenu destiné à alimenter le *Linked Infrastructure for Networked Cultural Scholarship* (LINCS) [73]. Effectivement, ce projet de création d'un *triplestore* canadien vient récemment d'être annoncé[74]. Il a pour objectif de promouvoir la culture canadienne par le biais de son intégration au web sémantique : il s'agit une mise en circulation de toutes les DOL que détendent les institutions canadiennes participantes au projet[75]. Le projet promouvra les données de la culture canadienne à l'international et, à travers elles, permettra de connecter les disciplines auxquelles elles seraient communes. LINCS se donne donc pour mission de convertir, connecter et rendre accessible des jeux de données immenses. Les chercheurs auront la possibilité d'explorer ce nouveau système de production scientifique en partenariat avec toutes les autres Humanités au sens large. Ils auront, à leur disposition, un moyen de confronter les données issues de multiples disciplines et d'explorer de nouvelles problématiques de recherches. Construire autant de liens entre les données et concevoir un outil permettant de dépasser les notions de compatibilité logicielle transformera la façon d'aborder bien des sujets. Les canadiens pourront donc examiner à l'aide de sources inédites leur Histoire culturelle, sociale, politique, etc. LINCS est un véritable tremplin destiné à relancer la productivité scientifique canadienne pour le web sémantique.

[72] Triplestore de Persée [en ligne, consulté le 17/07/2019, disponible à http://bit.ly/2k8IZYJ].
[73] CWRC, op. cit. voir : https://cwrc.ca/news/LINCS-CFI

[74] Le Canada est généralement un leader mondial dans le développement des Humanités Numériques. Toutefois, il a tardé à utiliser concrètement le potentiel des DOL.

[75] Les données disponibles pour le projet LINCS sont issues des collections de bibliothèques ou d'archives canadiennes : cela représente un montant total d'environ 30 millions de triplets, soit 300 TB de données.

L'un des plus grand défi de LINCS, parmi d'autres, vient de la gestion de la propriété intellectuelle des données. Certaines d'entre elles sont en effet verrouillées par des accès payants. A long terme, tout le monde pourra visualiser les données libres d'accès mais, lorsque les utilisateurs rencontreront des triplets générés à partir de données protégées par des accès payant, ils seront redirigés vers un site web dédié à ces dernières. C'est un véritable problème pour un projet qui se veut être l'infrastructure de Données Ouvertes Liées principale au Canada.

8. Conclusion de la partie 3

Les initiatives prises par les structures popularisant les données en accès libre sur le web est louable. Certes, elles se heurtent à de nombreux problèmes de gestion des données ou de détention de droits mais servent toutes le même objectif noble : pousser la recherche à évoluer vers un monde où les chercheurs travailleraient ensemble, toutes disciplines confondues.

Conclusion Générale

Lorsqu'en 1991 apparaît le premier serveur web, les chercheurs reçoivent en cadeau un moyen de publier le fruit de leur travail au même endroit et dépassent ainsi des contraintes matérielles qui, auparavant, restreignaient les projets à une faible visibilité. Cette époque voit aussi le développement des premiers outils destinés à donner une structure au web. Le succès du système entraine des conséquences à double tranchant. Plus le web devient populaire, plus la communauté s'agrandit et le nourrit de nouveaux documents et pages web. En revanche, plus elle le nourrit, plus il devient une bibliothèque monstre ayant avalé une quantité astronomique d'informations mais sans disposer d'étagères pour les ranger.

Dans ce contexte, Tim Berners-Lee sensibilise la communauté Internet à l'importance d'avoir un web structuré et l'interpelle sur un nouveau moyen, jusqu'alors ignoré, de le ranger : utiliser des données. La réaction qu'il provoque conduira, à la naissance du web sémantique qui enveloppera le traditionnel, afin de modifier sa structure organisationnelle. L'apparition de ces nouvelles technologies des Sciences de l'Informatique met à la disposition des usagers et chercheurs un moyen de manipuler les données contenues dans les documents qu'ils savent déjà utiliser. Celles-ci transforment les méthodes de travail dans les Humanités. Considérés comme de véritables atouts, ces outils enrichissent nos connaissances et nos collections notamment en appliquant aux documents un nouveau modèle de description des ressources numériques : le *Resource Description Framework* (RDF).

Ce modèle est au web ce qu'une bibliothèque est à ses livres : un moyen de stocker et d'organiser des données en fonction d'un schéma défini. Les Données Ouvertes Liées sont en effet, par essence, interopérables avec les outils et autres données de la communauté des Humanités Numériques au sens large. Faire en sorte qu'un projet rende accessible son contenu, dont les entités deviennent identifiables, permet alors aux données de communiquer entre elles. S'assurer que tout le monde comprenne le vocabulaire et la langue utilisés passe par la définition d'un dictionnaire, décrivant aux hommes et surtout aux machines, la signification des entités et la nature de leurs relations.

On appelle ces recueils de termes des « ontologies » auxquelles on a recours lors de la conversion des données. Effectivement, pour qu'une ressource soit compréhensible par une machine, elle doit être exprimée dans un format bien précis que l'on appelle RDF. Au cours de nos travaux, nous avons eu l'occasion de tester différentes méthodes de transformation. Les entités stockées à l'intérieur de document en XML/TEI ne sont ni ouvertes, ni liées, tant qu'elles ne sont pas converties en triplets RDF. Cette expression, triplet, désigne un sujet identifié par son URI unique, et la relation qu'il entretient avec un objet. L'ensemble des triplets créés pour faire des Données Ouvertes Liées est une ressource précieuse qu'il convient de stocker à l'intérieur de *triplestores*. Ces bases de données RDF qui leur sont dédiées sont parcourues par les utilisateurs qui y ont notamment accès via des interfaces de requêtage en SPARQL.

L'une des méthodes techniques permettant d'atteindre ce but est d'utiliser le langage XSLT pour convertir des document TEI en triplets RDF. Dans un contexte dominé par le langage XML, nous devons alors utiliser une sérialisation, une façon d'écrire, que l'on appelle XML/RDF. Cette solution offre l'avantage de pouvoir transformer rapidement l'ensemble des données d'un fichier textuel peu importe leur quantité. En effet, grâce à l'élaboration d'un modèle, seulement long de quelques dizaines de lignes, nous sommes en mesure d'obtenir plus de quatre mille triplets. Le principal désavantage de cette solution est d'être tributaire de la qualité de l'encodage TEI. Si un élément n'est pas caractérisé par l'attribut correct, l'exécution du modèle l'ignorera et nous raterons l'opportunité d'avoir des triplets complets. Le deuxième obstacle rencontré est que le XSLT et la sérialisation XML/RDF demandent des compétences en linguistique XML. La sérialisation *Turtle* est bien plus simple à utiliser mais n'offre pas la possibilité de traiter un aussi grand nombre de données que la méthode passant par le XSLT.

Finalement, le principal objectif du web sémantique est de rendre accessible des données de la recherche qui, sans lui, existeraient encore de manière isolée, sans moyen de communication entre elles. Si une donnée est identifiée sur le web par son URI et qu'elle est commune à plusieurs projets, ceux-ci seront en mesure d'échanger des informations.

Plusieurs projets motivés par des problématiques en Humanités Numériques s'attellent à la tâche de conversion des données en DOL. En France, Persée s'impose comme le plus grand *triplestore*. Au Canada, LINCS travaille actuellement à la mise en ligne de plusieurs *térabits* de triplets pour proposer son propre *triplestore*. D'une envergure jamais vue auparavant, il promet d'abolir les frontières entre les disciplines. Continuer cette exploration de projets, travaillant de concert avec des technologies du web sémantique, améliorera les méthodes de recherche en SHS, comme dans d'autres disciplines. Le projet *Scholastic Commentaries and Texts Archive*[76] se prête bien à la poursuite de la réflexion. Porté par des valeurs humanistes qui dénoncent la compétition au sein de la recherche, le projet exhorte tous les chercheurs le projet encourage l'utilisation et la réutilisation de ses données en libre accès pour les connexions scientifiques qu'elles permettent d'établir. Il exhorte également tous les chercheurs de la discipline scholastique à ne pas céder à la pression des éditeurs qui dissimulent les données derrière un péage virtuel afin de les monétiser, la plus grave maladie d'Internet. Au contraire, il encourage l'utilisation de ses données en libre accès pour établir les connexions scientifiques.

C'est de cette manière que l'on crée des ponts interdisciplinaires et que l'on dégage l'accès vers de nouveaux questionnements, confrontant ainsi les données issues de multiples domaines. Afin d'arriver à ce but, le langage XML/RDF se révèle un outil d'avenir au potentiel particulièrement intéressant, pour le projet LGLC comme pour d'autres, ce dont nous espérons avoir réussi à convaincre notre lecteur.

[76] *Scholastic Commentaries and Texts Archive*, accessible via https://scta.info/, consulté le 26/07/2019.

Références bibliographiques et ressources en lignes

9. Bibliographie

ALLEMANG D., HENDLER J., *Semantic Web for the working ontologist effective modeling in RDFS and OWL*, San Francisco, 2011.

ARP R., SMITH B., SPEAR A., *Building Ontologies with Basic Formal Ontology*, Londres, 2015.

BECKER B., "*Web of Data: A Primer of Linked Data and the Semantic Web"* in *Behavioral & Social Sciences Librarian 35*, 2016, pp. 42-45.

BERNERS-LEE T., "The next web*"* [conférence dans le cadre de TED2009], à Long Beach (CA), en Février 2009.

BERNERS-LEE T., HENDLER J., LASSILA O., "*The Semantic Web : a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities"* in *Scientific American*, 284(5), s.l., 2001, pp. 34–43 [en ligne, consulté le 29/07/2019, disponible à http://bit.ly/2kbCcxv].

BESHERO-BONDAR E., BIRNBAUM D., *XPath for processing XML and managing projects* [Support de cours], s.l., 2019 [en ligne, consulté le 10/06/2019, disponible à http://bit.ly/2IlfSS4].

BLACE R., FISHER M., HEBELER J., *et al*., *Semantic Web Programming*, Indianapolis, 2009.

BLANEY J., *Introduction to the Principles of Linked Open Data*, s.l., 2017 [consulté le 20/07/2019, disponible à https://programminghistorian.org/en/lessons/intro-to-linked-data].

BONEVA I., GAYO J., KONTOKOSTAS D. *et al.,* "*Validating RDF Data"* in *Synthesis Lectures on the Semantic Web: Theory and Technology 7*, 2017.

BREITMAN K., CASANOVA M., TRUSZKOWSKI W., *Semantic Web: Concepts, Technologies and Applications*, Londres, 2007.

BRUNETTI J., GIL R., GARCIA R., "*Facets and pivoting for flexible and usable linked data exploration"* in UNGER C., CIMIANO P., LOPEZ V. (eds.), *et al.*, *Proceedings of Interacting with Linked Data*, Heraklion, 2012.

CROMPTON C., LANE R., SIEMENS R., *Doing Digital Humanities,* Londres, 2016.

CROMPTON C., SCHWARTZ M., "*Lesbian and Gay Liberation in Canada: Representing the Dyke Dynamo"* in *Digital Studies/Le champ numériques 5*, 2016.

DUCHARNE B., *Learning SPARQL*, Sebastopol, 2013.

FITZGERALD M., *Learning XSLT: A Hands-On Introduction to XSLT and XPath*, Beijing, 2003.

HARDESTY J., "*Transitioning from XML to RDF considerations for effective moves towards Linked data and the Semantic Web"* in *Information Technology and Libraries 35*, Avril 2016, pp. 51-64.

HITZLER P., KROTZSCH M., RUDOLPH S., *Foundations of Semantic Web Technologies*, New York, 2009.

LUKASIK S., "*Why the Arpanet Was Built"* in *IEEE Annals of the History of Computing,* 1058-6180/11, 2011, pp. 4-20.

MCLEOD D., *Lesbian and gay liberation in Canada: a selected annotated chronology, 1964-1975*, Toronto, 1996.

MCLEOD D., *Lesbian and gay liberation in Canada: a selected annotated chronology, 1976–1981*, Toronto, 2016.

MAILLOT P., RAIMBAULT T., GENEST D. *et al.*, « *Aperçus de recherche : Interroger efficacement un ensemble de bases RDF »* in *Documents Numériques*, 2014, pp. 9-33 [consulté le 12/07/2019, disponible à http://bit.ly/2jQRzet].

POWERS S., *Practical RDF*, Sebastopol, 2003.

ROY-GAGNON M.-H., « Outils statistiques et bio informatiques pour l'infrastructure multisectorielle i-BALSAC *»* [conférence dans le cadre du DHSITE], à Ottawa, le 14/05/2019.

SEGARAN T., EVANS C., TAYLOR J., *Programming the Semantic Web*, Sebastopol, 2009.

TIDWELL D., *XSLT : 2nd Edition*, Sebastopol, 2009.

10. Ressources en ligne

Digital Dinah Craik [en ligne, consulté le 19/03/2019, disponible à http://www.digitaldinahcraikproject.org/].

Dossiers du projet Bouvard & Pécuchet [en ligne, consulté le 17/07/2019, disponible à http://www.dossiers-flaubert.fr/projet-originel].

DBPEDIA [en ligne, consulté le 14/05/2019, disponible à https://wiki.dbpedia.org/services-resources/interlinking].

BERNERS-LEE T., *Open, Linked Data for a Global Community*, Gov 2.0 Expo 2010 [vidéo en ligne, consultée le 17/07/2019, disponible à http://bit.ly/2jRSmf1].

GeoNames Authority Files [en ligne, consulté le 14/05/2019, disponible à http://www.geonames.org].

Lesbian and Gay Liberation in Canada (LGLC) [en ligne, consulté le 19/03/2019, disponible à https://lglc.ca/].

Lettre d'information de l'Université de Lille 3 [en ligne, consulté le 24/06/2019, disponible à https://www.univ-lille3.fr/actualites/?actu=15011].

Projet *5-star Open Data Plan* de Tim Berners-Lee [en ligne, consulté le 16/05/2019, disponible à http://5stardata.info/en/].

Ontologie du *Canadian Writing Research Collaboratory* (CWRC) [en ligne, consulté depuis mars 2019, disponible à http://sparql.cwrc.ca/ontologies/cwrc.html].

Ontologie du *Dublin Core Metadata Initiative* (DCMI) [en ligne, consulté le 17/04/2019, disponible à https://www.dublincore.org/].

Ontologie de *Schema* [en ligne, consulté le 7/04/2019, disponible à https://schema.org/docs/full.html].

Scholastic Commentaries and Texts Archive [en ligne, consulté le 12/08/2019, disponible à https://scta.info/].

TEI by Example [en ligne, consulté le 18/03/2019, disponible à https://teibyexample.org/].

Triplestore de Persée [en ligne, consulté le 17/07/2019, disponible à http://bit.ly/2k8IZYJ].

Virtual International Authority File (VIAF) [en ligne, consulté le 14/05/2019, disponible à https://viaf.org/].

Wikidata [en ligne, consulté le 23/03/2019, disponible à http://bit.ly/2lGCiNJ].

ANNEXE : Note de Stage

uOttawa

Etant donné que je possède déjà un master en Sciences Humaines et Sociales (Master Mondes Anciens, Spécialité Egyptologie, Université Lyon 2) ce mémoire est un peu particulier. Il est effectivement le fruit du travail que j'ai effectué lors de mon stage de recherches à l'Université d'Ottawa, dans le laboratoire *Humanities Data Lab.* (HDL)*, s*ous la supervision de Mme Constance Crompton, PI et CRC en Humanités Numérique[77]. C'est la raison pour laquelle nous avons convenu avec Mme Sabine Loudcher que cette note de stage serait brève.

En Mars 2019, j'ai pris mes fonctions d'assistante de recherche au laboratoire. Organisé autour d'un apprentissage avancé de la TEI, ce stage avait pour objectif de m'aider à développer les meilleures techniques de production de données pour le web sémantique selon la norme RDF. Le programme du stage comprenait aussi un entrainement aux transformations en XSLT. Finalement, nous avons rapidement survolé les concepts de base que je maitrisais déjà et 70% de mon travail a été de transformer les données au format TEI du projet LGLC en triplets RDF, dont les résultats sont expliqués dans le mémoire. Le reste du temps, je participais activement aux activités de recherches en Humanités Numériques de l'Université d'Ottawa.

[77] *Principal Investigator and Canada Research Chair.*

J'ai d'abord eu le plaisir d'animer une présentation, avec les autres membres du laboratoire, lors de la journée portes ouvertes, *Arts in April,* de la Faculté des Arts. Au mois de Mai, le laboratoire a ensuite été le pilote d'un évènement brièvement cité dans le mémoire : *The Digital Humanities Summer Institute Technologies East* 2019 (DHSITE2019). La semaine a été rythmée entre des conférences et des ateliers. Les conférences, aux heures des repas, étaient animées par des chercheurs qui venaient présenter leur projets en SHS et en quoi ils s'inscrivaient dans la dynamique des Humanités Numériques. Les ateliers étaient, eux, dirigés par des chargés de cours et se voulaient être une forme de stage intensif en Humanités Numériques. Le HDL a spontanément financé la participation de ses membres à quelques-uns des cours dispensés. Grâce à ce geste, j'ai pu assister à l'atelier ayant pour thème *Introduction to Linked Open Data* : une aubaine. Le dernier jour, j'ai eu l'agréable chance d'animer une table ronde autour du sujet des DOL, dans le cadre du projet LGLC, aux côtés d'autres chercheurs[78].

Lors d'une journée découverte de l'Université par une classe de jeunes élèves, j'ai aussi pu vivre ma première expérience officielle d'enseignante. En effet, nous avons donné un cours de vulgarisation des Humanités Numériques avec quelques amis candidats au doctorat du Département de la Communication. Cet exercice ludique, que nous avons tous apprécié, a demandé beaucoup d'agilité pour expliquer le fonctionnement et l'intérêt du web sémantique à nos élèves[79]. Finalement, cette journée a été un bon moyen de tester nos compétences en vulgarisation.

[78] Mon support de présentation est accessible en suivant ce lien.

[79] Voici le support de cours que j'ai tenté de rendre interactif en y ajoutant notamment un petit sondage sur les pratiques en ligne des élèves. La feuille de calcul, liée au diaporama, s'est transformée en graphique pendant ma présentation, sous leur yeux ébahis.

L'évènement, qui a le plus enrichi mes compétences et l'un des plus incroyables à vivre, est ma participation à la semaine du *Digital Humanities Summer Institute* (DHSI) proposé par l'Université de Victoria (Vancouver, Colombie-Britannique). Voyager jusqu'à Vancouver pour assister à l'un des plus grands évènements en HN du Canada n'était absolument pas prévu. Avec le soutien de Constance, j'ai effectué une demande de bourse d'études à l'Université de Victoria pour pouvoir me rendre à DHSI. Ma candidature a été retenue et j'ai été exonérée de tous les frais d'inscription à l'évènement. Cette réussite annonçait à peine la couleur fantastique de ce qui est arrivé tout au long de cette semaine. Officiellement, j'étais inscrite au cours intitulé *XPath for processing XML and managing projects* dispensé par Elisa Beshero-Bondar et David J. Birnbaum. Ils ont tous deux été fantastiques. En plus de nous permettre d'approfondir nos connaissances du XPATH, ils ont passé chaque soirée, après les cours, à aider leurs élèves en difficulté dans leurs projets respectifs. J'ai pu profiter de leur gentillesse et immense expertise lors d'une soirée où ils m'ont enseigné l'art des **<xsl:map>** des **<xsl:key>** et de la gymnastique entre des contextes différents grâce à des variables. Ces avancées ont largement débloqué mon travail que vous avez lu aujourd'hui.

Je garderai un souvenir impérissable de ce voyage au Canada qui a été d'une richesse professionnelle et personnelle immense. Les bénéfices de cette expérience semblent être partis pour durer quelques mois. Constance Crompton m'a effectivement proposé de continuer à travailler à distance avec l'équipe. Elle a aussi exprimé sa volonté, si je le souhaitais, d'entreprendre une thèse à ses côtés : éventualité dont nous sommes en train d'évaluer la faisabilité en cotutelle avec l'Université de Lyon 2.

Table des illustrations