Exploiter les données liées pour améliorer l'accès sujet dans les sources primaires en ligne – Étude de cas de la collection en ligne sur la Première guerre mondiale de l'Université du Colorado à Boulder

Thea Lindquist
University of Colorado Boulder
Boulder, CO, USA
E-mail: thea.lindquist[at]colorado.edu

Eero Hyvönen

Juha Törnroos

Eetu Mäkelä
Aalto University and University of Helsinki
Aalto, Finland
E-mail: first.last[at]aalto.fi

Traduction :
Jo-Anne Bélair
Bibliothèque de l'Université Laval
Canada
(jo-anne.belair[at]bibl.ulaval.ca)

Session: *117* — **Section de Classification et d'indexation** – ***L'accès sujet aujourd'hui : source d'inspiration, surprenant et garant d'autonomie***

Résumé :

Souvent, les usagers des institutions académiques trouvent le travail en ligne avec des sources primaires à la fois enrichissant et stimulant. L'amélioration de l'accès sujet dans ces sources est essentielle compte tenu que les collections numériques se propagent et que le travail avec des sources primaires devient de plus en plus important dans les programmes de sciences humaines. Une évaluation des besoins des usagers a été menée auprès des usagers du domaine des sciences humaines de l'Université du Colorado à Boulder (UC) pour faciliter l'interaction avec ces sources. Deux des principaux besoins identifiés sont l'amélioration des résultats de recherche et du contexte, en particulier pour les sujets historiques.

Les données liées peuvent répondre à ces besoins en liant les concepts connexes des sources à l'aide de termes tirés d'un vocabulaire spécialisé, en les enrichissant avec des ressources externes

et en offrant des services sémantiques riches qui rendent les usagers plus autonomes. Cet article présente un projet d'amélioration de l'accès sujet dans la collection en ligne de l'UC sur la Première guerre mondiale en établissant des liens profonds entre les données historiques sur l'expérience des civils en Belgique occupée. Ce travail est destiné à mener à une meilleure compréhension des forces qui ont façonné la période de la Première guerre mondiale.

I. INTRODUCTION

La richesse des collections numériques du patrimoine culturel en ligne offre aux usagers plus de possibilités que jamais auparavant d'interagir directement avec les sources primaires[i]. Les usagers des institutions académiques trouvent le travail à l'aide de ces sources à la fois enrichissant et stimulant. Bien que ces collections numériques offrent des ressources très précieuses, en particulier pour les disciplines des sciences humaines, des études indiquent qu'elles restent sous-utilisées[ii]. L'amélioration de l'accès pour ce groupe cible d'usagers universitaires est donc de plus en plus essentielle. Afin de comprendre les besoins de ces usagers et faciliter leur interaction avec les sources, une étude a été menée auprès de professeurs et d'étudiants en sciences humaines de l'UC. Deux des principaux besoins identifiés ont été l'amélioration des résultats de recherche et la disponibilité de contextes, en particulier pour les sujets historiques.

Cet article examine comment les données liées[iii] pourraient répondre à ces besoins. La collection en ligne de documents sur la Première guerre mondiale de l'UC a été employée comme banc d'essai. En plus de représenter les métadonnées de la collection en tant que données liées, notre objectif est d'insérer des liens profonds entre les différentes données historiques sur l'expérience des civils en Belgique occupée pour démontrer le type de questions complexes pour lesquelles une réponse peut être trouvée et démontrer les méthodes automatisées employées dans un domaine spécialisé. Cette approche peut répondre aux besoins des usagers en liant les concepts connexes des sources à l'aide de termes tirés d'un vocabulaire spécialisé, en les enrichissant avec des ressources externes et en offrant des services sémantiques riches qui rendent les usagers plus autonomes.

II. DEFIS DU TRAVAIL AVEC DES SOURCES PRIMAIRES EN LIGNE

Les sources primaires sont devenues une partie incontournable des études en sciences humaines aux premier et deuxième cycles, particulièrement en histoire.[iv] En effet, l'utilisation de sources primaires en ligne dans la salle de classe est considérée comme fondamentale aux approches pédagogiques actuelles qui favorisent la pensée critique et l'apprentissage constructiviste basé sur l'enquête. Lee et Clarke, par exemple, expliquent que "the nonlinear shape of the Web can serve as a lever to encourage students to deal with the multiple sequences, voices, outcomes, and implications of historical narrative."[v] Les sources primaires en ligne offrent aussi des avantages distinctifs pour la recherche par rapport aux formats analogues, principalement puisqu'ils sont plus accessibles, souples et faciles à manipuler pour la recherche.

Bien que les collections numérisées mises à disposition par les institutions du patrimoine culturel constituent une source très riche et précieuse de matériel pour l'enseignement, l'apprentissage et la recherche en sciences humaines, elles sont sous-utilisées par les professeurs et les étudiants pour diverses raisons. Comme la plupart de ces usagers se fient sur Google, la décontextualisation des sources, l'impénétrabilité des bases de données institutionnelles et la quantité de résultats représentent des obstacles majeurs à l'utilisation de ces ressources. Compte tenu des problèmes de résultats de recherche et de repérage des sources numériques primaires, il n'est guère surprenant que plusieurs études aient identifié une exigence croissante de simplification et de granularité de la recherche par sujet parmi et dans les documents et les collections de ces bases de données.[vi]

III. ÉVALUATION DES BESOINS DES USAGERS

Une étude sur les besoins des utilisateurs a été menée à la CU en janvier 2011[vii] afin de développer un outil éducatif numérique centré sur l'utilisateur pour le travail en ligne avec des sources primaires. Son but était de comprendre les besoins des usagers en sciences humaines et de faciliter leur interaction avec les sources primaires en fonction de leur rétroaction directe. L'étude a été réalisée en analysant plus de 20 entretiens semi-structurés avec des professeurs, des étudiants diplômés et des étudiants de premier cycle représentant un éventail de niveaux d'éducation et de disciplines ainsi que différents degrés de familiarité avec les sources primaires.[viii] Bien que l'étude ait été menée auprès d'usagers académiques, les besoins qu'ils ont exprimés sont représentatifs de ceux de tous les usagers de sources primaires en ligne.

L'étude a confirmé que les professeurs et les étudiants en sciences humaines font toujours face à des défis importants pour trouver et contextualiser les sources primaires en ligne.[ix] Ils ont tendance à ignorer toute la gamme des ressources disponibles et jugent inefficace de chercher dans plusieurs bases de données et sites web pour trouver des sources pertinentes. Une fois qu'ils repèrent une collection dans laquelle chercher, ils rencontrent des difficultés à trouver et à contextualiser les différentes sources et l'information qu'elles contiennent, en particulier pour les sujets historiques.

Les participants ont signalé que les métadonnées bibliographiques sont souvent insuffisantes pour révéler les sources pertinentes, particulièrement les sections selon le sujet, la période et la zone géographique avec la granularité souhaitée. Puisque des concepts similaires peuvent être exprimés de façon différente dans les textes, la recherche par mots-clés est aléatoire. Les sources primaires en ligne sont encore plus sensibles à la décontextualisation, car la recherche par mots-clés encourage les usagers à chercher des fragments d'un document dans lequel un terme donné est mentionné pour ensuite passer à un autre document plutôt que de lire le document dans son intégralité.[x] En outre, les moteurs de recherche et les collections de liens vers des sources en ligne peuvent contribuer à ce problème en désagrégeant des documents de leurs archives d'origine.

La contextualisation des sources primaires est souvent nécessaire pour les rendre suffisamment accessibles afin que les usagers, en particulier les étudiants et les néophytes, puissent interagir avec le matériel. Un contexte peut être composé de relations entre des documents individuels et des ressources qui aident à expliquer comment chaque document, et les informations qu'il contient, s'inscrit dans son contexte historique.[xi] Même avec des sources pertinentes et un

contexte adéquat, les usagers peuvent éprouver des difficultés avec les nouveaux défis inhérents à la recherche de sources primaires: les langues étrangères, le biais des documents, l'utilisation historique, l'orthographe, la grammaire, la paléographie/typographie, etc.[xii] Bien que le travail de recherche et l'utilisation des sources primaires en ligne soient difficiles et prennent un temps considérable, les participants ont convenu que ces sources présentent une occasion unique d'éducation et de recherche.

IV. AVANTAGES DES DONNEES LIEES

En évaluant les différentes options qui pourraient répondre à ces besoins des usagers, l'une des plus prometteuses a été de mettre en place des données liées et des services sémantiques riches pour accroître l'interopérabilité et la facilité d'utilisation des collections historiques numériques. Selon Tom Heath et Christian Bizer, les données liées "refers to a set of best practices for publishing and interlinking structured data on the Web."[xiii] En liant les concepts connexes dans et entre les documents de manière à ce que ce soit compréhensible par les ordinateurs, les données liées permettent (1) l'agrégation des données disponibles dans les sources primaires en ligne, (2) de nouvelles connexions entre elles et de les visualiser d'une façon qui n'était pas possible auparavant et (3) l'enrichissement des données grâce à des liens vers des ressources externes comme DBpedia.[xiv]

Le contenu des documents numériques du patrimoine culturel, le matériel historique en particulier, présente un niveau de complexité qui peut bénéficier de (méta)données sémantiques enrichies et de services utilisateurs intelligents, tant par l'amélioration du repérage que par la présence de contextes enrichis.[xv] Ils permettent la découverte de relations complexes, souvent non linéaires, entre les sujets, les personnes et les lieux enfouies dans les sources, en particulier en lien avec des ontologies et autres vocabulaires spécialisés qui donnent un sens à ces concepts et aux relations entre elles dans un domaine historique donné. Les applications de données liées peuvent également contribuer à atténuer les problèmes habituellement rencontrés lors de l'élaboration d'une stratégie de recherche basée sur les vedettes-matière et les mots-clés dans les sources historiques, par exemple en suggérant des termes de recherche pour affiner et mieux cibler leurs recherches. Par ces moyens, ils facilitent un accès sujet efficace au contenu historique.

Les données liées permettent en outre une plus riche contextualisation des sources en établissant des liens non seulement au sein des collections, mais aussi avec des sources extérieures pertinentes, permettant ainsi l'interopérabilité, le partage et la réutilisation des données des différentes collections historiques. Les données peuvent être présentées de manière à mettre en lumière la structure organisationnelle des collections, ce qui, non seulement préserve le contexte original des documents, mais révèle également les types de matériel disponible. Ces améliorations peuvent compenser les limites bien connues de vedettes-matière et l'imprécision de la recherche par mots-clés, faciliter la comparaison de données historiques à travers temps et espace et permettre le travail dans plusieurs collections. De plus, il est aisé de se servir de ces données liées, ce qui stimule le développement d'applications intelligentes faciles d'emploi qui offrent à l'usager une gamme d'options pour l'analyse et la visualisation des données.

V. PROJET DES DONNEES LIEES SUR LA PREMIERE GUERRE MONDIALE

Des informaticiens de l'Université d'Aalto[xvi] et un spécialiste sujet/expert de discipline de CU sont impliqués dans ce projet. La base de données utilisée est celle de la collection en ligne sur la Première guerre mondiale de CU, qui contient plus de 1 100 titres (55 000 pages) publiés de 1829 à 1922, la majorité des documents ayant été publiés entre 1914 et 1918.[xvii] La provenance de la collection n'est pas tout à fait claire, mais elle a probablement été ajoutée aux collections des bibliothèques universitaires CU dans les années 1920 ou 1930 par le Colorado in WWI Project, entrepris par le professeur d'histoire James Field Willard pour documenter les activités des citoyens et de l'État au cours de la guerre.[xviii] Les publications de la collection proviennent principalement des États-Unis et traitent d'une variété de sujets et de différentes régions géopolitiques, d'un conflit ethnique ou religieux, de l'Empire et des colonies. Plusieurs formats de publications sont représentés, brochures, livres, rapports, discours et cartes. Des négociations sont actuellement en cours pour publier le contenu en données ouvertes liées sous licence Creative Commons 2.0.

Un des principaux objectifs du projet est d'améliorer l'accès sujet dans la collection en ligne et de créer du contexte pour les documents en établissant des liens entre les données de la collection, les ensembles de données intégrées dans le projet et les sources de données externes comme DBpedia et Freebase.[xix] Un autre objectif est de faciliter l'annotation des concepts et d'établir des liens profonds entre les concepts des collections sur la Première guerre mondiale pour créer un sous-domaine historique spécialisé traitant de l'expérience civile en Belgique occupée. Ce sujet a été choisi non seulement parce qu'il était bien représenté dans la collection, mais aussi parce que l'impact de la «guerre totale» sur les populations civiles est un sujet qui intéresse aujourd'hui les chercheurs académiques. La majorité des publications de cette catégorie traitent des épreuves subies par les Belges lors de l'invasion et de l'occupation allemande, en particulier les atrocités telles que les meurtres et les déportations de travailleurs, et de l'impact du régime militaire sur la vie quotidienne. Dans les ensembles de données converties en RDF (Resource Description Framework - format de données liées) à ce jour, on retrouve les métadonnées de la collection (MARC), les vocabulaires normalisés de l'Imperial War Museum (IWM),[xx] de l'information sur les atrocités allemandes en Belgique et la hiérarchie de l'armée allemande.[xxi]

Établis à l'aide d'un vocabulaire spécialisé sur la Belgique au moment de la Première guerre mondiale combiné à une interface utilisateur intelligente, les liens profonds sont conçus pour démontrer les types de questions complexes auxquelles il est possible de répondre pour satisfaire aux besoins des usagers intéressés par ce sous-domaine, telles que: Est-ce que l'ampleur des atrocités perpétrées par les troupes allemandes en Belgique est fidèlement rapportée dans la documentation de la collection? Quelles divisions de l'armée allemande ont été impliquées dans la majorité des incidents? De quelles provinces belges provenaient les déportés? Ce type de fonctionnalité est installé pour permettre une meilleure compréhension des forces qui ont façonné cette période de l'histoire. Compte tenu de la nature très spécifique de ce sous-domaine et le manque d'ontologies existantes, un vocabulaire a dû être créé en adaptant la terminologie et les structures des documents imprimés sur ce sujet,[xxii] en ajoutant les commentaires des historiens du domaine[xxiii] et, dans la mesure du possible, en liant les termes pertinents d'autres ensembles de données inclus dans le projet, par exemple, les métadonnées et les vocabulaires de la collection de l'IWM.

Nous utilisons le système ontologique FinnONTO[xxiv], enrichi d'annotations sur les ensembles de données mentionnés ci-dessus et d'un cadre événementiel général pour la Première guerre mondiale actuellement à l'aide de l'outil d'annotation sémantique SAHA.[xxv] Nous utilisons aussi l'outil ARPA[xxvi] pour automatiser une partie du processus d'annotation. ARPA est un outil d'extraction d'information qui explore automatiquement les entités nommées et les mots-clés de documents textuels. Les annotations suggérées peuvent ensuite être validées et corrigées manuellement à l'aide de l'éditeur SAHA. Enfin, un portail Web sur la Première guerre mondiale est en cours de développement à l'aide du le moteur de recherche à facettes HAKO[xxvii]. Ce portail facilitera la recherche et la navigation par sujets, par noms de personnes, par noms de lieux et par périodes chronologiques et permettra de représenter les données de façon visuelle et et interactive.

Nous avons créé ce cadre de recherche et ces structures sur la Première Guerre mondiale afin qu'ils soient partagés, offrant ainsi la «colle sémantique» qui relie les différents ensembles de données et permet de chercher et de naviguer parmi eux. De plus, la stratégie développée pour ce projet est destinée à être adaptable à d'autres domaines et ensembles de données historiques, en particulier ceux ayant trait à des conflits tels que la guerre civile américaine ou la Seconde guerre mondiale.[xxviii]

VI. CONCLUSION

Nous espérons que la liaison des concepts connexes des ensembles de données sur la Première guerre mondiale en utilisant un vocabulaire spécialisé et en fournissant des services sémantiques riches permettra aux usagers de trouver et d'employer les sources primaires en ligne efficacement. Le centième anniversaire de la guerre générera sans aucun doute beaucoup d'intérêt pour ce sujet, en particulier dans les pays qui ont été impliqués. Nous pouvons profiter de cette occasion pour inciter les usagers à interagir activement avec le passé et avec la multitude de documents numériques que les institutions du patrimoine culturel mettent à leur disposition.

Remerciements Ce travail a été soutenu par le programme Fulbright U.S., la Fondation culturelle finlandaise et Tekes – l'Agence finlandaise de financement pour la technologie et l'innovation.

[i] Les sources primaires sont des documents, objets ou autres éléments de preuve du passé qui ont été créés au moment où les événements historiques ont eu lieu ou par ceux qui ont vécu ces événements. En voici quelques exemples: journaux, lettres, discours, documents gouvernementaux, livres, interviews, photographies, enregistrements audio et vidéo et artefacts.

[ii] D. Harley, "Use and Users of Digital Resources: A Survey Explored Scholars Attitudes about Educational
Technology Environments in the Humanities", *Educause Quarterly* 30, no. 4 (2007): 12-20.

[iii] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, Synthesis Lectures on the
Semantic Web: Theory and Technology, ed. J. Hendler and F. van Harmelen (San Rafael, CA: Morgan & Claypool,
2011). Freely available at: http://linkeddatabook.com/editions/1.0/.

[iv] See, e.g., J.K. Lee, "Digital History and the Emergence of Digital Historical Literacies", in *Technology in Retrospect: Social Studies in the Information Age, 1984-2009*, ed. R. Diem and M.J. Berson (Charlotte, NC: Information Age Publishing, 2010), 78-80, and D. Malkmus, "'Old Stuff' for New Teaching Methods: Outreach to History Faculty Teaching with Primary Sources", *portal: Libraries & the Academy* 10, no. 4 (2010): 414-416.

[v] J.K. Lee and W.G. Clarke, "High School Social Studies Students' Uses of Online Historical Documents Related to the Cuban Missile Crisis", *Journal of Interactive Online Learning* 2, no. 1 (2003): 3.

[vi] M.C. Pattuelli, "Modeling a Domain Ontology for Cultural Heritage Resources: A User-Centered Approach", *Journal of the American Society for Information Science & Technology* 62, no. 2 (2011): 314-342.

[vii] CU est une Carnegie Research University (activité de recherche très élevée) offrant une gamme de programmes de maîtrise et de doctorat en sciences humaines.

[viii] Les participants ont été sélectionnés dans les sept collèges et départements de sciences humaines sur le campus: Architecture et planification, Études classiques anglaises, françaises et italiennes, Histoire, Musique et Études religieuses qui offrent tous des programmes de doctorat.

[ix] Ce résumé n'inclut pas tous les besoins des usagers identifiés par cette étude, mais plutôt ceux qui ont trait au sujet de cet article. Pour plus de détails, voir T. Lindquist and H. Long, "How Can Educational Technology Facilitate Student Engagement with Online Primary Sources?: A User Needs Assessment", *Library Hi Tech* 29, no. 2 (2011): 224-241.

[x] J. Garrett, "KWIC and Dirty? Human Cognition and the Claims of Full-Text Searching", *Journal of Electronic Publishing* 9, no. 1 (2006), available at: http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;cc=jep;q1=garrett;rgn=main;view=text;idno=3336451.0009.106 (accessed 13 February 2012).

[xi] Une étudiante a donné l'exemple suivant du type de contexte qu'elle apprécierait trouver : "J'aimerais trouver plus d'information complémentaire et de contexte dans les sources primaires avec lesquelles je

travaille. [Plusieurs collections en ligne de sources primaires] présentent seulement la source sans fournir de contexte, par exemple la mention qu'une lettre a été livrée pendant une épidémie de choléra." (Lindquist and Long, 233).

[xii] T. Lindquist and H. Wicht, "Pleas'd By a Newe Inuention?: Assessing the Impact of Early English Books Online on Teaching and Research at the University of Colorado at Boulder", *The Journal of Academic Librarianship* 33, no. 3 (2007): 347-360.

[xiii] Heath and Bizer, chap. 2, "Principles of Linked Data", accessed 23 May 2012: http://linkeddatabook.com/editions/1.0/#htoc8.

[xiv] DBpedia est une version en données liées de Wikipedia (http://www.dbpedia.org/).

[xv] E. Hyvönen, "Semantic Portals for Cultural Heritage", in *Handbook on Ontologies*, 2d ed., ed. S. Staab and R. Studer, International Handbooks on Information Systems (Berlin: Springer, 2009).

[xvi] Semantic Computing Research Group, voir http://www.seco.tkk.fi/.

[xvii] Voir http://libcudl.colorado.edu/wwi/index.asp.

[xviii] Ce matériel représente la base de la collection historique de l'Université du Colorado ainsi que celle des Archives universitaires (David M. Hays, "The History of the Archives, University of Colorado at Boulder Libraries, 1917-2011" [texte non publié, Archives, University of Colorado Boulder Libraries], 1-2).

[xix] Voir http://www.freebase.com/.

[xx] Ces mots-clés désignant des événements relatifs à la Première guerre mondiale sont approuvés, les mots-clés géographiques relatifs au front occidental approuvés sont basés sur la taxonomie Getty TGN et agrémentés par des termes relatifs aux collections de l'IWM et des termes de la taxonomie thématique de l'IWM sur la Première guerre mondiale. Merci à l'Imperial War Museum de nous avoir permis d'utiliser ces vocabulaires.

[xxi] Nous désirons remercier sincèrement John Horne et Alan Kramer du Trinity College à Dublin, who gathered and analysé les données sur les atrocitéset qui nous a permis de les utiliser dans le cadre du projet (J. Horne and A. Kramer, *German Atrocities, 1914: A History of Denial* [New Haven: Yale University Press, 2001], Appendix 1, 435-439).

[xxii] **P. Lefèvre and J. Lorette, eds.,** ***La Belgique et la Première Guerre mondiale: Bibliographie*****, 2 vols. (**Bruxelles: Musée Royal de l'Armée, 1987-2001).

[xxiii] Nous remercions aussi sincèrement Martha Hanna (University of Colorado Boulder), Sophie de Schaepdrijver (Pennsylvania State University) and Tammy Proctor (Wittenberg University) pour leurs suggestions.

[xxiv] Voir http://www.seco.tkk.fi/projects/finnonto/.

[xxv] Pour consulter la documentation et trouver le code source de téléchargement, voir http://www.seco.tkk.fi/services/saha/.

[xxvi] Voir http://www.seco.tkk.fi/services/arpa/.

[xxvii] Voir http://www.seco.tkk.fi/tools/hako/.

[xxviii] La nature du contenu facilite les liens basés sur les concepts et les structures militaires tels que les différentes branches des forces militaires, des régiments et des batailles.