„Regards neufs et inédits sur les bases de données“, 10 July 2014, Digital Humanities Conference 2014

Paper Session 3
Chair: Mark Algee-Hewitt

Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie
Roe, Glenn; Gladstone, Clovis; Robert Morrisey
Lien vers le paper: http://dharchive.org/paper/DH2014/Paper-171.xml

Seeing the Trees & Understanding the Forest
Montague, John Joseph; Rockwell, Geoffrey; Ruecker, Stan; Sinclair, Stéfan; Brown, Susan; Chartier, Ryan; Frizzera, Luciano; Simpson, John
Lien vers le paper: http://dharchive.org/paper/DH2014/Paper-924.xml

Trading Consequences: A Case Study of Combining Text Mining & Visualisation to Facilitate Document Exploration
Hinrichs, Uta; Alex, Beatrice; Clifford, Jim; Quigley, Aaron
Lien vers le paper: http://dharchive.org/paper/DH2014/Paper-373.xml

Recensé par : Floraine Stauffer

A son arrivée, le premier jour de la conférence Digital Humanities, au Swiss Convention Center de l'EPFL, l'auditeur recevait une batterie de secours. L'objet laissait deviner le caractère technique des présentations à venir. Entre fascination et scepticisme, effervescence et réflexion, les cinq jours de colloque confirment le large intérêt que suscitent les humanités digitales. Des salles combles forcèrent même certains à assister aux conférences depuis l'embrasure des portes.
Les trois projets présentés en cet après-midi du mercredi 9 juillet remplissent facilement la salle 414 de l'Amphimax. Nous nous proposons, ici, de faire le compte rendu de ces trois présentations, portant toutes sur la fouille de texte (*text mining*) et sur les nouvelles informations que l'on peut tirer d'une base données.

Contrebande d'opinion dans l'Encyclopédie

La présentation débute avec un projet portant sur la recherche de discours clandestins dans l'Encyclopédie[1] de Diderot et d'Alembert. Menée par une équipe anglophone constituée de spécialistes de la littérature française, la recherche vise à révéler les discours subversifs contenus dans des articles à priori neutres. D'après les auteurs de la recherche, **GLENN ROE**[2] (The Autralian National University, Australie), **CLOVIS GLADSTONE** et **ROBERT MORRISEY**[3] (University of Chicago, USA) certains passages de l'Encyclopédie ou *Dictionnaire raisonné des sciences des arts et des métiers* auraient pu servir à la contrebande d'opinions controversées[4].
Pour localiser ces opinions dans la base de données extraite de l'Encyclopédie[5], les trois chercheurs ont choisi d'interroger la classification originale des articles. L'arbre de classification des connaissances humaines établit par Diderot et d'Alembert est supplanté par un outil numérique: le LDA (Latent Dirichlet Allocation). Cet outil probabiliste permet d'identifier, dans un corpus de texte, les discours gravitant autour d'un concept. C'est la technique du modèle de sujet (*topic modeling)*. En identifiant une série de mots-clés censés représenter un concept et en comparant les co-occurences de ces mots-clés dans un corpus de texte, les résultats du modèle de sujet proposent une nouvelle classification des articles en fonction de leurs liens au concept. Lorsque le LDA analyse un discours sur la biologie, la majorité des articles traitant

1 http://dharchive.org/paper/DH2014/Paper-171.xml
2 Site personnel : http://www.glennroe.net/
3 http://rll.uchicago.edu/faculty/morrissey Consulté le 12.08.2014
4 Termes utilisés pas les auteurs. Voir: http://dharchive.org/paper/DH2014/Paper-171.xml consulté le 19.07.2014
5 http://encyclopedie.uchicago.edu/http://encyclopedie.uchicago.edu/ Consulté le 13.08.2014

de ce thème se retrouve, sans surprise, dans la classification originale « biologie »[6]. D'autres discours, en revanche, se retrouvent liés à des catégories moins évidentes. Par exemple et selon les algorithmes, le discours sur le « *droit naturel* » est présent dans plus de 60 articles classés par Diderot et d'Alembert dans la catégorie « grammaire ». Il y aurait de ce fait une partie de la pensée des Lumières cachée dans des pages de grammaire.
Parmi les articles qui, selon les probabilités établies, se rattachent le plus au discours sur le « *droit naturel* », on trouve celui intitulé « inviolable ». Dans la définition d'inviolable, une phrase semble justifier ce rattachement : « La liberté de conscience est un privilège inviolable »[7]. L'on saisit, ici, la subtile allusion au fait que « la liberté de penser est un droit naturel parmi d'autres et qu'il est inaliénable »[8]. Autre exemple, le discours concernant la moralité est lui très présent dans l'innocente catégorie « géographie ancienne ». La pensée des Lumières trouve là un lieu où placer ses critiques face au mauvais gouvernement et à la tyrannie.
Autre fonctionnalité du *modèle de sujet:* la possibilité d'observer l'évolution de ces discours dans le temps long. Spécialité du Prof. Morrisey, l'idée est de comparer la présence d'un certain discours dans chaque volume. La contextualisation des résultats reste primordiale pour en comprendre la portée. Par exemple, sur les onze volumes qui constituent l'Encyclopédie, le discours traitant du « Commerce » diminue de publication en publication alors que le discours traitant de l'«Économie politique » augmente dans chaque volume. Quant à la question, posée par le public, du choix des discours analysés, les chercheurs reconnaissent ne pas avoir tout analysé mais s'être focalisés sur les sujets qui les intéressaient le plus. Libre donc aux historiens ou aux linguistes de choisir, à leur tour, les discours qui les intéressent et de partir explorer la face cachée de l'Encyclopédie.

Offrir aux chercheurs une carte à explorer

JOHN MONTAGUE (University of Alberta, Canada), professeur en communication visuelle et représentant d'un projet plus théorique, agrémente sa conférence de citations. Les textes synthétisent le propos. Au cœur du projet canadien[9], la question de la visualisation des données. Comment rendre les données massives (big data) pertinentes et intelligibles? Pour les chercheurs canadiens, les outils graphiques utilisés actuellement sont trop complexes pour être utiles. La quantité d'information que fournissent simulanément ces visuels les rend excessivement denses. Utilisés seuls, ils manquent de pertinence. Les dendrogrammes (arbres hiérarchiques) par exemple, ne forment qu'une masse illisible d'interconnexions. Les nuages de mots-clés (*word coulds*) eux, ne permettent pas de comparer les textes. Les moyens de visualisation actuels sont limités.
Pour éviter de perdre l'utilisateur dans une masse floue, deux pistes ont été suivies. La première est d'offrir la possibilité à l'utilisateur de se déplacer dans les données et de modifier lui-même le visuel. La deuxième consiste à assembler deux moyens de visualisation en un seul outil. Pour illustrer le propos, l'assistance découvre, sur l'écran de présentation, un espace en 3D parsemé de nuages de mots-clés. L'utilisateur est invité à se déplacer dans cet espace. Le déplacement faisant apparaître de nouveaux nuages et disparaître d'autres. « Plus nous encourageons les utilisateurs à explorer et à jouer avec les données, plus il est probable qu'ils développent des idées utiles »[10]. Pour dépasser la simple mise en image de ce que nous savons déjà et mettre à profit l'analyse en grande quantité que propose le numérique, il faut, selon les chercheurs, miser sur l'apprentissage par l'interactif. Utiliser un type de visualisation pour en explorer un autre et permet de donner sens à la forêt que sont les données massives.

6 Le topic modeling ramène ainsi la plus part des discours (75%) à leur classification initiale.
7 http://dharchive.org/paper/DH2014/Paper-171.xml conuslté le 21.07.2014
8 Idem
9 http://dharchive.org/paper/DH2014/Paper-924.xml Consulté le 19.07.2014
10 http://dharchive.org/paper/DH2014/Paper-924.xml Consulté le 19.07.2014

Testé par les historiens

« Nous voulions extraire et montrer de façon différente pour les historiens les informations contenues dans notre base de données »[11]. Elles sont deux, **UTA HINRICHS**[12] (University of St-Andrew, Écosse) et **BEATRICE ALEX** (University of Edinburgh, Écosse) à venir expliquer un projet pluridisciplinaire. Le but visé est de rendre accessibles et exportables les documents d'une base de données sur les ressources naturelles du Canada. Le résultat se présente sous la forme d'une interface. Celle-ci permet de choisir un produit commercial et de visualiser plusieurs informations à son sujet. La première est une carte des contextes dans lesquels les documents contenant le mot ont été produits. La deuxième est un graphique à barres représentant la répartition temporelle des documents dans la collection. La dernière information affichée est un classement des documents disponibles relatifs au sujet. L'exploration peut également être menée par région géographique.
L'outil a été testé par des historiens de l'environnement à Vancouver. Leur premier réflexe fut de vérifier si les informations sur les événements qu'ils connaissaient déjà étaient valables. Ils ont ainsi testé la fiabilité de l'outil avant d'avancer d'avantage dans sa découverte. Des améliorations ont ensuite été apportées pour répondre aux manquements constatés par les historiens. Pour éviter la confusion et les mauvaises interprétations, il a été décidé de donner d'avantage d'informations sur l'origine des données et de donner plus d'indications sur la façon dont l'outil fonctionne. Ces explications n'ont pas pour autant calmé les esprits. A la fin de la présentation, les questions fusent dans le public. Est-il possible de savoir si le lieu mentionné est le départ ou la destination du produit ? Peut-on comparer deux produits ? Les réponses laissent supposer que les développements futurs sauront convaincre les utilisateurs : la recherche des destinations et les comparaisons sont des pistes suivies par les chercheuses. Enfin, les présentatrices ont rappelé qu'il était avant tout important pour elles que le domaine scientifique visé bénéficie et « s'enrichisse » au contact d'un tel outil.

Suite

De quoi s'enrichit-on alors ? Que peut-on tirer, en tant qu'historien, de ces présentations ? Tout d'abord le fait que, jusqu'à présent, ce sont les textes et les bases de données qui servent de base au travail des chercheurs dans les domaines de la littérature et de l'histoire. Ensuite, que le développement de ces outils numériques est à ses débuts. Les questions du public sur les limites et les applications concrètes de ces outils démontrent bien le caractère embryonnaire de certaines applications. De plus, pour être sûr des explications extraites d'un travail en humanités digitales, il faut que les chercheurs des sciences humaines comprennent le fonctionnement de l'outil qu'ils utilisent. Enfin et surtout, retenons que les développements futurs des humanités digitales laisse grande place à l'imagination et à la nouveauté. Nous verrons donc, par la suite, le tournant que prendra l'aventure.

Floraine Stauffer
Université de Neuchâtel
floraine.stauffer@bluewin.ch

11 http://dharchive.org/paper/DH2014/Paper-373.xml Consulté le 19.07.2014
12 Site personnel http://www.utahinrichs.de/