



Google Books & Open Book Alliance :

Rapport méthodologique de recherche web



Réalisé par Aude Jurenak et Debora Orlando
Semestre d'hiver 2009-2010

Table des matières

1	Introduction	3
2	Présentation du sujet.....	3
3	Problématique.....	4
4	Déroulement du travail.....	4
4.1	Recherche d'informations générales sur le sujet	4
4.2	Elaboration d'un schéma heuristique	4
5	Méthodologie de recherche	6
5.1	Outils de recherche	6
5.1.1	Documents WWW	6
5.1.2	Documents multimédia	6
5.1.3	Flux RSS	7
5.2	Terminologie utilisée	7
5.3	Différences entre la recherche Web et la recherche dans les banques de données	8
6	Organisation du travail	8
6.1	Sélection des documents.....	8
6.2	Netvibes et Google Groupes.....	9
6.2.1	Mise en page sur Netvibes	9
7	Commentaires sur le travail effectué.....	10
8	Conclusion.....	11

1 Introduction

Dans le cadre d'un travail de recherche en Information documentaire réalisé à la Haute Ecole de Gestion de Genève, nous nous sommes penchées sur la problématique de la numérisation d'une partie des fonds de la Bibliothèque Nationale Suisse. Pour ce faire, nous avons analysé l'offre des deux grands acteurs de ce domaine : l'Open Content Alliance et Google Books, afin d'offrir un dossier qui fasse le point et qui permette de comprendre les objectifs et les conditions de ces deux sociétés. Nous avons réalisé une plateforme Netvibes pour mettre en avant et organiser les résultats de nos recherches sur internet, disponible à l'adresse www.netvibes.com/doaj.

2 Présentation du sujet

En décembre 2004, Google annonce la création de Google Print, une nouvelle fonction qui va permettre d'avoir accès en ligne à des documents numérisés par leurs soins. Pour cela, ils signent des accords avec différentes bibliothèques américaines qui donneront accès à une grande quantité d'ouvrages anciens (qui ne tombent plus sous le droit d'auteur), mais aussi des œuvres orphelines. A cause des droits d'auteurs, certains ouvrages ne sont disponibles que partiellement. De plus, les ouvrages entre 1839 et 1922 sont rarement accessibles en entier pour les internautes basés hors des Etats-Unis.

Google Print, désormais appelé Google Books, n'a cessé de se développer depuis sa création et parle aujourd'hui de 10 millions de livres qui ont été numérisés et qui sont mis à disposition des internautes. Ils souhaitent apporter l'accessibilité à la connaissance à tous et devenir la plus grande bibliothèque numérique au monde, porteuse du savoir humain.

En réponse à ce projet a été créée l'Open Content Alliance en 2005. Ce consortium d'organisations - dont font partie Yahoo ! et Microsoft - contribue à rendre accessible des livres numérisés par le biais de la plateforme Internet Archive. Cette dernière est une organisation non lucrative, fondée en 1996 dans le but de créer une bibliothèque numérique libre en ligne.

En 2007, Google Books est accusé d'avoir violé les droits d'auteurs de certains ouvrages numérisés et mis à disposition du public. Il va donc trouver un accord avec l'Author's Guild et l'Association of American Publishers en 2008. Cet accord n'est cependant pas au goût de tout le monde, et un groupement de professionnels, éditeurs, bibliothécaires et concurrents de Google (Microsoft, Yahoo ! et Amazon) s'est formé sous le nom d'Open Book Alliance afin de s'opposer à ce qu'ils appellent le monopole de Google dans la numérisation.

3 Problématique

Afin de savoir à qui confier le travail de numérisation des fonds, nous nous sommes intéressées aux conditions de numérisation que proposent Google Books et l'Open Content Alliance, mais aussi à leur politique de numérisation. Nous prenons également en compte ce qu'ils ont déjà effectué comme partenariats en Europe. Ceci nous permettra d'analyser au mieux la situation actuelle afin de conseiller la BNS à faire le meilleur choix possible.

4 Déroulement du travail

4.1 Recherche d'informations générales sur le sujet

Lors de la réception de notre sujet, nous avons dans un premier temps voulu analyser le thème choisi au mieux avant de nous lancer dans des recherches précises. Pour cela, nous avons redéfini notre problématique et nous en avons tiré les premiers mots-clefs. Nous avons ensuite effectué une recherche générale sur le sujet. Cela nous a permis de mieux cerner les acteurs de notre problématique et également de découvrir que notre sujet était vaste et l'actualité sans cesse renouvelée.

Pour planifier notre travail dans le temps, nous avons élaboré un petit calendrier comprenant les différentes étapes de recherche en lien avec les cours prévus à cet effet ainsi que la phase de rédaction et les dates butoirs.

Toutes ces informations nous ont permises de rédiger le rapport intermédiaire remis dans le cadre du cours sur les industries culturelles le 27 octobre 2009.

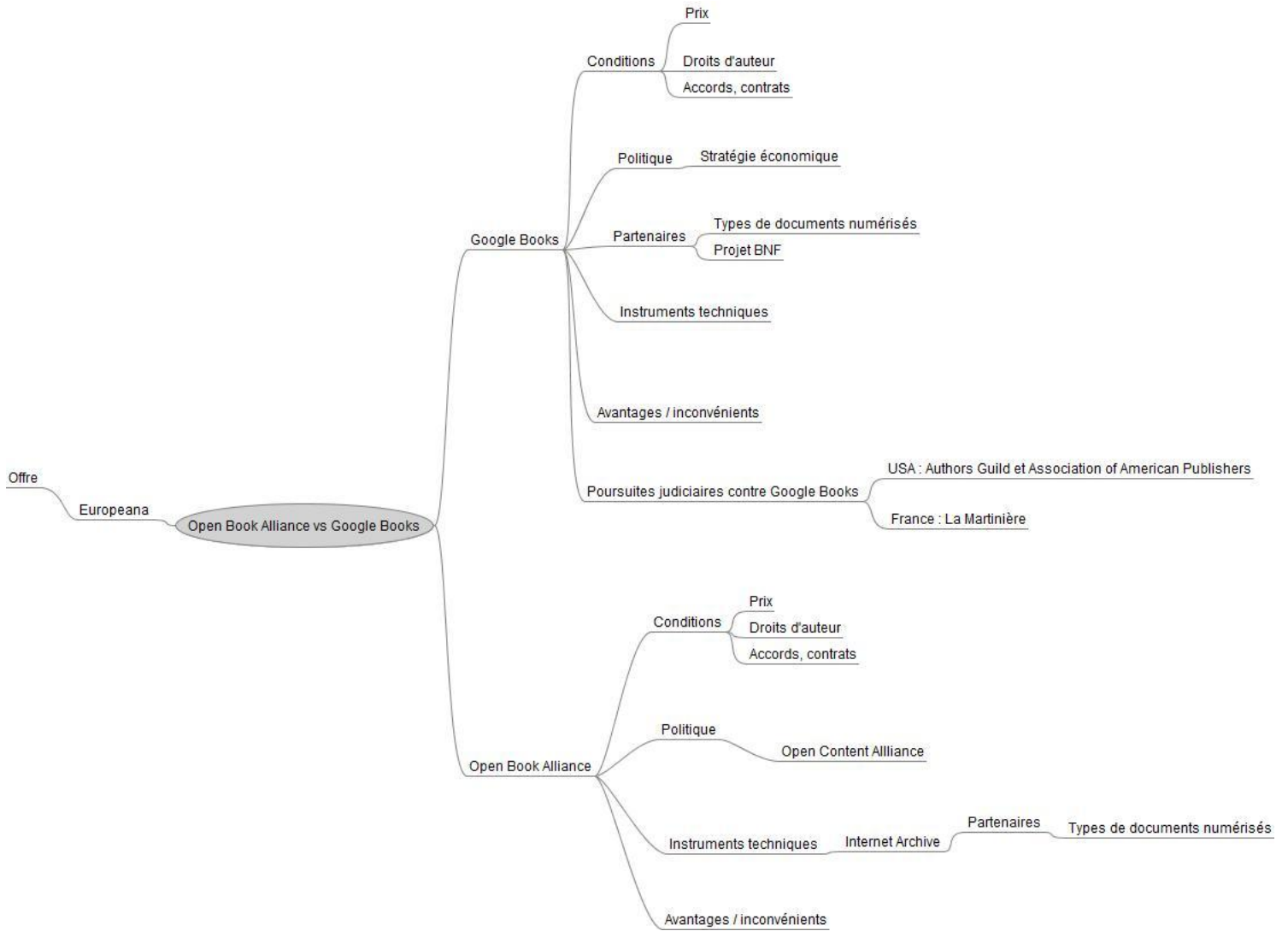
4.2 Elaboration d'un schéma heuristique

Afin de garder une certaine ligne directrice dans nos recherches et de rendre compte de notre orientation à Mme Rezzonico pour ce cours de recherche d'information sur le web 2.0 et internet, nous avons créé un schéma heuristique de ce que l'on a souhaité couvrir dans nos recherches de manière générale.

Pour cela, nous avons dans un premier temps ressorti les acteurs principaux : Google Books et l'Open Book Alliance. Nous avons également rajouté dans une moindre mesure Europeana qui est une alternative à la numérisation en Europe.

Nous avons ensuite créé différents axes de recherches qui nous semblaient judicieux d'explorer pour avoir une vision globale du sujet.

Schéma heuristique de nos recherches :



5 Méthodologie de recherche

5.1 Outils de recherche

Nous avons démarré nos recherches générales avec le moteur de recherche de Google. Nous sommes également parties des sites internet de l'Open Book Alliance et de l'Open Content Alliance ainsi que de Wikipédia pour récolter des informations afin de mieux cerner le sujet.

Dans un deuxième temps, nous avons utilisé plusieurs moteurs de recherche dans un souci d'exhaustivité, comme Bing, Yahoo !, mais aussi le métamoteur Metacrawler. Les outils de catégorisation (Clusty et Exalead) ont également été utiles pour approfondir nos recherches en fonction des sujets proposés. Il a alors été nécessaire de préciser plus finement les équations. Par exemple :

(open book alliance OR open content alliance OR internet archive) AND (digitiz* OR digitaliz* OR scan* OR numéris*) AND (library* OR bibliothe*)

5.1.1 Documents WWW

Dans les moteurs de recherche cités ci-dessus, nous avons parfois utilisé le mode avancé pour restreindre les dates, les régions et les domaines. Nous avons également beaucoup recherché dans Google Actualités.

La navigation de liens en liens nous a également permis de découvrir un grand nombre de documents.

Nous avons utilisé Google Scholar et l'Archive ouverte en Science de l'Information et de la Communication afin de repérer des textes scientifiques.

5.1.2 Documents multimédia

- **Images** : nous avons utilisé la recherche d'images sur Google, Flickr, Bing et Picsearch.
- **Vidéos** : nous avons recherché nos vidéos dans Google Vidéos, Exlalead (catégorie vidéos) et sur Youtube.
- **Podcasts** : n'ayant pas trouvé de podcasts très pertinents grâce aux moteurs de recherche spécialisés, nous avons opté pour une recherche dans l'émission médialogues¹ de la Radio Suisse Romande, qui traite des médias.

¹ <http://www.rsr.ch/la-1ere/medialogues>

5.1.3 Flux RSS

Nous avons retenu des flux RSS trouvés lors de notre recherche de documents (site de l'Open Content Alliance, blogs, site traitant du droit d'auteur, etc.). Lorsque nous avons trouvé des blogs intéressants, notamment via Google Blogs, nous avons préféré conserver les flux RSS de ces pages plutôt qu'un simple lien.

Nous avons également souhaité créer un flux à partir d'une recherche sur Google.com, afin de suivre l'actualité au sujet de Google Books et de l'Open Content Alliance.

5.2 *Terminologie utilisée*

Nous avons choisi d'utiliser des termes permettant la recherche de documents en français et en anglais.

Pour les deux **acteurs principaux**, il a été facile de déterminer des mots-clés puisqu'il s'agit de sociétés américaines qui ont en général le même nom dans toutes les langues :

- GOOGLE BOOK(S)* (SEARCH)/ GOOGLE LIVRE(S)*, GOOGLE PRINT (troncature utilisée pour exprimer des différentes possibilités)
- OPEN BOOK ALLIANCE, OPEN CONTENT ALLIANCE, INTERNET ARCHIVE

Pour exprimer le **contexte**, nous avons ressorti les notions clés de notre sujet de recherche :

- NUMERISER/NUMERISATION* = DIGITALIZE/DIGITALIZATION* = DIGITIZE/DIGITIZATION* = SCAN*
- BIBLIOTHEQUE = LIBRARY*
- DROIT AUTEUR = COPYRIGHT

* troncature utilisée pour exprimer des différentes possibilités

Au cours de nos recherches, nous avons complété ces termes afin de restreindre les résultats, avec des termes géographiques par exemple.

5.3 Différences entre la recherche Web et la recherche dans les banques de données

Même si les termes de recherche se rejoignent, la recherche web est très différente de la recherche dans les banques de données. Nous avons trouvé les informations plus facilement, notamment grâce à toutes les dernières actualités qui étaient disponibles de suite. Cependant, la masse d'informations est beaucoup plus grande. Nous avons donc pris plus de temps à l'évaluer et la trier.

Il a fallu également trouver des informations objectives. En effet, il est plus difficile à savoir si l'information trouvée est fiable. Dans les bases de données, l'information est déjà validée puisque publiée dans la presse ou dans des revues scientifiques.

Par contre, dans le web, nous avons trouvé plus de multimédia, par exemple des images, que nous n'avons pas du tout trouvé dans les bases de données.

6 Organisation du travail

6.1 Sélection des documents

Nous avons pris différents critères de sélection pour nos documents :

- **Date** : nous avons privilégié les sources récentes, car certaines datant de plusieurs années ne sont plus d'actualité. Cependant, d'autres documents restent à jour aujourd'hui. Nous avons donc dû évaluer chaque cas séparément.
- **Sujet** : beaucoup d'informations sont disponibles sur ce thème sur le Web, ce qui nous a aidées à comprendre notre sujet. Nous avons ensuite dû sélectionner les documents se rapportant précisément à la problématique de la numérisation des fonds de bibliothèques par Google Books et l'Open Content Alliance, en laissant de côté tout ce qui s'en éloigne.
- **Objectivité** : beaucoup d'articles prennent une position ferme en faveur ou contre Google. Nous avons tenté d'éliminer les contenus trop subjectifs afin d'arriver à une certaine objectivité des informations choisies.
- **Unicité des informations** : la documentation disponible est très abondante, mais les informations sont souvent répétitives. Afin d'éviter une certaine redondance, nous avons sélectionné des contenus uniques, représentatifs de plusieurs sources différentes.
- **Fiabilité des sources** : lorsque c'était possible, nous avons tenté de sélectionner des documents de sources sérieuses afin de garantir une certaine crédibilité et une fiabilité.

6.2 *Netvibes et Google Groupes*

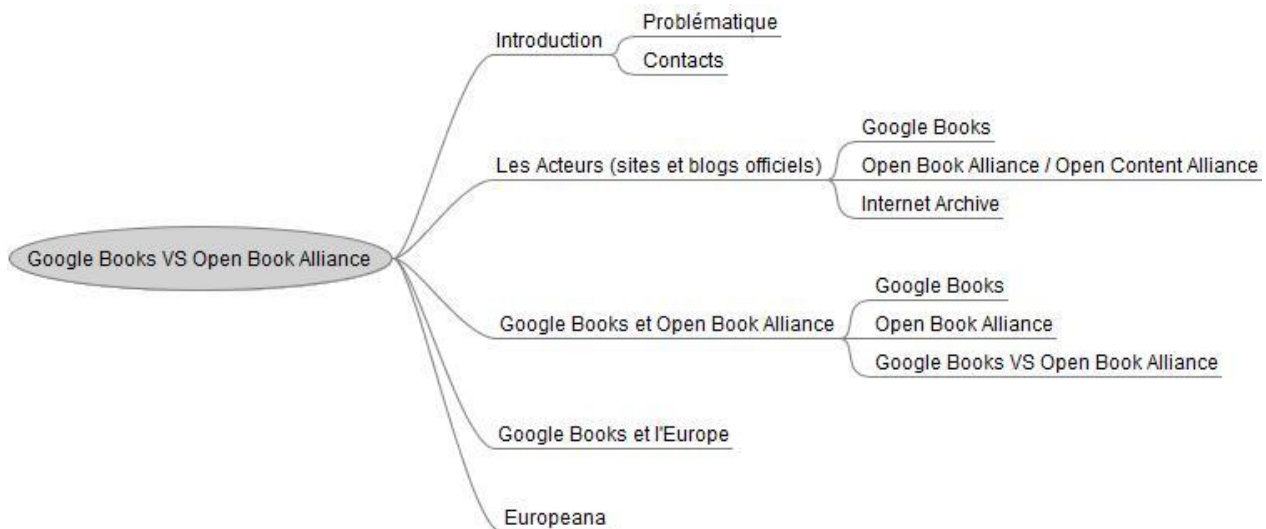
Pour mettre en avant notre travail de recherche sur internet, nous avons utilisé la plateforme Netvibes. Nous avons dans un premier temps décidé de garder tous les liens que nous trouvions intéressants sur le sujet, que nous avons réparti dans différents onglets.

Dans le même temps, nous avons créé un Google Groupe afin de se partager nos différents fichiers et permettre des discussions entre les membres du groupe. Cela nous a également permis de créer une page dédiée à nos recherches sur le web, où nous y avons intégré toutes nos équations de recherche et nos commentaires sur les différents articles trouvés.

6.2.1 Mise en page sur Netvibes

Dans un deuxième temps, nous avons dû procéder à une sélection de nos informations et réfléchir à notre page publique. Pour ce faire, nous avons décidé de créer cinq onglets principaux dont certains ont été partagés en sous-thèmes (qui sont l'équivalent des colonnes).

Voici le schéma heuristique de notre Netvibes :



Chaque onglet est composé d'une note explicative sur le thème. Chaque lien est ensuite commenté en quelques lignes.

Lors de nos premières recherches avons trouvé les différents sites et blogs officiels des acteurs principaux de notre problématique. Nous avons trouvé intéressant de créer un onglet qui leur est consacré (« Acteurs ») si l'on a besoin d'avoir des informations plus précises sur les activités de chacun.

Nous avons ensuite réalisé des recherches plus approfondies qui nous ont permis de créer deux nouveaux onglets. Le premier, « Google Books et Open Book Alliance », contient des sources sur ces deux acteurs et les enjeux qui se cachent derrière.

Dans l'onglet « Google Books et l'Europe », nous nous sommes intéressées aux difficultés qu'a pu rencontrer Google Books pour numériser en Europe, mais aussi les accords qu'il a pu passer avec des bibliothèques européennes et le marché que représente l'Europe pour lui.

Nous avons finalement créé un onglet « Europeana », qui ne rentre pas directement dans notre problématique, mais qui est une alternative intéressante que nous nous devons de citer et d'approfondir un peu.

7 Commentaires sur le travail effectué

La première difficulté que nous avons rencontrée a été de cerner la différence entre l'Open Book Alliance, citée dans notre sujet de recherche, et l'Open Content Alliance. Une fois nos premiers résultats analysés, nous avons pu inclure les deux termes à nos équations de recherche, puisqu'ils sont liés et complémentaires.

Grâce à une présentation de M. Alain Jacquesson au sujet de Google Books dans le cadre du cours sur les bibliothèques numériques, nous avons bénéficié de certaines pistes de recherche qui nous ont aidées dans notre travail.

Un des problèmes majeurs concernant la recherche d'informations sur ce sujet est la rapidité avec laquelle l'actualité change. Les récents événements de Google Books avec la justice aux Etats-Unis et en France ont contribué à cet effet. Il est alors difficile de faire le tri dans ce vaste sujet, où un seul aspect ne peut être séparé du reste et où les informations deviennent vite obsolètes.

Nous avons également rencontré des difficultés à trouver des informations sur l'Open Book Alliance en Europe, car cette organisation agit seulement aux Etats-Unis (sauf à la British Library par l'intermédiaire du projet Live Book Search de Microsoft qui a duré jusqu'en 2008).

8 Conclusion

Effectuer nos recherches pour notre travail sur le web a représenté un complément indispensable à la recherche dans les banques de données. En effet, même si la recherche demande plus d'imagination et d'efforts, internet contient un grand nombre de ressources indispensables. De plus, l'information prend tout type de formes et nécessite une évaluation en fonction des sources et de l'actualité.

Organiser nos résultats à l'aide d'une plateforme Netvibes nous a permis de les exploiter au maximum. En plus de regrouper et classer les informations en facilitant leur accessibilité, ce portail nous a permis de tirer profit des avantages du web 2.0, avec par exemple les différents flux RSS et les alertes, afin de suivre de près l'actualité sur notre sujet, qui bouge énormément.

Cette webographie nous a été utile tout au long de notre travail et nous permet aujourd'hui d'avoir une vue d'ensemble sur les questions de numérisation des fonds de bibliothèques par l'Open Content Alliance et Google.