Méthodes computationnelles pour analyser des collections numériques

Alexandre Honnis

Mémoire Master 1

Information Documentation

Méthodes computationnelles pour analyser des collections numériques.

Alexandre Honnis

Soutenance : 4 juillet 2023

Encadrement : Amel FRAISSE, Maîtresse de conférences en SIC, Université de Lille.
Tuteur professionnel : :Ismail TIMIMI, Maître de conférences en SIC, Université de Lille.

Remerciements

Je souhaite remercier les personnes qui m'ont aidé durant mon stage et ainsi que pour la rédaction de mon mémoire.

J'aimerais tout d'abord remercier Madame Fraisse, Maîtresse de conférences en SIC au sein de l' Université de Lille, pour avoir été ma tutrice de stage mais aussi pour son aide au cours de la rédaction de mon mémoire. Je vous remercie par la même occasion pour votre pédagogie, votre bienveillance et vos conseils durant le stage, mais aussi au cours de l'année.

Je souhaiterais remercier également mon tuteur professionnel Monsieur TIMIMI, Maître de conférences en SIC à l'Université de Lille et responsable du Master IDEMM, qui à toujours répondu présent au cours de l'année scolaire.

Pour terminer, j'aimerais remercier mes proches qui m'ont soutenues et conseillés pendant mon stage et lors de la rédaction de ce mémoire.

Résumé

Ce mémoire de stage porte sur les enjeux et les limites du traitement, de l'exploration et l'exploitation computationnelle des notices bibliographiques multilingues de The Adventure of Huckleberry Finn. Je vais d'abord réaliser un premier chapitre qui va être un Etat de l'Art. Je vais y expliquer le terme de notice bibliographique, le modèle FRBR, ses extensions et le traitement automatique. Ensuite je vais parler des référentiels d'indexation avec l'exemple de LCSH et RAMEAU et pour terminer je parlerais de l'indexation et du vocabulaire contrôlé. Ensuite je vais illustrer à travers quatres exemples le fonctionnement des catalogues en ligne de quatres échelles différentes, L'Université de Lille, national, européenne et mondial. Le troisième exemple servira de transition et va me permettre d'établir un lien avec ma problématique. Je vais évoquer les enjeux et les difficultés dans la rédaction de notices bibliographiques multilingues en prenant comme exemple le projet MACS. Je vais terminer mon mémoire en expliquant le travail réalisé durant le stage, le nettoyage et la constitution du corpus. Je conclurais par analyser les résultats obtenus et faire un parallèle avec la problématique de mon sujet.

Abstract

This internship dissertation deals with the challenges and limits of computational processing, exploration and exploitation of the multilingual bibliographic records of The Adventure of Huckleberry Finn. My first chapter will be a State of the Art. I'll explain the term bibliographic record, the FRBR model, its extensions and automatic processing. Then I'll talk about indexing repositories, using LCSH and RAMEAU as examples, and finally I'll talk about indexing and controlled vocabulary. Then I'll use four examples to illustrate how online catalogs work on four different scales : Lille University, national, European and global. The third example will serve as a transition and will enable me to establish a link with my problems. I will discuss the challenges and difficulties involved in writing multilingual bibliographic records, using the MACS project as an example. I'll conclude my thesis by explaining the work carried out during the internship, the clean-up and the constitution of the corpus. I'll conclude by analyzing the results obtained and drawing a parallel with the problematic of my subject.

Mots-clés :

Indexation, notice bibliographique, multilingue, méthode computationnelle, catalogage.

Keywords :

Indexing, bibliographic record, multilingual, computational method, cataloging.

Introduction générale

"Multilinguism is now an important issue is the field of bibliographic access. The ever increasing growth of Internet access has given widespread access to the catalogues of libraries." Cette phrase provient de la version révisée d'une présentation faite à l'IFLA (International Federation of Library Associations and institutions) à Jérusalem le 17 août 2000 lors d'un atelier sur l'indexation et la classification. Cependant, le texte se concentre sur la mise en place du projet MACS (Multilingual Access to Subjects). [12]. Déjà au début du 21e siècle le traitement des données multilingues est un sujet présent dans le monde des Sciences de l'Information et de la Documentation. L'idée de pouvoir naviguer dans les notices bibliographiques multilingues avec un seul programme qui permettrait de sélectionner un sujet peu importe la langue d'origine. Cette question est toujours d'actualité, avec des catalogues de documents qui sont de plus en plus volumineux et ont une plus grande diffusion sur le Web. Je n'ai pas choisi de commencer mon mémoire avec cette citation par hasard, en effet elle montre qu' il y a déjà plus de 20 ans que la question des notices et de l'émergence du web était abordée dès cette époque par la communauté scientifique.

Le sujet de mon stage porte sur les méthodes computationnelles pour analyser et explorer des collections. Dans le cadre de celui-ci j'ai étudié ces différentes questions, car elles avaient un lien avec mon travail durant ces deux mois. J'ai eu à ma disposition un corpus de données de notice bibliographiques provenant du Catalogue WorldCat. Mon travail consiste à traiter ces données, mais aussi comprendre les problématiques et enjeux de celles-ci. Pour associer le sujet de mon mémoire avec le travail que j'ai réalisé durant mon stage, je vais traiter tout au long de mon mémoire la question des enjeux et des limites du traitement, de l'éxploration et l'exploitation computationnelle des notices bibliographiques multilingues de The Adventure of Huckleberry Finn.

Pour répondre à cette interrogation je vais articuler le développement de mon mémoire autour de quatre chapitres. Tout d'abord un chapitre qui va se concentrer autour de l'État de l'Art des différents thèmes qui vont être traités dans mon mémoire. Je vais expliquer en détail les notices bibliographiques et leurs fonctions. Ensuite je m'attarderai sur le modèle FRBR et ses différentes extensions, expliquant ce que c'est et l'importance de sa mise en place dans le monde de la bibliothéconomie. Ensuite, je vais parler des référentiels d'indexation avec l'exemple de LCSH et RAMEAU et l'importance de leurs utilisations pour permettre une uniformisation des termes. J'aborderais aussi le traitement automatique des données en montrant comment il permet de faciliter le traitement des données. Pour finir dans ce premier chapitre je vais parler de l'indexation et le vocabulaire contrôlé, en expliquant la différence entre les deux et les avantages liés à leurs utilisations. Le second chapitre porte sur la consultation des notices bibliographiques numériques à différentes échelles. Je vais commencer par définir les catalogues et bibliothèques numériques en montrant qu'il y a beaucoup

de points communs, difficiles à différencier. Les quatres parties suivantes de ce chapitre vont décrire et analyser quatre catalogues en ligne à échelle différente. Nous commencerons par l'Université de Lille avec LillOA, puis nous passerons à l'échelle nationale avec Gallica. Nous passerons au niveau européen avec Europeana Enfin nous parlerons de WorldCat qui est le plus grand catalogue en ligne au monde et il est en lien direct avec mon corpus de documents. Mon troisième chapitre est très court et me permet surtout de faire une transition entre la partie théorique et la partie pratique de mon mémoire. Dans cette partie je vais soulever les enjeux et les difficultés dans la rédaction de notices bibliographiques multilingues à travers l'exemple du projet MACS. Le quatrième et dernier chapitre va se concentrer sur le déroulement de mon stage, commençant par une rapide présentation du projet ROSETTA. Ensuite, je vais parler du nettoyage et du prétraitement des données pour que je puisse constituer un corpus. Corpus que je vais décrire dans la deuxième partie pour mettre en lumière certains points. Pour finir je vais montrer les résultats et les visualisations graphiques que j'ai obtenues après avoir analysé les données de mon corpus et montrer le lien avec la problématique de mon mémoire.

Table des matières

Première partie

État de l'Art : Représentation des données bibliographique

1 Les Notices bibliographiques

Afin que cette étude soit au mieux affinée, il me semble indispensable d'expliquer les différentes représentations des données bibliographiques. Dans un premier temps nous allons aborder les notices bibliographiques, indispensables à l'identification et au référencement d'une multitude de documents. Ensuite nous allons poursuivre avec la modélisation des données bibliographiques et plus précisément le modèle FRBR et ses variantes. Pour terminer ce premier chapitre nous allons étudier des référentiels d'indexation à travers les deux exemples de RAMEAU et LCSH.

1.1 Origine et définition

L'origine des notices bibliographiques en France remonte au début de la production et de la diffusion des livres imprimés. Au cours du XVe siècle, avec l'invention de l'imprimerie par Johannes Gutenberg (XIVe,XVe siècle), il y a une multiplication des ouvrages, cela crée la nécessité de les identifier et de les référencer de manière précise.

Les bibliothécaires, les érudits et les libraires ont commencé à élaborer des listes d'ouvrages avec des informations telles que le nom de l'auteur, le titre de l'ouvrage, l'éditeur et l'année de publication. C'est à cette période que l'on commence à élaborer ces notices bibliographiques. Au fil du temps, l'organisation des collections de livres dans les bibliothèques et les besoins croissants en matière de recherche ont conduit au développement de normes et de systèmes de catalogage pour les Notices Bibliographiques, " Avec le développement de la littérature savante est apparue la nécessité de citer dans les publications les documents antérieurs." [6]

En France, l'une des étapes les plus importantes dans l'établissement de normes bibliographiques a été la création de la Bibliothèque nationale en 1537 par François Ier (1494-1547). La Bibliothèque nationale a joué un rôle central dans la conservation et l'organisation des ouvrages, ainsi que dans la création de références bibliographiques cohérentes. L'idée de notices bibliographiques était déjà présente à cette époque mais il n'y a avait aucune mise en place uniforme. Dans l'article de Arlette Boulogne il est expliqué que ; « Les premières règles apparaissent en France dès le XVIIe siècle, avec les recommandations de Dom Mabillon (1632-1701) sur l'emploi de la référence bibliographique (savoir authentifier et dater le document) en science de la diplomatique et, pour les

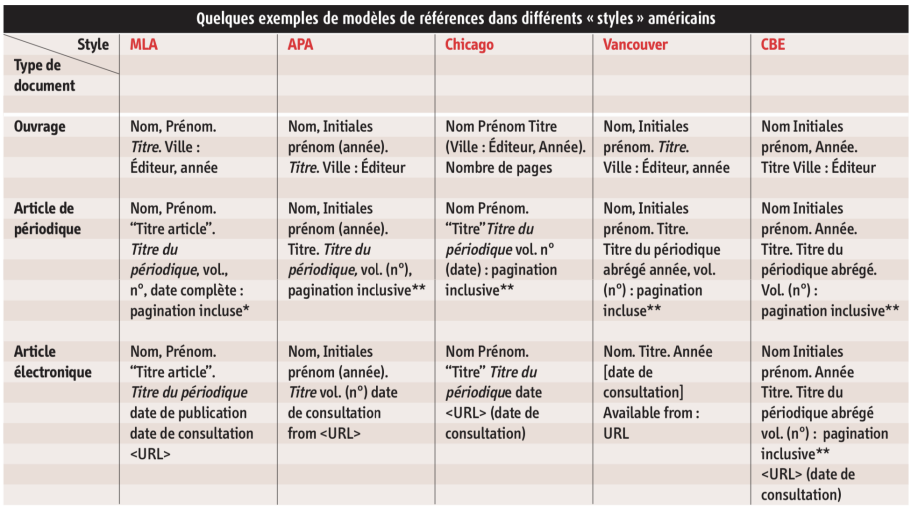
bibliothèques, L'Advis pour dresser une bibliothèque de Gabriel Naudé (1627). Les premiers principes de catalogage sont fixés à la fin du XVIIIe siècle. ». Durant les siècles qui ont suivi, des bibliographies et des bibliothécaires français ont continué à développer et à améliorer les pratiques de référence et de catalogage des ouvrages. Des règles et des formats ont été établis pour que les notices bibliographiques soient le mieux structurer possible, en fournissant des informations standardisées comme par exemple les titres, les auteurs, les éditeurs et les dates de publication. L'avènement de l'informatique et des technologies numériques à bouleverser les notices bibliographiques en France, qui ont connu une évolution majeure. utilisateur.

Aujourd'hui encore , la Bibliothèque nationale de France (BNF) joue toujours un rôle très important dans l'établissement de normes bibliographiques en France. Elle publie régulièrement des recommandations et des guides pour la création et la présentation des Notices Bibliographiques conformes aux normes françaises. Maintenant que nous avons réalisé un point historique sur les origines du terme de notice bibliographique, nous allons définir et expliquer quel est son rôle.

1.2 Utilisation et mise en pratique des notices bibliographiques.

La notice bibliographique se définit selon le Wikinotions INFODOC comme : « Une notice bibliographique comprend l'ensemble des éléments descriptifs d'un document (type de document, auteur, titre, auteur secondaire, édition, lieu de publication, maison d'édition, date de publication, volume, pagination, collection ; informations sur le document hôte pour les articles...) et peut aider à localiser un document. Elle suppose des règles de présentation des informations de chaque notice, associées à des normes (à l'international ISBD, ISO 690, en France AFNOR Z 44-***). La notice bibliographique peut prendre place à l'intérieur d'une bibliographie, d'une base de données ou d'un catalogue. Elle peut comprendre, après une analyse du document, des champs de résumé (dans une bibliographie analytique), de mots-clés, de classification et de localisation (dans le cadre d'une notice catalographique). » Cette définition de Wikinotions INFODOC est très dense mais complète et elle permet de comprendre le principe de la notice bibliographique. Nous avons défini les notices, il est désormais important de comprendre quels sont leurs rôles et leurs utilisations en France.

Les notices bibliographiques jouent un rôle essentiel dans la recherche et la documentation. Elles permettent de référencer et d'identifier de manière précise les sources utilisées dans un travail de recherche, qu'il s'agisse de livres, d'articles, de thèses, de rapports ou d'autres types de documents. Dans un monde ou les documents sont de plus en plus nombreux et leurs exactitudes parfois remises en cause. Il est indispensable de pouvoir référencer et identifier le mieux possible ces sources. Les notices bibliographiques françaises sont la plupart du temps établies conformément aux normes bibliographiques spécifiques utilisées



Quelques exemples de modèles de références dans différents « styles » américains

Style / Type de document	MLA	APA	Chicago	Vancouver	CBE
Ouvrage	Nom, Prénom. *Titre*. Ville : Éditeur, année	Nom, Initiales prénom (année). *Titre*. Ville : Éditeur	Nom Prénom Titre (Ville : Éditeur, Année). Nombre de pages	Nom, Initiales prénom. *Titre*. Ville : Éditeur, année	Nom Initiales prénom, Année. Titre Ville : Éditeur
Article de périodique	Nom, Prénom. "Titre article". *Titre du périodique*, vol., n°, date complète : pagination incluse*	Nom, Initiales prénom (année). Titre. *Titre du périodique*, vol. (n°), pagination inclusive**	Nom Prénom. "Titre" *Titre du périodique* vol. n° (date) : pagination inclusive**	Nom, Initiales prénom. Titre. Titre du périodique abrégé année, vol. (n°) : pagination incluse**	Nom Initiales prénom. Année. Titre. Titre du périodique abrégé. Vol. (n°) : pagination inclusive**
Article électronique	Nom, Prénom. "Titre article". *Titre du périodique* date de publication date de consultation <URL>	Nom, Initiales prénom (année). *Titre* vol. (n°) date de consultation from <URL>	Nom Prénom. "Titre" *Titre du périodique* date <URL> (date de consultation)	Nom. Titre. Année [date de consultation] Available from : URL	Nom Initiales prénom. Année Titre. Titre du périodique abrégé vol. (n°) : pagination inclusive** <URL> (date de consultation)

*Pagination incluse : seuls les chiffres significatifs du numéro de la dernière page sont donnés (exemple : pages 111 à 121 = 111-21, pages 231 à 239 = 231-9)
** Pagination inclusive : de la première page à la dernière page (exemple : pages 111 à 121 = 111-121)

FIGURE 1 – Exemple de modèles de références dans différents styles "Américain"

dans le pays, telles que les normes de l'Association française de normalisation (AFNOR) ou celles de la BNF (Bibliothèque Nationale de France. Il y a évidemment d'autres modèles de référence bibliographiques qui sont utilisés, tels que l'APA (American Psychological Association), la MLA (Modern Language Association) et le Chicago Style par exemple.

La présenations des notices bibliographiques peut varier en fonction du type de document cité. Par exemple, pour des livres, on peut trouver des informations sur l'édition, le nombre de pages, la collection et l'ISBN (International Standard Book Number). Pour les articles scientifiques, on mentionne souvent le titre du périodique, le volume et le numéro de la revue, ainsi que les pages concernées.

En France, la présentation des notices bibliographiques peut varier en fonction du type de document cité. Par exemple, pour des livres, on peut trouver des informations sur l'édition, le nombre de pages, la collection et l'ISBN (International Standard Book Number). Pour les articles scientifiques, on mentionne souvent le titre du périodique, le volume et le numéro de la revue, ainsi que les pages concernées.La citation des sources et la rédaction des notices bibliographiques sont tout de même soumises à des règles très encadrées pour qu'il y ait une certaine cohérence au sein de celle-ci. Ces règles permettent d'assurer une uniformisation des présentations des notices bibliographiques et facilitent la recherche et la consultation des sources citées. Cependant dans l'article [6] on nous explique qu'il y a une grande différence dans les pratiques de référencement entre pays qui posent problème lors de la rédaction ou la lecture de bibliographies.

Pour conclure il est important de comprendre que les notices bibliographiques sont des éléments indispensables dans le domaine de la recherche et de la documentation. Elles sont établies selon des normes spécifiques pour garantir la cohérence et la clarté des références. Les notices bibliographiques facilitent l'identification et la consultation des sources utilisées, renforcent la crédibilité de la recherche et témoignent du sérieux de l'auteur dans ses pratiques de recherche. Pour des étudiants en général mais plus particulièrement dans des cursus liés à l'information et la documentation, les notices sont essentielles dans les différents travaux. Ces notices permettent ensuite de réaliser des références bibliographiques qui sont un modèle d'écriture de ces notices. On peut prendre comme exemple la norme ISO 690 qui est la norme internationale pour la rédaction de références bibliographiques. Cependant il y a une volonté depuis la fin des années 90 de réaliser une nouvelle forme de modélisation bibliographique conceptuelle pour uniformiser les modèles et normes bibliographiques, ce modèle c'est le FRBR.

2 Le modèle FRBR, ses extensions et leurs fusions

Dans cette seconde partie je vais développer le modèle FRBR et ses différentes extensions. En effet cette modélisation conceptuelle présente dans le domaine de la bibliothéconomie q=pour organiser les ressources bibliographiques. Le modèle FRBR à des extensions qui permettent d'étendre son champ d'organisation.

2.1 Le modèle FRBR origine et fonction

Tout d'abord, nous allons parler dans cette seconde partie de ce qu'est le modèle FRBR et comment fonctionne t-il. Le modèle FRBR peut être défini comme une modélisation conceptuelle utilisée dans le domaine de la bibliothéconomie. Son objectif est d'organiser et représenter les ressources bibliographiques de la façon la plus efficace possible, mais aussi de surmonter les limites présentes dans les anciens modèles bibliographiques. [11] Elle à été développé par un groupe d'experts qui à travaillé en partenariat avec l'IFLA de 1992 à 1997. Ce modèle à été approuvé de manière officielle par le Comité permanent de la Section de catalogage de l'IFLA le 5 Septembre 1997 (Source BNF https ://www.bnf.fr/fr/modeles-frbr-frad-et-frsad). L'idée du modèle FRBR est donc de se concentrer davantage sur les besoins de l'utilisateur et moins sur la structure physique des documents.

Le modèle FRBR est d'ailleurs décomposé en quatre entités qui sont distinctes les unes des autres. Pour commencer, les œuvres qui sont les concepts intellectuels à l'origine même des ressources. Ensuite les expressions, elles sont la réalisation de ces concepts. Les manifestations sont les formes prises par ces

concepts, qu'elles soient numériques ou physiques. Enfin pour finir, nous avons les exemplaires qui désignent les exemplaires spécifiques qui se trouvent dans une bibliothèque ou une collection. Il y a évidemment un intérêt à utiliser le modèle FRBR plutôt que les autres modélisations bibliographiques. En effet, ce modèle à pour capacité de fournir une représentation précise et hiérarchisée des relations entre ces différentes entités. Il permet de mieux comprendre la structure d'une ressource bibliographique et facilite la navigation entre les différentes versions d'une œuvre ou les différentes manifestations d'une expression. Cela amène une meilleure accessibilité des ressources pour les utilisateurs et simplifie également la gestion des collections pour les bibliothécaires et les professionnels de l'information.

Le modèle FRBR à d'autres aspects qui sont essentiels et notamment sa reconnaissance de la diversité des ressources bibliographiques. En effet, il prend en compte la variété des formats, des éditions et des versions auxquelles une ressource peut être associée. Tout cela permet de représenter de manière beaucoup plus précise la richesse et la complexité du patrimoine bibliographique. Il est donc important de remarquer que le modèle FRBR offre une meilleure compréhension et contextualisation des ressources. Dans le rapport de Barbabra Tillet elle parle d'une :" proposition de notice bibliographique de niveau national pour tous les types de documents, et les tâches utilisateur associées aux ressources bibliographiques décrites dans les catalogues, bibliographies et autres outils bibliographiques."[17]

Pour terminer je pense qu'il est important de comprendre que le modèle FRBR est un outil fondamental dans le domaine de la bibliothéconomie pour organiser, décrire et représenter les ressources bibliographiques. Il apporte des nouvelles perspectives sur la structure et les relations des entre les notices bibliographiques et d'autorité. Il fournit un vocabulaire plus précis pour favoriser les futurs concepteurs de règles des catalogues et de systèmes à répondre aux besoins des utilisateurs [17].Son approche axée sur les besoins fonctionnels des utilisateurs, sa hiérarchie des entités et sa prise en compte de la diversité des ressources en font un modèle puissant et efficace. Il a un impact significatif sur la manière dont les bibliothèques et les institutions gèrent et rendent accessibles leurs collections, améliorant ainsi l'expérience des utilisateurs et favorisant la diffusion et la préservation du savoir. Bien que le modèle FRBR soit très complet, il y a tout de même des extensions qui permettent d'englober un plus grand nombre de données.

2.2 Les extensions du modèle FRBR : FRAD et FRSAD

La spécificité du modèle FRBR c'est qu'il prend en compte les notices bibliographiques et les parties d'exemplaires. Néanmoins, il manque beaucoup de données qui sont présentes dans les catalogues des bibliothèques. C'est en remarquant cette contrainte que l'IFLA à émis le souhait d'étendre la modélisation à

toutes les informations de ceux-ci. Ce sont deux groupes d'experts qui vont se charger de cette extension. Tout d'abord le groupe FRANAR (Functional Requirements and Numbering of Authority Records) [8] en avril 1999, qui va avoir pour mission de modéliser le contenu des notices d'autorité. Ensuite le groupe FRSAR (Functional Requirements for Subject Authority Records) en avril 2005, chargé de modéliser les relations entre données bibliographiques et fichiers d'autorité matière. [5]. En 2009 le groupe FRANAR va publier le modèle FRAD qui va traiter principalement des entités tel que les Collectivité, Famille, Personne et Oeuvre. Le modèle va se pencher sur les attributs et les relations de chacune de ces entités entre elles. Cette extension du modèle FRBR va être développée par le second groupe d'experts évoqué précédemment, le groupe FRSAR. Ce groupe va en 2010 terminer le modèle FRSAD, elle va comme le modèle FRAD le publier en ligne. Cette extension est différente car elle analyse les relations qui peut y avoir entre une Oeuvre, son sujet traité, la manière de nommer celui-ci et les informations qui sont contenues dans les systèmes d'indexation.

La position en France depuis 2014 est claire. En effet, la BnF et l'Abes ont exprimé la position nationale vis-à -vis du RDA (code de catalogage). La recommandation est de progressivement d'accès vers le modèle FRBR est réalisé une "FRBRisation" des catalogues. Depuis sa mise en place le modèle FRBR et ses extensions FRAD et FRSAD évoluent et vont permettre l'apparition d'un modèle réunissant les fonctions de chacun..Depuis 2010 l'IFLA veut que ces trois modèles se réunissent en un seul et c'est en 2017 qu'ils vont fusionner en une seule entité qui se nomme désormais IFLA LRM.

2.3 La mise en place de l'IFLA LRM

La science de l'information et du document est en évolution constante et même lorsque des nouveautés comme le modèle FRBR sont mises en place il y a des évolutions. En effet, des extensions à celui-ci sont apparus, tel que le FRAD et le FRSAD. Depuis 2010 l'IFLA veut que ces trois modèles se réunissent en un seul et c'est en 2017 qu'ils fusionnent en une seule entité qui se nomme désormais IFLA LRM. Sa version traduite en français est d'ailleurs publiée en septembre 2021. Tout comme le modèle FRBR et ses extensions, le modèle IFLA LRM sur les besoins des utilisateurs, il structure ces métadonnées autour de trois notions : l'entité , l'attribut et la relation. Le but de ce modèle est d'être un remplaçant aux notices bibliographiques actuelles par un rapport entre les entités.

Le monde est en constante évolution et les pratiques au sein de la bibliothéconomie n'échappent pas à cela. Depuis la création du modèle FRBR il y a une volonté de l'améliorer le plus possible pour qu'il devienne la référence en termes de catalogage. Les extensions FRAD et FRSAD sont venues se greffer au modèle de base pour l'améliorer et agrandir son champ de travail. Depuis 2010 l'idée de les assembler en un seul et même modèle voit le jour. Elle sera appliquée 7 ans plus tard en 2017 et portera le nom d'IFLA LRM. La FRBRisation est en marche

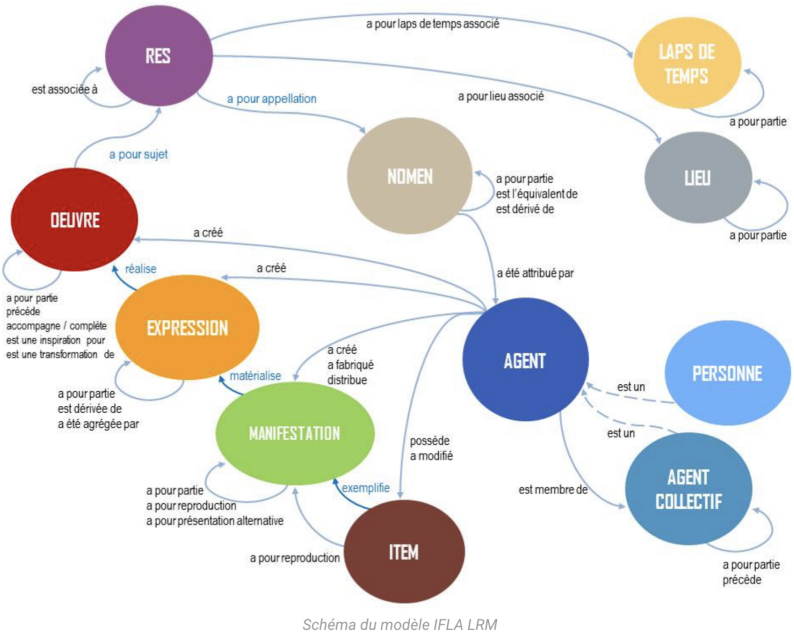


FIGURE 2 – Schéma représentativité du modèle IFLA LRM.

et un objectif clair, celui d'avoir une uniformisation du catalogage. C'est-à-dire de pouvoir créer un modèle universel compris par l'ensemble des pays et qui ne se porterait plus uniquement sur les informations des notices bibliographiques, mais plutôt dans un réseau de relation entre des entités. Dès 2004 dans son rapport sur le modèle FRBR Dr. Barbara Tillett exprime cette volonté de voir ce modèle évoluer : "This work within IFLA has spread worldwide and provides a conceptual model to guide us for many years to come. IFLA together with other interested parties will continue to encourage the application of this model to facilitate international standardization and reduce costs for cataloging on a global scale". [17]

3 Les référentiels d'indexation : LCSH et RAMEAU

Après avoir abordé les notices bibliographiques ainsi que le modèle FRBR et ses extensions. Nous allons nous intéresser à deux référentiels d'indexation, le premier LCSH et le second RAMEAU. A travers cette partie nous allons présenter, expliquer et exposer ce que les auteurs disent sur ces référentiels d'indexation.

3.1 LCSH : Library of Congress Subject Headings

Tout d'abord nous avons, le référentiel d'indexation LCSH (Library of Congress Subject Headings) est un ensemble de systèmes de recherches et de classifications qui est utilisé dans le monde de la bibliothéconomie, mais aussi dans la documentation. L'origine de sa création est vers la fin du XIXe siècle. En effet, la Bibliothèque du Congrès des États-Unis amorce l'idée de créer un système de classification qui permettrait de faciliter l'accès et l'organisation de leurs collections . C'est un choix logique contenu de la grandeur des collections qui ne font que d'augmenter. Au fur et à mesure, le LCSH à développé un vocabulaire contrôlé qui fournit des termes qui sont normalisés pour la description du contenu des documents (article,livre,périodique etc...) Elle devient par la suite une norme internationale dans le domaine de la recherche de documents et de l'indexation.

Ce qui fait la force du LCSH, ce sont les termes spécifiques et normalisés qui sont appelés des "sujets". Ces "sujets" représentent des concepts, des thèmes, des lieux, des événements, des personnes, des genres et toutes les autres informations qui peuvent être pertinentes pour un document. Ces sujets sont assignés à chaque document pour qu'il puisse être indexé et donc facilement repérables par les utilisateurs effectuant des recherches. Utiliser le référentiel LCSH présente beaucoup d'avantages pour ses utilisateurs. Il permet dans premier temps d'avoir des recherches beaucoup plus précises et affiner grâce aux termes normalisés et contrôlés. En effet l'utilisation de ces termes permet d'éviter ll'ambiguïté liés aux variations entre les langues. Ensuite, LCSH apporte une structure hiérarchique qui représente parfaitement les relations entre les différents sujets d'une œuvre. Cette représentation facilite l'exploration entre des ressources connexes. Pour terminer sur les avantages, il est important de voir que le monde de la documentation, de la bibliothéconomie est en perpétuelle mutation. C'est pout a que le LCSH est mise à jour régulièrement pour qu'il soit et est le plus efficace possible. Le LCSH est utilisé par les bibliothèques En plus de son utilisation, mais également dans plusieurs contextes liés à la recherche d'informations. En effet, il est utilisé dans les bases de données en ligne, les catalogues de bibliothèques et les moteurs de recherche. Il favorise l'interopérabilité entre toutes les institutions et facilite l'échange et le partage de données bibliographiques à l'échelle mondiale. Ce système de classification et de recherche qui est largement utilisé et qui facilite l'organisation, l'accès et la recherche de documents dans les bibliothèques et dans le monde de la documentation. Les termes normalisés et la structure hiérarchique permettent au LCSH améliore la précision et la cohérence de la recherche documentaire.

Des projets liés au système de référencement par indexation peuvent voir le jour pour développer des nouvelles options dans la recherche et l'indexation d'information. C'est le cas par exemple du projet MACS (Multilingual access to subjects) qui à pour objectif de développer un système qui offre un accès sujet multilingue aux catalogues bibliographiques en utilisant des langages d'in-

dexation existants comme le SHLs. Tout ça en lien avec LCSH et RAMEAU par exemple. LCSH pourra permettre aux utilisateurs de consulter les catalogues des bibliothèques partenaires d'une seule question écrite dans leur propre langue.[10]. Grâce à l'utilisation des termes normalisés et la structure hiérarchique, le LCSH est d'une précision très importante. L'ensemble de ces fonctions permettent de faciliter la classification, l'indexation et la recherche dans le monde des bibliothèques de documentation. Après avoir présenté un modèle d'indexation en provenance des Etats Unis, nous allons traiter de celui mit en place par la BnF.

3.2 RAMEAU : Répertoire d'Autorité Matière Encyclopédique et Alphabétique Unifié

Voyant le développement du système de référencement LCSH aux États Unis, la France et plus précisément la BnF (Bibliothèque nationale de France) à entrepris la mise en place de son propre système d'indexation et de référencement. Il se nomme RAMEAU (Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié)et à pour fonction de référencer et d'indexer et de référencer comme celui des voisins américains. La volonté de créer le système RAMEAU remonté aux années 1970. En effet, la BnF veut créer un système d'indexation français qui permet d'avoir des collections plus cohérentes et structurées . Son objectif est aussi assez similaire car il vise à faciliter la recherche et la récupération d'informations dans les catalogues et les différentes bases de données. Pour ça il va apporter des termes normalisés et contrôlées dans l'idée de pouvoir décrire le contenu d'un document grâce aux sujets, genres, auteurs...

Au sein du système RAMEAU, on peut retrouver des listes alphabétiques et des structures hiérarchiques permettant de décrire les concepts tels que le thème, le lieu, le genre, l'auteur et tout ce qui peut avoir une importance pour le document. L'ensemble de ces termes sont utilisés pour indexer les documents dans les bases de données et les catalogues de bibliothèques. L'ensemble de ces outils permettent ensuite de faciliter les utilisateurs lors de recherche de document. Tout comme son homologue américain, RAMEAU présente plusieurs avantages dans son utilisation. Pour commencer elle permet d'améliorer une recherche pour qu'elle soit la plus précise possible. Cette précision est permise grâce à l'utilisation de termes normalisés et contrôlés, qui permet d'éviter toute ambiguïté lors de la recherche. Tout comme le modèle LCSH, RAMEAU présente une structure hiérarchique qui permet de représenter les relations entre les différents termes. Les possibilités permises par RAMEAU sont utilisées dans les catalogues bibliographiques mais aussi dans les archives, les musées et tous les pôles qui nécessitent un système de référencement et d'indexation. Il est clair que RAMEAU évolue et son rapport avec la science de l'information et de la documentation est fortement lié. Cette science évolue et des outils sont mis en place pour faciliter et automatiser certaines tâches. C'est le cas des outils de traitement informatique des données.[1]

En conclusion, nous avons pu voir et comprendre le rôle de deux systèmes de référencement et d'indexation LCSH et RAMEAU. Les deux proviennent de pays différents mais ont plusieurs points communs dans leur utilisation. Leurs objectif est le même, pouvoir améliorer les recherches dans des catalogues bibliographies, bases de données et bien d'autres systèmes nécessité de l'indexation. LCSH et RAMEAU sont deux révolutions dans le domaine des sciences de l'information et de la documentation. Il est donc important de comprendre comment les documents sont indexés et quels sont les outils mis en place pour faciliter cette indexation.

3.3 L'utilisation du traitement automatique des données

Après avoir abordé les deux systèmes de référencements et d'indexation, il est important de voir comment et avec quel outil ils peuvent évoluer. Le traitement automatique des données est définie comme des opérations qui sont réalisées par des moyens automatiques, relatif à la collecte, l'enregistrement, l'élaboration, la modification, la conservation, la destruction, l'édition de données et, d'une façon générale, leur exploitation. En effet la fusion entre le traitement informatique et un système de référencement et d'indexation va permettre d'avoir des résultats encore plus rapides. Cependant, ceux-ci auront de plus grandes chances d'être moins pertinents. Il faut évidemment prendre en compte ce facteur, car l'utilisation d'un traitement automatique peut avoir des conséquences négatives sur la qualité du résultat obtenu. C'est un d'ailleurs un exemple que j'ai pu personnellement rencontré durant mon stage.[1]

Il y a aussi un intérêt à utiliser le traitement automatique d'information qui rejoint le sujet de mon stage. Ces systèmes d'indexation automatique doivent être capables de gérer des langues multiples et de prendre en compte les spécificités culturelles et linguistiques dans leur processus d'indexation. L'auteur souligne l'importance de la sensibilisation à ces enjeux et de l'adaptation des systèmes pour assurer une indexation précise et équitable. (citer livre). L'usage d'un système tel que RAMEAU avec les possibilités apportés par le traitement automatique d'information va permettre de traiter des données beaucoup vite et en plus grande quantité. Il ne faut néanmoins pas oublier que ce traitement à ses limites et les résultats peuvent manquer de précision.

4 L'indexation et le vocabulaire contrôlé

La classification des livres et des documents à été mise en place pour pouvoir naviguer à travers les catalogues et les allées des bibliothèques. Quand on se rend sur place les documents sont rangés selon la classification de Dewey qui est un système décimal divisé en dix catégories présentes dans ce tableau. Ces catégories sont elles-mêmes divisées en dix sous catégories qui peuvent avoir encore des sous catégories. Ce système permet de s'y retrouver lors de la recherche de documents. Cette partie va traiter de de deux sujets, tout d'abord

000	Généralités	500	Sciences
100	Philosophie	600	Techniques
200	Religion	700	Arts
300	Sciences sociales	800	Littérature
400	Langues	900	Géographie et Histoire

FIGURE 3 – Tableau des catégories de la classification Dewey.

l'indexation et ensuite le vocabulaire contrôlé. Le référencement et l'indexation à évoluer et ses méthodes aussi. Pour chaque partie je détaillerai les origines et le fonctionnement de ces éléments indispensables aux référencement. Je vais aussi parler des leurs évolutions dans le temps et le changement que ça peut apporter au domaine de la bibliothéconomie. [13]

4.1 L'histoire et le fonctionnement de l'Indexation

Précédemment nous avons vu que l'indexation est au cœur même de l'ensemble des systèmes de classifications et de référencement dans l'univers de la bibliothéconomie. Le principe d'indexation est de faciliter la recherche d'information grâce à l'attribution de termes d'index ou de mots-clés liés aux documents. Sa définition exacte tirée de la norme AFNOR Z 47-102(1978 :225) est : «l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document». Son origine remonte avant l'explosion de l'imprimerie et elle était réalisée majoritairement à la main. Au début le terme d'index n'existait pas, on parlait plus de tabula qui pouvait donner une lecture non linéaire qui était considéré comme de la paresse à l'époque. Ces tables permettait au final de pouvoir trouver plus rapidement les informations, c'est la première forme d'indexation. [2]

Avec le développement de l'imprimerie, la production des ouvrages et documents sont de plus en plus rapides et en plus grande quantité. Ce flot de documents à rendu l'indexation inévitable pour pouvoir référencer, trier et chercher des documents. Au fil du temps des dictionnaires vont être mis en place pour proposer des modèles d'indexation dans divers domaine tel que les sciences et les langues par exemple. Dans ces dictionnaires on retrouve déjà le principe de liste par ordre alphabétique qui facilite la recherche d'information précise grâce à des termes associés. Un élément va changer les méthodes d'indexation pour les rendre automatique, en effet, l'apparition du numérique et plus précisément de l'informatique va permettre l'avènement de l'indexation automatique. Pour le fonctionnement de l'indexation sur une méthode basée sur le linguistique qui vont permettre d'améliorer la précision des outils d'indexation. Cependant Muriel Amar à un avis plus nuancé sur le sujet. Elle explique qu'il faudra obligatoirement se pencher sur des pratiques professionnelles. Même si l'approche linguistique est très importante pour permettre de définir au mieux les inter-

actions entre les termes et les textes.[1] Le Web aussi à vu se développer une forme d'indexation pour l'ensemble des pages Web référencés.

Pour conclure, il est important de comprendre que l'indexation n'est pas un outil de référencement récent. En effet, avant l'imprimerie il y avait déjà une volonté de trier et organiser les ouvrages. Ce système à évolué avec les nouvelles technologies tel que le traitement automatique et le web par exemple. Il faut néanmoins comprendre que l'indexation ne serait rien sans la multitude d'outils qui la compose. C'est notamment le cas des sujets évoqués plus tôt mais aussi du vocabulaire contrôlé.

4.2 Le vocabulaire contrôlé (Thésaurus)

Dans les parties précédentes le terme de vocabulaire contrôlé à été introduit, mais il n'a pas été défini et expliqué en détail. C'est autour de la définition et l'utilisation générale de celui-ci que cette partie va se composer,en abordant aussi sa structuration à travers des Thésaurus.

Le vocabulaire contrôlé est défini d'après la COAR (Confederation of Open Access Repositories) comme : "un ensemble organisé de mots et expressions utilisés pour indexer du contenu et/ou le retrouver par navigation ou recherche. Typiquement, il inclut des termes préférentiels et leurs variantes et opère dans un périmètre défini ou décrit un domaine spécifique.". Ce vocabulaire est contrôlé ,car il a pour objectif d'uniformiser les mots entre chaque langue pour qu'un mot signifie la même chose dans l'ensemble des langages. Il provient d'une liste de termes qui est donc définie au préalable comme un Thésaurus.[3] Il permet en somme de pouvoir représenter un concept entier grâce à un seul mot ou une seule expression qui fait partie de la liste de vocabulaire définie. Néanmoins, il existe aussi un vocabulaire libre qui fait opposition à celui qui est contrôlé. Il peut être défini comme une liste de termes sélectionnés librement et de façon arbitraire par un chercheur ou alors un analyste lors de son indexation. Ces termes n'ont pas besoin d'être validés par un une liste de termes autorités. [3]

Le rassemblement de ce vocabulaire au sein d'un Thésaurus qui selon LeRobert, un répertoire structuré de termes (mots- clés) pour l'analyse de contenu et le classement de documents. Ces thésaurus permettent de faciliter l'indexation, la recherche et la récupération d'information.Dans l'article l'auteur aborde l'attribution de termes normalisés tel que le vocabulaire contrôlé qui permet de d'éviter les confusions qu'il peut y avoir sur des termes similaires.[15]On peut trouver pour des domaines précis tel que le thésaurus PACTOLS qui est créé par Frantiq (Fédération et Ressources sur l'Antiquité constituée en Groupement de Service du CNRS n°3378). PACTOLS est composé de vocabulaire contrôlé, normalisé et multilingue dans le domaine de l'archéologie et les sciences de l'antiquité. Il s'étale sur la période allant de la préhistoire à l'antiquité. Il existe des thésaurus pour autant de domaines qui nécessitent de répertorier des listes

de données structurées, permettant de faciliter l'indexation et la recherche de données. Dans l'article, l'auteur conclut en exprimant le fait que les thésaurus sont d'une grande utilité en tant qu'outils qui permettent de créer un langage commun et partagé. Celui-ci permet de favoriser la communication et l'échange d'information multilingue. Tout cela est possible grâce à sa structuration très organisée et ses relations sémantiques qui permettent une efficacité plus grande dans l'indexation et la recherche dans le domaine de la bibliothéconomie, mais pas seulement. [15]

Il est important de comprendre à travers cette partie l'importance des vocabulaires contrôlées et son impact sur les différents systèmes d'indexation. Il permet de représenter un thème grâce à un mot ou une expression, peu importe la langue. Le thésaurus sont eux des listes à grandes échelles et structuré qui peuvent accueillir le vocabulaire contrôlé. C'est le cas de PACTOLS et cela permet de mieux structurer et d'apporter une encore plus grande efficacité dans le domaine de l'indexation et de la recherche.

Deuxième partie

Consulter les notices bibliographiques à différentes échelles : les bibliothèques numériques et les catalogues numériques

Dans le cadre de mon stage j'ai été amené à traiter des données qui proviennent d'un catalogue numérique diffusé en ligne, Worldcat. Il est donc nécessaire qu'à travers une partie je propose une définition des catalogues numériques, les différences ou non avec la bibliothèque numérique. Également évoquer leurs origines et l'importance qu'ils ont dans une multitude domaine. Ensuite je présenterais différents catalogues en ligne sous différentes échelles, commençant par l'Université et finissant à l'échelle internationale.

5 Les catalogues et bibliothèques Numériques

Pour commencer je vais tout de suite définir le fait qu'un catalogue numérique fait partie intégrante d'une bibliothèque numérique. En effet, un catalogue numérique est simplement la version numérique d'un catalogue papier. La bibliothèque numérique est le regroupement numérique de documents écrits qui sont consultables et accessibles à distance. On peut aussi citer la définition proposée par Eric Hellman, cofondateur de OCLC « Any collection of digital resources managed with the primary goal of maximizing the collection's utility to a defined user community. » en effet, il explique que les ressources numériques avec l'objectif principal de maximiser l'utilité de la collection pour une communauté d'utilisateurs définie sont considérées comme des bibliothèques numériques.

A l'ère du numérique, les bibliothèques et les catalogues numériques sont essentiels pour accéder à l'information, notamment sur le Web. C'est pourquoi, durant de nombreuses années les bibliothèques ont été la grande source de connaissance grâce aux nombreux documents, ouvrages, articles présents au sein de leurs collections. Cependant, il est plus difficile de trouver les informations sur des sujets très précis avec des documents parfois difficiles à trouver ou alors inaccessibles à l'étude. Dans bon nombre de cas , les étudiants qui réalisent un mémoire n'ont pu avoir l'ensemble des informations sur leurs sujets. C'est d'ailleurs l'exemple que prend l'auteur Wallace Kirsop dans son texte, car il explique qu'avec le financement de sa thèse il n'a pas pu se déplacer comme bon lui semble. En effet, originaire d'Australie il à pu aller à Paris et rapidement au Pays Bas et Belgique dans le cadre de sa thèse qu'il à réalisé avec le corpus de

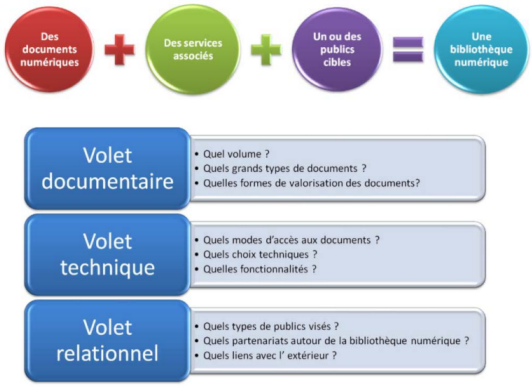


FIGURE 4 – Schéma de la BnF pour définir les bibliothèques numériques.

document qu'il avait à sa disposition. Après avoir soutenue sa thèse en 1960 à la Sorbonne, il assiste à une conférence qui lui soumet une réalité qu'il avait déjà constatée, sa thèse était imparfaite. Il explique à travers son expérience personnelle qu' à l'époque l'accès aux documents était beaucoup plus difficile. Alors qu'a l'ère du numérique il est désormais aisé de pouvoir consulter un document et éviter les demandes à des bibliothèques dans des pays plus ou moins éloignés de chez soi. [9]

C'est donc l'intérêt majeur de ces catalogues numériques qui permet aux bibliothèques de numériser et de rendre accessibles des milliers de documents tels que des livres, des revues et des journaux par exemple. Cet accès est ensuite utilisé dans le cadre de travaux de recherche, mais aussi dans le cadre de l'apprentissage et la découverte. La bibliothèque en elle-même ne permet pas de donner un accès à des utilisateurs. C'est grâce à la mise en place de certains portails qu'il est possible de consulter cette multitude de documents. Grâce à ces portails les utilisateurs pourront naviguer de manière plus simple dans le cadre de leurs différentes recherches. La mise à disposition de certaines options de filtrages qui permettent d'affiner les recherches et de pouvoir trouver exactement le ou les documents recherchés. Pour terminer il y a un point très important qu'il ne faut pas négliger, c'est l'accès aux bibliothèques numériques comparé aux bibliothèques traditionnelles, en effet, les bibliothèques numériques sont accessibles 24h/24 et 7j/7. Les utilisateurs peuvent les consulter depuis n'importe quel appareil tant qu'il y a un accès à internet. C'est important pour les populations qui n'habitent pas proche de certaines bibliothèques et qui grâce aux numériques peuvent tout de même consulter des ouvrages sans se déplacer.

L'arrivée de l'ère du numérique dans le domaine de la bibliothéconomie touche énormément de systèmes. Nous l'avons dans les parties précédentes avec les sys-

tèmes d'indexations, le FRBR etc.. Ici nous avons vu que les catalogues et les collections des bibliothèques aussi ont été numérisés. L'objectif de cette numérisation est de pouvoir donner un accès plus simple aux utilisateurs qui peuvent consulter les collections comme ils le souhaitent. Il faut néanmoins garder à l'esprit que les bibliothèques physiques sont toujours très importantes et apportent une expérience directe et un lieu plus adéquat à l'étude. Mais aussi, certains documents rares ne sont pas encore numérisés et leur accessibilité est uniquement sur place. J'ai abordé qu'il existait des portails mises en place pour les bibliothèques numériques. Dans la partie suivante je vais en présenter quatre en détails et chacun à une échelle bien précise.

6 Les catalogues en ligne, présentation à différentes échelles.

Dans cette partie je vais me concentrer sur les catalogues en ligne. Je vais les présenter sur quatre échelles différentes en commençant par l'Université de Lille et la plateforme LillOA. Ensuite à l'échelle national avec Gallica, à l'échelle européenne avec Europeana et pour finir à l'échelle mondial avec WorldCat. Pour chacun d'eux, je vais réaliser une présentation ainsi qu'une analyse du site pour voir ce qu'on y trouve et comment sont représentées les données.

6.1 À l'échelle de l'Université de Lille : LIllOA

6.1.1 Déscription et historique de LillOA

Pour commencer, LillOA (Lille Open Archive) est l'archive ouverte institutionnelle connectée à HAL de l'Université de Lille. Une archive ouverte est définie comme un réservoir des publications issues de la recherche scientifique et de l'enseignement. Elle est totalement gratuite et ouverte à toute la communauté des chercheurs de l'Université de Lille et aussi des étudiants. LillOA a été nommé ainsi car il a été créé à et pour la ville de Lille, plus particulièrement pour ses Universités. Il a été fondé en 2017 par l'Université de Lille qui n'avait pas encore fait de fusion entre les trois Universités de Lille 1, Lille 2 et Lille 3. Cette fusion à eu lieu le 1er janvier 2018. C'est la société Atmire qui s'occupe du développement en utilisant le logiciel Open Source D-Space. Les objectifs principaux de LillOA sont de donner accès aux différents travaux de recherches de l'Université de Lille, grâce à une classification et un moteur de recherche intégré. Mais aussi de l'utiliser pour la recherche surtout, il est aussi très important car il sert de plateforme qui permet aux différents chercheurs de pouvoir déposer leurs travaux et transmettre les résultats de leurs recherches.

Maintenant que j'ai réalisé la fiche de présentation de la plate forme nous allons nous pencher sur des aspects plus techniques est visuel. En effet, nous allons voir comment se déroule l'expérience utilisateur sur LillOA et quelles sont les

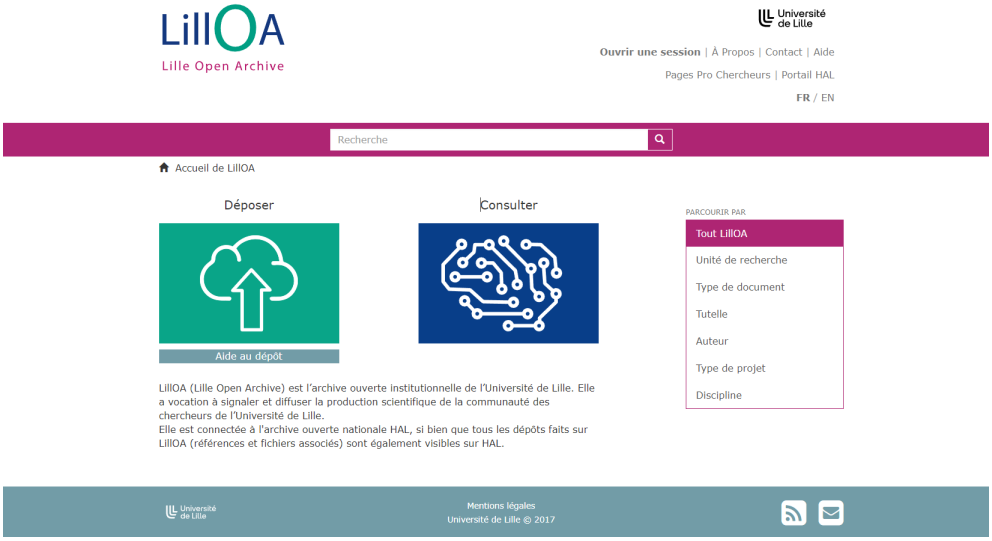


FIGURE 5 – Schéma de la BnF pour définir les bibliothèques numériques.

informations mise à disposition après une recherche.

6.1.2 Analyse du site et des données mises à disposition

Tout d'abord l'écran d'accueil qui est sobre et où on retrouve trois éléments qui vont nous intéresser et que je vais montrer par la suite. Tout d'abord la barre de recherche est disponible pour l'utilisateur qui sait le thème qu'il veut voir. Dans le cas contraire, il y a à droite de l'écran des propositions pour parcourir le catalogue en fonction de notre besoin. Ensuite on retrouve deux grands rectangles qui sont les deux fonctions principales de LillOA. Le premier en bleu c'est pour consulter des travaux publiés par des chercheurs de l'Université de Lille, ensuite à gauche on retrouve un rectangle vert qui est mis en place pour le dépôt de travaux de recherche de l'Université de Lille.

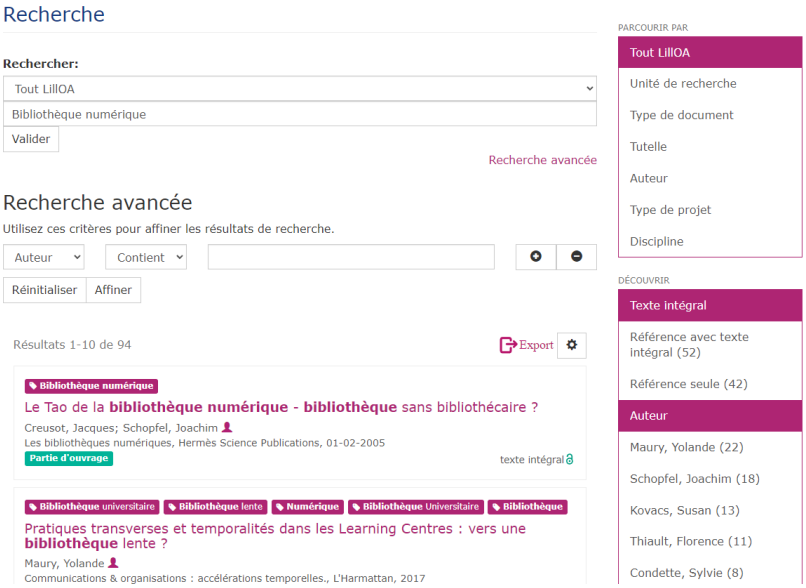


FIGURE 6 – Schéma de la BnF pour définir les bibliothèques numériques.

Tout d'abord lorsque que nous effectuons une recherche avec la barre mise à disposition, voici le résultat obtenu. On à une liste de résultats avec les mots-clés surlignés en rose, ensuite le titre de l'article ou l'ouvrage, le ou les auteurs et d'autres informations tel que le nom de la revue, l'éditeur et la date de parution. Sur le côté on retrouve des informations qu'on peut sélectionner pour affiner la recherche comme le nombre de références seules ou avec un texte intégrale, mais aussi le nombre de fois qu'un auteur à écrit un article sur le sujet recherché. On y retrouve d'autres informations qui ne sont pas présentes sur la capture d'écran comme le type de document, la langue et le titre de la revue. On peut d'ailleurs dérouler une menue recherche avancée pour sélectionner des options précises pour nos recherches.

Voici les informations que l'on obtient lorsque l'on clique sur un ouvrage qui nous intéresse. On retrouve des informations vues précédemment mais aussi de nouvelles qui ne pouvaient pas apparaître sur la page d'avant par manque de place. On y retrouve en plus l'éditeur ou le directeur scientifique, l'ISBN, la ou les disciplines HAL, le résumé, si c'est une vulgarisation ou non, sa source et sa collection. Toutes ces métadonnées sont très importantes dans le cadre de travaux de recherche pour pouvoir citer mais aussi analyser certaines données.

Je vais passer ensuite très rapidement sur l'onglet de consultation présent sur la page d'accueil car c'est une extension de l'onglet recherche. L'utilisateur peut choisir par liste d'unités et avec différents paramètres tel que l'auteur, le type de document, la tutelle, le type de projet et la discipline. Pour finir je vais

Le Tao de la bibliothèque numérique - ...

Export

Type de document :	Partie d'ouvrage
Titre :	Le Tao de la bibliothèque numérique - bibliothèque sans bibliothécaire ?
Auteur(s) :	Creusot, Jacques [Auteur] Institut de l'information scientifique et technique [INIST] Schopfel, Joachim [Auteur] Institut de l'information scientifique et technique [INIST]
Éditeur(s) ou directeur(s) scientifique(s) :	Fabrice Papy Gil-François Euvrard
Titre de l'ouvrage :	Les bibliothèques numériques
Éditeur :	Hermès Science Publications
Lieu de publication :	Paris
Date de publication :	2005-02-01
ISBN :	2-7462-1036-3
Mot(s)-clé(s) :	CNRS INIST Bibliothèque numérique Ressources électroniques Métier Formation Information scientifique
Mot(s)-clé(s) en anglais :	Digital library Electronic resources Job skills Training Scientific information
Discipline(s) HAL :	Sciences de l'Homme et Société/Sciences de l'information et de la communication
Résumé :	Le texte essaie de décrire quelques facteurs de la transformation

FIGURE 7 – Schéma de la BnF pour définir les bibliothèques numériques.

Liste des unités

PARCOURIR PAR

Type de document | Tutelle | Auteur | Type de projet | Discipline

Rechercher une unité de recherche :

Valider

Liste des unités de recherche

Advanced Drug Delivery Systems (ADDS) - U1008
Analyses Littéraires et Histoire de la Langue (ALITHILA) - ULR 1061
Autres travaux scientifiques
Cancer Heterogeneity, Plasticity and Resistance to Therapies (CANTHER) - UMR 9020 - UMR 1277
Centre d'Étude des Arts Contemporains (CEAC) - ULR 3587
Centre d'Études en Civilisations Langues et Lettres Étrangères (CECILLE) - ULR 4074
Centre d'Études et de Recherches Administratives, Politiques et Sociales (CERAPS) - UMR 8026
Centre d'Histoire Judiciaire (CHJ) - UMR 8025
Centre d'Infection et d'Immunité de Lille (CIIL) - U1019 - UMR 9017
Centre de Recherche "Individus Epreuves Sociétés" (CeRIES) - ULR 3589
Centre de Recherche Droits et Perspectives du droit (CRDP) - ULR 4487
Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL) - UMR 9189
Centre Interuniversitaire de Recherche en Éducation de Lille (CIREL) - ULR 4354
Centre Lillois d'Études et de Recherches Sociologiques et Économiques (CLERSE) - UMR 8019

FIGURE 8 – Schéma de la BnF pour définir les bibliothèques numériques.

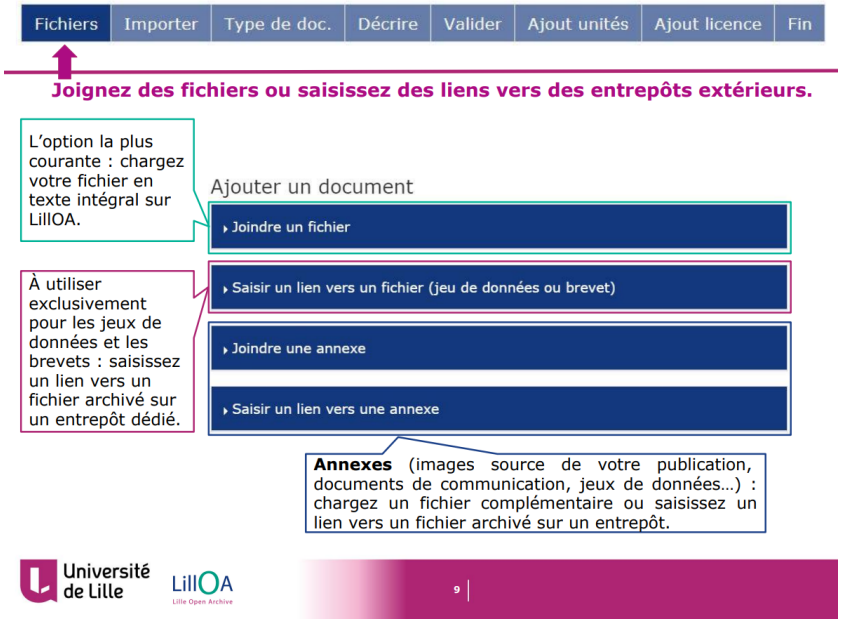


FIGURE 9 – Schéma de la BnF pour définir les bibliothèques numériques.

parler de l'onglet de dépôt de projet qui est destiné aux différents chercheurs de l'Université de Lille et leurs travaux. Il est en effet possible de déposer des travaux qui seront ensuite consultables sur le portail LillOA. Après avoir cliqué, on à la possibilité de télécharger un document pdf qui explique toute les étapes du dépôts. Je vais vous montrer la page sur laquelle on arrive après avoir accepté les conditions générales. Il y a 4 onglets pour ajouter les documents au milieu de l'écran et au dessus différents onglets pour qui sont les différentes étapes pour déposer un document. Après avoir déposé le document on doit ajouter sa description qu'on peut directement importer grâce au DOI (un identifiant numérique destiné aux objets scientifiques que l'on souhaite rendre citable). Ensuite on donne de type de document, sa description et on arrive sur l'onglet de validation qui est un récapitulatif du dépôt, pour terminer on peut ajouter des unités et une licence pour faciliter la réutilisation de son document.

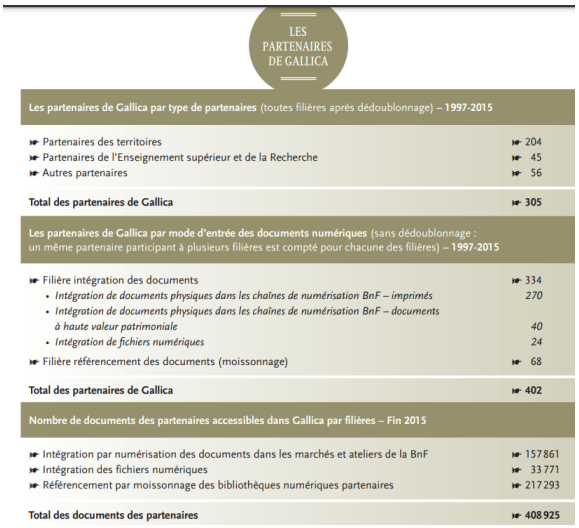
Après avoir présenté LillOA, j'ai décidé de me concentrer sur l'expérience utilisateur et le fonctionnement du site. LillOA est un cas un peu différent et pas vraiment dans le sujet du mémoire cependant, je voulais montrer comment fonctionne un portail à l'échelle de l'Université de Lille. Néanmoins, la façon de répertorier est intéressante. LillOA utilise un système basé en partie sur les mots-clés qui peuvent être similaires aux sujets utilisés dans les systèmes d'indexation de notices bibliographiques. Je vais désormais me pencher sur une échelle plus grande, celle de la France avec son catalogue numérique Gallica.

6.2 À l'échelle nationale : Gallica

Après s'être concentré sur une plateforme de l'Université de Lille, il est important de dézoomer pour aborder un catalogue en ligne à l'échelle national : Gallica. Je vais commencer par présenter Gallica, son origine et ses fonctions principales. Ensuite je vais étudier l'expérience utilisateur en analysant le fonctionnement des recherches de documents sur le site.

6.2.1 Déscription et historique de Gallica

Pour réaliser un portrait de Gallica j'ai utilisé le "à propos" présent sur le site qui explique en détail la création de ce projet. Le projet de bibliothèque d'un genre nouveau est envisagé dès 1988 par le président François Mitterrand. Cette nouvelle forme de bibliothèque pourrait permettre l'accès à son catalogue depuis les salles de lecture. Le développement du Web dans les années 1990 accélère la projet et elle le modifie car désormais cette bibliothèque numérique sera accessible de partout tant qu'il y a une connexion internet. Pour le projet, les œuvres devront être numérisées et c'est un problème car juridiquement parlant. Le choix sera fait de ne numériser que les œuvres libres de droits. C'est en fin 1997 que Gallica est lancé aux publics et la plateforme propose un accès à plusieurs milliers (20 000) de textes. Au fil des années il y aura de plus en plus de livres numérisés et les supports acceptés vont aussi s'agrandir. En 2015 la site comptabilise plus de 3.7 millions de documents qui proviennent majoritairement de la BnF. (source bertrand) Gallica va évoluer durant les années 2000. En 2000 Gallica propose une nouvelle version qui donne l'accessibilité à des documents en mode texte et des images. L'année 2004 à vu apparaître la première charte documentaire : "les quelque 100 000 documents imprimés, 80 000 images et 30 heures de son alors disponibles dans Gallica s'inscrivent dans une dominante disciplinaire en Histoire, Littérature, Sciences et Techniques." (source bertrand) . C'est la première fois qu'il y a un ensemble de documents aussi important est complet qui s'insère dans le catalogue. Gallica va encore évoluer car elle doit faire face à une concurrence très importante notamment de Google Books. La plateforme va donc continuer à numériser sur un rythme en 2007 de 100 000 imprimés par an. La dernière grande mise à jour se trouve en 2010 avec la numérisation dédiée aux documents précieux. A partir de 2010, en quatre Gallica va passer de 1 million à 3 millions de documents numérisés. Gallica est un projet français ambitieux et qui à désormais un catalogue très fourni, notamment



LES PARTENAIRES DE GALLICA

Les partenaires de Gallica par type de partenaires (toutes filières après dédoublonnage) – **1997-2015**

Partenaires des territoires	204
Partenaires de l'Enseignement supérieur et de la Recherche	45
Autres partenaires	56
Total des partenaires de Gallica	**305**

Les partenaires de Gallica par mode d'entrée des documents numériques (sans dédoublonnage : un même partenaire participant à plusieurs filières est compté pour chacune des filières) – **1997-2015**

Filière intégration des documents	334
• *Intégration de documents physiques dans les chaînes de numérisation BnF – imprimés*	*270*
• *Intégration de documents physiques dans les chaînes de numérisation BnF – documents à haute valeur patrimoniale*	*40*
• *Intégration de fichiers numériques*	*24*
Filière référencement des documents (moissonnage)	68
Total des partenaires de Gallica	**402**

Nombre de documents des partenaires accessibles dans Gallica par filières – Fin 2015

Intégration par numérisation des documents dans les marchés et ateliers de la BnF	157 861
Intégration des fichiers numériques	33 771
Référencement par moissonnage des bibliothèques numériques partenaires	217 293
Total des documents des partenaires	**408 925**

FIGURE 10 – Liste des partenaires de Gallica. (Source [4]

grâce à la BnF à l'initiative du projet. Cependant, il y a aussi des collections partenaires qui ne sont pas majoritaires mais bien présentes dans le catalogue. Ils sont très importants pour la plateforme est regroupe tout de même un grand nombre de documents référencés et indexés. On peut d'ailleurs voir cela sur la Figure 9 qui répertorie le nombre et le type des différents partenaires.

Gallica est un catalogue numérique qui référence et donne accès à plusieurs millions de documents. Cette plateforme est d'une grande utilité pour les chercheurs et les utilisateurs qui cherchent des documents précis. Gallica permet d'éviter le déplacement pour consulter différentes sources. Je vais maintenant me pencher sur ce qu'on peut trouver sur Gallica lorsque l'on réalise des recherches. Maintenant que j'ai réalisé un présentation de Gallica je vais montrer l'expérience utilisateur sur le site mais aussi, les différents résultats et options disponibles sur le site.

6.2.2 Analyse du site et des données mises à disposition

Commençons d'abord comme pour LillOA par la page d'accueil du site Gallica. (Figure 10) On voit directement sur la page d'accueil que Gallica à beaucoup plus de propositions que LillOA. En effet, Gallica est une plateforme à l'échelle nationale et donc ce n'est pas simplement un catalogue. Il y a différents contenus qui sont là pour animer et faire vivre le site. On remarque tout de suite au centre de l'écran une multitude d'onglets soit " A la une" ou "Blog". On peut cliquer sur ce qui nous intéresse pour avoir du contenu en lien avec l'image et la description de l'onglet. Ensuite en défilant la page on à plusieurs onglets que je ne vais pas montrer car ce n'est pas en lien avec mon sujet. Ces onglets sont

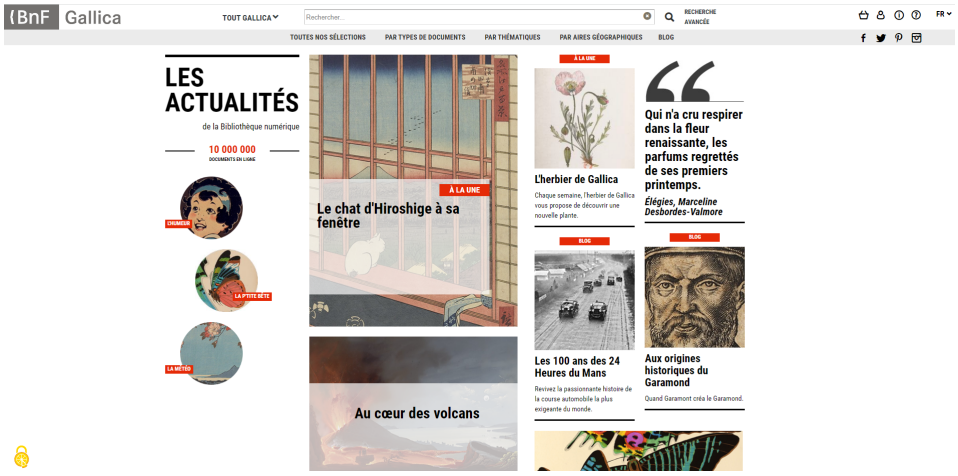


FIGURE 11 – Capture d'écran de la page d'acceuil du site Gallica

FIGURE 12 – Capture d'écran de la barre de recherche du site Gallica.

divers, ça peut aller du découpage pour les enfants, au tutoriel en passant par des aides pour le Bac. Nous allons nous concentrer sur la barre de recherche qui lorsque nous cherchons quelque chose elle suggère des réponses (Figure 11).

Passons désormais à la recherche (Figure 12), nous obtenons ce résulats avec les titres des ouvrages, les extraits ou le ou les mots recherchés sont présents et surligner en jaune. Ensuite sur le côté gauche on à la possibilité d'affiner la recherche en sélectionnant des paramètre précis tel que le site de consultation, le type de document (livre ;image, presse et revue etc..), l'auteur, la date d'édition, le thème, la version, la langue, le mode de texte et le type d'accès.

Lorsque l'on clique sur le document voulu on tombe sur une page (Figure 13) avec à droite le texte numérisé sur lequel on peut naviguer entre les pages et zoomer si besoin. L'utilisateur peut aussi choisir entre quatres façon de lire son document. Sur la gauche on retrouve l'onglet "En savoir plus" qui détaille les métadonnées du document. Toutes ces informations sont extrêmement utiles dans le cadre d'un travailleur qui se rapproche de celui que j'ai réalisé en stage. En effet toutes les informations permettant d'identifier le document même si

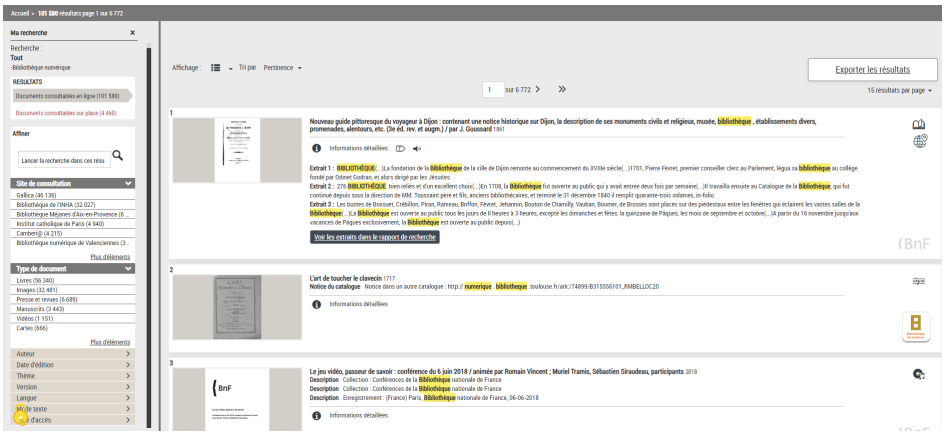


FIGURE 13 – Capture d'écran du résultat de la recherche sur le site Gallica.

dans le cas de la Figure 13 il n'y aucune information sur les mots-clés ou sujets liés à l'œuvre.

Pour terminer, sur Gallica il existe un onglet de recherche avancée qui permet d'avoir un plus grand choix d'options qui permettent d'avoir la recherche la plus affinée possible dès le début. (Figure 14) L'utilisateur peut choisir entre différents types d'options tel que le type de document, l'année d'édition et le format par exemple. Cependant il y a un une option pour choisir par proximité c'est à dire par terme choisis et en fonction de si il est plus ou moins proche d'un autre terme. Cette option permet de pouvoir chercher des groupes de mots dans certains textes et ça permet d'éviter le bruit lors de la recherche. Le bruit lors d'une recherche sur Google par exemple, peut être défini comme les résultats qui n'ont pas de rapport avec la recherche demandée. Par exemple si on utilisé sur Gallica l'option pour les termes et qu'on décide qu'on veut que les mot bibliothèque et numérique soient à la suite ça permet d'éviter tous les textes qui parle soit de bibliothèque ou de numérique. Cet outil est très utilisé dans le cadre de recherches précises sur un thème bien spécifique.

La plateforme Gallica est extrêmement importante et développée. Elle est le catalogue français de la BnF et regroupe d'après la BnF plus de 10 millions de documents numérisés. Créé sur une initiative de François Mitterrand et avec la volonté de de créer une bibliothèque numérique accessible depuis les bibliothèque traditionnelles. L'émergence du Web en à décidé autrement et c'est sur internet que la plateforme s'est développée et continue d'évoluer jusqu'en 2015. Le site est très complet et permet de faciliter la recherche de documents sur un catalogue en ligne très fourni et avec des documents très bien indexé. L'outil de recherche avancée permet aussi de chercher dans un cadre plus restreint grâce à plusieurs options et notamment la recherche par proximité de termes. Gallica est un catalogue numérique qui jouit à la fois du catalogue de la BnF mais aussi de ses

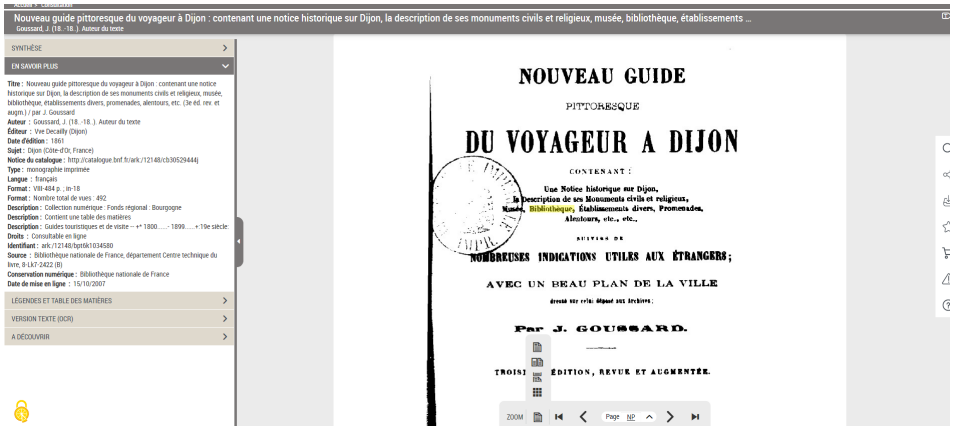


FIGURE 14 – Capture d'écran d'un document numérisé sur le site Gallica.

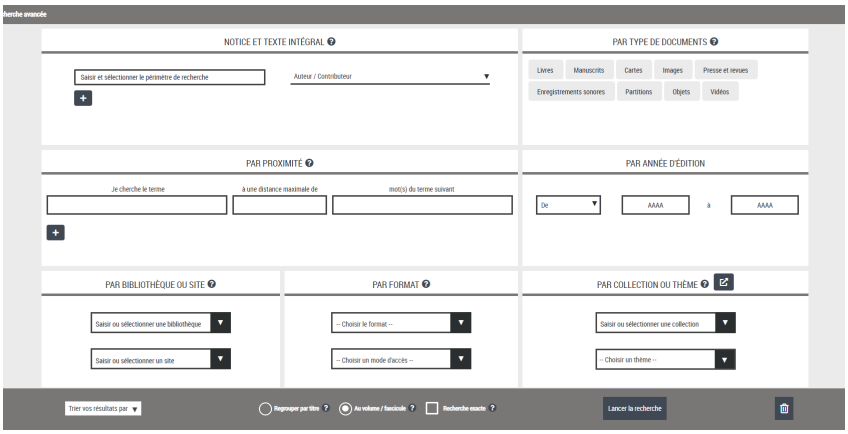


FIGURE 15 – Capture d'écran de l'option de recherche avancée sur le site Gallica.

partenaires qui sont très nombreux. La différence avec LillOA qui est destiné à un public plus restreint est très visible. Son échelle nationale fait de Gallica une catalogue beaucoup plus visité et il faut donc assumer cette lourde charge. C'est pour ça qu'il est intéressant dans la prochaine partie d'analyser un catalogue en ligne avec une échelle encore plus grande, celle de l'Europe.

6.3 À l'échelle Européenne : Europeana

Comme pour les deux parties précédentes, je vais dans un premier temps présenter la catalogue en ligne européen, Europeana. Je vais réaliser un point historique ainsi que définir quels sont les objectifs principaux de Europeana. Ensuite je vais réaliser une étude plus approfondie du site, d'abord dans sa forme et pour finir analyser les informations que la plateforme propose lors des différentes recherches. On passe encore à une échelle supérieure, celle de l'Europe et donc je vais analyser s' il y a des grandes différences avec Gallica par exemple.

6.3.1 Déscription et Historique de Europeana

Je vais d'abord parler de l'histoire derrière la création d'Europeana et qui sont à l"initiative de ce projet. Son histoire commence en 2005 grâce à une lettre du président français de l'époque Jacques Chirac avec le soutien de l'Espagne, des Pays-Bas, de l'Allemagne, des Etats membres de l'UE et de la commission européenne. Tous les membres à l'origine de sa création se donnent l'objectif de mettre en place une bibliothèque numérique qui donne un accès en ligne aux collections de l'ensemble des pays des Etats membres. Cette décision vient répondre au projet annoncé par Google en , "Google Search Books". Le géant américain avait comme objectif de numériser plusieurs millions de livres. Des prototypes sont réalisés et c'est le 20 novembre 2008 que Europeana est lancé. Le jour de son lancement, la plateforme donnait accès à plus de 2 millions d'objets numériques tels que des textes, des images,des sons et des vidéos et tout ça dans les 21 langues officielles de l'UE. [7] L'objectif qui était de passer la barre des 10 millions de documents numériques en 2010 à été plus qu'atteint, car Europeana à dépassé les 15 millions.

La commission européenne est toujours derrière le projet qui ne cesse de se développer depuis son lancement en 2008. L'enrichissement du catalogue d'Europeana est mené depuis toutes ces années à travers projets de numérisations. On peut prendre comme exemple le projet Europeana Regia qui est un projet dirigé par la BnF et qui s'étale de janvier 2010 à juin 2012. Ce projet à réuni cinq bibliothèques européennes (Munich,Valence,Bruxelles, Wolfenbüttel et Paris) et à pour objectif la numérisation de 900 manuscrits provenant des collections royales. On retrouve plusieurs autres projets de numérisation comme Europeana Collections 1914-1918 sur le patrimoine de la Première Guerre mondiale. Europeana est incontestablement le catalogue numérique le plus enrichit d'europe. Ce travail en collaboration avec les pays de l'Union Européenne et financé par la Commission Européenne s'est parfaitement développé. Son objec-

tif de proposer un catalogue qui regroupe un ensemble de documents numérisés provenant du patrimoine culturel des pays d'Europe est un franc succès. Pour faire un point sur les chiffres mis à dispositions sur le site, Europeana c'est : 31 millions d'images, 24 millions de textes 634 000 audio, 356 000 vidéos et 5000 documents en 3D. Ça fait plus de 56 millions de documents numérisés et recensés sur le site, cinq fois plus que Gallica.

Maintenant que j'ai réalisé une présentation de l'histoire et du fonctionnement de Europeana, je vais passer à l'analyse du site et surtout de la recherche de document dans le catalogue.

6.3.2 Analyse du site et des données mises à disposition

Commençons par la première vision que l'on à en arrivant sur le site, la page d'accueil. (Figure 15) Elle est au premier abord très sobre avec la barre de recherche qui est centrale, un peu comme un moteur de recherche et le fond de la page représente des œuvres européennes sur fond bleu comme le drapeau de l'UE. On à des onglets déroulant en haut qui permettent d'accéder à l'accueil, aux collections qui sont regroupés par thème. L'onglet Histoire qui sont des expositions en lignes ou des blogs qui se penchent sur des périodes ou moments précis de l'Histoire. Ensuite un onglet pour rejoindre le programme Europeana Pro et pouvoir échanger des connaissances. En rejoignant ce programme, il est aussi possible de publier des travaux qui sont en lien avec l'Europe. Pour finir, il y a un dernier onglet destiné à la connexion sur le site. Lorsqu'on défile la page d'accueil nous avons d'autres informations qui sont les mêmes que le bandeau en haut de la page que j'ai précédemment décrit. Ce qui nous intéresse c'est la barre de recherche qui permet de naviguer dans le catalogue. Il faut également noter qu'on ne reçoit pas de recherches suggérées par cette barre contrairement à Gallica.

Passons maintenant à ce que l'on obtient après avoir fait une recherche sur le site Europeana. On obtient une liste de documents associés aux sujets demandés, ici j'ai pris le même sujet que sur les autres plateformes. Plusieurs éléments sont présents sur la page pour consulter les données, tout d'abord on à la possibilité de choisir la visualisation de celles-ci. En effet, on peut voir les données avec le titre et l'image comme on peut voir sur la Figure 16, mais on peut aussi mettre les images beaucoup plus en valeur avec les deux autres options. Ensuite on à des informations qui sont données sur chaque document, en haut il y la provenance de l'article, juste en dessous son titre et à droite une image qui illustre le document. Tout en bas on peut voir quatre informations sur le document. Tout d'abord le In et No Copyright, le premier peut être utilisé par toutes les façons autorisées par la législation du droit d'auteur et des droits voisins. Cependant une autre utilisation demande l'autorisation de l'auteur. Pour le second le il faut automatiquement une autorisation si on veut utiliser cet objet. Il y a aussi sur chaque document son type (texte,vidéo,image etc...). A côté on à la possibilité de sauvegarder l'article et aussi de mettre un j'aime comme sur les réseaux sociaux.

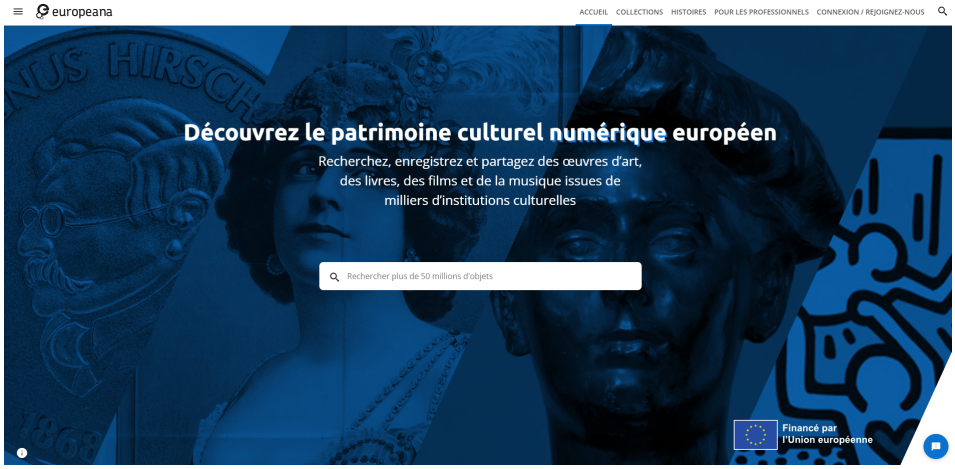


FIGURE 16 – Capture d'écran de l'écran d'acceuil du site Europeana.

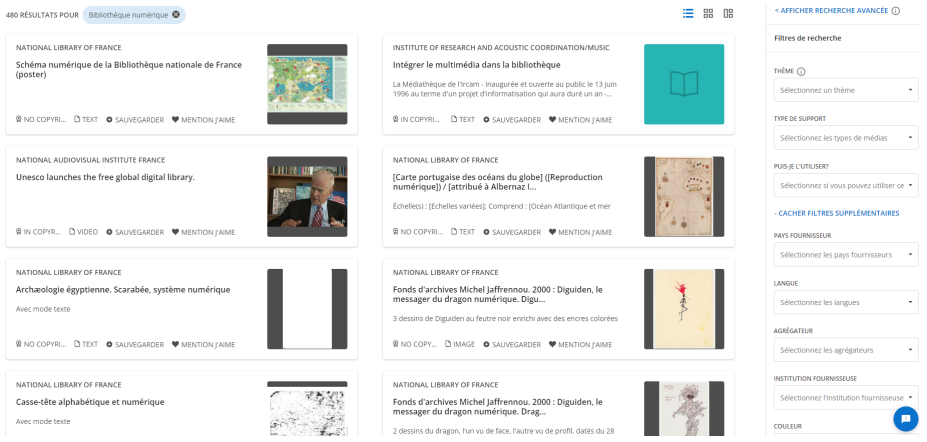


FIGURE 17 – Capture d'écran du résultat de la recherche sur le site Europeana.

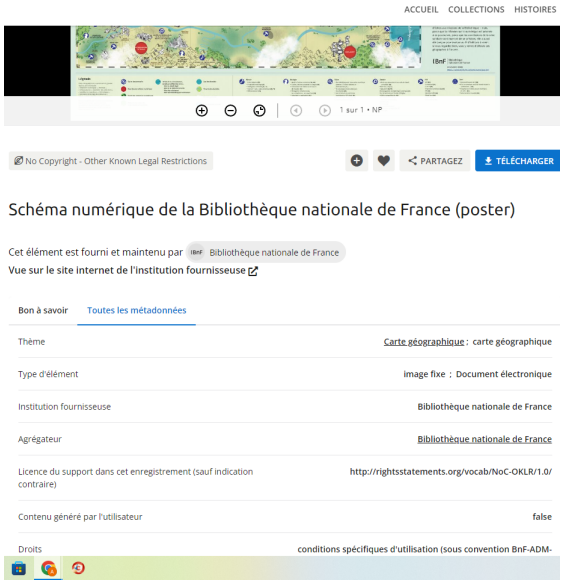


FIGURE 18 – Capture d'écran des métadonnées d'un document sur le site Europeana.

Pour terminer sur cette page à droite, il y a la possibilité d'utiliser des options de filtrages sur la recherche en cours. Je ne vais pas citer les filtres classiques tel que le thème qui sont présents dans les autres catalogues. Forcé de constater que Europeana propose des options très précises tel que le pays d'origine du document, la possibilité de savoir si on peut utiliser le document et avec quelle condition. L'utilisateur peut choisir la couleur, l'orientation de l'image, sa taille ainsi de son Rights Statement. Cette multitude de choix dans les options de filtrages est très intéressante dans le cadre de travaux de recherche sur des documents très précis. On pourrait par exemple analyser les différents types de documents sur un sujet donnés en fonction de son pays de provenance et ainsi voir les différences qu'il peut y avoir. Mise à part le titre et le type du document, Europeana ne donne pas l'accès à plus d'informations sur la page de recherche. Pour les avoir on doit cliquer sur le document et on obtient l'ensemble des métadonnées Figure 17). On à la visualisation du document en haut, ensuite son titre et le fournisseur de celui-ci, ensuite on à l'ensemble des métadonnées du document avec des spécificités à Europeana comme L'Agrégateur et l'institution qui à fourni le document. Mais aussi le pays fournisseur et un Horodatage qui permet de voir quand le document à été enregistré sur le site. Le type d'élément est mis à disposition et à pour objectif d'indexer le document en fonction de son type, si c'est une image fixe, une monographie etc... Néanmoins, il n'y a aucun mots-clés ou sujets qui permettent de donner des informations sur le contenu du document. C'est ce que j'ai cru avec l'exemple que j'ai utilisé, cependant en

regardant d'autres exemples ont peut voir que le thème de chaque documents réunis des informations sur le contenu de celui-ci. Cependant ces informations restent très globales et ne donnent pas autant d'informations que les sujets d'un document.

En conclusion, Europeana est un catalogue en ligne financé par la Commission Européenne qui à su s'imposer comme le plus important en Europe. Depuis sa création, c'est une idée collaborative et l'ensemble des pays de l'UE ont participé à étayer ce projet. La mise en place de plusieurs projets tel que Europeana Regia à participer à permis d'améliorer son catalogue. La plateforme propose plus de 50 millions d'objets numérisés en 2022 et dans toutes les langues officielles de l'UE. Le site est très facile dans sa navigation et il donne accès à beaucoup de documents mais aussi, des métadonnées très détaillées sur chacun d'eux. Il y a une volonté à travers le thème de chaque document, de vouloir ressembler aux fonctionnements des sujets sur WorldCat. Cependant les deux plateformes ne sont pas à la même échelle et surtout n'ont pas exactement le même objectif.

6.4 À l'échelle Internationale : Worldcat

Comment faire une partie sur les catalogues ou les bibliothèques en ligne sans parler de la plus grande de toute, WorldCat. Cette dernière partie va se dérouler de la même manière que les trois précédentes avec d'abord une présentation de WorldCat et ensuite je vais analyser le site et les différentes informations qu'ils proposent. Sur les quatre catalogues étudiés, WorldCat est celui qui est en lien direct avec le travail que j'ai réalisé en stage, en effet, l'entièreté des donnés traiter et analyser proviennent de ce catalogue.

6.4.1 Déscription et Historique de WorldCat

Connu comme étant le plus grand catalogue en accès public du monde , Worldcat à été créé par l'Online Computer Library Center (OCLC). Cette organisation est à but non lucratif et travaille avec les bibliothèques du monde entier. Leur objectif principal est de donner l'accès au plus grand nombre et le moins onéreux possible à l'information. WorldCat est la contraction entre World et Catalog, ce choix de nom montre déjà la volonté au départ d'être le plus grand catalogue mondial. Dès sa création en 1967, l'idée de développer un catalogue collaboratif était déjà là. C'est en 1971 que le catalogues verra ses premiers documents être enregistrés. Néanmoins le WorldCat que nous connaissons aujourd'hui n'était pas encore envisagé car les technologies de l'époque ne pouvaient pas permettre sa mise en place. Comme pour les autres catalogues que j'ai abordé précédemment, WorldCat à aussi eu une période de test sur le Web avec Open WorldCat. Ce programme mettait à disposition des notices abrégées aux sites qui étaient partenaires. Depuis 1996, il était possible d'accéder au catalogue de WorldCat mais uniquement pour les bibliothèques abonnés. Dix ans plus tard, en 2006, WorldCat.org est lancé et peut être utilisé par tous ceux qui ont un ordinateur et internet. WorldCat est un précurseur dans son

domaine et il va continuer à se développer en mettant en place au fil des années des nouvelles options disponible sur son catalogue. Il va par exemple mettre à disposition des métadonnées sur les documents de son catalogue dès 2007. La dernière mise à jour en date de WorldCat date de moins d'1 an en août 2022, en effet cette mise à jour apporte une plus grande accessibilité aux collections. Cette mise à jour a aussi remplacé les critiques des livres réalisés par les utilisateurs par les critiques provenant de GoodReads. [14]

L'objectif principal de WorldCat et de l'OCLC est de donner l'accès le plus ouvert possible aux informations de milliers de bibliothèques. En effet, Worldcat est une plateforme qui met l'accent sur la collaboration des bibliothèques. C'est grâce à elles que le catalogue continue de se développer avec le temps, le site WorldCat donne les chiffres du mois d'avril. Durant cette période il y a eu plus de 217 000 notices Worldcat enrichies par les bibliothèques membres du programme, 2,4 millions en 2022. Les bibliothèques en partenariat avec WorldCat sont elles aussi différentes. On peut retrouver des bibliothèques nationales, universitaires, publiques et spécialisées. Chacune d'entre elles contribue à enrichir le catalogue de WorldCat. L'utilisation du catalogue est très intuitive et peut être utilisée dans le cas de travaux de recherche mais aussi lors d'une utilisation personnelle. Pour conclure je dirais que WorldCat à su comprendre dès le départ les enjeux de la création d'une base de données. En effet, la volonté de donner accès aux informations aux plus grand nombre coïncide avec la création de cette base de données. WorldCat à su créer des partenariats avec des bibliothèques du monde entier, ce qui à permit au catalogue de se développer au fil du temps. Aujourd'hui WorldCat répertorie sur son site plus de 405 millions de livres, 440 millions d'articles, 25 millions d'enregistrements sonores, 10 millions de partitions musicales, 6 millions de cartes géographiques et 30 millions de Mémoire/Thèses. Ces chiffres montrent la différence très nette entre l'échelle européenne avec Europeana et l'échelle mondial de WorldCat. La description de l'histoire de WorldCat et son fonctionnement étant terminé, je vais me pencher sur ce que le site propose lorsque l'on recherche des documents.

6.4.2 Analyse du site et des données mises à disposition

Lorsqu'on arrive sur la page d'accueil de WorldCat (Figure 18) on peut voir que c'est une fusion entre Europeana et Gallica, en effet on retrouve la barre de recherche au centre de l'écran avec un fond bleu comme Europeana et on peut aussi voir des informations sur différents thèmes ici sur le Jazz, l'automobile et la cuisine. On peut trouver en haut de l'écran les différents onglets tels que l'accès à l'accueil, la liste des bibliothèques proche de chez nous, une liste de sujets qui sont déjà rangés par thèmes. Il y a aussi l'onglet liste qui permet à l'utilisateur de créer une liste avec les documents qu'il souhaite. Ensuite il y a l'onglet à propos qui explique le rôle de WorldCat et enfin un onglet pour bibliothécaire qui explique les avantages à devenir une bibliothèque WorldCat. Lorsqu'on défile l'écran d'accueil on retrouve des informations sur comment utiliser WorldCat, ainsi que des sujets mis en avant. On retrouve aussi les chiffres

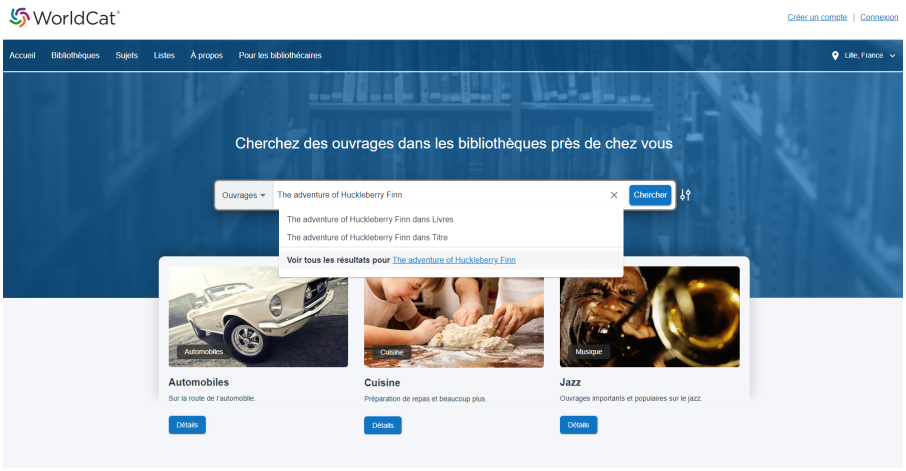


FIGURE 19 – Capture d'écran de la page d'acceuil du site WorldCat.

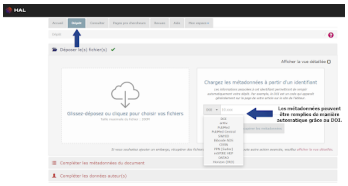


FIGURE 20 – Capture d'écran de la recherche avancée du site WorldCat.

des ressources disponibles sur le catalogue. Ce qui nous intéresse ici c'est la barre de recherche qui à plusieurs fonctionnalités. Tout d'abord on peut directement choisir à gauche de la barre entre des ouvrages, des bibliothèques et des listes. En fonction du choix, la barre change et apporte des fonctionnalités propres aux choix précédents.

On retrouve aussi à droite la possibilité de réaliser une recherche avancée (Figure 19) avec la possibilité de chercher en fonction de un ou plusieurs mots-clés, du titre, de l'auteur, de l'année, du format et de la langue.Lorsque qu'on déroule les onglets de la recherche avancée on peut voir une multitude d'autres options d'affinage. En plus de ceux énoncés précédemment on à l'ISBN, l'ISSN, le nom du périodique, le n° OCLC, l'éditeur et le sujet. Chaque option est très détaillée car elle permet d'utiliser le ET, OU et SAUF pour les mots-clés et les titres. Cette option permet d'affiner le plus possible sa recherche avant de l'effectuer et ça permet d'éviter d'être inondé par un flot de document qui ne serait pas intéressant. On peut aussi choisir une période précise au niveau de l'année. La barre de recherche de WorldCat est la plus complète des quatres catalogues étudiés. Elle permet de faire une sélection extrêmement précise en amont et ça permet d'éviter le bruit qu'ils peut y avoir lors de la recherche.

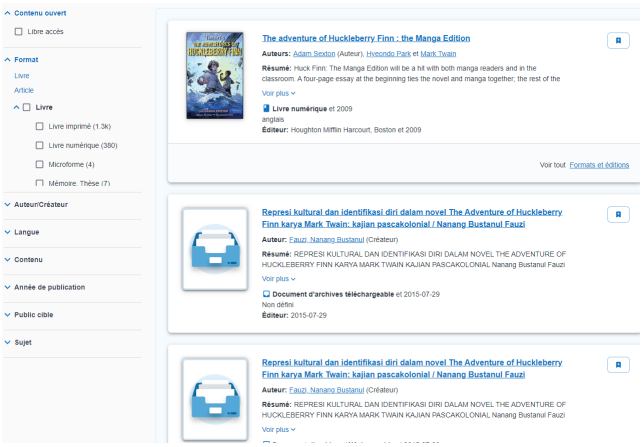


FIGURE 21 – Capture d'écran du résultat de la recherche sur le site WorldCat.

Après avoir entré la recherche, l'utilisateur accède à cette page qui ressemble à celle des autres catalogues (Figure 20). On à au centre le résultat de notre recherche avec des informations sur le document tels que son titre, son ou ses auteurs, un résumé qu'on peut défiler pour avoir plus de détails, le type de document, la date de publication et l'éditeur. On à tout de même plusieurs informations sur cette page ce qui permet pour des recherches qui ne seraient pas approfondies, de ne pas avoir besoin de cliquer sur chaque document. Sur la gauche on peut voir les différentes options de filtrages avec le même principe que sur Europeana. Un filtrage très détaillé qui permet d'affiner le plus possible sa recherche. On peut par exemple choisir le format si on veut les livres imprimés ou numériques. Les données que j'ai traitées dans le cadre de mon stage provenaient exclusivement des livres imprimés de The Adventure of Huckleberry Finn. Il y a d'autres options comme l'auteur, la langue, le contenu (fiction, documentaire,biographie etc...), l'année de publication et le sujet. On à même une option qui n'était présente dans les autres catalogues, le public cible par l'œuvre. Il est possible d'ajouter les documents à une liste comme sur Europeana.

Lorsque l'on clique sur le document souhaité, on arrive sur toutes les métadonnées du document qui sont beaucoup plus détaillées qu'à la page précédente (Figure 21). Il y a cependant, des informations qui sont propres au catalogue WorldCat. En effet, il y a la mise à disposition d'information tel que l'OCLC number qui est unique à l'instar du ISBN qui ne l'est pas forcément. On a aussi un élément que j'ai présenté dans le Chapitre 1, les sujets. Ces termes déjà existants permettent de décrire le contenu d'une œuvre. C'est d'ailleurs sur ces sujets que j'ai accès mon analyse qu'on abordera dans la dernière partie destinée au stage. Pour finir en bas de la page WorldCat propose la possibilité d'obtenir ce livre dans une bibliothèque et il référence les bibliothèques qui possèdent

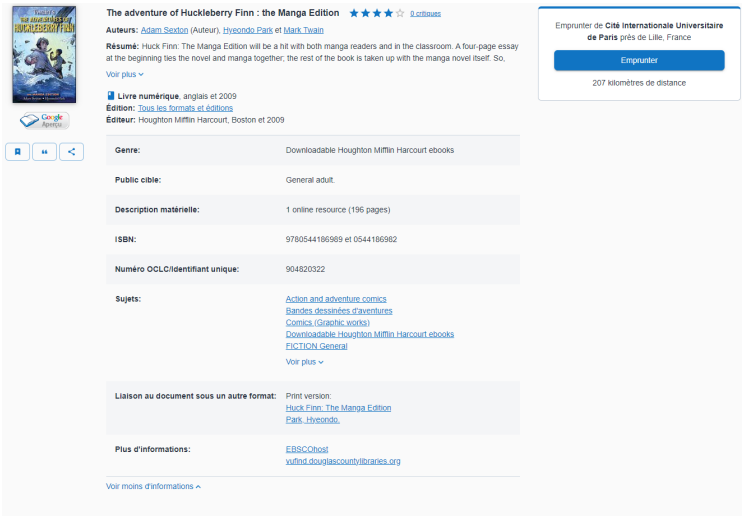


FIGURE 22 – Capture d'écran des métadonnées d'un document sur le site WorldCat.

cette œuvre. Avec la localisation, il met en évidence les bibliothèques les plus proches de chez soi par la localisation.

Il ne faut pas oublier que WorldCat est un catalogue qui référence une multitude d'œuvres avec des notices bibliographiques composées de métadonnées très détaillées. Cependant, beaucoup des ouvrages référencés de son pas consultable en ligne. Par exemple, The Adventure of Huckleberry Finn est beaucoup référencé mais 1 300 exemplaires sont des livres qui ne sont pas forcément consultables depuis WorldCat. C'est le catalogue le plus grand du monde et il permet de pouvoir consulter des millions de documents depuis chez soi. Les données que j'ai analysées durant mon stage proviennent de WorldCat, la raison et la mise à disposition des sujets. Ces termes permettent de décrire avec des termes, le contenu du document. Néanmoins, même s' ils permettent de décrire le contenu, ces sujets peuvent varier en fonction des langues et c'est cette question qui nous à intéresser durant ce stage. Après avoir décrit quatre catalogues en ligne avec des échelles différentes, je vais aborder les problèmes et les enjeux liés aux notices bibliographiques multilingues.

Troisième partie

Les notices bibliographiques multilingues : problématique et enjeux

7 Les enjeux et les difficultés dans la rédaction et l'exploitation des notices bibliographiques multilingues : L'exemple du projet MACS

Le contenu des parties précédentes m'a permis d'expliquer l'ensemble des procédés destinés au référencement, ainsi que les nouveaux outils utilisés pour l'indexation. J'ai ensuite proposé une présentation et une analyse détaillée des différents catalogues avec des échelles très différentes. Avec l'ensemble des idées traitées dans les deux parties précédentes, il est temps d'aborder le sujet en lien direct avec ma problématique, celui des notices bibliographiques multilingues. Nous l'avons vu avec WorldCat, le catalogue propose des notices bibliographiques dans plus d'une centaine de langues et celles-ci sont (sauf exception) toutes similaires. En effet chacune à un titre, une date, un genre et toutes ont des sujets qui lui sont propres. C'est cette dernière catégorie qui va nous intéresser dans cette partie car les sujets peuvent être différents pour une même œuvre en fonction des langues.

Les enjeux à travers la rédaction de notices bibliographiques multilingues c'est d'uniformiser l'ensemble de ces notices pour que toutes les notices veulent dire la même chose peu importe la langue de traduction de l'œuvre. [16] Un projet nommé MACS (Multilingual Access to Subjects) a d'ailleurs été commandé à CoBRA+ en 1997. [12]L'objectif de ce projet est de solutionner le problème à l'accès des sujets multilingues des notices bibliographiques. Pour se faire, le projet à réuni quatres bibliothèques qui ont accepté de rejoindre le projet : la Bibliothèque nationale suisse, la Bibliothèque nationale de France, la Bibliothèque allemande et la British Library. Ces bibliothèques sont à l'origine et la maintenance des vedettes matières tel que RAMEAU en français, SWD en allemand et LCSH en anglais. L'enjeu du projet est d'offrir la possibilité pour les utilisateurs de recherchés par sujet au sein des catalogues en fonction de la langue souhaitée. Une première version du projet est sortie début 2000. Les informations précédentes proviennent d'un article qui date de plus de 20 ans. [12], Il n'y a pas eu de suite au projet, on a une autre source provenant de la Bibliothèque nationale suisse datant de 2007 qui aborde le projet MACS. L'objectif reste le même, avoir un système qui permet la consultation d'un catalogue avec une seule question et ça peut importe la langue.

Dans le cadre de mon stage j'ai aussi dû analyser les sujets des notices bibliographiques de The Adventure of Huckleberry Finn. L'objectif était de voir s' il pouvait y avoir des différences entre les sujets. C'est la problématique majeure des notices multilingues, les sujets peuvent varier d'une langue à l'autre alors que c'est la même œuvre. Le problème vient du fait que chaque pays à son histoire et certains thèmes seront plus ou moins tabou ou alors selon les termes disponibles dans le langage d'indexation matière. Les mots présents dans RAMEAU ne sont pas les mêmes que dans LCSH et vice versa. Cet élément peut être la cause de problèmes au niveau des sujets qui seront présents dans la notice bibliographique. Un autre problème est que la langue des sujets est majoritairement en anglais et cela peut importe la langue de la traduction. C'est un point que je mettrais en valeur dans la partie suivante qui va se concentrer sur le travail réalisé en stage.

Ce troisième chapitre est l'encart théorique de mon mémoire. En effet, elle me permet de progressivement passer de l'aspect plutôt théorique des deux premières parties afin d'arriver à la dernière partie qui se concentre sur le déroulé de mon stage. Durant celui-ci j'ai à gérer un corpus de notices bibliographiques multilingues que j'ai dû traiter, décrire, analyser et en extraire des graphiques.

FIGURE 23 – Capture d'écran de la carte des traductions de The Adventure of Huckleberry Finn du projet ROSETTA.

Quatrième partie

Traitement d'une des notices multilingues d'une œuvre patrimoniale littéraire : The Adventure of Huckleberry Finn.

Toutes les étapes que j'ai réalisées précédemment je vais les expliquer dans cette partie, mais avant je vais présenter le projet Rosetta en lien direct avec mon stage. Ainsi que comment s'est déroulé mon stage, dans quel lieu et avec quel matériel. On retrouve sur le site de ROSETTA un résumé du projet que j'ai utilisé pour cette introduction. Le projet ROSETTA s'inspire de la pierre de Rosette qui à permit de déchiffrer les hiéroglyphes grâce aux traductions grecques et ne pas laisser ce langage disparaître. Le projet partage la même volonté, préserver les langues menacées de disparition. Pour faire ça, ce projet d'humanité numérique et collaboratif va avoir comme objectif de préserver le patrimoine culturel et à soutenir la diversité de connaissance grâce aux disciplines des Sciences de l'Information et la Communication, la Linguistique Computationnelle, La Littérature Américaine et la Traductologie. Le premier test réalisé ce déroule sur l'œuvre de Mark Twain, The Adventure of Huckleberry Finn. Le projet permet aussi une visualisation des traduction par nombres grâce à une carte du monde interactive (Figure 22).

J'ai donc réalisé mon stage avec Madame Amel Fraisse, maîtresse de conférence en Science de l'Information et de la Communication à l'Université de Lille et membre de laboratoire Gériico (Groupement d'Étude et de Recherche Interdisciplinaire en Information et Communication) . C'est de manière hybride que j'ai travaillé durant mon stage, une partie du temps à l'Université de Lille et

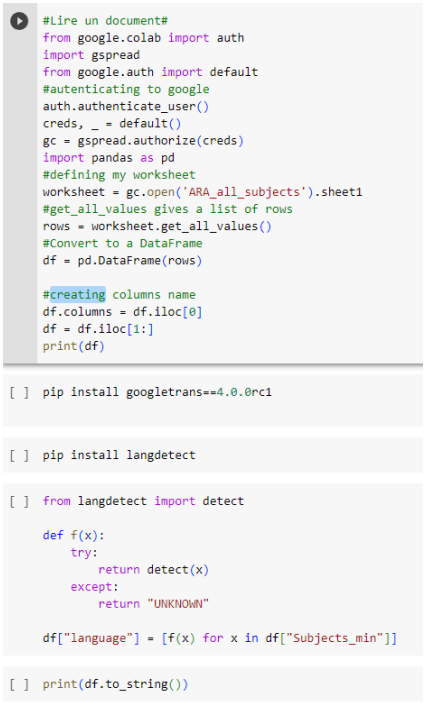
l'autre partie du temps à mon domicile. Le travail à effectuer durant le stage m'a permis une flexibilité au niveau de mon lieu de travail car j'avais uniquement besoin d'un ordinateur et d'une connexion Internet. Chaque semaine j'ai fait un point avec ma tutrice de stage, Madame Fraisse et nous nous échangions des mails lorsqu'il y avait des problèmes. Durant l'ensemble de mon stage j'ai reçu un très bon encadrement de la part de Madame Fraisse, que ça soit au niveau du travail effectué mais aussi au niveau de la rédaction de mon mémoire.

8 Nettoyage et prétraitement des données

Mon stage à débuté le lundi 03 avril dans le bureau de Madame Fraisse à l'Université de Lille. Nous avions déjà parlé du thème du mémoire mais je n'avais pas encore de vision concrète sur ce que j'allais faire pendant le stage. La première étape de mon travail va être de réceptionner, nettoyer et prétraiter un jeu de données.

Les données proviennent d'un dossier partagé sur Google Drive qui se nomme "L2SID_HSI_Data" et qui regroupe 86 documents. Ces documents sont des devoirs réalisés par des élèves de Licence 2 HSI. Ils avaient comme consigne de réaliser un document qui répertorie les notices bibliographiques de The Adventure Of Huckleberry Finn sur une langue choisie. Au sein de ce dossier les document peuvent être divisés tout d'abord en fonction de langues traitées dans le documents. Ensuite on à deux types de documents par langue, les données brut et les données analysées. Mon premier travail à été de vérifier l'ensemble des fichiers pour voir s' il n'y avait pas d'erreurs entre les données brut et les données analysées. Après avoir vérifié l'ensemble des documents je l' ai renommé en "Raw_Data_" pour les données brutes et "Data_Analysis" pour les données analysées par les élèves. Les problèmes rencontrés ont été le manque de données pour 1 langue (Deutsch) et aussi les différences entre les données brut et les données analysées.Parfois les chiffres étaient différents. Il à fallu donc tout regarder et porter une grande attention pour ne pas se tromper et changer si nécessaire.

La seconde mission a été de centraliser les données brutes autour d'un document complet par langue nommé "Raw_data_all". Pour ces données j'ai créé une nouvelle feuille de calcule Google Sheets et j'ai mis l'ensemble le contenues des fichier "Raw_Data" par langue. Ensuite,j'ai appliqué une formule (= =NB.SI(B :B ;B1)>1) sur la colonne "oclc-number" (Le numéro OCLC est une chaîne numérique de longueur variable avec des zéros en tête pour les numéros de moins de 8 chiffres et sans zéros en tête pour les numéros de 8 chiffres ou plus.) pour mettre en valeur les doublons avec une couleur choisie(Figure 23). J'ai ensuite sélectionné les doublons et je les ai supprimés. Avant ça j'avais eu l'idée d'utiliser la fonction sur Google Sheets pour supprimer les doublons d'une colonne choisie. Cependant je me suis rendu compte de mon erreur car les ça



```
#Lire un document#
from google.colab import auth
import gspread
from google.auth import default
#autenticating to google
auth.authenticate_user()
creds, _ = default()
gc = gspread.authorize(creds)
import pandas as pd
#defining my worksheet
worksheet = gc.open('ARA_all_subjects').sheet1
#get_all_values gives a list of rows
rows = worksheet.get_all_values()
#Convert to a DataFrame
df = pd.DataFrame(rows)

#creating columns name
df.columns = df.iloc[0]
df = df.iloc[1:]
print(df)
```

```
[ ] pip install googletrans==4.0.0rc1
```

```
[ ] pip install langdetect
```

```
[ ] from langdetect import detect

    def f(x):
        try:
            return detect(x)
        except:
            return "UNKNOWN"

    df["language"] = [f(x) for x in df["Subjects_min"]]
```

```
[ ] print(df.to_string())
```

FIGURE 24 – Script Python pour analyser la langue d'une colonne Google Sheets.

supprimer les doublons de la colonne et ça mettait mon fichier en désordre. J'ai vérifié s' il n'y avait pas d'erreurs et j'ai enregistré le document dans un dossier All Data. Pour les données analysées j'ai aussi réalisé un document qui regroupe toutes les données et plus précisément les sujets de chaque document que j'ai nommé "all_subject". En effet, ces données nous intéressent dans le cadre de nos analyses. En ce qui concerne les données analysées, j'ai récupéré l'ensemble des sujets que j'ai récupéré sur la deuxième feuille des documents "Data_analysis" de chaque langue. J'ai mis l'ensemble en commun dans un document Google Sheets que j'ai nommé "all_subject" et j'ai utilisé la même formule que pour la data brutes qui m'a permis de mettre en évidence les doublons et les surprimes. Cependant il y avait un problème car la formule ne détectait pas les doublons s' il y avait un majuscule ou une minuscule au début de chaque mot. Pour contrer ça j'ai appliqué une formule (=LOWER) et j'ai créé une colonne "Subjects_min" avec l'ensemble des sujets en minuscule. J'ai appliqué de nouveau la formule pour mettre en valeur les doublons et encore plus affiner les sujets. J'ai préféré utiliser la formule (= =NB.SI(B :B ;B1)>1) sur Google Sheets car j'ai expliqué que le document était en désordre après l'utilisation de l'option "supprimer les doublons". Cela a été le cas, j'avais utilisé cette fonction de Google Sheets et lors d'un entretien avec Madame Fraisse nous avons remarqué que les sujets

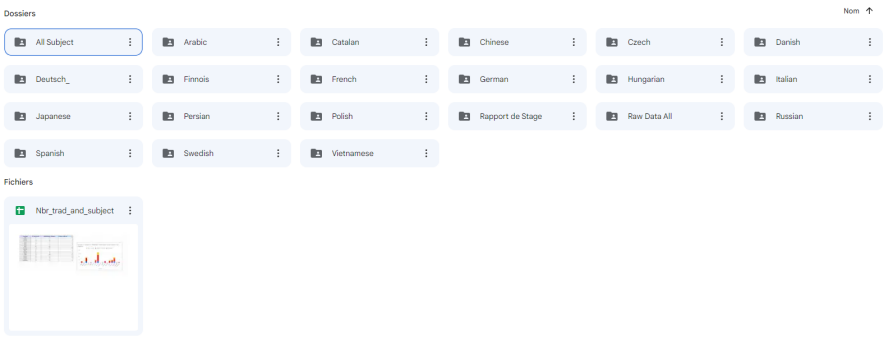


FIGURE 25 – Organisation des dossiers du corpus utilisé durant le stage

n'étaient pas en accord avec la langue qui lui était assignée. Nous avons décidé ensemble d'utiliser un script Python qui me permettrait de détecter la langue de l'ensemble des sujets pour vérifier. J'ai donc cherché pendant plusieurs jours et trouvé comment réaliser cette opération (Figure 23), cependant je me suis rendu compte que l'erreur venait de moi en amont et après avoir rectifié ce script ne me serait plus utile. C'était ma première erreur que j'ai su corriger mais qui m'a tout de même fait perdre du temps. Aparté fini, voici le résultat final des données trier et ranger dans les dossiers pour chaque langue et un dossier pour les données brutes et un autre pour les sujets(Figure 24). J'ai en plus de ça réalisé un tableau qui récapitule le nombre de traduction, les sujets multilinguals et les sujets cibles.

Le traitement des données m'a pris du temps car il à fallu affiner le plus possible les données pour commencer le travail d'analyse.Après avoir expliqué le nettoyage et le prétraitement des données je vais passer à la description du corpus sur lequel je vais étudier.

9 Description du corpus

Le travail de nettoyage et prétraitement des données étant réalisés, j'ai un corpus prêt à être analysé. Cependant, avant ça je vais décrire le corpus sur lequel je vais travailler. J'ai expliqué dans la partie précédente comment j'ai traiter les données pour constituer mon corpus, mais je n'ai pas parlé du contenu de celui-ci.

Le corpus provient des élèves de L2 HSI qui ont réalisé par groupe un projet de référencement dans deux documents des notices bibliographiques de The Adventure of Huckleberry Finn. Toutes ces notices ont été récupérées sur le catalogue en ligne WorldCat en fonction de la langue traiter. Je vais montrer comment avec ce corpus j'ai réalisé un autre corpus qui réunit toutes les informations au même endroit en le plus affiné possible. Tout d'abord, je vais commencer dans un premier temps par les données de "Raw_Data" qui est constitué de 17 documents par langue de traduction. Je vais d'ailleurs énoncer toutes ces

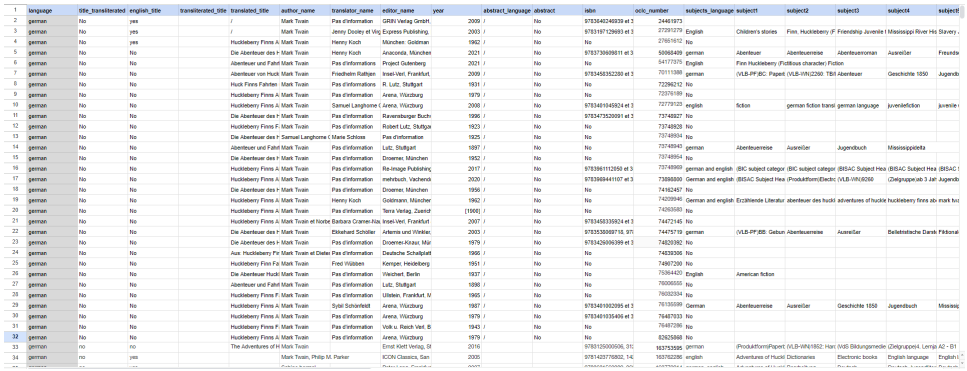
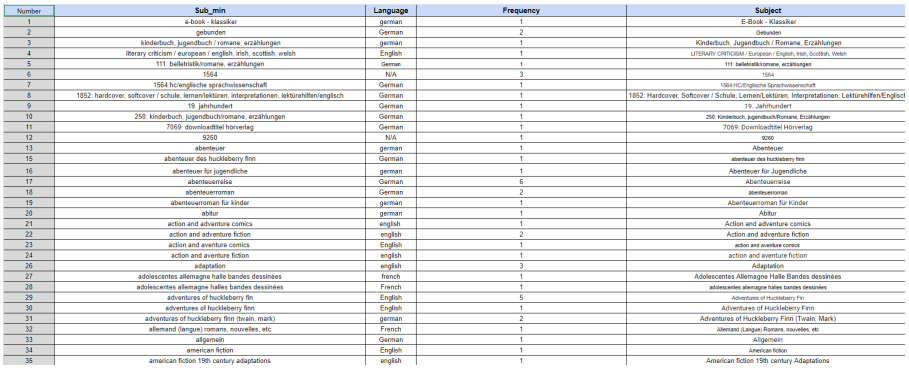


FIGURE 26 – Tableau des données de "GER_raw_data_all.

langues pour situer quelles sont celles que j'ai traitées durant mon stage. Avant ça, nous avons fait le choix avec Madame Fraisse de ne pas utiliser le Deutch car il manquait des documents dans le corpus de base et il valait mieux éviter de travailler avec des données incomplètes. Nous avons donc comme langue traité : Arabe, Allemand, Catalan, Chinois, Danois, Espagnol, Finnois, Français, Hongrois , Italien, Japonais, Persan, Polonais, Russe, Suedois, Tchèque et Vietnamien. Pour décrire le corpus nous allons utiliser comme exemple le document sur les traductions allemandes, "GER_raw_data_all". On peut voir sur la Figure 25, les différentes colonnes présentes dans l'ensemble des documents de "Raw_Data_All", Je commence tout à gauche pour finir sur la droite de la feuille de calcule. Tout d'abord tout à gauche la langue de la notice exploités, à côté la colonne si oui ou non le titre est translittéré. Ensuite si le titre est en anglais ou non, si le titre est translittéré la quatrième colonne sert à mettre ce titre. Après ça il y a les colonnes qui répertorie les titres traduits, le nom de l'auteur, celui du traducteur, de l'éditeur et enfin la colonne avec l'année. Par la suite on retrouve deux colonnes pour le résumé et la langue de celui-ci. Pour terminer, il y a les colonnes les plus importantes qui permettent d'identifier le document. D'abord l'ISBN, mais ici il n'est pas indispensable car c'est plutôt sur l'OCLC number qu'on va s'appuyer pour trier les documents,en effet, les notices viennent de WorldCat qui utilise ce numéros unique, qui permet d'identifier avec certitude une oeuvre. Les dernières colonnes sont ensuite la langue et les différents sujets de chaque notice. J'ai centralisé toutes les données brutes de chaque notice par langue et cet ensemble de données est mon corpus. Toutefois, j'ai aussi réuni dans un deuxième temps les données des documents de "Data_Analysis" pour en faire la aussi un corpus globalisé.

Les élèves devaient en effet réaliser deux dossiers bien distincts avec d'un côté des données brutes et de l'autre les données analysées. Dans ce second ils se sont penchés plus précisément sur des données présentes dans le "Rax_Data" pour en faire un document analytique. Dans celui-ci on retrouve deux feuilles,



Number	Sub_min	Language	Frequency	Subject
1	e-book - klassiker	german	1	E-Book - Klassiker
2	gebunden	German	2	Gebunden
3	kinderbuch, jugendbuch / romane, erzählungen	german	1	Kinderbuch, Jugendbuch / Romane, Erzählungen
4	literary criticism / european / english, irish, scottish, welsh	English	1	LITERARY CRITICISM / European / English, Irish, Scottish, Welsh
5	111: belletristik/romane, erzählungen	German	1	111: belletristik/romane, erzählungen
6	1564	N/A	3	1564
7	1564 hc/englische sprachwissenschaft	German	1	1564 HC/Englische Sprachwissenschaft
8	1852: hardcover, softcover / schule, lernen/lektüren, interpretationen, lektürehilfen/englisch	German	1	1852: Hardcover, Softcover / Schule, Lernen/Lektüren, Interpretationen, Lektürehilfen/Englisch
9	19. jahrhundert	German	1	19. Jahrhundert
10	250: kinderbuch, jugendbuch/romane, erzählungen	German	1	250: Kinderbuch, Jugendbuch/Romane, Erzählungen
11	7069: downloadtitel hörverlag	German	1	7069: Downloadtitel Hörverlag
12	9260	N/A	1	9260
13	abenteuer	german	1	Abenteuer
15	abenteuer des huckleberry finn	German	1	abenteuer des huckleberry finn
16	abenteuer für jugendliche	german	1	Abenteuer für Jugendliche
17	abenteuerreise	German	6	Abenteuerreise
18	abenteuerroman	German	2	abenteuerroman
19	abenteuerroman für kinder	german	1	Abenteuerroman für Kinder
20	abitur	german	1	Abitur
21	action and adventure comics	english	1	Action and adventure comics
22	action and adventure fiction	english	2	Action and adventure fiction
23	action and aventure comics	English	1	action and aventure comics
24	action and aventure fiction	english	1	action and aventure fiction
26	adaptation	english	3	Adaptation
27	adolescentes allemagne halle bandes dessinées	french	1	Adolescentes Allemagne Halle Bandes dessinées
28	adolescentes allemagne halles bandes dessinées	French	1	adolescentes allemagne halles bandes dessinées
29	adventures of huckleberry fin	English	5	Adventures of Huckleberry Fin
30	adventures of huckleberry finn	English	1	Adventures of Huckleberry Finn
31	adventures of huckleberry finn (twain, mark)	german	2	Adventures of Huckleberry Finn (Twain, Mark)
32	allemand (langue) romans, nouvelles, etc	French	1	Allemand (Langue) Romans, nouvelles, etc
33	allgemein	German	1	Allgemein
34	american fiction	English	1	American fiction
35	american fiction 19th century adaptations	english	1	American fiction 19th century Adaptations

FIGURE 27 – Tableau des données de "GER_all_subjects.

une première qui répertorie le nombre de traduction, le nombre de résumé, le total de sujets en anglais et le nombre de sujets cibles, c'est-à-dire de la même langue que les traductions, ensuite le tableau continuait avec les sujets et leurs langues. La seconde page est encore plus importante car c'est avec celle-ci que j'ai centralisé les données pour chaque langue. On y retrouve quatre colonnes, la première qui sert juste de numérotation, la seconde ce sont les sujets, la troisième leurs langues et pour finir la fréquence d'apparition dans les notices bibliographiques traitées. J'ai utilisé le même pattern pour mon document qui réunit l'ensemble des sujets trier. La seule différence c'est le rajout de la colonne "Sub_min" qui représente tout simplement les sujets en minuscule (FIgure 26). Au final mon corpus est composé au total de 603 traductions provenant de 17 langues différentes. On retrouve après affinage, 1291 sujets 902 multilingue et 389 sujets cibles.

Avec madame Fraisse, nous avons décidé d'accès la réflexion du stage autour des notices bibliographiques multilingues et plus précisément sur l'étude des sujets présents au sein de celles-ci. C'est avec cette idée en tête que j'ai centralisé toutes les informations du corpus de base avec les notices bibliographiques mais aussi sur les sujets. Maintenant que mon corpus de données est présenté, je vais dans la partie suivante analyser les données pour en tirer des conclusions sur les notices bibliographiques multilingues.

10 Analyses et visualisation des données

Toutes les étapes ont été réalisées sur le corpus de données et il faut désormais analyser les résultats trouvés. A travers ces analyses et ses visualisation de données je vais mettre en valeur les points importants des notices multilingues et de l'utilisation des sujets. En effet, c'est le sujet majeur de mon mémoire de comprendre les méthodes computationnelles des notices bibliographiques multilingues mais aussi de voir quelles peuvent être leurs problématiques. Je vais

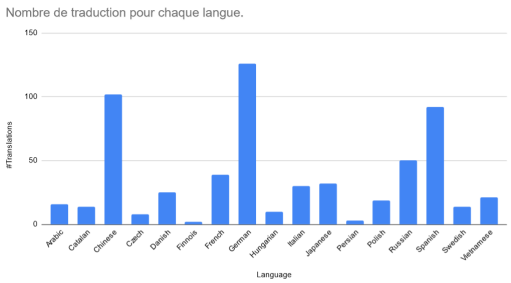


FIGURE 28 – Graphique du nombre de traductions par langue.

commencer par analyser les données dans leurs globalités et ensuite je vais me pencher sur les sujets cibles et multilingues. Je terminerais par un point que nous n'avons pas pu étudier au cours du stage mais qui serait important de traiter par la suite.

Je vais commencer par analyser le nombre de traductions par langues et la différence entre les sujets multilingues et cibles. Au préalable je vais analyser le nombre de traductions recensées par langue, en effet sur le premier graphique (Figure 27), on peut voir que c'est totalement disproportionné. Il y en à trois qui sortent du lot, le chinois, l'allemand et l'espagnol, elles sont toutes au dessus de 70 traductions et même plus de 100 pour le chinois et l'allemand. On peut retrouver en dessous des langues telles que le français, le russe, le japonais et l'italien. On remarque qu'en fonction de la langue il y a plus ou moins de traductions. Les langues "fortes" avec une grande population et une langue reconnue mondialement ont plus de traduction. Ces langues sont aussi surement plus référencés et ont une demande plus importante de traduction. Si je prends l'exemple du persan, il y a très peu de traduction et ça peut s'expliquer par le manque de demande et aussi le manque de référencement sur le catalogue WorldCat. Néanmoins, le fait d'avoir plus ou moins de traductions n'est pas forcément très grave dans le cadre de notre étude. En effet, nous avons décidé avec Madame Fraisse de nous pencher sur les sujets pour chaque langue et d'analyser la différence de nombre entre sujet multilingue et sujet cible. J'ai réalisé un graphique qui permet de le mettre en valeur, le résultat est presque sans appel.

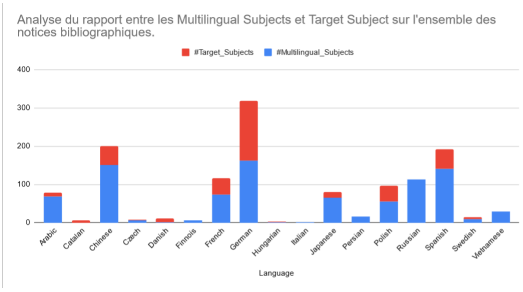


FIGURE 29 – Graphique du rapport entre les Target Subjects et les Multilinguals Subjects.

Les sujets sont en grande majorité multilingues. La seule exception provient des notices bibliographiques allemandes qui ont presque autant de sujets cibles que de multilingues. Pour le reste du corpus ont voit nettement la domination des notices multilingues. En voyant ce résulats on comprend les problèmes des notices bibliographiques multilingues, c'est qu'ils n'utilisent même pas des sujets propre à leur à leur langue. Les sujets sont des termes qui permettent de décrire une œuvre et son contenu. Ces termes sont pour la plupart au sein de thésaurus de langage d'indexation matière tel que LCSH en Anglais ou RAMEAU en français. C'est référentiel servent dans le cadre du choix de sujets pour les notices bibliographiques et c'est pour ça que pour le vietnamien, le persan ou encore le finnois, les sujets utilisés sont multilingues car ils proviennent de ces langages d'indexation. Je me suis penché plus précisément sur 7 langues qui avait une grande quantité de sujet recensé. Je me suis concentré sur le français, l'allemand,le chinois, l'espagnol, le polonais, le japonais et le russe. J'ai réalisé des graphiques pour chacune des sept langues choisies. Voici les résultats obtenues :

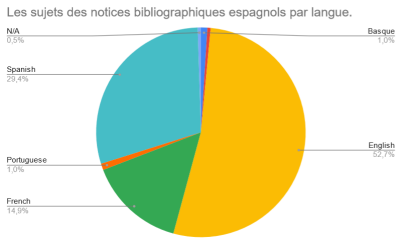


FIGURE 30 – Graphique sur les sujets des notices bibliographiques espagnols par langue.

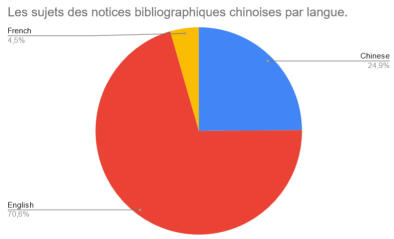


FIGURE 31 – Graphique sur les sujets des notices bibliographiques chinoises par langue.

Lorsque l'on voit les résultats on remarque que les sujets en anglais sont beaucoup plus présents que tous les autres, en effet, il y a presque 60% des sujets (59.3% exactement) qui sont anglais. La raison est l'utilisation pour écrire des sujets de langages d'indexation est ici c'est l'anglais LCSH. Il est utilisé pour trouver des termes qui sont déjà pré enregistrés et qui vont permettre lors de la rédaction de la notice bibliographique de les utiliser. L'analyse de ces données m'a permis de mettre en valeur le manque de représentativité des sujets cibles au sein des notices bibliographiques. On voit clairement l'impact des langages d'indexation car même le nombre de sujets en allemand et en français est aussi

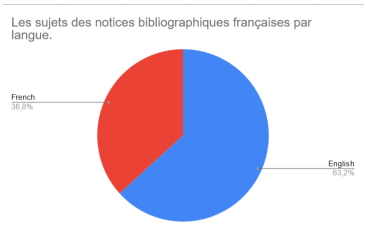


FIGURE 32 – Graphique sur les sujets des notices bibliographiques françaises par langue.

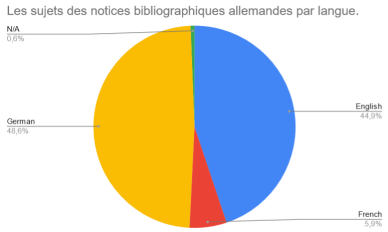


FIGURE 33 – Graphique sur les sujets des notices bibliographiques allemandes par langue.

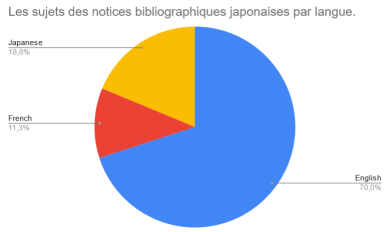


FIGURE 34 – Graphique sur les sujets des notices bibliographiques japonaises par langue.

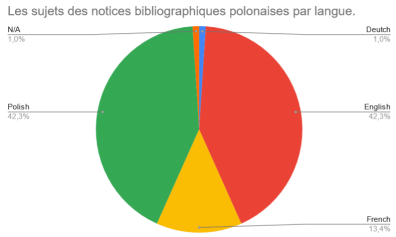


FIGURE 35 – Graphique sur les sujets des notices bibliographiques polonaises par langue.

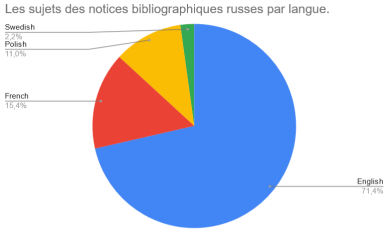


FIGURE 36 – Graphique sur les sujets des notices bibliographiques russes par langue.

très important. Mes visualisations graphiques m'ont permis de mettre en lumière ce point au cours de mon stage.

Il y a un troisième thème que nous n'avons pas pu traiter durant le stage, c'est les différences de sujets en fonction de la langue, j'ai soumis l'idée à Madame Fraisse d'étudier si les sujets sont similaires entre les langues ou s' il y a des différences. Mais aussi si des sujets sont surreprésentés ou sous représentés en fonction du pays. Malgré les langages d'indexation, certaines œuvres peuvent être vues différemment en fonction du pays et c'est ce que j'aurais voulu mettre en lumière dans cette dernière analyse. Pour le faire j'ai dû réaliser un script python qui lit et traite la colonne "Sub_min" de chaque langue pour diviser les sujets de plusieurs mots en plusieurs sujets de un seul mot. J'ai réussi à trouver un script qui fonctionne à la fin du stage (Ajouter Photo script), mais malheureusement nous n'avons pas réussi à l'appliquer alors que tout était bon. Mon stage est terminé mais mon travail sur les notices bibliographiques multilingues ne l'est pas. C'est un travail qui est encore en cours de traitement et ce mémoire de stage sera voué à évoluer lui aussi.

Le chantier des notices bibliographiques multilingues est encore en cours et des tentatives comme le projet MACS vont peut être permettre l'évolution dans le domaine. Le traitement des données m'a permis de mettre en lumière les différentes problématiques liées aux notices bibliographiques. On comprend que les sujets utilisés dans ces notices proviennent des langages d'indexation tels que le RAMEAU, LCSH par exemple. Je n'ai malheureusement pas pu terminer l'analyse de mes données et plus précisément sur les sujets utilisés en fonction de la langue. Ce n'est pas une fin en soi car je vais pouvoir dans le futur continuer le projet au-delà du stage et de mon mémoire.

Conclusion

Les bibliothèques sont les temples de l'esprit et de la connaissance pour les Hommes, le besoin d'organiser et conserver les textes sont à l'origine de leurs créations. Déjà existantes au IIIe siècle avant notre ère avec la Bibliothèque d'Alexandrie, ces lieux de savoirs sont encore au cœur des sociétés, cependant les besoins ont évolué avec le temps et les contenus bibliographiques sont de plus en plus nombreux. Pour pallier ce besoin il à fallu trouver un moyen de catégoriser les livres en notice bibliographique qui reprend les informations principales de l'œuvre pour faciliter son référencement. Ces notices ont évolué et les méthodes de référencement avec elles. Le flot de notices bibliographiques à poussé les chercheurs dans le domaine des Science de l'Information et la Documentation à mettre en place des systèmes permettant l'organisation et la représentation efficace de ces notices.

J'ai abordé ce thème dans mon premier chapitre qui est l'État de l'Art de mon mémoire, c'est pour cette raison que j'ai commencé mon développement en expliquant ce que sont les notices bibliographiques et quels sont leurs rôles. Le monde de la bibliothéconomie évolue et avec lui ses outils. La mise en place du modèle FRBR et ses extensions, mais aussi des référentiels d'indexation tel que RAMEAU, LCSH et du traitement automatique des données et pour finir les systèmes d'indexation et le vocabulaires contrôlées. L'installation de ces systèmes permet de transformer le monde de la bibliothéconomie et de faciliter le référencement, le traitement et la navigation des notices bibliographiques, c'est le sujet que j'ai abordé dans le deuxième chapitre en expliquant et analysant les catalogues en ligne à quatre échelles différentes. J'ai commencé avec l'Université de Lille et LillOA, ensuite Gallica un catalogue national, ensuite je me suis attardé à l'échelle européenne avec Europeana et pour terminer le WorldCat le plus grand catalogue en ligne du monde. Cette partie m'a permis de comparer ces quatres catalogues pour voir les points communs et les différences. Mon troisième chapitre est court mais il m'a permis de réaliser une transition entre la partie théorique et la partie pratique de mon développement. J'ai abordé les enjeux et les problématiques liés aux notices bibliographiques multilingues. Je me suis appuyé sur un article qui parle du projet MACS qui montre que la volonté de traiter ces données de manières uniformisées était déjà présente il y a plus de 20 ans. Mon dernier chapitre s'est concentré sur le travail que j'ai effectué durant mon stage, le corpus que j'ai traité et nettoyé sont de notices bibliographiques multilingues de The Adventure of Huckleberry Finn provenant de WorldCat. J'ai ensuite constitué mon corpus et j'ai réalisé des visualisations de données qui m'ont permis de répondre en partie à ma problématique, en effet, il existe de nombreuses méthodes computationnelles pour explorer et traiter des notices bibliographiques multilingues. Le corpus de documents que j'ai traité et analysé le montre, trouver des notices bibliographiques multilingues et les répertorier est possible. C'est notamment grâce à WorldCat qui met à disposition des centaines de millions de ces notices. Un de mes constat c'est la difficulté à analyser ces

données en détails avec des questions très précises. Les raisons sont diverses, les sujets qui sont majoritairement multilingues, des notices bibliographiques de “plus petit” pays qui sont moins fournis par exemple.

Le traitement des notices bibliographiques multilingues est un sujet qui est complexe. Je ne suis pas parvenu à trouver la solution finale à ma problématique mais j’ai apporté des hypothèses qui peuvent être des éléments de réponse. La mise en place de projets tels que MACS très ancien et plus récemment le projet ROSETTA, montrent la volonté de travailler sur ce sujet des notices bibliographiques multilingues. Ce dernier à un objectif très clair, mettre en valeur l’ensemble des langues et cultures du monde qui ne bénéficie pas des moyens techniques liés à l’information. C’est une cause qui est très importante à mes yeux et j’espère qu’à travers mon stage et mon mémoire j’ai pu mettre une petite pierre à ce très grand édifice.

Références Bibliographiques

[1] Étienne Cavalié (DIR.) *L'indexation matière en transition : de la réforme de Rameau à l'indexation automatique*. Bibliothèques. Éditions du Cercle de la librairie, 2019. ISBN : 978-2-7654-1623-4.

[2] Muriel AMAR. *Les Fondements théoriques de l'indexation : Une approche linguistique*. Paris : Association des professionnels de l'information et de la documentation (ADBS), 2000. ISBN : 2-84365-042-9.

[3] Claude ARSENAULT. "L'utilisation des langages documentaires pour la recherche d'information". In : *Documentation et bibliothèques* 52.2 (2006), p. 139-148. URL : https://doi.org/10.7202/1030017ar.

[4] Sophie BERTRAND et Aline GIRARD. "Gallica (1997–2016) : de la bibliothèque de «l'honnête homme» à celle du Gallicanaute". In : *Bulletin des bibliothèques de France (BBF)* n° 9 (2016), p. 48-59. ISSN : 1292-8399. URL : https://bbf.enssib.fr/consulter/bbf-2016-09-0048-005.

[5] BIBLIOTHÈQUE NATIONALE DE FRANCE. *Modèles FRBR, FRAD et FRSAD*. https://www.bnf.fr/fr/modeles-frbr-frad-et-frsad.

[6] Arlette BOULOGNE. "L'usage des références et des notices bibliographiques : historique et pratiques actuelles". In : *Documentaliste-Sciences de l'Information* 39.4-5 (2002), p. 174-180. DOI : 10.3917/docsi.394.0174. URL : https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2002-4-page-174.htm.

[7] Élisabeth FREYRE. "Les bibliothèques nationales et l'Europe". In : *Bulletin des bibliothèques de France (BBF)* (2011), p. 56-59. ISSN : 1292-8399. URL : https://bbf%20.enssib.fr/consulter/bbf-2011-02-0056-011.

[8] IFLA WORKING GROUP ON FUNCTIONAL REQUIREMENTS AND NUMBERING OF AUTHORITY RECORDS (FRANAR) et PATTON, GLENN E. *Functional Requirements for Authority Data - A Conceptual Model*. IFLA Publications 34. Berlin/Munich : De Gruyter Saur, jui 2013. ISBN : 9783598242823. URL : https://repository.ifla.org/handle/123456789/757.

[9] Wallace KIRSOP. "Bibliothèques numériques, catalogues en ligne et bibliographie matérielle". In : *Réforme, Humanisme, Renaissance* 88.1 (2019), p. 207-220. DOI : 10.3917/rhren.088.0207. URL : https://www.cairn.info/revue-reforme-humanisme-renaissance-2019-1-page-207.htm.

[10] Steven A. KNOWLTON. "Three Decades Since Prejudices and Antipathies : A Study of Changes in the Library of Congress Subject Headings". In : *Cataloging & Classification Quarterly* 39.3/4 (2005), p. 123-146. DOI : 10.1300/J104v39n03_06.

[11] "La transition bibliographique : le modèle FRBR David Forfait Sous la direction de Monsieur Philippe Bourdenet". Mém. de mast.

[12] Patrice Landry. “Le projet MACS : Accès multilingue aux sujets (LCSH, RAMEAU, SWD)”. In : *Catalogage international et contrôle bibliographique* 30 (2001), p. 46-49.

[13] Olivier Le Deuff. “Utopies documentaires : de l’indexation des connaissances à l’indexation des existences”. In : *Communication et Organisation* 48 (2015). Consulté le 16 juin 2023. doi : 10.4000/communicationorganisation.5082. url : http://journals.openedition.org/communicationorganisation/5082.

[14] Gary Price. “OCLC annonce officiellement le lancement d’un WorldCat.org « repensé et réimaginé »”. In : *Library Journal* (2022). url : https://www.infodocket.com/2022/08/24/wc/.

[15] Marina Rennesson et al. “Le thésaurus, un vocabulaire contrôlé pour parler le même langage”. In : *Médecine Palliative* 19.1 (2020). Documentation et pratiques documentaires en soins palliatifsCoordonné par Caroline Tête, p. 15-23. issn : 1636-6522. doi : https://doi.org/10.1016/j.medpal.2019.09.003. url : https://www.sciencedirect.com/science/article/pii/S1636652220300052.

[16] Pat Riva. “The multilingual challenge in bibliographic description and access”. In : *JLIS.it* 13.1 (jan. 2022), p. 86-98. url : https://doi.org/10.4403/jlis.it-12737.

[17] Barbara Tillet. *What Is FRBR ? : A Conceptual Model for the Bibliographic Universe*. Rapport technique 8. Library of Congress, février 2004.