

Axelle Lecroq

Licenciée en histoire

Licenciée en LLCER Allemand

Maître en Histoire médiévale

Entre enrichissement et développement de projets

**l'utilisation de données externes pour la
correspondance d'Alexander von Humboldt**

Mémoire pour le diplôme de master

Technologies numériques appliquées à l'histoire

sous la direction d'Ariane Pinche

2021



Résumé

Ce mémoire a été réalisé à la suite d'un stage effectué à la *Berlin-Brandenburgische Akademie der Wissenschaft* (BBAW) du mois d'avril à fin juillet 2021 dans le cadre du Master Technologies numériques appliquées à l'histoire de l'École nationale des chartes. Il présente diverses réalisations menées au sein du projet académique *Alexander von Humboldt auf Reisen - Wissenschaft aus der Bewegung* qui met en oeuvre l'édition numérique et imprimée des carnets de voyage et de la correspondance du scientifique Alexander von Humboldt. Le premier projet a été le développement de fonctions de recherche et de visualisations de la correspondance d'Alexander von Humboldt au sein d'un Jupyter Notebook. Le second avait pour objectif d'enrichir les données de correspSearch, portail web initié par la BBAW et proposant des lettres issues de correspondances éditées. Puisqu'au sein de ces deux missions les données utilisées sont des données externes, ce mémoire revient également sur les notions de web sémantique, de l'*open source* et du *linked open data*.

Mots-clés : correspondances, sciences du XIX^e siècle, Alexander von Humboldt, histoire scientifique allemande, histoire scientifique européenne, échanges épistolaires, carnets de voyage, Python, Jupyter Notebook, Application Flask, correspSearch, BBAW, humanités numériques, visualisation de données

Informations bibliographiques : Axelle Lecroq, *Entre enrichissement et développement de projets : l'utilisation de données externes pour la correspondance d'Alexander von Humboldt*, mémoire de master Technologies numériques appliquées à l'histoire, sous la direction d'Ariane Pinche, École nationale des chartes, 2021.

Remerciements

Le stage duquel est issu ce mémoire n'aurait pu avoir lieu sans l'aide financière du programme Erasmus+ ainsi que de l'Organisation franco-allemande pour la jeunesse (OFAJ). Les bourses que j'ai reçues ont été une aide précieuse pour effectuer ce stage dans les meilleures conditions possibles. Dans ce sens, je remercie particulièrement monsieur Alexis de Canck, responsable des relations européennes et internationales de l'École nationale des chartes ainsi que madame Jennifer Lauer, responsable des échanges universitaires et des volontariats au sein de l'OFAJ pour leur disponibilité et leur accompagnement.

Je voudrais remercier toute la communauté scientifique et administrative de l'Académie des sciences de Berlin-Brandebourg qui m'a accueillie et intégrée dès mon arrivée et ce, tout au long de mon séjour me permettant de profiter pleinement de cette expérience. Mes pensées vont particulièrement à monsieur Ulrich Päßler, chercheur et responsable adjoint du projet *Alexander von Humboldt auf Reisen - Wissenschaft aus der Bewegung* qui a également eu le rôle de tuteur de stage. Merci pour son encadrement et tous les échanges riches au cours de ces quatres mois d'apprentissage. Je remercie également monsieur Tobias Kraft, chercheur et responsable de ce projet académique ainsi que monsieur Christian Thomas, chercheur en son sein pour leur accompagnement, leur disponibilité et leur intérêt. Un remerciement particulier se tourne vers monsieur Gordon Fischer qui m'a accompagné techniquement sur les deux projets réalisés. Je tiens à témoigner ma reconnaissance à toutes ces personnes citées pour l'autonomie et la liberté qu'elles m'ont accordées. J'ai beaucoup appris et j'en ressors riche de cette expérience professionnelle.

Pour son encadrement, je tiens à remercier très chaleureusement madame Ariane Pinche. Son expertise, sa bienveillance, son écoute ainsi que ses conseils ont été très précieux au cours de cette année scolaire et lors de la rédaction de ces pages.

Pour son investissement et ses corrections particulièrement justes, je tiens à apporter toute ma gratitude à Louise Delaporte, amie de longue date, qui a pris le temps de relire et de corriger ces pages. Son soutien et son amitié m'ont été particulièrement importants sur la fin de la rédaction de ce mémoire.

Enfin, je tiens à témoigner de mon amitié sincère à Léa Périssier, étudiante également au sein du master Technologies numériques appliquées à l'histoire à l'École nationale des chartes, pour son soutien tout au long de cette année scolaire, pour sa grande camaraderie et son humour. Ces mois de cours en ligne auraient été beaucoup plus fades sans elle. Je souhaite de tout coeur que ces derniers mois représentent le début d'une longue amitié.

Table des matières

| | |
|--|-----------|
| Résumé | iii |
| Remerciements | v |
| Table des matières | vii |
| Introduction | 1 |
| I Apports de données externes | 5 |
| 1 L'édition, ses données et leur interopérabilité | 7 |
| 1.1 Données structurées en TEI | 7 |
| 1.1.1 Le TEI et le DTABf | 7 |
| 1.1.2 L'environnement logiciel oXygen Author et l'interface ediarum | 9 |
| 1.1.3 L'encodage général | 10 |
| 1.2 Interopérabilité et web de données | 11 |
| 1.2.1 Entre théorie... | 11 |
| 1.2.2 ...et pratique | 13 |
| 2 Les données externes et leur accessibilité | 17 |
| 2.1 L'utilisation au sein de l' <i>edition humboldt digital</i> de données et services web tierces | 17 |
| 2.1.1 Définitions | 17 |
| 2.1.2 Les services web et données externes : l'édition numérique au sein d'un large réseau | 19 |
| 2.2 Les données manipulées au cours du stage | 22 |
| 2.2.1 Kalliope-Verbundkatalog | 22 |
| 2.2.2 Société Américaine de Philosophie (APS) | 24 |
| 2.2.3 Catalogue général de la Bibliothèque nationale de France | 25 |
| 3 Des données brutes aux données traitées | 27 |
| 3.1 Nettoyer | 27 |

| | |
|---|-----------|
| 3.2 Fusionner | 29 |
| 3.3 Enrichir | 30 |
| II Développer de nouvelles fonctionnalités avec des données tierces | 33 |
| 4 Entre recherche et ingénierie | 35 |
| 4.1 Le corpus archivistique | 35 |
| 4.1.1 La correspondance d'Alexander von Humboldt | 35 |
| 4.1.2 Reconstituer la correspondance d'Alexander von Humboldt : les archives de la <i>Berlin-Brandenburgische Akademie der Wissenschaft</i> | 37 |
| 4.2 L'aide à la recherche numériquement retranscrite | 38 |
| 4.3 <i>Humboldt Chrotonopographie</i> : un outil expérimental | 39 |
| 5 Développement d'outils de recherche et des visualisations | 45 |
| 5.1 L'environnement technique et les librairies choisis | 46 |
| 5.1.1 Le choix de l'environnement technique : Jupyter Notebook | 46 |
| 5.1.2 Les librairies utilisées | 47 |
| 5.2 Algorithmie et réalisations | 49 |
| 5.2.1 Fonctions de recherche | 50 |
| 5.2.2 Visualisations | 53 |
| 6 Accessibilités | 57 |
| 6.1 Accès aux données | 57 |
| 6.2 Accès au projet | 60 |
| 6.2.1 Livrable : un dépôt github | 60 |
| 6.2.2 Communiquer autour du projet | 60 |
| III Enrichir des projets numériques de la BBAW grâce aux données externes | 62 |
| 7 correspSearch : collaborer autour des échanges épistolaires | 64 |
| 7.1 correspSearch : le projet et ses objectifs | 64 |
| 7.2 Le format CMI (CMIF) | 66 |
| 8 De la conception au développement | 68 |
| 8.1 Les objectifs du <i>cS matching tool</i> | 68 |
| 8.2 Développements back end et front end | 70 |
| 8.2.1 Back end : algorithmie | 70 |
| 8.2.2 Front end : interface utilisateur | 74 |

| | |
|--|------------|
| 9 Apports et livrables | 78 |
| 9.1 Résultats et perspectives de l'outil développé | 78 |
| 9.1.1 Enrichissement des données de correspSearch : les résultats | 78 |
| 9.1.2 Améliorations et perspectives | 79 |
| 9.2 Cible et accessibilité du projet | 80 |
| Conclusion | 82 |
| Bibliographies thématiques | 86 |
| Acronymes | 93 |
| Table des figures | 95 |
| Liste des tableaux | 97 |
| Annexes | 100 |
| A Le modèle de données en réseau de l'<i>edition humboldt digital</i> | 100 |
| B Différence entre le format MODS et DC. : <i>Exemple d'une lettre de Abich envoyée à Humboldt en 1852 ou 1853.</i> | 101 |
| C Fonction en Python retournant les informations géographiques d'un lieu donné | 103 |
| D Aide à la recherche dactylographiée : <i>Exemples de pages tirées de l'aide à la recherche des archives Humboldt de la BBAW</i> | 105 |
| E L'outil <i>Humboldt Chronotopographie</i> : <i>Captures d'écran de l'outil expérimental</i> | 107 |
| F Visualisations cartographiques | 110 |
| G Comparaison du design des sites web des projets de la BBAW | 112 |

Introduction

En histoire, comme ailleurs, ce qui compte, ce n'est pas la machine, mais le problème. La machine n'a d'intérêt que dans la mesure où elle permet d'aborder des questions neuves, originales par les méthodes, les contenus et surtout l'ampleur¹.

L'article "La fin des érudits" d'Emmanuel Le Roy Ladurie, publié dans le *Nouvel Observateur* le 8 mai 1968, fait encore sens aujourd'hui et particulièrement pour les ingénieurs de recherche en humanités numériques. Il met en exergue l'analyse de vastes corpus de documents dont les données sont capitales mais dont les dimensions ont défié les efforts des chercheurs. Il s'agit d'une référence incontournable dans les débats et dans la réflexion des historiens autour du rôle de l'ordinateur et des nouvelles technologies en général. Le Roy Ladurie exprime l'importance qu'il a accordé aux nouvelles méthodes proposées par les sciences dans son propre travail d'historien. Le recours aux technologies en sciences humaines et sociales ont profondément modifié les procédés d'analyse des sources primaires de la recherche. Les documents qui fournissaient auparavant l'unique matière des études historiques subissent dorénavant des étapes de numérisation ou d'encodage en vue d'un traitement informatique. Le matériel qui est le document archivistique se transforme en données manipulables par machine et exploitables dans des projets numériques.

Le partage de ces données historiques donne l'occasion d'accroître la visibilité et les dimensions d'un corpus de recherche. Cet enjeu de la valorisation des données brutes est le cœur du mouvement de l'*open data*. Ce mouvement prône l'ouverture des données c'est-à-dire leur libre accès et leur gratuité en vue d'un traitement et/ou d'une exploitation par toute personne ou projet de recherche intéressés. L'*open data* vise à la diffusion des sources numériques auprès d'un large public ainsi qu'à l'enrichissement des ressources de la communauté scientifique dans le but de favoriser l'innovation. Le recours à des standards et normes permet à des projets scientifiques de collaborer autour de corpus communs par la production de données interopérables. En sciences humaines et sociales, la recherche a recourt à une expertise pluridisciplinaire. Ces données standardisées constituent un socle commun pour cette expertise et l'étude de corpus transverses.

C'est dans cette lignée que s'inscrit le projet académique *Alexander von Humboldt*

1. Emmanuel Le Roy Ladurie, « La fin des érudits », *Le Nouvel Observateur*, 8 (1968), p. 38-39

*auf der Reisen - Wissenschaft aus der Bewegung*² initié par la *Berlin-Brandenburgische Akademie der Wissenschaft* (en français : Académie des sciences de Berlin-Brandebourg) (BBAW). Ce projet d'édition hybride des carnets de voyage et de la correspondance d'Alexander von Humboldt se situe dans la continuité de la reconstitution et l'édition de documents hérités du scientifique par la BBAW. Alexander von Humboldt (1769-1859) est un scientifique et plus particulièrement naturaliste, géographe et explorateur allemand. Il est difficile de lui accorder un seul métier ou activité. En effet, sur sa page Wikipédia française, plus de vingt-cinq activités différentes lui sont accordées témoignant de l'intérêt de Humboldt pour toutes les sciences naturelles et de sa curiosité sans faille.³. Toutefois, ses relevés topographiques et prélèvements de faune et flore lors de ses expéditions ont contribué à sa renommée. Humboldt a laissé derrière lui un important héritage scientifique, culturel et archivistique. Rédacteur de nombreux livres et notamment du *Cosmos*, son œuvre majeure et résultat de cinq années de travail présentant une description physique de l'univers, il est également un épistolarier particulièrement productif. Lors de ces nombreuses expéditions sur le globe, en plus d'être resté en contact avec son réseau scientifique grâce à une correspondance constante, il a en outre tenu des carnets de voyage contenant les itinéraires de ses périples, des mesures scientifiques et des observations ethnologiques et sociologiques.

Entre instrument de travail et carnet de voyage, les manuscrits de ses expéditions américaine et sibéro-russe se trouvent au centre du projet *Alexander von Humboldt auf der Reisen - Wissenschaft aus der Bewegung*. Leur édition comprend onze volumes qui sont publiés à la fois sous forme d'éditions imprimée et numérique⁴. Le projet de recherche et d'édition remplit ses tâches en étroite collaboration avec l'université de Potsdam, la bibliothèque d'État de Berlin, l'Université technique de Berlin et d'autres institutions de recherche de la région Berlin-Brandebourg. En collaboration avec l'université de Potsdam, le journal web *HIN - Humboldt im Netz* est publié deux fois par an et apporte des éclairages sur l'avancée des études humboldtiennes en allemand, anglais, espagnol et français. *Alexander von Humboldt auf Reisen* a débuté en 2015 et a été estimé à 18 ans d'activité. Faisant partie du programme servant à préserver le patrimoine culturel allemand, il est financé par le gouvernement fédéral et les Länder.

J'ai participé à ce projet en qualité de stagiaire d'avril à juillet 2021 au sein de la BBAW grâce aux financements du programme Erasmus+ et de l'Organisation franco-

2. En anglais, le projet se nomme *Travelling Humboldt – Science on the Move*. Dans ce mémoire, il s'agit plutôt du terme allemand qui jalonne ses pages.

3. Parmi ses activités celles-ci sont mentionnées : géologue, explorateur, botaniste, géographe, *Geheimer Rat*, chambellan, océanographe, démographe, volcanologue, écrivain voyageur, écrivain scientifique, météorologue, polymathe, mécène, zoologiste, naturaliste, essayiste, minéralogiste, astronome, climatologue, ethnologue, scientifique, collecteur de plantes, ornithologue, globe-trotteur, économiste, homme politique « Alexander von Humboldt », *Wikipedia* (, 27 juil. 2021), URL : https://fr.wikipedia.org/wiki/Alexander_von_Humboldt (visité le 28/07/2021)

4. L'édition numérique est disponible sur le site *edition humboldt digital (ehd)*.

allemande pour la jeunesse (OFAJ). À mon arrivée, les missions du stage n'étaient pas entièrement définies. La conception des outils à réaliser a évolué en fonction des besoins et des attentes de l'équipe de l'édition ainsi que de celle du pôle des humanités numériques de la BBAW. Ainsi, j'ai été amenée à traiter des données externes au projet académique dans le but notamment de développer de nouveaux outils. J'ai été accueillie en présentiel au sein de l'équipe d'*Alexander von Humboldt auf Reisen* comprenant six personnes. Dr. Ulrich Päßler, chercheur, responsable adjoint du projet académique et tuteur de mon stage, a supervisé mes missions conjointement avec Dr. Tobias Kraft, chercheur et responsable du projet. Techniquelement, j'ai été soutenue par une tierce personne, Dr. Gordon Fischer, ingénieur d'étude au sein *The Electronic Life Of The Academy* (TELOTA), pôle des humanités numériques de la BBAW.

Ce mémoire, réalisé en vue de la validation du Master Technologies numériques appliquées à l'histoire (TNAH) de l'École nationale des chartes, a pour but de présenter les missions menées en examinant les choix effectués d'une part et les méthodes mises en place pour répondre aux attentes de l'équipe de recherche d'*Alexander von Humboldt auf Reisen* d'autre part. Il ne s'agit donc pas d'un mémoire de recherche sur l'histographie d'Alexander von Humboldt ou des sciences au XIX^e siècle, ni d'un rapport de stage sur mes activités au sein du projet académique qui m'a accueilli, mais davantage d'une présentation critique et analytique des missions réalisées.

Afin de mieux cerner la nature des données de l'*edition humboldt digital* (ehd), la première partie de ce mémoire examinera la production de ces données dans des formats standards et au sein d'un environnement logiciel adapté. Cette partie sera également l'occasion de revenir sur l'interopérabilité des données produites pour l'édition et leur insertion au sein du web sémantique de par l'emploi d'identifiants pérennes. La mise en réseau et l'utilisation des services web externes par l'ehd seront détaillées de façon précise. Aussi, les données externes utilisées au cours des missions du stage seront présentées ainsi que leur accessibilité. L'explication des traitements effectués sur ces données permettra de mieux comprendre la transformation des données brutes en données traitées. Ces données traitées sont le matériel manipulé pour la réalisation de deux missions effectuées au cours du stage. Chacune des missions sera l'objet d'une partie de ce mémoire.

Le développement de nouveaux outils implique une forte compréhension du corpus à valoriser. Dans la seconde partie du mémoire, il s'agira de revenir sur le corpus archivistique et l'état actuel de la reconstitution de la correspondance d'Alexander von Humboldt avant de présenter la première mission du stage. Cette dernière consiste à la mise en place des fonctions de recherche et de visualisations afin d'explorer et de découvrir la correspondance du scientifique grâce à de nouvelles méthodes. Tout comme Emmanuel Le Roy Ladurie l'a particulièrement bien exprimé⁵, la machine ici permet "d'aborder des questions neuves, originales par les méthodes, les contenus et surtout l'ampleur". Il s'agira

5. Voir la citation en chapeau de cette introduction, Id., « La fin des érudits »...

ainsi de présenter le travail réalisé autour du corpus important qu'est la correspondance d'Alexander von Humboldt au sein d'un environnement technique adapté.

Dans une troisième et dernière partie, la seconde mission du stage sera analysée. Cette mission consiste à enrichir les données d'un projet initié par la BBAW, correspSearch, grâce à des ressources externes. CorrespSearch est un portail en ligne qui répertorie les correspondances éditées de manière numérique ou imprimée. Dans un premier temps, une présentation du projet correspSearch et de ses objectifs sera proposée. L'analyse des développements *back end* et *front end* de l'outil développé par mes soins, *cS matching tool*, permettant l'enrichissement des données de correspSearch, s'effectuera dans un second temps. Ces chapitres permettront de mieux situer le *cS matching tool* dans son environnement technique et de justifier les moyens et les contraintes qui ont déterminé son développement.

Ce travail est ponctué de considérations techniques permettant de mettre en lumière mon raisonnement et mes méthodes algorithmiques. Il ne s'agit cependant que d'éclairages représentatifs ; les deux projets menés au cours de mon stage sont disponibles ouvertement sur la plateforme GitHub accompagnés de contextualisation et de commentaires supplémentaires apportant des compléments profitables à la parfaite compréhension de l'analyse technique. Plusieurs documents utiles à la contextualisation et la mise en oeuvre des outils produits sont également joints en annexe de ce mémoire.

Première partie

Apports de données externes

Chapitre 1

L'édition, ses données et leur interopérabilité

1.1 Données structurées en TEI

Sont considérées comme données structurées toutes les données organisées et classées selon un modèle préétabli permettant ainsi de faciliter leur traitement. Ces schémas de données fournissent une structure claire et permettent aux machines de comprendre leur contenu.

1.1.1 Le TEI et le DTABf

Créé en 1987, le *Text Encoding Initiative* (TEI) est le produit d'un consortium et a pour objectif de "faciliter la création, l'échange et l'intégration des données textuelles informatisées"¹. Il propose un ensemble de *guidelines* qui spécifient les méthodes d'encodage des textes qui deviennent de cette manière lisibles par machine. Il est principalement utilisé dans le domaine des sciences humaines et sociales car il répond bien aux besoins de l'édition de texte et en particulier des manuscrits. En effet, les protocoles proposés par le TEI sont surtout adoptés dans le cas de projets d'édition de textes anciens, de manuscrits ou de dossiers génétiques lorsqu'il s'agit de reconstituer le processus de création d'une version définitive d'un texte mais aussi la manière dont il a été structuré. Ces informations sont propres à ce type de documents. En plus de structurer et de modéliser, le TEI permet l'intégration d'informations sur la mise en forme et sur les caractéristiques d'origine d'un document. Les documents produits sont par conséquent structurés, interopérables et peuvent être échangés et réutilisés.

L'édition humboldt est une édition hybride des carnets de voyages d'Alexander von

1. Marcello Vitali-Rosati et Michael E. Sinatra, *Pratiques de l'édition numérique*, Les Presses de l'Université de Montréal, Montréal, 2014 (Parcours Numériques), URL : <http://www.parcoursnumeriques-pum.ca/>, chap. 10

Humboldt, encodée en *eXtensible Markup Langage* (XML)-TEI. Elle propose une version numérique qui est actualisée tous les ans sur edition-humboldt.de ainsi qu'une édition imprimée. Cette dernière se concentre sur la reconstitution des itinéraires de voyage tandis que l'édition numérique transcrit et commente de la manière la plus complète possible les manuscrits. Ces transcriptions et commentaires sont accompagnés d'une utilisation des services web judicieuse, insérant l'édition dans le web de données². L'édition numérique a pour ambition de dépasser les possibilités de l'édition imprimée en terme de présentation et de recherche. Elle permet de présenter l'état actuel du travail de l'équipe effectué sur les manuscrits et les données. L'utilisation du TEI est particulièrement appropriée dans le cas d'une édition hybride. Il faut noter que ces deux types d'éditions se complètent et il serait malheureux de les positionner en concurrence.

Le modèle de données de l'ehd est issu du *Deutsches Textarchiv – Basisformat* (DTABf)³. Il suit par conséquent une norme recommandée par la *Deutsche Forschungsgemeinschaft* (DFG) pour le balisage et l'archivage de textes en XML-TEI⁴. Ce format de document suit les *guidelines* P5 du TEI. Ces *guidelines* publiées en novembre 2007 sont la version la plus récente des *guidelines* du TEI. Elles sont actualisées tous les six mois permettant de fixer des problèmes et d'apporter des améliorations mineures à des fonctionnalités⁵. Toutefois, étant donné que les *guidelines* du TEI visent à offrir des solutions en matière d'édition pour toutes les typologies de textes et manuscrits elles nécessitent une spécification pour les cas concrets. L'objectif du DTABf est de proposer une norme dans le traitement des textes historiques permettant ainsi une analyse complète du corpus du *Deutsches Textarchiv* et une interopérabilité entre eux⁶. Suivre ces recommandations permet à l'édition de s'insérer dans le corpus des textes du *Deutsches Textarchiv* et limite les possibilités de variations dans l'annotation des textes tout en garantissant la cohérence du corpus entier. Par conséquent, l'édition numérique s'insère dans un corpus général des éditions numériques encodées en XML-TEI et plus précisément dans le corpus du *Deutsches Textarchiv* avec l'utilisation du DTABf, plus adapté à la spécificité des documents encodés. Afin d'encoder les carnets de voyage d'Alexander von Humboldt, l'équipe du *Alexander von Humboldt auf Reisen* et toutes les équipes des éditions numériques de la BBAW⁷ utilisent le logiciel oXygen Author accompagné de l'interface ediarum.

2. cf. section 1.2

3. Pour plus d'informations sur ce format, veuillez vous diriger vers le site internet : deutschtextarchiv.de

4. Deutsche Forschungsgemeinschaft (2015) : Förderkriterien für wissenschaftliche Edition in der Literaturwissenschaft. Bonn, S.6

5. Voir le site des TEI *guidelines* : tei-c.org/guidelines

6. Voir la rubrique "Ziele und Fokus des DTA-Basisformats" du site du *Deutsches Textarchiv* : deutschtextarchiv.de

7. Les projets d'éditions numériques sont nombreux à la BBAW. En plus d'*Alexander von Humboldt auf Reise - Wissenschaft aus Bewegung*, on peut également mentionner l'édition de la correspondance de Jean Paul, l'édition des chansons médiévales du *Langen Ton* de Regenbogen, les archives dramaturgiques et administratives d'August Wilhelm Iffland ou encore l'édition numérique de la correspondance linguistique de Wilhelm von Humboldt parmi d'autres.

1.1.2 L'environnement logiciel oXygen Author et l'interface ediarum

Contrairement à oXygen XML Editor qui contient toute une partie développement et débugage de schémas, oXygen XML Author a été spécialement créée pour les auteurs de contenu et se base sur une édition visuelle. Ainsi, si les besoins s'arrêtent à l'édition visuelle de documents XML et leur publication dans divers formats de sortie, alors la version oXygen Author est adaptée. Cependant, si le but du projet est de développer des feuilles de styles *eXtensible Stylesheet Language* (XSL) ou des schémas, il vaudra mieux dans ce cas privilégier l'utilisation de la version Editor du logiciel qui offre de plus amples fonctionnalités. Le logiciel propose dans tous les cas trois modes différents de travail : le mode texte qui permet de travailler directement avec l'encodage brut, le mode grille exposant l'arborescence du fichier sous forme de tableau et le mode auteur qui apporte un aperçu du fichier avec une *Cascading Style Sheets* (CSS) minimale.

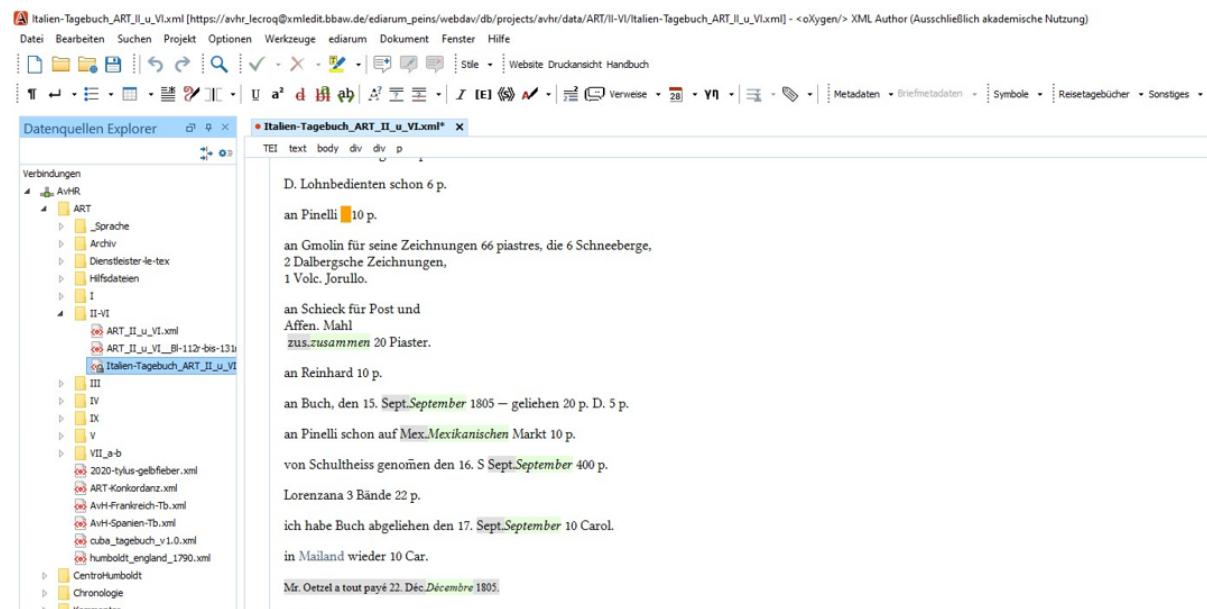


FIGURE 1.1 – Extrait du carnet de voyage d'Italie (1805) dans l'environnement ediarum.

Afin de s'adapter aux équipes d'éditions numériques de la BBAW, TELOTA a développé un ensemble d'extensions pour ce logiciel, il s'agit d'ediarum. Cette extension vise à fournir un environnement optimisé pour travailler avec des transcriptions et leur balisage. Il permet à l'utilisateur de travailler avec des données XML spécifiquement TEI en utilisant une interface utilisateur graphique qui propose un affichage *what you see is what you get* (WYSIWYG). Ces éditeurs ou traitements de texte WYSIWYG offrent une interface utilisateur qui permet de composer visuellement le résultat voulu. Ce sont des interfaces plus intuitives et par conséquent plus accessibles puisque l'utilisateur peut directement voir à l'écran ce à quoi ressemblera le résultat final. Grâce à cet environnement logiciel

l'encodage devient plus simple et permet aux personnes des équipes d'éditions numériques n'étant pas ou peu familières avec le code brut XML de pouvoir malgré tout travailler sur l'encodage des textes.

Les diverses fonctionnalités apparaissent sur une barre graphique semblable à celle que les utilisateurs ont l'habitude d'utiliser sur d'autres logiciels. Il est ainsi possible d'encoder de manière intuitive les annotations, notes, ajouts, rayures, abréviations et autres spécificités du document de manière simple. Une icône apporte un accès aux index de lieux, de personnes et des œuvres mentionnés permettant de rajouter les balises adaptées ainsi que les identifiants correspondants. L'interface graphique ediarum offre la possibilité d'ajouter des balises et des métadonnées sans avoir à modifier manuellement les fichiers XML.

Bien que certaines parties des modules ne sont actuellement disponibles qu'en langue allemande et que la documentation doit être améliorée, ediarum offre déjà aux chercheurs un moyen flexible et ouvert d'éditer leurs données TEI. Il s'agit également d'une ressource sur laquelle d'autres développeurs peuvent s'appuyer puisqu'ediarum est accessible en open source. En effet, après avoir été utilisé au sein des projets internes de la BBAW pendant plusieurs années⁸, les modules ont été successivement mis à disposition du public sur un dépôt GitHub qui comprend toutes les informations nécessaires concernant ediarum. Les modules d'ediarum peuvent ainsi être utilisés comme point de départ ou bien comme boîte à outils pour d'autres développeurs qui souhaiteraient créer des cadres personnalisés et spécifiques à des projets d'édition numérique. C'est au sein de cet environnement logiciel que les carnets de voyage d'Alexander von Humboldt ont été encodés par l'équipe de *Alexander von Humboldt auf Reisen - Wissenschaft aus der Bewegung*.

1.1.3 L'encodage général

La version adaptée d'oxYgen XML Author accompagnée d'ediarum est utilisée afin de saisir et d'éditer les données. Les données sont stockées dans la base de données XML gratuite eXistdb. Avec le serveur web Jetty, il sert également de base à l'édition numérique qui a été réalisée avec les langages XQuery, XSLT et XPath.

L'*edition humboldt digital* suit un modèle d'édition numérique qui met en œuvre une approche visuelle, textuelle et liée au contenu des sources manuscrites. Visuellement, les modifications textuelles apportées par l'auteur sont reproduites telles que le surlignage, les ratures, les annotations et corrections.



FIGURE 1.2 – Les différents index directement accessibles dans ediarum

⁸. *Ediarum. A toolbox for editors and developers – RIDE*, URL : <https://ride.i-d-e.de/issues/issue-11/ediarum/> (visité le 07/07/2021)

**Voyage d’Espagne aux Cā
Juin à Oct. 1799 [= Tageb**

H: Staatsbibliothek zu Berlin – Preußischer Kulturbesitz, Han
Edited by Carmen Götz und Ulrike Leitner

Critical text Reading text Text with facsimile

Open all text notes Close all text notes

[u1] |
I.
Voyage d’ Espagner [sic] aux Cana
et à Cumaná.

FIGURE 1.3 – Les diverses approches textuelles accessibles sur le site de l'édition.

Tous les textes édités sont également accompagnés de leur facsimilé numérisé dans la mesure du possible. L'état de conservation, de la numérisation et du cadre juridique sont des principaux freins à l'affichage d'un facsimilé des documents au sein de l'interface web. Le logiciel digilib développé au Max-Planck Institut pour l'histoire des sciences est utilisé pour afficher les facsimilés. L'approche textuelle, quant à elle, se compose en trois parties : un texte critique, un texte avec facsimilé qui contient tous les commentaires critiques des éditeurs et un texte de lecture qui vise à une présentation grand public et qui simplifie la typographie du document. L'approche liée au contenu se concentre sur

la préparation d'un maximum d'informations contenues dans les textes sous forme de données structurées telles que les noms de personnes et de lieux, les titres d'oeuvre, les indications de mesures effectuées par Humboldt. Ces différentes approches textuelles des manuscrits sont directement accessibles sur le site de l'édition numérique avec un système d'onglet.

Des registres ont été également créés permettant d'indexer divers éléments au sein de l'édition : les personnes, les plantes et les lieux. La création de ces registres est concevable grâce notamment à un balisage constant de ces éléments par les éditeurs au sein de l'encodage ainsi qu'à l'utilisation d'identifiants pérennes vers des notices d'autorités. L'utilisation de ces identifiants dans l'encodage insère l'édition au cœur du web de données.

1.2 Interopérabilité et web de données

1.2.1 Entre théorie...

L'interopérabilité d'un document est un terme informatique désignant des documents capables de s'adapter et d'être utilisés facilement par d'autres systèmes. Synonyme de compatibilité, cette capacité permet notamment de faciliter la création d'un réseau et le transfert de données provenant de documents ou programmes différents.

L'interopérabilité de la TEI est en grande partie due au fait qu'il s'agisse d'un langage issu du XML. Ce dernier est un métalangage et est spécialement conçu pour les données. Un des objectifs de la création de XML était de favoriser les échanges et le partage des données entre machine et utilisateur. Ainsi, un fichier XML est autodescriptif, extensible, interopérable et pérenne et est convertible en plusieurs autres formats grâce à des feuilles de style définies en XSL. Il existe trois types d'interopérabilité : l'interopérabilité tech-

nique, l'interopérabilité sémantique et l'interopérabilité syntaxique. La première concerne la capacité des technologies à communiquer entre elles et à se comprendre et permet par exemple d'échanger des données issues de normes bien définies ; tandis que la seconde assure le fait que la signification exacte des informations échangées soit compréhensible par n'importe quelle autre application. Pour assurer une interopérabilité sémantique, les deux côtés de l'échange doivent se référer à un modèle de référence d'informations communs. L'interopérabilité syntaxique concerne, quant à elle, la façon dont sont codées et formatées les données en définissant notamment la nature, le type et le format des informations échangées.

Le XML s'insère dans ce qu'on appelle le web sémantique puisqu'il offre la création de documents composés de données structurées. Le web sémantique permet aux machines de comprendre la sémantique c'est-à-dire la signification de l'information sur le web. Dans le web sémantique, les informations sont publiées accompagnées de métadonnées, fournissant ainsi un contexte dit sémantique. Le contenu de l'édition est alors enrichi avec des données interactives. Son sens est renforcé par des balises sémantiques et des liens directs vers des ressources externes ou vers les références citées. On parle alors du concept de *semantic publishing* (éditorialisation sémantique), c'est-à-dire de l'enrichissement sémantique des publications et éditions scientifiques.

Le web sémantique met en oeuvre le web de données, appelé *linked data* en anglais, qui consiste à lier et structurer l'information pour accéder simplement à la connaissance qu'elle contient déjà. Tim Berners-Lee le définit : "Les données du *linked data* peuvent être traitées directement ou indirectement par des machines pour aider les utilisateurs à créer de nouvelles connaissances"⁹. Ses objectifs sont multiples :

- mettre à disposition des données en utilisant des techniques standardisées qui garantissent l'interopérabilité
- relier les données entre elles et les rendre interprétables par les machines
- permettre aux données d'être partagées et réutilisées

Le web de données permet de relier non pas les documents (pages HTML par exemple) mais les données entre-elles et de les rendre exploitables par des machines. Il s'appuie notamment sur des standards d'informations qui sont ceux du web sémantique. Par conséquent, le web de données vise à favoriser la publication de données structurées sur le web dans le but de constituer un réseau global facilitant la navigation des utilisateurs au sein d'un espace d'information.

Il faut toutefois noter que le XML ne constitue pas un modèle qui est entièrement adapté au contexte du web sémantique comme le souligne Grégory Fabre¹⁰. L'image

9. Tim Berners-Lee, James Hendler et Ora Lassila, « The Semantic Web : a new form of Web content that is meaningful to computers wil unleash a revolution of new possibillities », *Scientific American Magazine* (, 17 mai 2001), URL : <http://web.archive.org/web/20081114135540/http://www.sciam.com/article.cfm?id=the-semantic-web&print=true> (visité le 25/07/2021)

10. Fabre Grégory, Marcotte Sophie, "L'organisation des métadonnées", in : M. Vitali-Rosati et M.

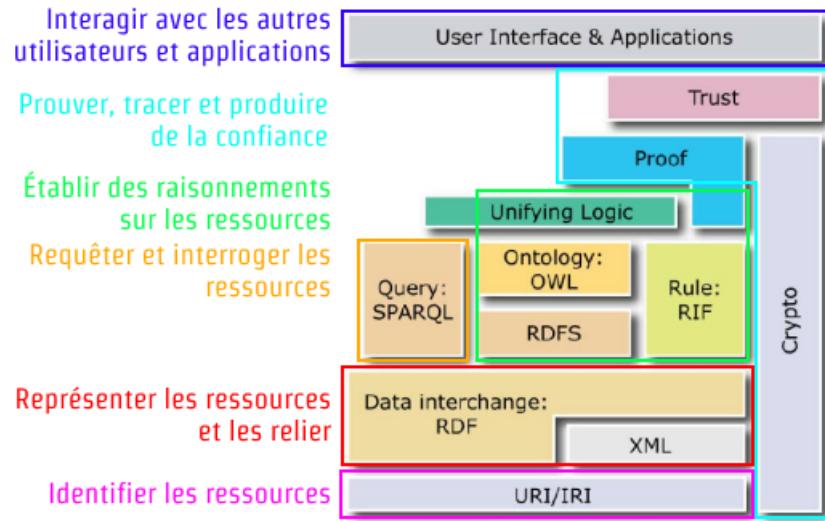


FIGURE 1.4 – L’architecture du Web sémantique ou *semantic stack*. Source : Wikipedia, Semantic Stack

ci-dessous permet de visualiser que le XML doit impérativement être accompagné du *Resource Description Framework* (RDF) afin d’offrir une mise en relation des sources. Ce dernier est un modèle de graphe qui décrit les ressources utilisées ainsi que les métadonnées. Toutefois, l’utilisation du XML est une des portes d’entrée possible pour une édition historique de faire partie du web sémantique et donc du web de données.

1.2.2 ...et pratique

Une fois les concepts définis, il est intéressant de voir comme cela s’applique au sein d’une édition numérique scientifique. En tant que scientifique et botaniste, la diversité des plantes et des spécimens biologiques fait partie intégrante des notes écrites par Alexander von Humboldt dans ses carnets de voyage. En ce qui les concerne, aucune entrée distincte n’est enregistrée au sein d’un index stocké dans la base ediarum.BASE comme c’est le cas pour les personnes, lieux et institutions. Les noms scientifiques des plantes sont normalisés dans l’édition numérique. Pour chaque plante balisée dans l’encodage, elle est automatiquement ajoutée à la liste des références la concernant et est également automatiquement liée à diverses bases de données taxonomiques externes en utilisant divers services web et *Application Programming Interface* (API). Le service web *Global Name Resolver* permet de relier de manière automatique un nom scientifique à ces bases externes. Parmi ces dernières peuvent être mentionnées le *catalogue of life*, la *biodiversity*

heritage library ainsi que le *international plant names index*. L'index des plantes proposés sur le site de l'édition numérique est donc créée de manière entièrement dynamique.

Pour ce qui est des personnes, lieux et institutions, les registres proposés sur le site de *edition humboldt digital* se forment d'une manière différente et s'alignent sur des référentiels externes. Au sein de l'encodage, une personne est balisée dans un `<persName>` dans laquelle l'attribut `@key` stocke l'identifiant interne de l'édition commençant forcément par un H, comme par exemple : `<persName key="H0012737">James Arnold</persName>`. Cet identifiant se réfère directement à une entrée au sein de l'index de personne. Les index des personnes, des lieux et des institutions sont traités et stockés en XML-TEI dans l'environnement ediarum.BASE. Chaque entrée possède un identifiant unique et permanent. Elle est également pourvu d'un ou plusieurs identifiants provenant de divers services externes afin de permettre une interopérabilité de l'édition. Une courte description concernant l'objet ou la personne encodée est rédigée et accompagne les données de base.

Prenons donc plus précisément l'exemple d'une entrée au sein de l'index des personnes. Chaque nouvelle entrée dans l'index est encodée dans une balise `<person>`. Un identifiant est automatiquement généré dans l'attribut `@xml-id` et qui sera par la suite normalisé. S'il s'agit d'un personnage fictif ou mythologique, alors l'attribut `@role` est ajouté et accepte les valeurs "fictional" et "mythological". Le nom et le prénom de la personne sont enregistrés dans la balise `<persName type="reg">`, la valeur de `@type` signifiant *regular*. Les variantes ou les noms alternatifs sont chacun stockés dans une balise `<persName>` distincte où les attributs `@type="alt"` pour alternatif et `@subtype` sont acceptés. Les dates de naissance et décès sont indiquées dans les balises adaptées à savoir `<birth>` et `<death>`. Une note, plus ou moins longue en fonction des besoins d'explications de l'édition, peut également accompagner cette entrée. Ainsi, des informations détaillées concernant leur vie et activité professionnelle y sont mentionnées. L'édition utilise des référentiels de données normés divers. Une personne est liée à son identifiant *Gemeinsame Normdatei* (GND) ou en anglais *integrated authority file* du site de la *Deutsche Nationalbibliothek* ou bien à un identifiant *Virtual International Authority File* (VIAF) quand cette personne ne possède pas d'identifiant GND. L'identifiant GND est pour l'Allemagne ce que l'identifiant de la Bibliothèque nationale de France (BnF) est pour la France. Particulièrement employés par les bibliothèques, les identifiants GND sont de plus en plus utilisés par les centres d'archives, les musées et autres institutions culturelles et scientifiques. Son emploi permet de relier les données de diverses origines entre elles. Elles forment un réseau de données interopérables par les machines. Une entrée complète liée à un identifiant GND au sein du registre des personnes se présente comme suit :

```

<person xml:id="H0004159" role="mythological">
    <idno type="uri">http://d-nb.info/gnd/118557513</idno>
    <persName type="reg">
        <name>Jesus Christus</name>
    </persName>
    <persName type="alt">
        <name>Jesus von Nazareth</name>
    </persName>
    <note>Religionsstifter, Sohn Gottes (Gottvaters), sein Wirken wird im Neuen
        Testament der Bibel geschildert</note>
</person>

```

L'identifiant GND est balisé au sein `<idno>` et est stocké sous forme d'*Uniform Resource Identifier* (URI). Encoder un lieu au sein de l'édition fonctionne sur le même principe que l'encodage d'une personne : avec un registre, un identifiant interne et une ressource externe liée au sein d'une balise `<idno>` stockée également sous forme d'URI. Le référentiel auquel les lieux de l'édition s'alignent est GeoName qui proposent des identifiants pérennes.

Au sein du web sémantique, l'URI se situe à la base de son architecture¹¹ car tous les hyperliens du Web sont exprimés sous cette forme. Quatre piliers du web de données définis par Tim Berners-Lee sont fondés sur les bonnes pratiques de l'utilisation de ces URI¹² :

- utiliser des adresses URI uniques pour identifier les choses
- utiliser des adresses URI *Hypertext Transfer Protocol* (HTTP)
- fournir à travers l'adresse URI des renseignements exploitables, lisibles par les humains et machines en s'appuyant sur des formats ouverts
- mailler l'adresse URI initiale en lui associant des adresses URI externes et ce, pour améliorer la navigation de l'utilisateur sur le web

Ainsi, l'URI identifie une ressource physique ou abstraite dont la syntaxe respecte une norme d'Internet et doit pouvoir l'identifier de manière permanente. Il faut différencier *Uniform Resource Locator* (URL) et l'URI. En effet, l'URL identifie ce qui existe sur le web et permet d'identifier une ressource web comme une page. Une URI est, quant à elle, attribuée à toute ressource et permet de l'insérer dans le web de données. Pour résumer, l'URI est l'identifiant unique qui faire référence à une ressource tandis que l'URL est le chemin d'accès pour obtenir une représentation de cette ressource. En effet, dans le *linked data*, l'URI doit être interrogable via le protocole HTTP. Par conséquent, l'URI doit pouvoir s'exprimer par une URL.

Il faut toutefois préciser que l'ajout de références par l'hyperlien doit être faite de manière judicieuse comme le mentionne Romain Wenz¹³. Si les références sont évidentes

11. Voir Figure 1.4.

12. Voir l'article Wikipédia sur le Web de données.

13. Romain Wenz, « Hypertextualisation », *Revue de la BNF*, n° 42–3 (2012), p. 36-41, URL : <https://doi.org/10.3917/rbnf.042.0036>

alors elles sont sans intérêt. Elles deviennent inutiles si elles sont trop complexes ou nombreuses. La référence d'une ressource doit impliquer une forme d'éclaircissement. Elle doit être explicative mais doit également pouvoir aller au-delà de la simple citation de sources. Les références doivent offrir de nouvelles possibilités de découverte sous une forme d'association d'idées. Le réseau formé par cet ensemble de liens ne relève pas du hasard. En effet, les liens correspondent à chaque fois à des citations volontaires de la part des éditeurs de l'édition numérique. Ils sont également explicites et significatifs puisqu'ils ont pour but d'apporter plus de sens à l'utilisateur.

L'alignement des données à un référentiel externe est possible grâce à l'utilisation d'identifiants de notice d'autorité au sein de l'édition, comme l'identifiant GND ou VIAF ou encore de service web tierces comme *Global Name Resolver*. Ces identifiants permettent de proposer des données interopérables et de s'insérer au sein du web de données. Cette démarche d'alignement à des référentiels vise à favoriser la recherche d'information, le partage et la citation. Les liens qui sont créés entre les différents référentiels permettent d'exposer des données plus riches, plus fiables, pérennes et interopérables. Les référentiels mentionnées dans ce chapitre sont loin d'être les seules ressources externes utilisées au sein de l'*edition humboldt digital*(ehd). Cette dernière s'insère dans un véritable réseau en utilisant un grand nombre de services web, de données et référentiels externes divers.

Chapitre 2

Les données externes et leur accessibilité

2.1 L'utilisation au sein de l'*edition humboldt digital* de données et services web tierces

2.1.1 Définitions

Sont qualifiées de données externes toutes les données qui ne sont pas générées par l'organisation et qui sont accessibles via des sources extérieures à l'organisation-même. Les données peuvent être achetées, échangées ou mises à disposition de tous gratuitement. Ces dernières sont considérées comme ouvertes à tous dans le cadre des politiques d'*open data*.

L'importance de l'*open data*

L'*open data* est l'ouverture des données, c'est-à-dire la mise à disposition gratuite dans des formats techniques réutilisables. Les données sont diffusées de manière structurée. La loi Valter précise également que les données doivent être diffusées publiquement en ligne dans un standard ouvert qui est aisément réutilisable. Les données, quant à elles, doivent être complètes et de qualité c'est-à-dire qu'elles doivent être brutes et accompagnées de leurs métadonnées¹. Le concept de l'*open data* comprend deux notions complémentaires² :

- le droit de réutiliser les données en question (ouverture juridique garantie par une licence ouverte)
- la possibilité technique qui rend possible les échanges avec d'autres projets ou

1. Céline Castets-Renard et Nathalie Gandon, « Open data des données de la recherche publique : entre réformes législatives et retour d'expérience sur un guide pratique à destination des chercheurs », *LEGICOM*, N° 56-1 (8 mars 2016), p. 67-75, URL : <https://www.cairn.info/revue-legicom-2016-1-page-67.htm> (visité le 08/07/2021)

2. R. Wenz, *L'open data, un levier pour l'évolution des catalogues*, 2016, URL : <https://www.cairn.info/vers-de-nouveaux-catalogues--9782765415138-page-13.htm> (visité le 07/07/2021)

communautés conduisant ces entités à vérifier la qualité de leur production de données et par conséquent à améliorer son propre processus grâce à l'*open data*

Les licences ouvertes et gratuites qui garantissent l'ouverture juridique des données sont diverses. Elles sont en mesure de répondre aux besoins d'une ouverture des données dans le domaine des humanités numérique et dans le cadre de l'*open data*. Les plus courantes sont la Creative Commons 4.0 et l'*Open Database Licence* (ODbL) de l'Open Knowledge Foundation. La première est centrée sur le droit d'auteur tandis que la seconde peut être appliquées aux bases de données.

Il faut cependant dissocier le *linked data* c'est-à-dire le web de données et l'*open data*, les données ouvertes. Ces deux concepts sont néanmoins particulièrement liés puisque les institutions qui publient des informations en ligne le font pour permettre leur diffusion large. Si elles le font sans apporter des liens vers les informations d'origine alors elles deviennent invisibles pour le public. De même, si elles le font sans ouvrir les données en question, elles créent une insécurité juridique pour les utilisateurs et seront soit contournées soit considérées comme inutiles³. La mise à disposition des données d'une institution par l'intermédiaire d'API permet de créer des applications ou des fonctionnalités que le site d'origine n'aurait pas nécessairement implémentées. Il faut néanmoins noter un risque : des applications peuvent être dépendantes de données externes et de service web dont on ne peut garantir la pérennité et parfois leur bon fonctionnement.

L'interface de programmation applicative ou API

La mise à disposition des données favorise la création d'API. Dans le web de données, une API permet à un système source d'exposer ses données à des fins d'exploitation par d'autres systèmes. Il s'agit d'un ensemble normalisé de classes, de méthodes, de fonctions et de constantes qui sert de façade par laquelle un logiciel offre des services à d'autres logiciels. Dans ce domaine, les bibliothèques se sont longtemps contentées de mettre en place un serveur *Search/Retrieve via URL* (SRU) pour rendre interrogables leurs données bibliographiques⁴. Elles sont bien souvent accompagnées d'une description qui spécifie comment les clients peuvent se servir des fonctionnalités du serveur⁵. Afin d'utiliser les API et d'avoir accès aux données de l'institution source, il faut souvent déclarer son projet de développement auprès de la plateforme dans le but d'obtenir une clé d'authentification appelée *consumer key*. Cette déclaration est de nature contractuelle et engage le client à

3. *Ibid.*

4. Yves Tomic, « De l'usage des API », *Documentaliste-Sciences de l'Information*, Vol. 51–3 (25 sept. 2014), p. 17-18, URL : <https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2014-3-page-17.htm> (visité le 08/07/2021)

5. La description de l'API SRU du catalogue de Bibliothèque nationale de France est un bel exemple, étant particulièrement détaillée et fournie. Son interface permet de guider au mieux l'utilisateur non confirmé dans le domaine des API.

suivre les conditions générales de la politique de développement⁶.

2.1.2 Les services web et données externes : l'édition numérique au sein d'un large réseau

Au sein de l'ehd de nombreuses API sont utilisées, insérant l'édition numérique dans un véritable réseau de services web et d'utilisation de données tierces. Une infographie a été réalisée par le projet *Alexander von Humboldt auf der Reisen* afin de faire voir ce réseau et rend compte des services exploités⁷. En plus d'exposer les diverses ressources externes employées au sein de l'édition numérique, plusieurs captures d'écran de l'interface seront introduites dans cette description afin de présenter leur affichage dans l'interface web et la manière dont l'utilisateur peut y accéder.

Dans les parties précédentes ont été évoqués les identifiants externes utilisés au sein de l'édition permettant de lier les données de l'édition avec des données externes et ainsi de faire partie du web de données. L'identifiant GND permet de faire référence à des personnes physiques. Ce même identifiant est utilisé afin d'importer les portraits de personnes issues de Wikimedia Commons. Ces portraits sont directement visibles pour l'utilisateur dans l'interface web. Toutes les informations relatives à la personne sélectionnée et tous les liens externes qui s'y réfèrent sont accessibles sur l'interface

The screenshot shows a digital library entry for Joseph Louis Gay-Lussac. At the top, it says "Joseph Louis Gay-Lussac" and "1778-1850". Below this is a portrait of him. The text next to the portrait reads: "französischer Physiker und Chemiker; Freund A. von Humboldts, mit dem er chemische Versuche durchführte; ab 1809 Professor u.a. an der Sorbonne". It also mentions "Source: Wikimedia Commons". On the right side, there's a "Links to external resources" section with links to "Hidden Kosmos", "CERL Thesaurus", and "BNF data". Below that is a "GND-ID: http://d-nb.info/gnd/118716". Further down, there are sections for "Mentioned in travel journals" (with a link to "Voyage d'Espagne aux Canaries et à Cumaná Obs. astron. de Juin à Oct. 1799 [= Tagebücher der Amerikanischen Reise I] 72r, 96r") and "Mentioned in editor's notes in travel journals" (with a link to the same document). At the bottom, there's a "Mentioned in letters" section with a table:

| Date | Correspondent | Place |
|------------|--|---------|
| 10.10.1811 | From Aimé Bonpland To Karl Ludwig Willdenow | Paris |
| 29.09.1848 | To Christian Gottfried Ehrenberg | Potsdam |

FIGURE 2.1 – Gay-Lussac dans le registre des personnes. Capture d'écran de l'interface de l'édition numérique.

6. Jean-Marc Francony, « L'éditorialisation des données aux bornes des API : enjeux et perspectives pour une analyse empirique », *Les Enjeux de l'information et de la communication*, N° 19/2-2 (2018), p. 69-79, URL : <https://www.cairn.info/revue-les-enjeux-de-l-information-et-de-la-communication-2018-2-page-69.htm> (visité le 09/07/2021)

7. N'hésitez pas à vous référer à l'annexe en question, Figure A.1.

de l'édition numérique. La création de cette page résulte de la combinaison d'informations issues de sources diverses : le portrait provenant de Wikimedia Commons, la présentation de la personne tirée du registre de l'édition, les liens situés à droite vers des ressources externes. Tous ces liens apportent davantage d'informations sur le personnage présenté. En dessous de la présentation de la personne sont listées ses mentions au sein de l'édition. Dans ce cas précis, il s'agit des mentions de Gay-Lussac au sein des carnets de voyage écrits par la main de Humboldt, ainsi que ses mentions dans les commentaires des éditeurs. Cette liste est donc générée grâce au fait que le nom de cette personne a été systématiquement balisé par les éditeurs⁸ dans l'encodage.

Cumaná

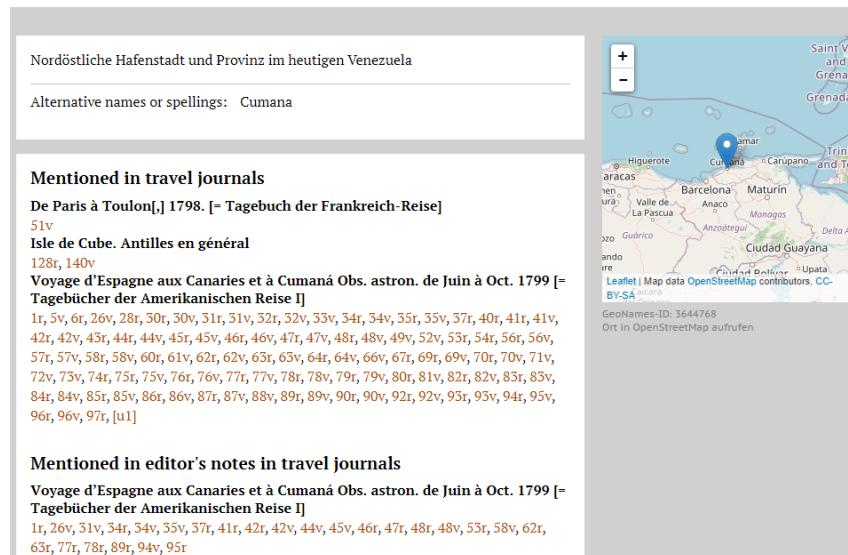


FIGURE 2.2 – La ville de Cumaná dans l'index des lieux de l'édition numérique

Par ailleurs, deux applications sont nécessaires afin de pouvoir présenter sur une carte un lieu sélectionné dans l'interface web. L'emploi de l'identifiant GeoName dans l'encodage permet également de récupérer ses geopoints qui sont ses points de géolocalisation. En plus d'une note présentant la lieu, le pays dans lequel il se situe et ses noms alternatifs, le lieu sélectionné dans l'index des lieux est présenté aux lecteurs de l'édition

numérique sur une carte qui provient de l'application Open Street Map. Tout comme pour les personnes, la présentation du lieu et ses noms alternatifs proviennent directement du document XML-TEI encodé par l'équipe du projet *Alexander von Humboldt auf der Reisen*. Les mentions de ce lieu au sein des divers carnets de voyages et des notes des éditeurs sont listées sous sa présentation générale.

En plus de ces données externes, l'emploi de services web et de standards permet à l'édition numérique de s'insérer dans des réseaux spécifiques. Grâce à l'utilisation du standard DTABf⁹, l'édition s'insère dans le corpus des institutions qui ont recourt également ce format. Les lettres éditées au sein de l'*edition humboldt* sont encodées dans le *Correspondence Metadata Interchange Format* (CMIF) qui est issu du XML-TEI. Ce format est

8. Voir la sous-section 1.2.2

9. Voir la sous-section 1.1.1

notamment employé dans le projet correspSearch¹⁰ initié par la BBAW. La plateforme correspSearch met à disposition la correspondance des personnes de diverses époques qui ont été éditées en version imprimée ou numérique. Il est ainsi possible de retrouver les lettres de l'ehd sur le site de correspSearch comme toutes les correspondances éditées au sein des éditions numériques de BBAW. Par ailleurs, la ressource archivistique est indiquée dans l'*edition humboldt digital*. Le lien vers le Kalliope-Verbundkatalog est indiqué au sein de l'encodage dans la balise <sourceDesc>. Le lien vers le catalogue en ligne est directement cliquable au sein de l'interface web. En cliquant sur ce lien, l'utilisateur accède davantage d'informations concernant le document physique et peut également avoir accès à la numérisation de ce document quand celle-ci existe.

```

<sourceDesc>
    <msDesc rend="manuscript">
        <msIdentifier>
            <institution>Biblioteka Jagiellońska</institution>
            <collection>Nachl. Alexander von Humboldt (Königliche Bibliothek)</collection>
            <idno>
                <idno type="shelfmark">Bd. 3/1, Bl. 127-149</idno>
                <idno type="URLImages">http://jbc.bj.uj.edu.pl/Content/350048</idno>
                <idno type="uri">http://kalliope-verbund.info/DE-611-HS-2877719</idno>
                <idno type="URLCatalogue">https://jbc.bj.uj.edu.pl/dlibra/docmetadata?showContent=true&id=350048</idno>
            </idno>
        </msIdentifier>
        <physDesc>
            <p>Manuskript in Folio, 23 Blatt</p>
        </physDesc>
    </msDesc>
</sourceDesc>
```

FIGURE 2.3 – Le lien vers le Kalliope-Verbundkatalog dans le <sourceDesc>

Isle de Cube. Antilles en général

H: Biblioteka Jagiellońska, Nachl. Alexander von Humboldt (Königliche Bibliothek) - Bd. 3/1, Bl. 127–149 Catalogue ↗

Edited by Ulrike Leitner, Piotr Tylus und Michael Zeuske

[Critical text](#) | [Reading text](#) | [Text with facsimile](#)

FIGURE 2.4 – Le lien vers le Kalliope-Verbundkatalog cliquable dans la version numérique de l'édition

Les données de la correspondance accessibles sur ce catalogue archivistique en ligne font par ailleurs partie des données qui ont été exploitées au cours du stage pour les projets menés. Ensemble, étudions maintenant ces données.

10. Le projet correspSearch et le CMIF font l'objet d'un chapitre complet, voir le chapitre 7

2.2 Les données manipulées au cours du stage

2.2.1 Kalliope-Verbundkatalog

Présentation du Kalliope-Verbundkatalog

Le Kalliope-Verbundkatalog est un catalogue numérique collectif et se situe dans la continuité du *Zentralkartei der Autograph* (en français : répertoire central des autographes) (ZKA), fondé en 1966 à Berlin Ouest. Les données à l'origine Kalliope étaient les 1,2 millions de fiches du ZKA ainsi que les références de près de 450 institutions allemandes¹¹. La base de données *ZKA online* a été développée de 2001 à 2003¹². Depuis octobre 2004, les fonds de base du catalogue concernant les fiches du ZKA sont consultables en ligne. Le portail a été renommé par la suite Kalliope. Ce dernier est un outil de référence nationale Outre-Rhin pour les références archivistiques manuscrites. Aujourd'hui ce sont environ 500 institutions comprenant des bibliothèques, des centres d'archives et des musées d'Allemagne mais aussi d'autres pays qui sont répertoriés comme institutions propriétaires et partenaires. La plupart des données hors Allemagne proviennent d'Autriche, Suisse et de personnes d'origine allemande ou germanophones situées dans d'autres pays tels que les États-Unis¹³.

| Année | Documents | Institutions partenaires ¹⁴ |
|-------|--------------------|--|
| 2010 | 1 530 600 | 54 |
| 2012 | 1 610 156 | |
| 2014 | 2 200 443 | |
| 2015 | 2 350 000 | 102 |
| 2021 | plus de 3 millions | 950 |

TABLE 2.1 – Quantité des documents et des institutions partenaires du Kalliope-Verbundkatalog

L'ensemble de la base de données a augmenté d'environ 20% par an ces quatre dernières années¹⁵. Les données sont mises à disposition en format *XML-Encoded Archival Description* (EAD) grâce à une API et plus particulièrement un serveur SRU.

11. *Kalliope / Historie*, URL : <https://kalliope-verbund.info/de/ueber-kalliope/historie.html> (visité le 26/07/2021)

12. *Ibid.*

13. *Ibid.*

14. Ce tableau a été réalisé à partir des chiffres présentés sur : *Ibid.* ; *KALLIOPE OPAC*, 26 oct. 2014, URL : <https://web.archive.org/web/20141026122746/http://kalliope.staatsbibliothek-berlin.de/> (visité le 26/07/2021)

15. *Kalliope / Historie...*

Des requêtes aux données

Sur le serveur SRU du Kalliope-Verbundkatalog, l'utilisateur peut récupérer les données qui l'intéresse sous format Dublin Core (DC) ou Metadata Object Description Schema (MODS). Le DC sert à décrire des documents de manière standardisée permettant une interopérabilité minimale. Le MODS, quant à lui, est issu du métalangage XML. Il a été conçu à la croisée du format MARC21, considéré comme complexe, et le DC qui propose une simplicité de jeu de métadonnées¹⁶. MODS structure des données bibliographiques en XML notamment dans le cadre de projet de catalogues bibliographiques ou de portails documentaires. Il contient en effet de nombreux éléments qui permettent d'indiquer les données au sujet de la description de la version numérique d'un document.

Les deux types de format ont été récupérés et utilisés au sein des projets menés au cours de mon stage. Dans un premier temps, le DC proposait suffisamment d'informations afin de mettre en place des fonctions de recherche et des visualisations. Néanmoins, ce format n'apportait pas assez de métadonnées pour le second projet, à savoir le développement d'un outil de corrélation entre deux sets de données. En effet, dans le format MODS, il est possible de récupérer les identifiants GND des correspondants épistolaires¹⁷ ce que ne propose pas le DC. Ainsi, le choix du format implique de savoir ce que l'on souhaite faire de ces données. Il est superflu de récupérer le maximum d'informations quand le projet ne nécessite pas toutes ces données. Dans la pratique et plus particulièrement dans la requête, le choix du format doit y être mentionné afin de récupérer les données dans le format souhaité.

Décortiquons une requête sur le serveur SRU du Kalliope-Verbundkatalog afin de comprendre son fonctionnement. Prenons cet exemple :

```
https://kalliope-verbund.info/sru?version=1.2&operation=searchRetrieve&
query=ead.addressee.gnd%3D%3D22118554700%22+AND+ead.genre%3D%3D%22Brief
%22&maximumRecords=4000&recordSchema=mods
```

- **version=1.2** indique la version de la demande du client et constitue une déclaration de ce dernier selon laquelle il souhaite que la réponse reçue soit égale ou inférieure à cette version. Ce paramètre **version** est obligatoire.
- **operation=searchRetrieve** est le type d'opération effectuée sur le serveur. L'opération **searchRetrieve** est la principale opération du SRU. Elle permet au client de soumettre une demande de recherche et d'extraction de données correspondantes auprès du serveur. Cet élément est obligatoire dans une requête.
- **query=** indique le début de la requête.
- **maximumRecords=5000** permet de préciser le nombre d'enregistrements à renvoyer. La valeur doit être égale ou supérieure à 0. En outre, si le nombre contenu dans

16. « Metadata Object Description Schema », *Wikipédia* (, 13 déc. 2020), URL : https://fr.wikipedia.org/wiki/Metadata_Object_Description_Schema (visité le 27/07/2021)

17. Voir l'annexe B.

la base est inférieur au nombre indiqué, le serveur renvoie un nombre d'enregistrements inférieur à celui demandé mais il ne doit pas en renvoyer plus. Ce paramètre est optionnel. Quand il n'est pas indiqué alors le serveur envoie le nombre d'enregistrements définit par défaut. Cette valeur varie en fonction du paramétrage du serveur par l'institution. Dans cette requête exemple, il est indiqué 4000 enregistrements maximum afin de récupérer en une seule requête les enregistrements de toutes lettres correspondantes. Or, le serveur renvoie 3236 entrées.

- `recordSchema=mods` indique le schéma dans lequel les enregistrements doivent être renvoyés. Ici, nous désirons un format MODS. Si le format DC est souhaité alors `dc` devra être indiqué à la place de `mods`.

On requête sur deux éléments distincts : `ead.addressee.gnd%3D%3D%22118554700%22` et `ead.genre%3D%3D%22Brief`. Ces deux éléments sont articulés par un opérateur logique AND. On recherche par conséquent tous les enregistrements qui sont du genre "lettre" (*Brief*) et dont le destinataire (*addressee*) a comme identifiant GND le nombre 118554700 qui est l'identifiant GND d'Alexander von Humboldt. Une seconde requête a ensuite été effectuée afin de récupérer toutes les lettres écrites par Alexander von Humboldt.

2.2.2 Société Américaine de Philosophie (APS)

Originellement, l'*American philosophical Society* (en français : Société Américaine de Philosophie) (APS) était un cercle de discussions fondé au milieu du XVIII^e par Benjamin Franklin dans la ville de Philadelphie. Débats scientifiques, publications, création d'une bibliothèques, les activités y étaient diverses¹⁸. Alexander von Humboldt en était l'un des membres en ayant été élu à la fin de son expédition américaine¹⁹.

Aujourd'hui la bibliothèque de l'APS abrite plus de treize millions de manuscrits, 350 000 volumes et périodiques, 250 000 images et des milliers d'heures de bandes audio²⁰. Grâce à ces fonds, l'APS est considérée comme l'une des premières institutions à documenter l'histoire naturelle aux XVIII^e et XIX^e siècles. Outre les collections de manuscrits remarquables telles que celle des documents de Benjamin Franklin, les journaux de Lewis et Clark ou la correspondance de Charles Darwin, l'APS abrite notamment la plus grande collection de documents manuscrits et imprimés d'Alexander von Humboldt des États-Unis.

L'APS propose une bibliothèque en ligne qui permet aux utilisateurs de découvrir les collections numérisées. En son sein, seulement quelques dizaines de manuscrits d'Alexander von Humboldt sont disponibles car la plupart des documents du scientifique conservés

18. « Société américaine de philosophie », *Wikipédia* (, 25 avr. 2020), URL : https://fr.wikipedia.org/wiki/Soci%C3%A9t%C3%A9_am%C3%A9ricaine_de_philosophie (visité le 28/07/2021)

19. « Alexander von Humboldt »...

20. *Use the APS Library*, American Philosophical Society, URL : <https://www.amphilsoc.org/library> (visité le 28/07/2021)

aux archives de l'APS ne sont pas numérisés. Les données de ces numérisations sont téléchargeables dans un fichier *Comma-separated values* (CSV), accessible directement par un bouton sur l'interface de la bibliothèque numérique. Une API est également proposée aux utilisateurs et les données sont récupérables en XML-EAD. La collection d'Alexander von Humboldt contient environ 250 documents. Le détail de son contenu est visible sur le site search.amphilsoc.org. Pour chaque document, diverses informations sont mentionnées : la personne qui est productrice de ce dernier, la date, la description physique, le lieu de conservation de l'original quand il s'agit d'une copie, le lieu de production ainsi que la typologie du document. Parfois quelques lignes résument le contenu du document concerné. En étudiant les documents de cette collection, on remarque que plusieurs dizaines de lettres y sont conservées. Celles-ci sont intéressantes pour le projet que nous devons mener.

2.2.3 Catalogue général de la Bibliothèque nationale de France

La France entre officiellement dans l'*open data* en 2011 avec l'ouverture du site national.data.gouv.fr. La BnF y intègre toutes les données structurées dans les formats du web sémantique et les diffuse dans le projet data.bnf.fr. Les objectifs de ce projet sont multiples²¹ :

- accroître l'exposition des données de la BnF sur le web
- fédérer les données de la BnF au sein et au-delà des catalogues
- contribuer à l'échange de métadonnées par la création de liens entre les ressources structurées et de référence
- faciliter la réutilisation des métadonnées par des tiers avec la garantie que ces dernières se trouvent sous licence ouverte.

C'est à partir de 2014 que toutes les métadonnées issues des catalogues de la BnF sont devenues librement réutilisables²². Cette ouverture se limite aux données descriptives c'est-à-dire aux métadonnées issues des catalogues et inventaires qui sont plus simples à extraire et diffuser. Il est devenu impératif d'exposer les données numériques dans des formats qui permettent aux utilisateurs de les réutiliser. Ce projet intègre des données produites en Intermarc pour les catalogues de livres, XML-EAD pour les inventaires d'archives et DC pour la bibliothèque numérique²³. Les données sont modélisées et regroupées par des traitements automatiques puis sont publiées dans les standards RDF.

Le site data.bnf.fr met en avant des informations précises et structurées. Ces informations qui sont des ressources pertinentes sont regroupées autour de concepts comme celui d'auteurs, d'oeuvres ou de thèmes. Le navigateur évolue parmi les données provenant

21. Voir le site de Data BnF.

22. R. Wenz, *L'open data, un levier pour l'évolution des catalogues...*

23. « Semantic Web and data model », *Data BnF* (), URL : <https://data.bnf.fr/en/semanticweb> (visité le 21/08/2021)

de sources diverses et accède au contenu des catalogues de manière immédiate. Cela est possible car la BnF fournit des URI pour les ressources grâce à des identifiants pérennes attribués selon le mécanisme *Archival Resource Key* (ARK) qui permet d'accéder à toutes les ressources que la bibliothèque propose.

La documentation accessible à l'utilisateur sur le site de data.bnf est riche que ce soit sur l'insertion du site au sein du web sémantique, sur les identifiants ARK ou encore sur la manière de se repérer dans le site. Tout est particulièrement bien documenté permettant aux utilisateurs de s'approprier de manière confortable les outils proposés par data.bnf.

Le service SRU du Catalogue général de la BnF permet d'interroger le catalogue général via des requêtes HTTP. L'ensemble des données sont récupérées dans différents formats bibliographiques encapsulés dans du XML. Le service est libre d'accès et ne demande aucune authentification de la part des utilisateurs. Le SRU est interrogable manuellement dans n'importe quel navigateur Internet et peut également être interrogé dans l'interface proposé sur le site de api.bnf.fr. Les divers paramètres utilisables dans le SRU sont explicitement détaillés sur le site permettant à chacun d'interroger le serveur selon ses besoins.

Ces données du Catalogue général sont différentes dans le format et leur niveau de détail de celles issues du Kalliope-Verbundkatalog et de l'APS. Afin de pouvoir manipuler toutes ces données comme un ensemble unique, il faut les traiter dans le but de les exploiter.

Chapitre 3

Des données brutes aux données traitées

La définition du traitement des données de Wikipédia permet de bien comprendre ce concept¹ :

En informatique, le terme traitement de données renvoie à une série de processus qui permettent d'extraire de l'information ou de produire du savoir à partir de données brutes. [...] Si la finalité n'est pas de présenter des résultats à un utilisateur humain, l'objectif du traitement de données est généralement d'offrir une information de plus haut niveau ou une information de meilleure qualité à un autre outil de traitement ou d'analyse. Ce traitement de l'information peut alors relever de la fusion de données, de l'extraction d'information ou de la transformation de la représentation.

Dans la phase de traitement, la donnée est nettoyée, compilée, croisée et analysée afin d'être enrichie.

3.1 Nettoyer

Interroger les API permet de récupérer des données formatées et structurées mais qui sont considérées comme brutes car elles n'ont pas encore été traitées. Le nettoyage est une étape essentielle dans le processus du traitement des données. Il s'agit de nettoyer les doublons, les données obsolètes ou incomplètes, les erreurs. Chaque set de données récupéré doit ainsi être nettoyé dans le but poursuivre le processus du traitement des données et d'être exploitables dans les fonctions de recherche.

Par exemple, les données récupérées sur l'API de l'APS contiennent des documents de divers typologies : y sont présentes des lettres, des poèmes, une déclaration des bagages

1. « Traitement de données », *Wikipédia* (, 8 juin 2021), URL : https://fr.wikipedia.org/wiki/Traitement_de_donn%C3%A9es (visit  le 30/07/2021)

FIGURE 3.1 – Entrée au sein du le fichier XML-EAD des données de l'APS

```

<---->
- <c02 level="item" id="ref14">
  - <did>
    <unittitle>Letter to Karl Ludwig Willdenow, Berlin</unittitle>
    - <origination label="Creator" audience="internal">
      <persname source="naf" rules="aacr">Humboldt, Alexander von, 1769-1859</persname>
    </origination>
    <unitdate>March 4, 1801</unitdate>
    <physdesc label="General Physical Description note" id="aspace_90fd2a84ad573a03d7d97854f1b595f8">9x7-1/4</physdesc>
  </did>
  - <scopecontent id="aspace_06e62921098deead4718f9dd0cae76a6">
    <head>Scope and Contents note</head>
    <p>Havana, A.L.S. 2p. and.add. In German. Old Call Number: B H88.23</p>
  </scopecontent>
</c02>

```

FIGURE 3.2 – Entrée au sein du le fichier DC issu du Kalliope-Verbundkatalog

```

-<srw:record>
  <srw:recordSchema>info:srw/schema/1/dc-v1.1</srw:recordSchema>
  <srw:recordPacking>xml</srw:recordPacking>
  <srw:recordData>
    <srw_dc:dc>
      <dc:identifier>DE-611-HS-2849533</dc:identifier>
      <dc:identifier>http://kalliope-verbund.info/DE-611-HS-2849533</dc:identifier>
      <dc:identifier>Fasc. germ. 268.8.</dc:identifier>
      <dc:publisher>DE-611</dc:publisher>
      <dc:title>Brief von Karl von Abel an Alexander von Humboldt</dc:title>
      <dc:created>20150609</dc:created>
      <dc:modified>20150622</dc:modified>
      <dc:contributor>Bayerische Staatsbibliothek</dc:contributor>
      <dc:language>ger</dc:language>
      <dc:language>ger</dc:language>
      <dc:type>item</dc:type>
      <dc:date>1840-10-22</dc:date>
      <dc:coverage>München</dc:coverage>
      <dc:creator>Abel, Karl von (1788-1859)</dc:creator>
      <dc:subject>Humboldt, Alexander von (1769-1859)</dc:subject>
      <dc:format.extent>1 eh.Br.m.U.m.Adr.</dc:format.extent>
    </srw_dc:dc>
  </srw:recordData>
-<srw:record>

```

FIGURE 3.3 – Entrées du CSV composé des données du Catalogue général de la BnF

| Identifiant | n° notice BnF | Type de notice | Type de document | Localisation | Exemplaire n° | Titre | Auteur | Contributeur |
|--|---------------|---|---|--------------|---------------|-------|--------|--------------|
| http://catalogue.bnf.fr/ark:/12148/cb38794301p | 38794301 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb38794301p | 38794301 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb387942947 | 38794294 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb387942947 | 38794294 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb44900472j | 44900472 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb38794293w | 38794293 | notice de recueil factice recueil de pièces | manuscrit moderne o | | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb38794293w | 38794293 | notice de recueil factice recueil de pièces | manuscrit moderne o | | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb387942978 | 38794297 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb387942978 | 38794297 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb38794298n | 38794298 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb38794298n | 38794298 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb38794299z | 38794299 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |
| http://catalogue.bnf.fr/ark:/12148/cb38794299z | 38794299 | monographie | manuscrit moderne ou document d'archive | Richelieu | | | | |

d'Alexander von Humboldt entrés sur le territoire des États-Unis... Dans le projet conçu, seules les lettres sont attendues. En outre, au sein de ce set de données, il faut extraire uniquement les données concernant les lettres, données qui seront ajoutées au set final c'est pourquoi 261 lettres ont été extraites du set de données source.

Par ailleurs, les données issues du Catalogue général de la BnF contenaient vingt-quatre entrées différentes. L'étude de ces données met en évidence le fait que seuls douze identifiants ARK sont à définir et non vingt-quatre. Chaque ensemble de documents est conservé dans un carton qui a été également microfilmé. Ainsi, chaque document présente deux entrées dans les données issues de l'API du Catalogue général : une entrée pour le document physique et une seconde pour son microfilmage. Un seul enregistrement est suffisant et est d'ailleurs intéressant pour notre projet. Par conséquent, les entrées

représentant les documents microfilmés ont été retirés des données sources afin d'éviter les doublons dans le set de données traitées.

Les données issues du Kalliope-Verbundkatalog se distinguent, quant à elles, en deux sets : les lettres reçues par Alexander von Humboldt et celles envoyées par ce dernier. En tout, ce set contient exactement 4566 lettres : 3235 rédigées par Alexander von Humboldt et 1331 lettres reçues par ce dernier. Ces deux sets étaient particulièrement propres et seule une transformation du XML-EAD vers *JavaScript Object Notation* (JSON) a été nécessaire. Cette conversion de format s'insère dans la seconde étape du traitement de nos données : la fusion.

3.2 Fusionner

Nous sommes à présent en possession de trois sets de données nettoyés : des données structurées dans un CSV, d'autres en XML-EAD et les dernières en DC. Chacune des structures est différente des autres. Or, les données doivent former un ensemble homogène et propre afin d'être exploitables au sein de notre projet. On parle alors de fusion de données. Cette dernière consiste à combiner plusieurs données issues de sources différentes, les compiler en un même set de données produisant une information plus sûre.

Afin de fusionner les différents sets les uns avec les autres pour former un unique ensemble, ces sets de données doivent tous être dans un format identique. Ce format peut être utilisé uniquement pour l'étape de la fusion et ne plus être employé au sein du projet final. Choisir la manière dont les données seront stockées dans une application ou un projet informatique n'est pas une étape anodine et mérite réflexion. Conceptualiser et mettre en place une base de données très structurée comme une base de données relationnelle prend du temps. Toutefois utiliser une base de données relationnelle n'est pas toujours le plus adapté en fonction des besoins. Pour diverses raisons, l'équipe du projet a décidé de stocker toutes ces données dans un fichier commun JSON qui présente plusieurs avantages :

- Stocker en JSON demande un faible besoin de place de stockage.
- JSON est un format largement connu et les données y sont facilement manipulables.
- JSON est un format également utilisé dans le projet développé en parallèle *Humboldt Chronotopographie*².

Le format JSON est un format de données textuelles et qui structure des informations. Une des caractéristiques du document JSON est qu'il comprend deux types d'éléments structurels : des ensembles de paires nommés clé/valeur et des listes ordonnées de valeur imbriquées les unes dans les autres. Le document JSON est plus facile à interpréter dans du code qu'un document XML. Ce dernier impose le recours à des techniques souvent plus lourdes. En effet, pour interpréter un document XML dans le langage Python il faut par exemple parcourir hiérarchiquement l'arbre *Document Object Model* (DOM).

2. Voir la section 4.3

Cet arbre DOM a d'ailleurs dû être parcouru afin de récupérer les données souhaitées dans le document XML-EAD et DC pour pouvoir les stocker dans le fichier JSON. Des courts algorithmes rédigés en Python ont été programmés afin de convertir les divers formats en JSON. Pour cela, la librairie `jxmlease` est particulièrement adaptée pour les conversions XML vers JSON ou vice versa. Une fonction de quelques lignes permet de convertir facilement les documents possédant une structure XML vers JSON :

```

1 def xml_to_json(file:str, data: list):
2     """
3         Convert a XML file to a JSON file
4     :param file: str
5     :param data: list
6     """
7     with open(file, "w+") as f:
8         d = jxmlease.parse(data)
9         json.dump(d, f)
10        f.close()

```

Pour le set de données en CSV de la BnF, la fonction `DictReader` permet de le convertir en JSON d'une manière tout aussi simple que ce que le permet la librairie `jxmlease` pour les fichiers XML :

```

1 def csv_to_json(file: str, delimiter : str, outputfilename: str):
2     """
3         Convert a CSV file to a JSON file
4     :param file: str
5     :param delimiter: str
6     :param outputfilename: str
7     """
8     data = {}
9     with open(file) as csv_file:
10        reader = csv.DictReader(csv_file, delimiter)
11        data = list(reader)
12        writeJSON(outputfilename, data)
13        csv_file.close()

```

Une fois que les trois sets de données ont été nettoyés et convertis en JSON, il est alors possible de les fusionner dans un fichier commun. Ce dernier, nommé `records.json` au sein de notre projet contient, après fusion, 4931 enregistrements de lettres provenant de trois sources différentes.

3.3 Enrichir

Enrichir ses données est fortement lié à ce qu'on souhaite en faire. Parfois les enrichir est inutile mais bien souvent un enrichissement est nécessaire afin d'obtenir les résultats escomptés. Lors de cette étape, on peut créer des liens avec des ressources externes. On

parle alors d'alignement avec des référentiels³. Ce type de lien permet un rapprochement avec des entités similaires et augmente ainsi la visibilité du jeu de données au sein de web sémantique.

Puisqu'un des objectifs du projet est de produire une visualisation cartographique, les geopoints des lieux des données doivent nécessairement venir compléter notre set. Afin que ce projet puisse s'insérer au sein d'*Alexander von Humboldt auf Reisen*, l'identifiant interne de l'édition doit également être ajouté. L'API GeoName permet de récupérer l'identifiant GeoName ainsi que les coordonnées géographiques d'un lieu. GeoName est une base de données géographiques mondiale. Elle contient plus de onze millions de noms de lieux. Afin de pouvoir utiliser son API, l'utilisateur doit s'assurer d'être identifié. GeoName et son API font déjà partie du réseau des services web utilisés au sein de l'édition numérique. J'ai ainsi pu profiter de l'identifiant de l'édition afin de récupérer les données dont j'avais besoin pour étoffer le set de lettres. Les données qui viennent enrichir notre set sont multiples. Pour chaque enregistrement au sein du set de données, doivent être rajoutés :

- l'identifiant GeoName, l'identifiant interne de l'édition, la géolocalisation du lieu de conservation (clé **contributor** dans le fichier JSON)
- l'identifiant GeoName, l'identifiant interne de l'édition, la géolocalisation du lieu d'envoi ou de réception de la lettre (clé **coverage_place**)

Ainsi, l'algorithme a été rédigé afin de compléter les données concernant les lieux de conservation des lettres, puis un second afin d'enrichir les données des lieux d'envoi de ces mêmes lettres. Afin de limiter le nombre de requêtes envoyé à l'API GeoName, le programme recherche dans un premier temps si le lieu recherché n'est pas déjà enregistré dans l'index de lieu de l'édition. Un second index a été créé localement afin de pouvoir stocker les lieux qui ne sont pas dans l'index de lieu de l'édition et qui ont déjà été requêtés sur l'API GeoName. Cela permet de ne pas effectuer deux requêtes différentes sur l'API GeoName pour un même lieu. La fonction a pour paramètre le nom du lieu dont les informations sont incomplètes⁴ et doivent être enrichie.

Ainsi, un enregistrement de lettre au sein du fichier JSON comprenant les informations géographiques ajoutées se présente comme suit :

3. Cécilia Fabry, Clotilde Roussel, Alain Collignon, François Parmentier, Elise Moreau et Nicolas Thouvenin, « Publier des données liées et ouvertes en sept étapes », *I2D - Information, donnees documents*, Volume 54-1 (1^{er} avr. 2017), Bibliographie_available : 0 Cairndomain : www.cairn.info Cite Par_available : 0 Publisher : A.D.B.S., p. 12-14, URL : <https://www-cairn-info.proxy.chartes.ps1.eu/revue-i2d-information-donnees-et-documents-2017-1-page-12.htm> (visité le 09/08/2021)

4. Voir l'annexe C.

```
{  
  "identifier": ["DE-611-HS-1318487", "http://kalliope-verbund.info/DE-611-HS-1318487",  
    "Nachl. Alexander von Humboldt, gr. Kasten 11, Nr. 130"],  
  "publisher": "DE-611",  
  "title": "Brief von Karl Degenhardt an Alexander von Humboldt",  
  "created": "20071026",  
  "modified": "20150804",  
  "contributor": "Staatsbibliothek zu Berlin. Handschriftenabteilung",  
  "language": "ger",  
  "type": "item",  
  "date": "1838-02-08",  
  "coverage": "Clausthal-Zellerfeld",  
  "creator": "Degenhardt, Karl (-1844)",  
  "subject": "Humboldt, Alexander von (1769-1859)",  
  "format.extent": "4 Bl.",  
  "coverage_location": { "geoname_id": 2939995, "address": "Clausthal-Zellerfeld",  
    "coordinates": ["10.33821", "51.80949"] },  
  "contributor_location": { "key": "H0005221", "geoname_id": 2950159,  
    "address": "Staatsbibliothek zu Berlin. Handschriftenabteilung",  
    "coordinates": ["13.41053", "52.52437"] }  
}
```

Deuxième partie

**Développer de nouvelles
fonctionnalités avec des données
tierces**

Chapitre 4

Entre recherche et ingénierie

Ce dialogue entre recherche et ingénierie est particulièrement important. Le rôle du chercheur est crucial dans la modélisation des données des projets numériques en sciences humaines et sociales. En effet, cette modélisation repose sur l'explicitation de ce que veut faire le chercheur de ces données. Créer ou développer un outil qui n'aura aucune utilité au chercheur ne fait pas sens. Le chercheur doit faire comprendre aux ingénieurs ses besoins et ses attentes afin que les outils numériques soient pleinement utilisés par l'équipe de recherche.

4.1 Le corpus archivistique

Afin que les ingénieurs puissent développer des outils adéquats, il est impératif que les chercheurs explicitent leurs buts et pour cela l'ingénieur doit également comprendre le corpus archivistique qui sera mis en valeur dans l'outil numérique qu'il développera. En effet, la singularité du corpus va guider et contraindre les ingénieurs dans la modélisation des données et la conception des outils.

4.1.1 La correspondance d'Alexander von Humboldt

Alexander von Humboldt est considéré comme l'un des épistoliens les plus prolifiques toutes époques confondues. Il est estimé que ce dernier a écrit plus de 30 000 lettres entre 1787 et 1858¹. Avant sa mort, il a lui-même déclaré qu'il recevait en moyenne entre 1600 et 2000 lettres par an². Néanmoins, il est fort probable qu'il ait reçu bien plus de lettres à la fin de sa vie qu'au début. Sa renommée et son mode de vie devenu alors

1. Ingo Schwarz, « Die Korrespondenz », dans *Alexander von Humboldt, Handbuch : Leben-Werk-Wirkung*, dir. Ottmar Ette, Springer-Verlag, Stuttgart, 2018, p. 80-91

2. Kurt-R. Biermann et Fritz G. Lange, « Die Alexander-von-Humboldt-Briefausgabe », *Forschungen und Fortscritte*, 36, 8 (1962), p. 225-230

sédentaire³ sont des raisons possibles de ces quantités de lettres reçues. Tous ces chiffres restent cependant des estimations difficiles à établir puisque la plus grande partie de la correspondance d'Alexander von Humboldt ne nous est pas parvenue. En effet, lors de ces voyages et expéditions à l'étranger, le scientifique n'avait pas pour habitude de conserver toutes les lettres qu'il recevait et en gardait seulement des morceaux, des bribes. Il est alors difficile d'estimer le nombre de lettres qu'il a envoyé et il est encore plus difficile d'estimer le nombre de lettres qu'il a reçu au cours de sa vie, riche d'échanges épistolaires.

Le nom d'environ 2800 correspondants d'Alexander von Humboldt est à ce jour connu⁴. Dès les années 1980, l'historien des sciences et spécialiste des travaux et de la vie d'Alexander von Humboldt, Kurt-Rheinhardt Biermann, a identifié 182 correspondants qu'il considère être les plus proches du scientifique voyageur⁵. Un correspondant proche est une personne ayant reçu plus d'une dizaine lettres écrites de la main d'Alexander von Humboldt, lettres connues et conservées aujourd'hui. Les astronomes, physiciens et chimistes, géoscientifiques et mathématiciens, géologistes et botanistes occupent la première place de ces partenaires désignés comme proches et appartenant à la sphère des spécialistes des sciences naturelles. Néanmoins, les spécialistes des sciences humaines sont tout aussi nombreux à correspondre avec Alexander von Humboldt. Parmi ces derniers, ce sont les historiens et linguistes qui dominent. Dans les domaines artistique et littéraire, on trouve des écrivains, des peintres et un compositeur. Alexander von Humboldt a également correspondu avec divers chefs d'État dont notamment le roi de Prusse, Friedrich Wilhelm IV, ainsi qu'avec des ministres et des officiers. Le cercle familial prend aussi une part non négligeable au sein de la correspondance du scientifique qui écrivait régulièrement à son frère Wilhelm von Humboldt et ses nièces. Des lettres destinées à des banquiers sont également conservées, environ 250 lettres sont parvenues aux yeux des chercheurs et du public intéressé⁶. Ces échanges nombreux avec les banques sont les témoins de sa situation financière qui a pu parfois être précaire mais également de son intérêt pour les questions économiques.

Il existe cependant aucune institution qui conserve l'entièreté des lettres et documents manuscrits d'Alexander von Humboldt. Sa dense correspondance est dispersée dans les archives, les bibliothèques et collections privées au quatre coins du monde. Reconstituer la correspondance du scientifique, ou du moins une partie de cette dernière, a donc été une importante mission débutée en Allemagne de l'Est au début des années 1950.

3. Alexander von Humboldt a effectué de nombreuses expéditions sur plusieurs mois voire années et a donc vécu de nombreuses années de manière plus ou moins nomade.

4. I. Schwarz, « Die Korrespondenz »...

5. *Ibid.*

6. *Ibid.*

4.1.2 Reconstituer la correspondance d'Alexander von Humboldt : les archives de la *Berlin-Brandenburgische Akademie der Wissenschaft*

Le projet de reconstitution de la correspondance d'Alexander von Humboldt a débuté avec la création d'une commission Alexander von Humboldt à la *Deutsche Akademie der Wissenschaften zu Berlin* (Académie allemande des sciences de Berlin) en 1956. Deux ans plus tard, le bureau de Berlin de la nouvelle Commission est fondée et les premiers membres du personnel comprennent notamment Kurt-R. Biermann, historien des sciences, et Johannes Eichhorn, bibliothécaire⁷.

En 1960, les académies des sciences d'Allemagne de l'Est et de l'Ouest ainsi que l'académie autrichienne des sciences ont envoyé aux académies, archives, bibliothécaires et collectionneurs une demande commune de soutien international dans le but d'éditer la correspondance de Humboldt. Cet appel à soutien, rédigé en allemand, anglais, français, russe et espagnol, a été diffusé en 322 exemplaires et dans seize pays différents⁸. Il était également signé par les représentants des diverses académies, ce qui a permis d'accroître la visibilité internationale de cette édition. La diffusion de cet appel à l'échelle internationale a mis en réseau les institutions participantes. Cela a permis de décloisonner, du moins dans le cadre de ce projet, la République Démocratique Allemande (RDA) dans le domaine de la recherche et de l'édition scientifique⁹.

Deux ans plus tard, des copies d'environ 7 600 lettres provenant du monde entier ont été reçues à Berlin, constituées de 7000 lettres écrites par Humboldt et 600 lettres reçues par ce dernier. L'objectif principal de cette collection matérielle était et reste à ce jour l'édition historico-critique complète des manuscrits du scientifique voyageur¹⁰.

| Année ¹¹ | Lettres d'Humboldt | Lettres à Humboldt | Autres documents |
|---------------------|--------------------|--------------------|------------------|
| 1962 | 7000 | 600 | |
| 1965 | 8800 | 1400 | |
| 1974 | 10500 | 2700 | |
| 2001 | 12500 | 3000 | |
| 2021 | 8690 | 2215 | 2175 |

TABLE 4.1 – Tableau du nombre de documents d'Alexander von Humboldt conservés à la BBAW

7. Gregor Schuchardt, *Fakt, Ideologie, System. Die Geschichte der ostdeutschen Alexander von Humboldt-Forschung*, Franz Steiner Verlag, Stuttgart, 2010, p.50

8. *Ibid.*, p.56

9. *Ibid.*, p.57

10. Cette expression est particulièrement utilisée par Marie-Noëlle Bourget dans son ouvrage : Marie-Noëlle Bourget, *Le monde dans un carnet : Alexander von Humboldt en Italie (1805)*, Édition du félin, Paris, 2017.

11. K.R. Biermann et F. G. Lange, « Die Alexander-von-Humboldt-Briefausgabe »..., p. 227-8;K.R. Biermann, « Der Zugang an Briefen Alexander von Humboldts hält an », *Spektrum. Mitteilungsblatt für*

Comme le montre ce tableau, la collection a été en constante augmentation. Il s'agit de la base qui a permis à la *Alexander-von-Humboldt-Forschungsstelle* (centre de recherche Alexander von Humboldt) de commencer ses travaux en 1970. Ainsi, de 1973 à 2014, une équipe de chercheurs a publié quarante-deux volumes de correspondances, monographies et anthologies. La collection de manuscrits photocopiés et la bibliothèque de référence, qui s'est également enrichie depuis les années 1950, constituent le cœur du travail d'édition du projet actuel *Alexander von Humboldt auf Reise - Wissenschaft aus der Bewegung* qui poursuit ce travail de longue haleine dans l'esprit des humanités numériques depuis 2015.

Cette collection de manuscrits est conservée au sein de la bibliothèque Humboldt de la BBAW et fonctionne toujours sur un système de classement non informatisé. Il existe au sein de la bibliothèque plusieurs documents auxquels se référer pour effectuer des recherches dans les archives conservées. Tous les noms sont répertoriés sur des cartes conservées dans des tiroirs de classements. Afin de trouver un document, il suffit de rechercher un nom dans la collection des cartes sur laquelle est inscrite la boîte de classement contenant le document recherché ou bien de se référer à l'aide à la recherche.



FIGURE 4.1 – Exemple du tiroir contenant les cartes des noms commençant par la lettre L.

4.2 L'aide à la recherche numériquement retranscrite

L'aide à la recherche, sorte de catalogue, appelé *Findbuch* en allemand, est un petit carnet bleu unique existant en un seul exemplaire et conservé sous clé au sein de la bibliothèque Humboldt de la BBAW. Ce carnet contient un index de lieux de conservation des documents, le nombre de documents qui y est conservé ainsi que le numéro du tiroir de classement¹².

Cette aide à la recherche a été entièrement retranscrit numériquement dans un tableau

¹² *die Mitarbeiter der Deutschen Akademie der Wissenschaft zu Berlin*, 11, 2 (1965), p. 55-58, p. 55 et p.58; Id., « Die Alexander-von-Humboldt-Forschung an der Akademie der Wissenschaften der D.D.R. - Ergebnisse und Ziele », dans *Boston Studies in the Philosophy of Science*, 1974 (15), p. 295-305, p.296 ; I. Schwarz, « Zur Geschichte der Alexander-von-Humboldt-Forschung und der Berlin-Brandenburgischen Akademie der Wissenschaft », dans *Die Berliner und Brandenburger Lateinamerikaforschung in Geschichte un Gegenwart. Personen und Institutionen*. Dir. Gregor Wolff, Wissenschaftlicher Verlag Berlin, Berlin, 2001, p. 107-127, p. 112. Pour ce qui est de l'année 2021, l'inventaire a été effectué par Anne McKinney, ancienne stagiaire au sein du projet d'édition. Le tableau a d'ailleurs été réalisé par cette dernière.

12. Voir Figure D.1 et Figure D.2 des annexes.

FIGURE 4.2 – Aide à la recherche en version numérique

| Alexander von Humboldt Forschungsbibliothek Briefarchiv: Findbuch | | | | | | | | | |
|---|-----------|--|--|---|--|--------------|--------|---|--|
| K. Nr. | Stadt | Register ID (https://www.geonames.org/) | geonames ID (https://www.geonames.org/) | Besitzende Institution | GND-Nr. (http://dbpedia.org/resource/) | von H. an H. | sonst. | Webseite | Bemerkungen |
| 1 | Aberdeen | H0017270 | | University of Aberdeen, University Library | 113996-4 | 1 | | https://www.abdn.ac.uk/library/ | |
| 1 | Altenburg | 2957773 | | Landesarchiv Thüringen, Staatsarchiv Altenburg | 1119367980 | 1 | | https://landesarchiv-thueringen.de/altenburg | |
| 1 | Amsterdam | H0002435 | | Koninklijke Nederlandse Akademie van Wetenschappen | 37531-7 | 2 | 2 | https://www.knaw.nl/nl | |
| 154, 188 | Ann Arbor | 4984247 | | University of Michigan Library | 63035-4 | 1 | 1 | https://elements.umich.edu/ | |
| 157,2 | Aurich | 6557459 | | Ostfriesische Landschaftsbibliothek | 5142607-9 | 1 | 1 | https://www.ostfriesischelandschaft.de/de/4.html | |
| 1 | Avignon | H0002629 | | Calvet | 5045755-X | 2 | | http://www.bibliotheques-calvet.org/ | |
| 1 | Baltimore | H0005091 | | Maryland Center for History & Culture | 1048183-7 | | | https://www.mdhistory.org/ | |
| 1 | Baltimore | H0005091 | | Enoch Pratt Free Library | 502755-X | | 1 | https://www.prattlibrary.org/ | |
| 1 | Bamberg | H0005094 | | Staatsarchiv Bamberg | 200518-3 | 53 | 12 | https://www.gda.bayern.de/bamberg/ | |
| 2 | Barnaul | H0005133 | | Staatliches Heimatmuseum des Altai (Алтайский государственный краеведческий музей/ Altajski Gosudarstvennyi Krajewedtcheskiy Muzej) | k.A. | | | https://myagkm.ru/ | nicht zu verwechseln mit Gosudarstvennyi Muzei Iстории, Искусства и Культуры Altaja (16032782-9, http://gmlklik22.ru/) der ca. 70 km außerhalb Barnaus Das Altajski Gosudarstvennyi Krajewedtcheskiy Mi weist dagegen explizit auf Humboldt als Besucher hin zusammengeführte und aktualisierte Einträge von Autographen(s), der Universitätsbibliothek Basel (Autographensel., Menzel=1 Brief von H: Autograph Geigy-Hagenbach=1 Brief von H [ehemals 2 Briefe, 1 im Jahr 1971 an die SBB PK Berlin verkauft]; Autogra Brüderlein=2 Briefe von H) sowie 5 weitere Handschr. Hs. die in anderen Nachlässen und Abteilungen der liezen. |
| 2 | Basel | H0005157 | | Universität Basel, Universitätsbibliothek | 2023655-4 | 9 | | https://ub.unibas.ch/de | |

Excel par Anne MacKinney, ancienne stagiaire au sein du projet d'édition *Alexander von Humboldt auf Reisen - Wissenschaft aus der Bewegung* en mars 2021. Ainsi, les données de l'aide à la recherche, à savoir côté du tiroir de classement, ville, institution propriétaire, détails des documents conservés, y sont retranscrits. En plus de ces données ont été rajoutés l'identifiant GeoName et l'identifiant interne à l'édition des lieux de conservation, le site internet de l'institution conservant les documents ainsi qu'une colonne commentaire. Si l'institution concernée a déposé ses données sur le site du Kalliope-Verbund¹³, alors la requête a été copiée dans le document afin d'avoir un accès direct à la correspondance d'Alexander von Humboldt conservée dans cette institution-ci.

Ce tableau a permis d'effectuer l'inventaire des documents conservés¹⁴ au sein de la BBAW. En plus d'être le point de départ de la première mission de mon stage, il a également été utile pour l'outil expérimental développé par Dr. Gordon Fischer pour le projet *Alexander von Humboldt auf Reisen*.

4.3 *Humboldt Chrotonopographie* : un outil expérimental

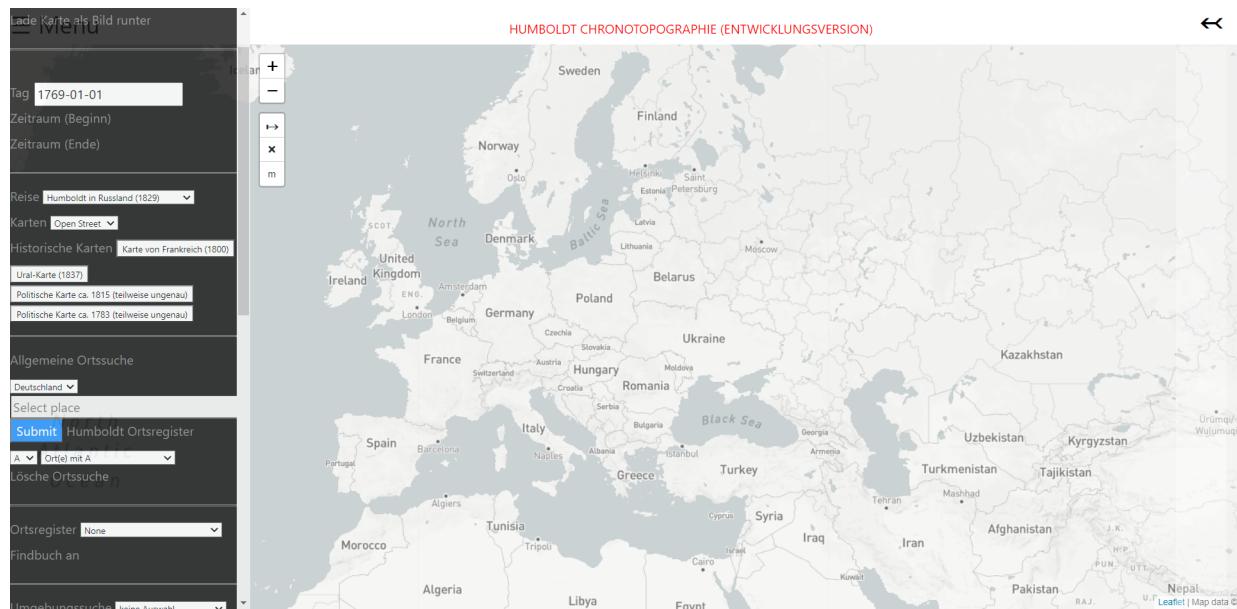
En effet, les données de l'aide à la recherche ont été insérées au sein du projet *Humboldt Chronotopographie*. Ce projet a pour but de visualiser grâce aux cartes les itinéraires empruntés lors des divers voyages et expéditions d'Alexander von Humboldt. Il a pour vocation de proposer des données ouvertes téléchargeables par tous et de devenir un projet collaboratif. Ce travail, à la croisée d'outil de travail scientifique et de visualisations presque tout public, a débuté à l'automne 2020. Il est appuyé par l'équipe du projet

13. Voir la sous-section 2.2.1.

14. Voir tableau présentant le nombre de documents d'archives d'Alexander von Humboldt conservés à la BBAW, Tableau 4.1.

Alexander von Humboldt auf Reisen et est développé par Dr. Gordon Fischer, ingénieur au sein du service *The Electronic Life Of The Academy* (TELOTA) qui est le pôle des humanités numériques de la BBAW.

FIGURE 4.3 – Copie d'écran générale de l'outil *Humboldt Chronotopographie*



Le menu de gauche¹⁵ de cet outil permet de sélectionner ce que l'utilisateur souhaite visualiser : définir une période, choisir son fond de carte, choisir un voyage ou bien des données particulières. Dans ce menu, il suffit de cliquer sur *Findbuch an* (en français : afficher le catalogue) afin de visualiser tous les lieux de conservation des documents manuscrits sur une carte du monde. Un menu à droite s'affiche alors avec la ville et l'institution conservatrice. Le détail des documents conservés au sein de l'institution est également affiché. Ces données forment une visualisation en elles-mêmes et ne sont pas reliées à d'autres données¹⁶. Les données ont directement été extraites du tableau Excel et ont été insérées dans l'outil expérimental au cours de son développement.

Dans les premiers temps du développement de l'outil, l'accent a été mis sur les données chronotopographiques de l'*edition humboldt digital*. Ces données sont la combinaison de la datation et la géolocalisation des informations provenant des carnets de voyage du scientifique. À noter que ces données sont elles-mêmes un ensemble de données reliées à des ressources externes¹⁷.

Les itinéraires de voyages de Humboldt, que ce soit la grande expédition américaine (1799-1804) ou encore l'expédition en Russie et Asie centrale entreprise en 1829, n'ont encore jamais été visualisées de manière dynamique. Dans l'édition numérique, chaque

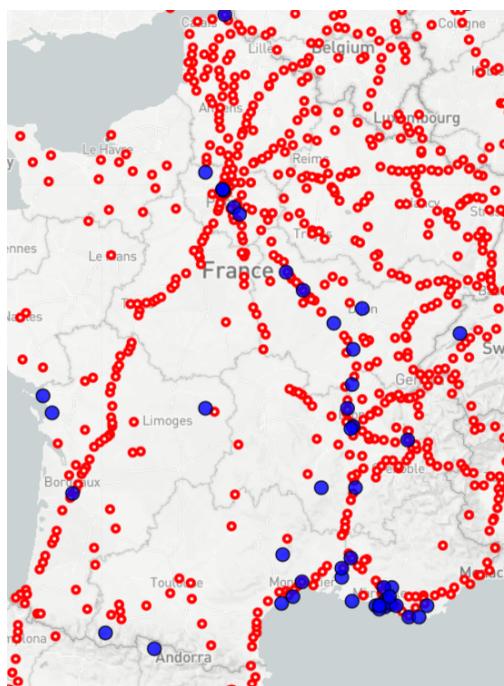
15. cf Figure E.1.

16. Voir la copie d'écran de cette visualisation, Figure E.1.

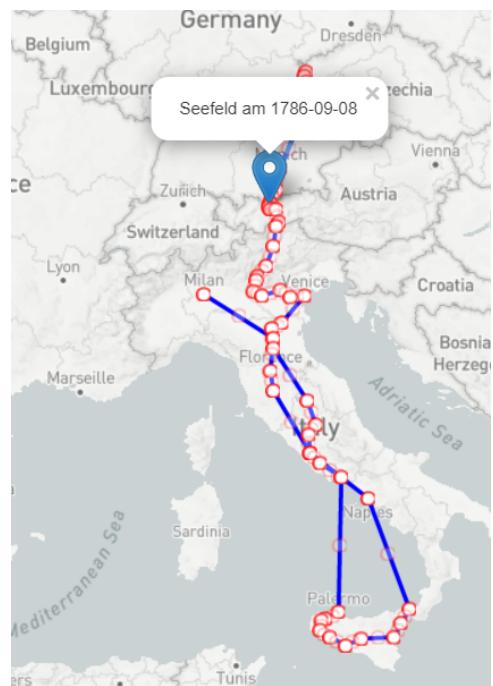
17. cf. chapitre 1

lieu qui a été balisé dans un `<placeName>` au sein de l'encodage XML est relié à son identifiant GeoName et à son identifiant interne à l'édition¹⁸. Environ 3000 lieux sont balisés dans un `<placeName>` au sein de l'*edition humboldt digital*. Afin de visualiser les divers itinéraires de voyage de Humboldt, les données sont extraites directement des documents encodés en XML-TEI de l'édition et plus particulièrement des registres de lieux et des entrées chronologiques. Ces données alors agrégées forment un nouvel ensemble de données exploitable par l'outil chronotopographique et stockées dans une base de données eXist-DB. Elles sont visualisables via une transformation XSLT vers *Hypertext Markup Language* (HTML) sur l'application web *Humboldt Chronotopographie*.

FIGURE 4.4 – Copies d'écran de *Humboldt Chronotopographie*



(a) Itinéraire de Humboldt en France en 1798 (en bleu) superposé aux relais de poste (en rouge).



(b) Animation de l'itinéraire de Goethe lors de son voyage en Italie en 1786.

L'un des objectifs de cet outil expérimental est donc de visualiser les voyages et lieux de séjour d'Alexander von Humboldt en exploitant les données de l'*edition humboldt digital*, enrichies par des données tierces et en les liant à des services web divers. Les données du projet ont été complétées par des données externes comme les routes postales définies par Giacomo Casanova. Toutes les villes possédant un relais postal sont listées et sont accessibles de manière ouverte sur le site de giacomo-casanova.de. Dr. Gordon Fischer a, pour chaque lieu listé, ajouté sa géolocalisation et son identifiant GeoName. Il est ainsi possible de faire apparaître tous les relais de poste listés par Casanova vers 1750 et de les comparer avec les routes empruntées par Humboldt lors de ses itinéraires comme le

18. cf section 1.2.

présente la capture d'écran (a) ci-dessous. Humboldt a donc suivi une des routes principales où les relais de poste sont particulièrement nombreux afin de descendre de Paris vers le sud de la France. Dans l'avenir, il s'agirait de pouvoir visualiser les routes de voyages empruntés par d'autres figures de l'histoire contemporaine. Un voyage en Italie effectué par Goethe en 1786 a d'ailleurs été pris en exemple et inséré au sein de l'outil chronotopographique, comme le présente la copie d'écran (b) ci-dessus.

Représenter les itinéraires de voyage n'est pas une idée inédite. En effet, il existe déjà divers projets d'humanités numériques qui proposent de visualiser les routes de voyage ou les voyages entrepris de certaines grandes figures lettrées de l'époque des Lumières. *Mapping of the Republic of letters*¹⁹, développé à l'université de Stanford, propose de cartographier la République des Lettres avec notamment des visualisations sophistiquées et interactives de données sur les réseaux de correspondances et des cartes d'itinéraires de voyages de ces érudits de l'époque contemporaine. Il est ainsi possible de découvrir le réseau de correspondances de Voltaire, Condorcet, Benjamin Franklin mais aussi des voyageurs effectuant ce qu'on appelle le Grand Tour²⁰. Une visualisation cartographique des villes italiennes visitées par des architectes au cours de leur Grand Tour permet de rendre compte des villes les plus visitées au cours du XVIII^e siècle. Chaque point expose le nom du lieu et le nombre d'architectes qui l'ont visité. La taille du point représenté est proportionnelle au nombre de visites du lieu. Par ailleurs, le site *Encyclopedia of Romantic Nationalism in Europe* (ERNIE) est un site collaboratif, proposant des données ouvertes de plus de 1700 articles analytiques sur divers thèmes et figures emblématiques de cette époque. Il se propose de retracer la montée internationale de la construction de la culture nationale dans l'Europe du XIX^e siècle, construction qui s'est établie au sein du mouvement romantique-nationaliste²¹ à travers tout matériel produit à cette époque à savoir les lettres, la musique, les œuvres peintes, mais aussi d'autres événements comme les expositions ou les voyages individuels. Cette plateforme d'études académiques destinée à l'étude critique des documents historiques est un projet collaboratif avec la participation de plus de 361 contributeurs principalement situés en Europe et en Amérique du Nord²². ERNIE propose à ses utilisateurs de visualiser une quinzaine d'itinéraires de voyage. Tous apparaissent sur une même carte et le menu de droite contenant la liste des itinéraires

19. Les visualisations et plus d'informations sont disponibles sur le site internet.

20. Le Grand Tour, aussi appelé *Junkerfahrt* ou *Cavalierstour* dans les pays du Saint-Empire romain germanique, avaient diverses fonctions éducatrices : approfondir ses connaissances dans les arts et sciences sociales, acquérir une formation politique approfondie avec notamment la comparaison des systèmes politiques des lieux visités, rencontrer des jeunes aristocrates étrangers ou du moins vivant à l'étranger. Ces voyages avaient une fonction sociale une fois que les jeunes aristocrates rentraient au pays puisqu'ils constituaient un élément de reconnaissance ou d'ascension sociale en affirmant les moyens financiers et la culture du jeune aristocrate. Alexander von Humboldt a effectué un tour d'Italie avec Gay-Lussac quelques mois après son expédition en Amérique du sud en 1805. L'historienne, Marie-Noëlle Bourguet, revient sur ce voyage de plusieurs mois en analysant notamment le carnet de voyage du scientifique, voir M.N. Bourguet, *Le monde dans un carnet : Alexander von Humboldt en Italie (1805)*...

21. Voir la documentation du site.

22. Voir la carte du monde des contributeurs.

disponibles permet de désélectionner ceux que l'utilisateur ne souhaite pas visualiser. Cliquer sur un point apparaissant sur la carte permet d'avoir accès à des informations supplémentaires dont notamment le site internet du projet collaborateur.

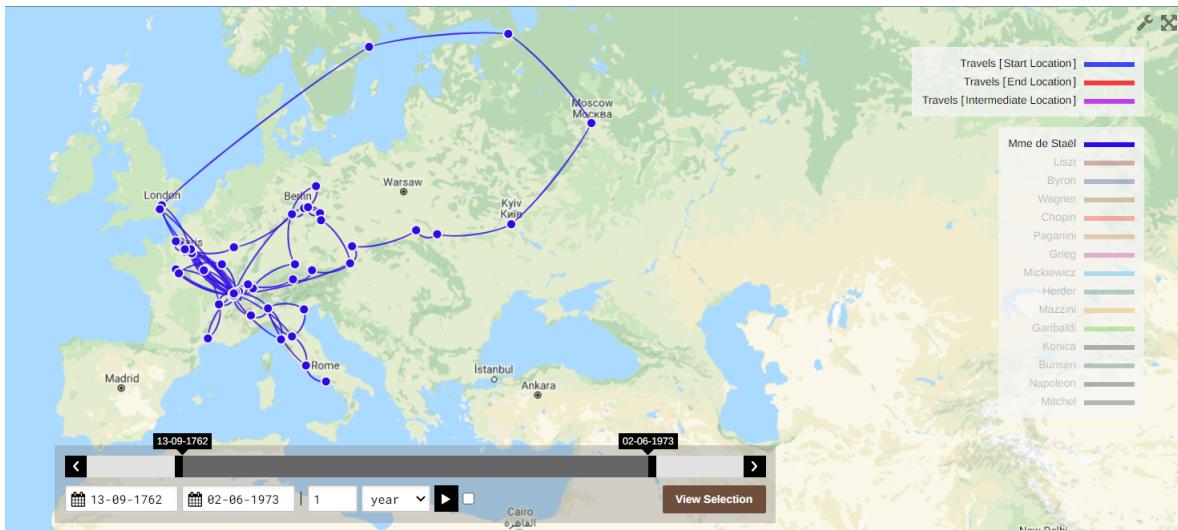


FIGURE 4.5 – Copie d'écran du site ERNIE présentant l'itinéraire du voyage de Mme de Staél, itinéraire sélectionné dans le menu à droite.

À la différence de tous ces autres projets d'humanités numériques, l'outil développé par la BBAW a pour vocation, en plus de présenter les données de l'édition, d'accompagner les chercheurs dans leur recherche scientifique et de combler les lacunes des données. En effet, dans ses carnets de voyage, Humboldt n'a pas toujours précisé toutes les villes-étapes où il a séjourné. Par conséquent, il existe des lacunes de plusieurs jours dans ses itinéraires.

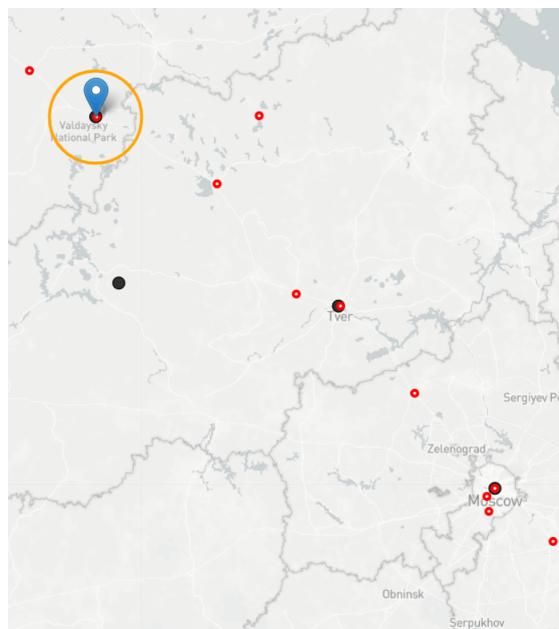


FIGURE 4.6 – Définir un itinéraire : l'exemple de Valtaï à Moscou

Prenons l'exemple, représenté ci-dessus, d'un itinéraire effectué lors du voyage en Russie et Asie centrale de 1829. Humboldt écrit dans son journal "*De Pét à Moscou 113–27 (Waldai II 59)*²³". Il va donc de Saint-Pétersbourg à Moscou en passant par la ville de Valdaï. Cependant, quelles ont été les étapes entre toutes ces villes ? Moscou se situe à plus de cinq jours de voyage de la ville de Valdaï. Grâce à l'outil développé, il est possible de choisir une zone de recherche à savoir entre un et cinq jours de voyage. En cliquant sur un lieu, une zone s'affiche sous forme de cercle à partir du lieu sélectionné et permet au chercheur de définir les possibles lieux de séjour de Humboldt. En croisant ces données avec les relais de poste listés par Casanova et intégrés à l'outil, il est également possible de définir les routes les plus empruntées et ainsi d'émettre des hypothèses quant aux lieux où Humboldt a pu séjourner. Sur la copie d'écran, le cercle orange représente la zone autour de la ville de Valdaï accessible en un jour de voyage. À savoir qu'un jour de voyage en cheval a été estimé à une trentaine de kilomètres. C'est en tous cas l'estimation utilisée dans l'outil *Humboldt Chronotopographie* afin d'établir ces zones. Les points rouges représentent les relais de poste appartenant à la liste de Casanova et les points noirs sont les villes évoquées par Humboldt dans son carnet de voyage. L'évocation d'une ville ou d'un lieu ne signifie pas que Humboldt y a séjourné. Il n'existe aucune différence dans l'encodage pour les villes où Humboldt a séjourné et celles dont il fait une simple référence. Par conséquent et puisque les données représentées dans l'outil sont directement issues de l'encodage, l'outil ne peut effectuer une différence dans l'affichage de ces points qui sont alors représentés de la même manière.

L'outil permet également de superposer divers fonds de cartes qui apportent des informations supplémentaires à l'utilisateur. Trois fonds de cartes additionnels sont disponibles sur *Humboldt Chronotopographie* : la carte mondiale des frontières politiques vers 1783 et celle vers 1815 ainsi qu'une carte de la France vers 1800²⁴. Les lieux tirés des carnets de voyage de Humboldt se superposent au fond de carte choisi. Il est ainsi plus aisément de situer les itinéraires entrepris par le scientifique. Cela a permis de remarquer par exemple que Humboldt a suivi la frontière kazakhe lors de son voyage en Russie et en Asie centrale de 1829, comme le présente la carte ci-contre.

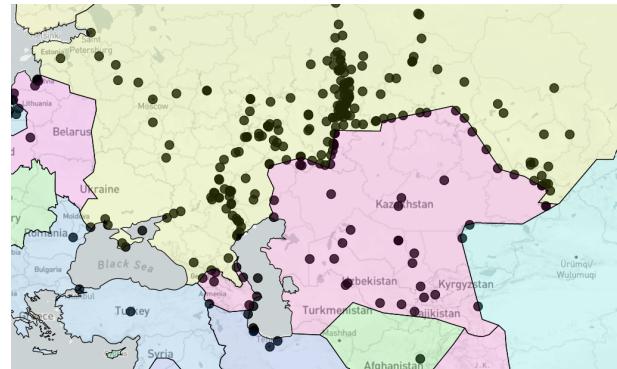


FIGURE 4.7 – Superposer les cartes : exemple de la frontière kazakhe

23. *Fragmente des Sibirischen Reise-Journals 1829*, 9r. L'édition de cette page accompagnée de son facsimilé est disponible dans l'*edition humboldt digital*.

24. Voir les différents fonds de cartes, Figure E.2

Chapitre 5

Développement d'outils de recherche et des visualisations

Dans les traces de l'outil *Humboldt Chronotopographie*, ma première mission de stage a été de reconstruire, du moins en partie, la correspondance d'Alexander von Humboldt et de proposer des fonctions de recherche afin de l'explorer. Le développement de ces fonctions a été soutenu techniquement par Dr. Gordon Fischer, ingénieur au sein du pôle TELOTA de la BBAW et développeur de l'outil expérimental *Humboldt Chronotopographie* sur les carnets de voyages d'Alexander von Humboldt pour le projet *Alexander von Humboldt auf Reisen*. Dr. Ulrich Pässler, chercheur et responsable du projet de recherche et d'édition ainsi que tuteur de mon stage, en a supervisé la conception.

Numériser toutes les photocopies conservées au sein des archives Humboldt de la BBAW est un travail de plusieurs années. Afin de reconstruire de manière numérique les archives de la correspondance d'Alexander von Humboldt, il a été choisi, à partir de l'aide à la recherche retranscrit numériquement, d'aller chercher les données correspondantes aux lettres manuscrites écrites et reçues par Alexander von Humboldt directement sur les sites d'archives conservatrices de ces lettres proposant des données accessibles. Les données utilisées au sein de ce projet proviennent du Kalliope-Verbund¹, du Catalogue général de la Bibliothèque nationale de France² et de la Société américaine de philosophie³. Afin d'explorer et de découvrir les données de cette nouvelle collection numérique de la correspondance de Humboldt, des fonctions de recherche ainsi que des visualisations de données ont été réalisées.

1. Voir la sous-section 2.2.1.

2. Voir la sous-section 2.2.3

3. Voir la sous-section 2.2.2.

5.1 L'environnement technique et les librairies choisies

Les visualisations de données constituent un moyen d'expression modulable afin d'exposer les données de la correspondance d'Alexander von Humboldt. Visualiser sur une carte les échanges épistolaires de ce dernier permet de rendre compte de l'ampleur de ceux-ci et du réseau mondial dans lequel s'insère le scientifique.

5.1.1 Le choix de l'environnement technique : Jupyter Notebook

Il existe de nombreuses technologies qui permettent la réalisation de visualisations de données. Néanmoins, ces technologies ne sont pas toutes adaptées et n'exigent pas le même niveau de compétence technique. L'environnement technique dans lequel le projet se développe est également important pour la conception des visualisations : une visualisation sera différemment présentée sur une application Android ou sur une application web, de même, si elle a pour vocation d'être interactive ou statique. L'environnement technique choisi présente des configurations réduites aux fonctionnalités qu'il propose. Ces possibilités de configuration ont un impact direct sur les réalisations potentielles.

Le choix de l'environnement technique dans lequel ont été développées les fonctions de recherche et les visualisations s'est porté sur les Jupyter Notebook. Il s'agit d'une application web et open-source qui permet de créer des documents compartimenté par cellule. Chacune d'entre elle peut avoir une typologie de contenu différente telle que du code ou du Markdown. Cela permet au programmeur de documenter de manière riche ses cellules de code. Ces dernières peuvent être lancées indépendamment des unes des autres ce qui est un réel avantage afin de pouvoir tester quelques lignes de code ou une unique fonction par exemple.

Les Jupyter Notebook présentent de nombreux avantages :

- Ils proposent des sorties de cellule, appelés *outputs*, interactifs avec l'utilisation possible de HTML mais aussi l'affichage d'images et de visualisations interactives.
- Ils supportent plus de quarante langages de programmation différents.
- Ils permettent d'exploiter et d'explorer les données à l'aide de nombreux outils et librairies utilisables directement en leur sein.

Les utilisations possibles des Jupyter Notebook sont multiples : le nettoyage et la transformation des données, la modélisation statistique, la visualisation des données, le machine learning... En effet, les Jupyter Notebook sont particulièrement puissants dans le traitement de données et notamment de larges sets de données. Dr. Gordon Fischer a notamment utilisé des Jupyter Notebook afin d'extraire des données affichées sur l'application web *Humboldt Chronotopographie*⁴.

4. cf section 4.3.

Un des avantages des Jupyter Notebook est le fait qu'il n'y ait pas besoin de programmer une application web complète pour visualiser des données. Néanmoins, il faut avoir quelques connaissances techniques afin de pouvoir utiliser un Jupyter Notebook : ce n'est en effet pas un outil grand public. Si on souhaite que le grand public ait accès aux données et au projet alors il serait plus judicieux de développer une application web à laquelle les utilisateurs auraient directement accès par leur moteur de recherche. Sans être un projet à part entière, l'utilisation des Notebooks peut également être une étape au sein de la conception d'un site internet en rendant compte des données et des visualisations qui y seraient proposées.

5.1.2 Les librairies utilisées

Les Notebooks supportent de nombreux widgets interactifs qui permettent aux utilisateurs de visualiser et contrôler les changements au sein de leurs données. Ces widgets sont proposés par diverses librairies. Certaines d'entre elles ont été nécessaires pour le développement de fonction de recherche et de visualisations. Toutes les librairies qui ont été utilisées pour leur élaboration sont open source.

Ipyleaflet

Ipyleaflet permet de créer des cartes interactives au sein des Jupyter Notebooks. Cette librairie prend en charge les annotations, les divers marqueurs de lieux et propose différents fonds de cartes. Elle est particulièrement adaptée afin de réaliser des visualisations cartographiques de la correspondance d'Alexander von Humboldt.

La documentation est bien fournie sur le site qui lui est dédié⁵. Chaque fonctionnalité de la librairie est présentée avec des exemples de code afin de la prendre en main aisément.

Ipywidgets

Ipywidget est également connu sous le nom de jupyter-widgets. Cette librairie propose de nombreux widgets HTML interactifs au sein des outputs de cellules d'un Jupyter Notebook. Un widget est un objet du langage de programmation, ici Python. Ils ont une représentation dans le navigateur web de l'utilisateur. Ils servent notamment à créer des interfaces graphiques interactives. La librairie met à disposition environ une quinzaine de widgets et fournit pour chacun une documentation explicite⁶. Le programmeur peut avoir la parfaite maîtrise de l'apparence des widgets grâce aux nombreux paramètres disponibles pour chacun d'entre eux.

5. Voir le site ipyleaflet.readthedocs.io

6. Voir la documentation sur ipywidgets.readthedocs.io

Pandas

Pandas est une librairie proposant des outils d'analyse et de manipulation de données. Elle est particulièrement adaptée pour les structures de données. Parmi elles, DataFrame est une des principales structures de données gérées par cette librairie. Il s'agit d'un stockage des données en deux dimensions, c'est-à-dire sous forme de tableau. Cette structure de donnée a régulièrement été utilisée au sein de notre projet notamment pour présenter les résultats des fonctions de recherche.

| | | identifier | publisher | title | created | modified | contributor | language | type | date | coverage |
|------|---|------------|--|-------|----------|----------|--|----------------|-------------|------------|----------|
| 0 | [DE-611-HS-1650160, http://kalliope-verbund.in... | DE-611 | Brief von Hermann Abich an Alexander von Humboldt | | 20100505 | 20100505 | Universitätsbibliothek Freiburg | ger | item | 1852/1853 | Tiflis |
| 1 | [DE-611-HS-2945920, http://kalliope-verbund.in... | DE-611 | Brief von Pierre Jean François Turpin an Alexa... | | 20151208 | 20170615 | Universitätsbibliothek Leipzig | fre | item | NaN | o. O. |
| 2 | [DE-611-HS-2946346, http://kalliope-verbund.in... | DE-611 | Brief von Louis Nicolas Vauquelin an Alexander... | | 20151209 | 20170616 | Universitätsbibliothek Leipzig | fre | item | 1819-10-05 | Paris |
| 3 | [DE-611-HS-1783480, http://kalliope-verbund.in... | DE-611 | Brief von Gonzalo O'Farrill y Herrera an Alexa... | | 20120302 | 20120302 | Universitätsbibliothek Leipzig | fre | item | 1828-05-04 | Paris |
| 4 | [DE-611-HS-1783988, http://kalliope-verbund.in... | DE-611 | Brief von Joseph Barclay Pentland an Alexander... | | 20120306 | 20120328 | Universitätsbibliothek Leipzig | eng | item | 1826-06-01 | Lima |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4666 | http://catalogue.bnf.fr/ark:/12148/cb387942916 | NaN | [Alexander von Humboldt] Lettre à Edme-François... | | NaN | NaN | Bibliothèque nationale de France (Paris) | fre français | monographie | 1811 | NaN |

FIGURE 5.1 – Structure de données de la librairie pandas : dataframe de toutes les données de la correspondance d'Alexander von Humboldt

Pandas⁷ propose également des fonctionnalités multiples dont la table pivot qui est un tableau croisé dynamique. La table pivot regroupe les données selon un ou plusieurs critères et les présente sous forme de sommes, de moyennes ou encore de comptage. Cette fonctionnalité se présente dans l'output d'une cellule d'un Jupyter Notebook et permet de manipuler les données de manière confortable et de les croiser. Une table pivot de la librairie pandas se présente comme ci-contre. Les

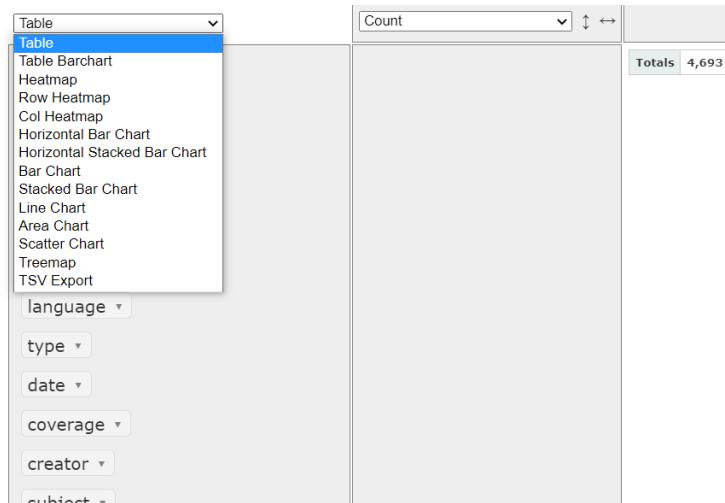


FIGURE 5.2 – Table pivot de pandas avec toutes les données de la correspondance d'Alexander von Humboldt

7. La documentation est disponible sur pandas.pydata.org

données sont disponibles dans la colonne de gauche et sont manipulables au sein de l'output. En fonction de ce que l'on souhaite, il suffit de faire glisser les données de la colonne de gauche vers celle de droite ou sur la ligne du haut. La table pivot produit elle-même le calcul et présente les résultats sous forme de tableau.

Matplotlib et NumPy

Matplotlib est une librairie destinée à tracer et visualiser des données sous forme de graphiques. La documentation de matplotlib⁸ est particulièrement riche et détaillée. Toutes les possibilités de réalisation ou presque sont présentées et accompagnées de visuels afin de les prendre en main.

Matplotlib est régulièrement combinée avec la librairie NumPy qui sert à effectuer des calculs scientifiques. Ces deux librairies forment un combo complet afin de réaliser des graphiques les plus complexes et avoir la totale maîtrise de leur apparence.

5.2 Algorithmie et réalisations

L'algorithmie du projet a été rédigée en Python au sein de Jupyter Notebook. Les visualisations et les fonctions de recherche sont disponibles dans les cellules de sortie de ces documents une fois que les cellules de code ont été lancées. Un premier Jupyter Notebook a été créé pour les fonctions de recherche puis un second pour les visualisations. Ils ont pour vocation d'être utilisés par l'équipe du projet d'*Alexander von Humboldt auf der Reisen* dont beaucoup sont chercheurs. Pour cette raison, la plus grande partie du code est "cachée" dans des fichiers .py. Dans les Notebooks, seulement quelques lignes sont visibles dans les cellules de code permettant de lancer l'algorithme des fonctions de recherche et des visualisations. Un des avantages des documents Jupyter Notebook est le fait d'associer des cellules de code et des cellules Markdown, utile à la rédaction d'une documentation qui apporte des informations sur les données utilisées, la présentation des résultats, des chiffres à propos des données. Cela ajoute du contexte à l'utilisateur autour des sets de données et du travail réalisé.

Présentons quelques informations chiffrées autour du set de données afin de comprendre l'ampleur de celui-ci. Pour rappel⁹, les données sont stockées sous format JSON et contient 4932 entrées de lettres reçues ou envoyées par Alexander von Humboldt. Parmi elles, se sont pas moins de 580 personnes ou institutions ayant envoyé une lettre au scientifique et 570 personnes ayant reçu une lettre écrite de la main de ce dernier. En tout, 351 lieux d'envoi ou de réception sont enregistrés. Dans ce set de données, y sont présentes les lettres de soixante-douze institutions. En effet, soixante-dix institutions conservatrices de ces lettres ont déposé leurs données sur le Kalliope-Verbundkatalog. Ces institutions

8. Voir la documentation sur Matplotlib.org

9. Voir chapitre 3.

ne sont pas toutes localisées en Allemagne puisqu'il y a notamment la Bibliotheka Jagiellonska située à Krakow ou encore Has-Sifriya hal-Le'ummit située à Jérusalem, parmi d'autres. À ces plusieurs dizaines d'institutions s'ajoutent la Bibliothèque nationale de France et la Société américaine de philosophie pour lesquelles les données ont été récupérées via leur API respectives¹⁰.

5.2.1 Fonctions de recherche

Certaines fonctions de recherche permettent de rechercher au sein des institutions conservatrices de lettres. Pour ces fonctions, les données issues de l'aide à la recherche numériquement retranscrit sont venues enrichir le set de données de la correspondance d'Alexander von Humboldt. Cela augmente le nombre d'institutions conservatrices à 310 institutions au lieu de soixante-dix. Ce chiffre important témoigne du travail qu'il reste à effectuer afin de reproduire la collection analogue des archives Humboldt de la BBAW. Malgré les données récupérées sur les diverses API, le nombre de lettres enregistrées au sein de notre base de données reste inférieur au nombre de lettres enregistrées dans les archives Humboldt : 10 905 lettres à l'inventaire de 2021¹¹ contre 4932 dans notre base de données.

Les fonctions de recherche effectuent des recherches au sein de divers éléments : expéditeurs, destinataires, lieu de réception ou d'envoi et institutions conservatrices. Pour chacune de ces fonctions un dropdown menu permet à l'utilisateur de sélectionner la valeur voulue. Ce dropdown menu est accompagné d'un bouton *New search* donnant la possibilité à l'utilisateur à tout moment d'effectuer une nouvelle recherche en relançant la cellule du Notebook. La fonction qui crée le bouton et permet relancer la cellule quand celui-ci est cliqué par l'utilisateur se présente comme ci-dessous.

```

1 def btn_new_search():
2     """
3     Create a new search button.
4     :return: btn
5     :rtype: button
6     """
7     # Create the button
8     btn = createButton('New search', 'info')
9     output = wgt.Output()
10
11    def new_search(b):
12        # This function clears the output of a jupyter cell
13        with output:
14            display(Javascript('IPython.notebook.execute_cell()'))
15
16    # When the button is clicked, then the output of the jupyter
17    # cell will be clean.
18    btn.on_click(new_search)
19    return btn

```

10. cf sous-section 2.2.2 et sous-section 2.2.3.

11. En effet, 8690 lettres écrites par Humboldt et 2215 reçues par ce dernier ont été inventoriées. Voir Tableau 4.1

Search by Sender letters (to AvH)

Senders Buschmann, Johann Carl Eduai

Search by Coverage place

Places Berlin

Search by Stockholding institution

Institutions Deutsches Literaturarchiv Marb

New search

FIGURE 5.3 – Exemple des dropdown menus apparaissant de manière dynamique par la fonction récursive.

Une fonction de recherche appelée dynamique a également été réalisée. Cette dernière autorise l'utilisateur à choisir dans un premier temps par quel élément de recherche il souhaite débuter son exploration et la fonction continue de proposer un nouvel élément de recherche jusqu'à ce qu'il n'y en ait plus ou bien jusqu'à ce que les résultats ne contiennent plus qu'une unique lettre. Ainsi, l'utilisateur peut tout d'abord sélectionner le nom d'un destinataire, puis l'année de la lettre, puis l'institution conservatrice de la lettre qu'il

cherche, puis le destinataire jusqu'à ce que la lettre qu'il recherche apparaisse en tant que résultat. Cette recherche est possible par le développement d'une fonction récursive. Une fonction récursive est une fonction qui fait appel à soi-même au sein de son propre algorithme. Ainsi, notre fonction de recherche s'appelle soi-même et renvoie les éléments de recherches restants. En effet, l'utilisateur peut effectuer une recherche parmi quatre éléments, à savoir le destinataire, l'expéditeur, le lieu d'envoi et l'institution conservatrice. Si l'utilisateur choisit dans un premier temps le lieu d'envoi alors la fonction va effectuer un appel vers soi-même avec une liste des éléments de recherche restant, à savoir une liste contenant les éléments expéditeurs, destinataires et institutions de recherche. Le paramètre *flag* est une variable booléenne qui permet à la fonction de reconnaître si c'est la première fois qu'elle est appelée ou bien s'il s'agit d'un appel récursif.

Toutefois, pour chacun des éléments énoncé ci-dessus à savoir les expéditeurs, les destinataires, le lieu de conservation et le lieu d'envoi des lettres, une fonction de recherche a été créée indépendamment des autres. Ainsi, l'utilisateur peut effectuer une recherche pour un élément précis sans avoir à utiliser la recherche dynamique. Chacune de ces fonctions de recherche indépendante des autres se présente sous deux éléments : un dropdown menu contenant les valeurs sélectionnables et un bouton *New search*.

Prenons l'exemple de la création d'une fonction de recherche et notamment du dropdown menu pour la recherche à travers les dates. Afin de ne pas apporter trop d'informations à l'utilisateur mais aussi dans le but de réduire la quantité des valeurs sélectionnables, seules les années sont proposées au sein du dropdown menu et non toutes les dates entières sous le format YYYY-MM-DD. Afin de récupérer toutes les valeurs correspondantes aux dates de création au sein de la base de données JSON, la fonction `nested_lookup()` de la librairie `nested-lookup` effectue une recherche par clé sur un document. Cette fonction prend comme argument la clé recherchée au sein du document et le document dans lequel la fonction doit rechercher. La fonction `nested_lookup()` renvoie sous forme de liste les valeurs correspondantes à cette clé donnée. Cette liste retournée est ensuite nettoyée de différentes manières avant d'être envoyée, en tant qu'argument, à la fonction permettant de créer le dropdown menu.

D'autre part, certaines lettres ont été rédigées par plusieurs personnes. C'est un cas fréquent notamment au sein de la famille Mendelssohn dont plusieurs membres sont en contact avec Humboldt. Cette famille rédige régulièrement des lettres à plusieurs. Tous les différents rédacteurs ont été encodés par les institutions conservatrice sous forme de tuples¹². Pour les fonctions de recherche réalisées, il a été décidé de retirer ces tuples afin de proposer une liste plus claire de noms. C'est la fonction `avoidTupleInList()` qui retire ces tuples de la liste de noms. Ensuite, la fonction `getYears()` récupère pour chacune des dates les quatre premiers éléments de la date d'envoi, c'est-à-dire l'année. Une lettre au sein de la base de données se doit d'être envoyée ou reçue au cours des années de vie d'Alexander von Humboldt. Toutefois, des erreurs d'encodage sont présentes dans notre base de données. Au sein de celle-ci, certaines dates enregistrées en tant que date de création des documents se situent bien au delà de la date de naissance ou de décès du scientifique. On retrouve notamment 1937-02-18 pour une lettre d'Alexander von Humboldt à Johann Carl Eduard Buschmann¹³ ou encore 1938-02-08 pour une lettre de Karl Degenhardt à Humboldt. Peut-être s'agit-il de copies des manuscrits. Afin d'éviter que ces éléments erronés apparaissent au sein de la liste du dropdown menu et qu'ils apportent de la confusion à l'utilisateur, ceux-ci sont retirés grâce à la fonction `getHumboldtYears()` qui retourne les lettres envoyées ou reçues entre 1769 et 1859. Ces explications faites, voyons à quoi ressemble à présent cette fonction qui permet de créer la fonction de recherche par date :

```

1 def search_date(data:dict):
2     """
3     Create a dropdown menu with all years when a letter
4     (to and by AvH) has been sent
5     :param data: dict
6     :return: dropdown menu
7     :rtype: widget
8     """
9     years = getHumboldtYears(getYears(avoidTupleInList(nested_lookup('date', data))))
10    dropdown = createDropdown('', years)
11    dropdown.observe(onChangeDate)
12    return dropdown

```

Letters to Alexander von Humboldt

Actual count of senders to AvH : 582

In [18]: display(HBox([search_creators(d), btn_new_search()]))

In []:

- Burckhardt, Johann Karl (1773-1825)
- Burkart, Josef (1798-1874)
- Burmeister, Hermann (1807-1892)
- Buschmann, Johann Carl Eduard (1805-1880)
- Bustamante, José María
- Butakow, Alexei Iwanowitsch (1812-1869)
- Buyß-Ballet, Christoph Heinrich Diedrich (1817-1890)
- Böckn, August (1785-1867)
- Bülow, Heinrich von (1792-1846)
- Calmberg, Adolf
- Camphausen, Otto (1812-1896)
- Cancrin, Georg (1774-1845)
- Candolle, Augustin Pyramus de (1778-1841)
- Candolle, Augustin Pyramus de (1778-1841) [vermutlich]
- Canning, George (1770-1827)
- Caroline Amalie (Dänemark, Königin) (1796-1881)
- Carus, Carl Gustav (1789-1869)
- Chevrel, Michel E. (1786-1889)
- Cichacev, Petr A. (1809-1890)
- Clausijs, Rudolf (1822-1888)

FIGURE 5.4 – Exemple du dropdown menu pour les expéditeurs de lettres.

FIGURE 5.4 – Exemple du dropdown menu pour les expéditeurs de lettres. Le dropdown menu affiche une liste de noms et leurs dates de naissance ou de décès. Il existe également un bouton de recherche et un bouton pour effectuer une nouvelle recherche.

Une fois que la liste de valeurs est nettoyée et stockée dans la variable `years`, cette

12. Un tuple est un ensemble ordonné de valeurs. Le séparateur de chaque valeur est, en Python, une virgule.

13. Buschmann est linguiste, bibliothécaire et est également le secrétaire privé des frères Humboldt. Il assiste également Alexander von Humboldt dans l'élaboration d'une de ses œuvres principales *Kosmos*. Par conséquent, les échanges épistolaires entre les deux hommes sont particulièrement riches comme le présente l'histogramme, Figure 5.5.

même-liste devient l'argument de la fonction qui crée le dropdown menu. La fonction `observe()`, quant à elle, gère l'accès aux données et leur présentation une fois qu'une valeur est sélectionnée par l'utilisateur au sein du dropdown menu.

5.2.2 Visualisations

La façon dont l'utilisateur a accès aux données grâce aux visualisations sera abordée dans le chapitre prochain. Il s'agit ici de présenter la manière dont les visualisations ont été programmées et peuvent apparaître dans la cellule de sortie du Jupyter Notebook. Certaines visualisations sont directement liées aux fonctions de recherche, c'est-à-dire qu'elles sont le résultat d'une valeur sélectionnée dans un dropdown menu par l'utilisateur. En plus de la présentation des résultats par un DataFrame de la librairie pandas, les visualisations sont une autre représentation des résultats de recherche. Aussi, ces visualisations ne présentent qu'une partie, celle sélectionnée par l'utilisateur, des données de notre base de données. Parmi ces visualisations, deux types sont à distinguer :

- les visualisations cartographiques. Ces dernières sont interactives c'est-à-dire que l'utilisateur peut zoomer, cliquer sur les différents points de la carte pour obtenir des informations supplémentaires.
- les visualisations représentant des histogrammes qui sont statiques.

Les histogrammes

Les histogrammes permettent de visualiser de manière graphique la répartition d'une variable. Cette répartition s'affiche sous forme de colonne. Les histogrammes sont particulièrement adaptés afin de représenter une variable dans la durée par exemple. Ces visualisations graphiques ont permis de représenter, au sein de notre projet, les échanges épistolaires entre une personne sélectionnée par l'utilisateur et Humboldt.

Tout comme pour les fonctions de recherche, un dropdown menu est disponible afin que l'utilisateur sélectionne une personne. Ce menu est également accompagné d'un bouton *New search*. Une fois qu'une valeur est choisie, l'histogramme des échanges de lettres apparaît.

Il est toutefois possible que rien ne s'affiche sur l'histogramme pour certains partenaires épistolaires. Cela signifie que pour toutes les lettres enregistrées dans la base et échangées entre cette personne et Humboldt aucune date n'est connue ou n'a été enregistrée par les catalogues en ligne

Anzahl des Briefwechsels zwischen AvH(1769-1859) und Buschmann, Johann Carl Eduard (1805-1880)

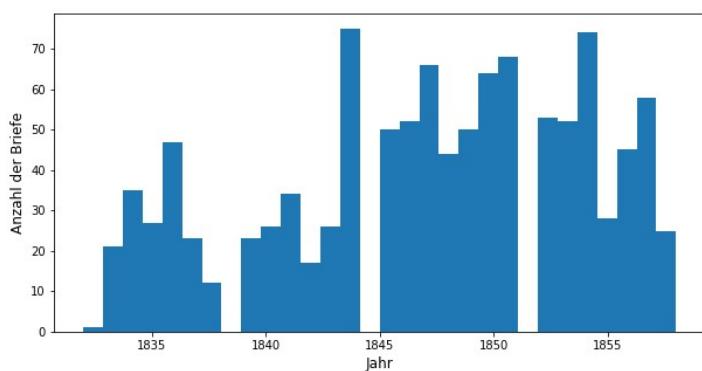


FIGURE 5.5 – Échanges épistolaires entre Buschmann et Humboldt

sur lesquels les données ont été récupérées. Par conséquent, apparaissent seulement les lettres dont la date est connue. Afin d'avoir accès à toutes les données correspondantes à la personne sélectionnée, l'utilisateur est convié à se diriger vers les fonctions de recherche qui affichent toutes les données disponibles sans condition c'est-à-dire sous forme de DataFrame. Les visualisations des données sont par conséquent dépendantes des données enregistrées. Certaines de ces visualisations sont peu pertinentes quand les données sont incomplètes. Ce point mériterait des améliorations de code afin que l'histogramme s'affiche seulement quand il a des données à visualiser.

Les cartes

De même que les histogrammes, les visualisations cartographiques sont dépendantes des données et plus particulièrement des données géographiques. Seulement les lettres accompagnées d'un lieu d'envoi ou de réception suffisamment précis et pour lesquelles les coordonnées correspondantes sont venues compléter notre set de données¹⁴ au moment de son enrichissement sont représentables.

Bien que des visualisations cartographiques sont une représentation possible du résultat d'une fonction de recherche, deux cartes ont été créées afin d'exposer l'entièreté de la base de données¹⁵. Ces cartes apportent une vue d'ensemble sur la correspondance de Humboldt. Grâce à elles, l'utilisateur peut prendre connaissance de l'ampleur de ces échanges épistolaires qui touchent tous les continents. En effet, à l'étude de ces cartes, il est devenu remarquable qu'au moins une lettre avait été reçue ou envoyée sur chacun des continents.

La première de ces cartes expose tous les lieux où une lettre a été envoyée ou reçue par Alexander von Humboldt au cours de sa vie. Les points sur la carte sont des cercles dont la taille est proportionnelle au nombre de lettres reçues ou envoyées en ce lieu. En vérité, la taille des points n'est pas complètement proportionnelle pour des raisons de lisibilité. Pour chacun des lieux, le nombre de lettres envoyées ou reçues est calculé. Toutefois, si ce chiffre est supérieur à dix lettres alors le rayon du point sera de 12. Cela signifie qu'à partir de onze lettres, la taille du point représenté n'est plus proportionnel. Ces points, bien que non proportionnels au chiffre qu'ils représentent, sont malgré tout plus grands que ceux

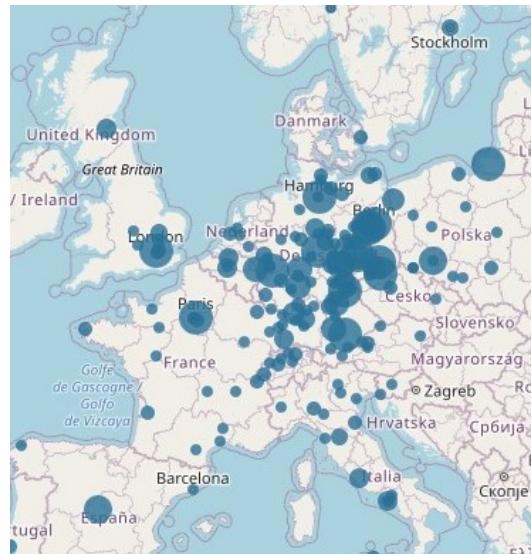


FIGURE 5.6 – Zoom sur l'Europe de la carte représentant l'ensemble de la correspondance de Humboldt.

14. Voir la partie **Enrichir** de chapitre 3.

15. Voir les cartes Figure F.1 et Figure F.2

représentés proportionnellement permettant d'apporter un ordre de grandeur suffisant à l'utilisateur afin de comprendre l'information. Cela est important car si chaque rayon d'un point représenté restait proportionnelle au nombre de lettres qu'ils représente alors la carte deviendrait illisible. Prenons l'exemple de la ville de Berlin où 1134 lettres¹⁶ y ont été reçues ou envoyées. Si le point représentant Berlin était proportionnelle à ces plus de milles lettres alors le point sera plus grand que la carte elle-même et elle deviendrait incompréhensible. L'opacité des points est de 80% ce qui permet également une meilleure lisibilité de l'information puisque tous les points peuvent apparaître, même en transparence sous des points au rayon plus important et proches des uns des autres.

La seconde visualisation¹⁷ présente également toute la correspondance de Humboldt. Néanmoins, les points représentés ne sont pas proportionnels au nombre de lettres mais possèdent un code couleur. Ce code couleur est produit de manière aléatoire par l'algorithme et définit six grandes périodes dans la vie d'Alexander von Humboldt :

- la période de ses études de 1792-1798
- le voyage en Amérique centrale de 1799 à 1804
- le voyage en Italie en 1805
- une période de sédentarisation parisienne de 1806 à 1828
- le voyage en Russie et Sibérie en 1829
- son retour à Berlin jusqu'à son décès de 1830 à 1859

Une légende rudimentaire informe du code couleur de chacune de ces périodes. Une explication ou du moins des informations supplémentaires pourraient accompagner cette légende afin d'apporter du contexte. Il serait également fort pratique que cette légende soit interactive : l'utilisateur cliquerait sur l'une des périodes permettant de faire disparaître les points des autres périodes de la carte. Cette fonctionnalité n'a pas été implémentée mais pourrait l'être si le développement d'une application web se concrétise.

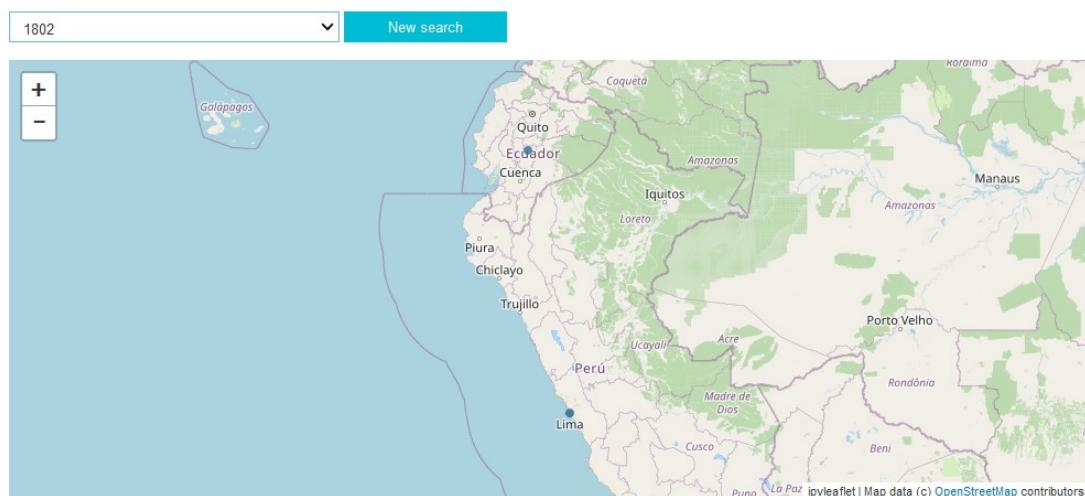


FIGURE 5.7 – Visualisation cartographique des résultats pour les lettres de l'année 1802 avec un zoom optimal.

16. Voir la carte, figure 6.2

17. Cette seconde carte est disponible dans les annexes, voir Figure F.2

Pour ce qui est des visualisations liées aux fonctions de recherche, elles se présentent sensiblement de la même manière que les deux visualisations expliquées précédemment. Toutefois seule une partie des données, celle sélectionnée par l'utilisateur, y est exposée. Un niveau de zoom optimal a été implémenté, c'est-à-dire que l'algorithme calcule le niveau de zoom le plus bas permettant d'afficher tous les points sur la carte. Cela apporte une dynamique dans les cartes et améliore l'expérience utilisateur. Il serait malheureux d'afficher une mapmonde si un seul point est visualisé sur la carte.

Toutes ces cartes sont interactives : elles sont zoomables et cliquables. Cette interactivité permet à l'utilisateur d'avoir accès à des informations supplémentaires sur les données représentées : le nombre de lettres reçues ou envoyées en ce lieu, le nom du lieu, parfois le lien vers le catalogue en ligne quand il s'agit des visualisations liées aux fonctions de recherches. Malheureusement la taille d'affichage des visualisations cartographiques est contrainte par le format des Jupyter Notebook.

Chapitre 6

Accessibilités

Ce projet s'est construit au fil du temps, en adaptant les idées et besoins de l'équipe de *Alexander von Humboldt auf der Reisen*. La conception du projet et des réalisations est loin d'avoir été linéaire. Le but premier était de reconstituer, du moins en partie, la collection analogue de photocopies des manuscrits issus de la correspondance du scientifique conservées au sein des archives Humboldt à la BBAW. Les données concernant la correspondance d'Alexander von Humboldt ont été récupérées sur diverses API conservatrices de lettres¹. Ce projet a permis à ce que l'équipe de chercheurs (re)découvrent ces archives grâce aux nouvelles technologies qui leur sont, pour certains, peu familières. Cela a apporté des réflexions nouvelles sur des perspectives pour l'exploration de ces archives. Ce projet, à vocation originellement expérimentale, pourrait devenir une étape au sein d'une réalisation plus large : la conception d'une application web permettant aux chercheurs et autres utilisateurs de découvrir et d'étudier la correspondance du scientifique.

Bien que l'exploration des données de la correspondance ne soit pas disponible sur une application web entièrement grand public et accessible via un moteur de recherche, il est important que le projet jusqu'alors mené soit disponible sur une plateforme accessible afin que tous puissent avoir accès aux données qu'il contient.

6.1 Accès aux données

La cellule de sortie affiche les différentes visualisations créées ainsi que les résultats des fonctions de recherche de diverses manières. Si une seule lettre est présente dans les résultats de recherche alors la page web du catalogue de l'institution conservatrice s'affiche dans l'output. L'avantage de présenter directement cette page au sein du Jupyter Notebook est d'apporter un accès direct à l'utilisation de toutes les informations concernant la lettre ainsi qu'à sa numérisation quand celle-ci est disponible. Dans la base de données, aucun lien vers la numérisation du document n'est stocké. C'est pour cela que

1. Voir la section 2.2.

FIGURE 6.1 – Accès aux catalogues en ligne.



(a) Affichage de la page du Kalliope-Verbundkatalog dans l'output.
 (b) Fenêtre PopUp d'une lettre envoyée à Paris en 1840.

l'accès à la page web du catalogue est intéressante. Toutefois, cette représentation de la page web n'est plus possible quand le résultat de la recherche contient plusieurs lettres. Dans ce cas, les résultats s'affichent sous forme de DataFrame de la librairie Pandas² permettant à l'utilisateur d'avoir une vue d'ensemble de toutes les informations stockées pour les lettres résultantes.

D'autre part, l'utilisateur a également accès aux informations des lettres grâce à l'interactivité des cartes. Comme évoqué précédemment, ces dernières sont cliquables et la sélection d'un point par l'utilisateur permet d'afficher une fenêtre PopUp. Ainsi, s'y affiche le lieu d'envoi de la lettre et sa date d'envoi, le nom de l'expéditeur ainsi que du destinataire, le lieu de conservation accompagné du lien vers le catalogue en ligne.

L'utilisation des Jupyter Notebook apporte ici des contraintes dans la présentation de ces informations. En effet, les cartes sont affichées en format panoramique, c'est-à-dire qu'elles sont moins hautes que longues. Ce format d'affichage n'est pas optimal afin de présenter toutes les informations pour chacune des villes. Seuls les éléments de trois lettres peuvent être affichés sans contrainte. L'utilisateur est invité à se tourner vers le DataFrame pour avoir un accès sans limite aux informations contenues dans la base de données. Toutefois, les DataFrame ne permettent pas à l'utilisateur d'avoir accès au catalogue en ligne de manière aisée puisque les liens qui s'y affichent n'y sont pas cliquables.

Les données disponibles ne sont pas seulement les données externes récupérées via des API. En ef-

FIGURE 6.2 – Exemple d'affichage : les 1134 lettres de Berlin

2. Voir Figure 5.1.

fet, les données extraites de l'aide à la recherche ont été introduites dans les fonctions de recherche et l'utilisateur peut également y avoir accès. Toutefois, seule une recherche par institutions conservatrices permet d'accéder aux données de la BBAW puisqu'il s'agit de la seule information disponible au sein du document Excel : aucune information sur le contenu des lettres conservées par les institutions n'y est informé. Par conséquent, l'utilisateur sélectionne l'institution et dans la cellule de sortie apparaissent deux types de résultats :

- les données de l'aide à la recherche présentées sous forme de string³. Ici sont présents les tiroirs où se trouvent les lettres correspondantes au sein des archives mais aussi le nombre de lettres inventoriées
- les données des catalogues en ligne.

Les données des catalogues en ligne sont présentées sous deux formes afin de permettre à l'utilisateur de comparer les deux collections. En effet, la collection de la BBAW peut être obsolète et non à jour.



In the analogue collection of the BBAW (K. Nr. 54c, 153, 155, 160, 171):
49 letters by AvH
Today in the online catalogue : 20 results
{'document': 4, 'letter': 28, 'other': 229}

| | identifier | notice_id | type | document_type | contributor | location |
|---|---|-----------|-------------|---|--|---|
| 0 | http://catalogue.bnf.fr/ark:/12148/cb38794301p | 38794301 | monographie | manuscrit moderne ou document d'archive | Bibliothèque nationale de France (Paris) | Richelieu Société de Géographie Tolbiac ... |
| 1 | http://catalogue.bnf.fr/ark:/12148/cb38794294 | 38794294 | monographie | manuscrit moderne ou document | Bibliothèque nationale de France | Richelieu Société de Géographie |

FIGURE 6.3 – Exemple : les résultats pour la BnF

Cet exemple ci-contre présente les résultats pour la Bibliothèque nationale de France. Les premières lignes correspondent aux données de l'aide à la recherche montrant que 49 photocopies de lettres provenant de la BnF sont stockées dans les archives de la BBAW. Ces lignes sont suivies du nombre d'entrées dans notre base de données pour la Bibliothèque nationale de

France puis le détail de ces entrées. En effet, une entrée ne représente pas forcément une lettre mais un dossier, un ensemble de documents ou bien une ou plusieurs lettres. Ainsi, le nombre d'entrée pour la Bibliothèque nationale de France est de vingt mais ces vingt résultats représentent finalement 28 lettres et 233 autres documents. Ici, on comprend bien que la collection de la BBAW n'est pas à jour et ces lignes permettent de comparer l'état actuel des collections.

Toutes ces données sont accessibles gratuitement dès lors que l'utilisateur a installé le projet sur son ordinateur. Le projet est open source et est disponible sur un dépôt GitHub.

3. Une string en Python est une chaîne de caractères.

6.2 Accès au projet

6.2.1 Livrable : un dépôt github

La donnée se doit d'être suffisamment contextualisée pour que l'utilisateur puisse en comprendre le sens. Par exemple, le chiffre « 19 » peut représenter une température, un nombre de clients ou un montant. Dès lors, il faut que des informations complémentaires accompagnent les données pour permettre leur contextualisation. Le format du Jupyter Notebook autorise la rédaction d'une documentation riche et explicite accompagnant les cellules de code. Plus le projet sera fourni en documentation et en contextualisation, plus ce dernier sera facilement pris en main par les utilisateurs et sera compris.

Le dépôt GitHub fait donc office de portail permettant à l'utilisateur d'accéder au projet facilement. Il est accompagné d'un document README, document Markdown, contenant le contexte de développement du projet et quelques lignes de contextualisation historique. Ces informations sont suivies des commandes d'installation pour une installation sur une machine Linux. L'utilisateur doit copier les commandes directement dans son terminal et suivre les instructions afin de lancer le projet et de découvrir la riche correspondance épistolaire d'Alexander von Humboldt directement sur sa propre machine.

Néanmoins, installer le projet sur sa machine reste une opération technique. Cela entend que l'utilisateur sache utiliser un terminal et comprenne le fonctionnement d'un Jupyter Notebook. Une utilisation via le terminal est loin d'être une opération que le grand public sera capable d'effectuer de manière aisée et sans connaissance préalable. Il serait bien plus confortable pour tous si ces fonctions de recherche et ces visualisations seraient disponibles sur une application web, accessible via les moteurs de recherche. À ce stade du projet, l'accessibilité reste par conséquent limitée à des techniciens à l'aise avec les technologies GitHub et Jupyter Notebook.

6.2.2 Communiquer autour du projet

Communiquer autour du projet c'est apporter de la visibilité, de la compréhension et du contexte supplémentaires aux utilisateurs potentiels. Cette action de communication fait également partie de l'accessibilité du projet puisqu'elle le rend beaucoup plus accessible, beaucoup plus visible auprès des chercheurs à défaut du grand public.

Au cours de mon stage, j'ai eu la chance de présenter mon projet lors de deux conférences. Ces deux conférences ont été des étapes importantes mais aussi de réels défis : prendre la parole en allemand devant du public intéressé, recevoir des retours et des avis sur ce que j'avais réalisé, répondre aux questions afin de lever des incompréhensions ou bien afin d'approfondir des points.

La première conférence à laquelle j'ai participé a eu lieu au cours d'un événement du vDHd21. Le vDHd21 est un événement communautaire de humanistes numériques qui

proposent de nombreuses conférences, des publications et des rencontres tout au long de l'année. Le vDHd est axé sur les expérimentations et les formats alternatifs. Il conçoit une forme d'échanges décentralisée et virtuelle. En plus d'apporter de la visibilité au projet, cela permet à ce dernier de s'ancrer dans la communauté scientifique. La seconde conférence fait partie du colloque organisé par la BBAW elle-même. Ce colloque se concentre sur les humanités numériques dans le but d'intensifier le dialogue interdisciplinaire dans le domaine. En tant que principale institution de recherche non universitaire de Berlin, ce colloque inscrit la BBAW dans cette communauté des *Digital Humanities*. L'événement se concentre sur des sujets pratiques et des exemples d'applications mais aussi sur une réflexion critique de la recherche en humanités numériques. Ainsi, tous les premiers vendredis du mois, une conférence est proposée et se déroule depuis mars 2020 exclusivement de manière virtuelle⁴. La participation n'est pas exclusivement réservée aux membres de la BBAW et à ses partenaires, toutes personnes peuvent présenter son sujet et proposer une conférence.

Lors de ces deux participations, j'ai été accompagnée par Dr. Gordon Fischer et Christian Thomas. L'idée était de présenter l'extraction des données de l'édition numérique ainsi que leur visualisation cartographique. L'outil *Humboldt Chronotopographie*⁵ a ainsi été exposé pour la première fois depuis son développement au public. En tant que projet expérimental et dont le développement est particulièrement récent, il est important de lui apporter une certaine visibilité afin qu'il puisse prendre de l'ampleur. De plus, il a pour vocation à être collaboratif. Au cours de la conférence, j'ai eu le plaisir de présenter mon projet et particulièrement les visualisations de données réalisées puisqu'elles s'adaptaient particulièrement au sujet de la conférence. Pour ces deux présentations, environ une trentaine d'auditeurs ont répondu à l'appel.

Cela a été, pour ma part, une occasion de présenter mes réalisations, d'expliquer ce qui a été produit mais aussi de le diffuser. De cette manière, les participants ont eu accès au GitHub où est disponible tout le projet et savent le chemin à emprunter afin d'avoir accès aux données de la correspondance.

4. Voir le programme sur le site de la BBAW.

5. Voir section 4.3.

Troisième partie

**Enrichir des projets numériques de la
BBAW grâce aux données externes**

Chapitre 7

correspSearch : collaborer autour des échanges épistolaires

Le portail web correspSearch est développé depuis avril 2014 par TELOTA de la BBAW et en coopération avec le TEI Correspondence *Special Interest Group* (SIG) ainsi que d'autres chercheurs participants. En tant qu'initiative de la BBAW, ce projet est financé par des tiers. Le modèle de données en réseau de l'*edition humboldt digital*¹ permet de rendre compte du lien entre l'édition et correspSearch : l'édition numérique met à disposition les métadonnées de la correspondance d'Alexander von Humboldt et, en contrepartie, elle appelle les métadonnées des lettres éditées par d'autres éditions afin d'enrichir son propre corpus. Toutes ces données échangées sont structurées en CMIF. Le CMIF ainsi que le projet correspSearch sont les sujets de ce chapitre.

7.1 correspSearch : le projet et ses objectifs

CorrespSearch est un projet initié par la BBAW et dirigé par Stefan Dumont, chercheur et ingénieur d'étude au sein de TELOTA, le pôle des humanités numériques de la BBAW. Le but du projet correspSearch est de mettre à disposition des chercheurs et du grand public un répertoire de diverses éditions de lettres sur une interface web.

Le projet est parti d'un constat : les lettres comptent parmi les sources importantes de la recherche historique. Elles abordent et commentent tous les sujets possibles qui ont intéressé leur rédacteur. Témoins de leur époque, elles permettent également d'étudier les réseaux épistolaires entre diverses personnes. Néanmoins, pour des raisons éditoriales, la correspondance historique n'est éditée que par extraits. En effet, seul l'échange de lettres entre deux personnes est éditée ou bien toute la correspondance d'une seule et même personne. Il est alors difficile d'explorer un large corpus épistolaire sans effectuer des recherches de longue haleine sur plusieurs éditions de lettres. Le service correspSearch a

1. Voir le modèle de données détaillé, Figure A.1

pour but de pallier à ces problèmes en effectuant le premier pas vers une mise à disposition de toutes les lettres éditées sur un même service et en redirigeant le chercheur vers les publications originales.

Les métadonnées des lettres provenant de diverses éditions imprimées et numériques ainsi que d'index numériques sont centralisées et distribuées sur le site ouvertement sous une licence CC-BY 4.0, licence libre. Le service web agrège les métadonnées de correspondance de ces index numériques qui sont hébergés ailleurs sur le web, créées et fournies par divers projets d'éditions savantes.

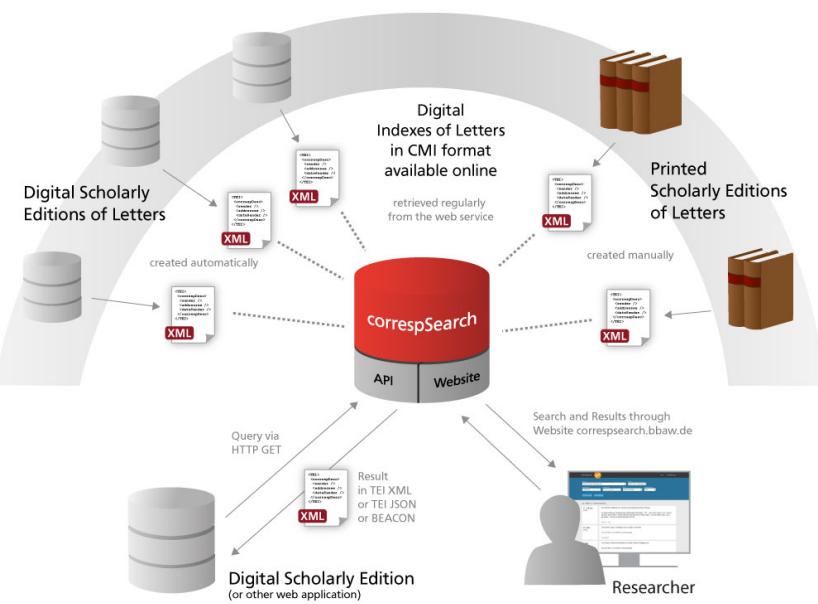


FIGURE 7.1 – Fonctionnement de correspSearch.

Les métadonnées, sous format XML-TEI, fournies par ces index de correspondances sont ensuite récupérées par correspSearch à intervalles réguliers dans l'année permettant une actualisation fréquente. L'utilisateur peut effectuer une recherche dans le répertoire de correspSearch et avoir accès en une seule recherche à toutes ces éditions. Grâce à la recherche avancée, le service web propose d'effectuer des recherches par nom d'expéditeur, nom de destinataire, lieu et date de rédaction de la lettre.

La base de données correspSearch représente une collection constamment enrichie : les éditions savantes de lettres numériques sont de plus en plus nombreuses. Le service web ne fixe pas de limites concernant la période historique ou bien le lieu de production de la correspondance. Toute édition intéressée a la possibilité de déposer ses données en CMIF par le biais du service web. Il existe déjà une variété d'organisations numériques mettant à disposition des données issues de correspondances épistolaires. Le service correspSearch est conçu pour s'intégrer dans ce paysage déjà varié et chercher à le compléter. Pour les ressources purement archivistiques, il existe déjà divers formats et services numériques comme le format EAD et le catalogue du Kalliope-Verbund². C'est pour cela que correspSearch met à disposition non pas des données archivistiques mais des données issues d'édition numérique ou bien imprimée, c'est-à-dire la correspondance ayant fait l'objet d'édition scientifique accompagnée de commentaires critiques.

CorrespSearch est un projet qui sans la collaboration des éditions et des chercheurs

2. Voir sous-section 2.2.1

ne fonctionnerait pas ou moins bien. C'est l'une des raisons pourquoi le site invite à la collaboration³ et propose de nombreux outils afin d'aider à convertir les données de chacun en CMIF. Fournir un fichier CMIF présente de nombreux avantages pour les projets de recherche qui participent à l'enrichissement de correspSearch. En effet, cela leur permet de rendre leurs lettres éditées plus accessibles à la communauté des chercheurs mais également de lier leur édition avec d'autres éditions. D'autre part, ces projets contribuent à faciliter la recherche puisque les correspondances sont centralisées dans un seul et même service.

Le chercheur, qu'il participe au sein de correspSearch ou bien qu'il soit un simple utilisateur du service web, a la possibilité d'accéder aux données via l'API mise à disposition. Cette API fonctionne par des requêtes HTTP GET et renvoie un résultat sous divers formats⁴ dont le CMIF.

7.2 Le format CMI (CMIF)

Afin que correspSearch puisse agréger les métadonnés de lettres provenant de divers sources, les métadonnées doivent être fournies dans un format normalisé et lisible par machine. Le format XML-TEI semble approprié dans ce contexte car il est utilisé dans les éditions numériques depuis plusieurs années déjà et est devenu une norme. De plus, il est possible de s'appuyer sur les travaux du Correspondance SIG qui a développé l'élément `correspDesc` pour les TEI *guidelines*. Cette extension de la TEI fournit un jeu de balises pour enregistrer des métadonnées spécifiques à la correspondance dans le `teiHeader`, telles que l'expéditeur, le destinataire et le lieu d'écriture. Après avoir subi quelques modifications `correspDesc` a été intégré dans les *guidelines* de la TEI en avril 2015⁵.

Le TEI Correspondance SIG a également développé le CMIF et rend possible une standardisation des métadonnées de lettres. L'essence du document CMIF consiste en de multiples éléments `correspDesc`, chacun correspondant à une lettre. Cependant les éléments `correspDesc` sont utilisés de manière très restreinte afin de permettre un traitement automatique par la suite. Par exemple, la balise `correspAction` n'accepte que deux éléments dans son attribut `@type` qui sont `sent` et `received`. Cette balise accepte seulement les balises `persName`, `placeName` et `date` en tant que balises enfants. L'emploi très restreint de balises est nécessaire afin de permettre une interopérabilité sans aucune intervention humaine. Cette restriction impose également l'encodage d'informations limitées aux métadonnées de bases concernant les lettres, à savoir : l'expéditeur, le destinataire, le ou les lieux de réception et les dates. L'édition de laquelle la lettre est issue est encodée dans l'attribut `@source` de la balise `correspDesc` sous forme d'identifiant.

3. Voir l'onglet **participate** sur le site de correspSearch : correspsearch.net/en/participate.html

4. Voir l'onglet API de correspSearch.

5. Stefan Dumont, « correspSearch – Connecting Scholarly Editions of Letters », *Journal of the Text Encoding Initiative* (, Issue 10[2016]), Number : Issue 10 Publisher : Text Encoding Initiative Consortium, DOI : 10.4000/jtei.1742

La spécification CMIF recommande fortement un alignement à des référentiels grâce à l'emploi d'identifiants provenant de notice d'autorité afin de distinguer les personnes et les lieux. Pour chaque entité, il existe une notice avec un identifiant unique, indépendant du projet et permanent. Cet identifiant qui peut être référencé par tous est largement utilisé sur le web comme référence⁶. Dans le contexte de correspSearch, l'utilisation de fichiers d'autorités permet d'éviter l'utilisation de chaînes de caractères encodées par des humains et dont les erreurs, variantes orthographiques et homonymes sont fréquents. Ainsi, l'outil de recherche de l'interface correspSearch effectue une recherche sur les identifiants des notices d'autorité encodés au sein des fichiers CMIF. CorrespSearch acceptent diverses notices d'autorité pour les personnes : l'identifiant GND mis en place par la Bibliothèque nationale allemande, l'identifiant de la BnF, la *New Diet Library* (NDL), notice d'autorité japonaise ainsi que l'identifiant VIAF. Pour les lieux, ce sont les identifiants GeoName qui sont pris en charge. Dans le contexte des technologies du web sémantique, l'utilisation d'identifiants uniques et normalisés pour les entités communes est cruciale afin d'assurer le partage des données de recherche en tant que données ouvertes et liées.

Maintenant que les diverses balises employées au sein d'un fichier CMIF et que les notices d'autorité acceptées ont été présentées, voici ce à quoi ressemble l'encodage d'une lettre en format CMI⁷ :

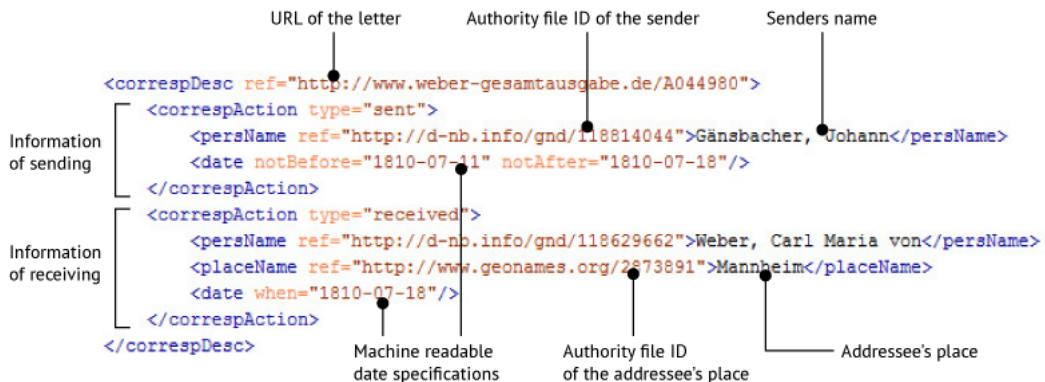


FIGURE 7.2 – Exemple d'une lettre en CMIF.

Depuis peu, la possibilité de faire référence au document archivistique d'une lettre a été implémentée. Cette référence sous forme d'URI est stockée au sein de **correspDesc/@corresp**. C'est exactement dans ce contexte que le *cS matching tool* s'est développé.

6. À ce propos, je renvoie le lecteur vers la Partie I

7. Cet exemple est celui présenté sur le site correspSearch au sein de la documentation pour le format CMI.

Chapitre 8

De la conception au développement

8.1 Les objectifs du *cS matching tool*

Jusqu’alors il était possible de faire référence à l’édition d’une lettre, qu’elle soit imprimée ou numérique, au sein du `correspDesc/@ref` dans un document CMIF. La référence vers le document manuscrit conservé dans un centre d’archives s’insère dans l’attribut `@corresp` de cette même balise. Cette possibilité de référence a été implémentée il y a peu. Les lettres déjà versées dans `correspSearch` avant l’implémentation de cette fonctionnalité ne sont donc pas pourvues de cette référence. L’idée, initiée par Stefan Dumont, chercheur et ingénieur de recherche au sein du pôle TELOTA de la BBAW, est donc de créer un outil permettant d’insérer les liens vers la référence archivistique, c’est-à-dire au sein de `correspDesc/@corresp` pour chacune des lettres correspondantes.

Les données utiles à ce projet sont celles déjà versées sur `correspSearch` ainsi que celles du Kalliope-Verbundkatalog où les références aux lettres de la correspondance von Humboldt sont les plus importantes. Le but est donc de pouvoir faire correspondre une lettre des données du Kalliope-Verbunkatalog à celles de `correspSearch`. Si les données sont les mêmes au sein des deux sets de données alors il s’agit de la même lettre. Par conséquent, le lien vers le Kalliope-Verbundkatalog peut être ajouté à la balise correspondante au sein des données de `correspSearch`. Afin d’identifier une lettre, quatre éléments sont nécessaires :

- l’expéditeur
- le destinataire
- la date d’envoi
- le lieu d’envoi

Néanmoins, ces informations ne sont pas toujours complètes. Beaucoup de lettres ne possèdent pas d’information concernant la date d’envoi ou le lieu d’envoi. C’est particulièrement fréquent quand deux correspondants s’échangent très régulièrement des lettres et vivent dans la même ville. Bien souvent les partenaires de correspondance d’Alexander

von Humboldt sont plus méticuleux dans la rédaction de ces informations. La date ainsi que le lieu d'envoi sont d'une manière générale plus complets. Humboldt, quant à lui, ne prend pas souvent la peine d'indiquer la date exacte sur ses lettres quand il vivait à Berlin et se suffisait d'un simple "ce matin" ou bien "hier". Difficile ensuite d'identifier pleinement cette lettre et par conséquent, il existe deux matches possibles :

- ceux qu'on appelle les *full matches* c'est-à-dire les lettres pour lesquelles les quatre éléments cités ci-dessus correspondent
- les *possible matches* pour lesquels l'expéditeur, le destinataire ainsi que la date d'envoi correspondent mais pas le lieu.

On peut se demander si l'expéditeur, le destinataire ainsi que la date d'envoi ne sont pas suffisants pour considérer un match comme complet. Il est possible qu'Humboldt ait envoyé une lettre le même jour à la même personne. Cela arrive et c'est pour cela que l'utilisateur à la possibilité ensuite de confirmer ou non ces matches considérés comme possibles. Aussi, les données issues de correspSearch sont bien souvent plus complètes que celles du Kalliope-Verbundkatalog. En effet, les données de correspSearch proviennent des éditions imprimées et/ou numériques des correspondances, c'est-à-dire que des chercheurs ont ajouté des commentaires critiques et ont ajouté des informations à propos de ces lettres. Ils ont identifié, pour certaines lettres, des lieux d'envoi qui n'apparaissent pas explicitement dans les documents d'archives. C'est pour cela qu'il est important de laisser l'utilisateur décider s'il s'agit d'un match ou non. Ces matches sélectionnés par l'utilisateur sont ensuite ajoutés au fichier CMIF et seront téléchargeables.

Par conséquent, l'outil doit intégrer diverses fonctionnalités :

- prendre en charge des fichiers CMIF
- présenter les possibles matches afin que l'utilisateur puisse les sélectionner ou non
- ajouter l'URI correspondante au sein de `correspDesc/@corresp`
- les matches, une fois intégrés au document correspondant, doivent être téléchargeables en CMIF directement sur l'ordinateur de l'utilisateur

Il serait également utile que l'utilisateur puisse décider s'il souhaite télécharger le fichier contenant toutes les lettres de correspSearch avec les URI ajoutées aux balises correspondantes ou bien s'il télécharge un fichier CMIF contenant uniquement les matches et au sein duquel les lettres qui n'ont trouvé aucune corrélation auront été supprimées.

Cet outil est destiné aux techniciens de correspSearch afin d'enrichir leurs données avec l'ajout d'une référence supplémentaire. La cible n'est pas le grand public. Toutefois, l'expérience utilisateur doit être suffisamment intuitive afin que le technicien ou bien la personne qui s'occupera d'effectuer cette tâche puisse la réaliser de manière aisée sans avoir à lire une documentation complète sur la manière dont fonctionne cette application. Étant donné que de nombreuses personnes effectuant leur stage au sein du projet *Alexander von Humboldt auf der Reise* sont plutôt familières à la recherche qu'à l'ingénierie et qu'il s'agit d'une tâche qui pourrait être réalisée par ces dernières, il est important qu'une

interface utilisateur soit développée. C'est pour cela qu'il a été décidé de créer une application web qui pourrait être installée localement sur l'ordinateur prêté aux stagiaires. Le développement de l'outil est donc mixte : un développement *back end* contenant toute la logique de l'application et un développement *front end* pour l'interface utilisateur.

8.2 Développements back end et front end

On dissocie le développement *back end* du développement *front end*. Le *back end* fait référence à la partie du code qui est invisible de l'utilisateur contrairement au *front end* qui représente la partie visible, l'interface utilisateur c'est-à-dire ce qui est affiché dans le navigateur de l'utilisateur. Au sein de ce projet, la partie *back end* est développée en Python tandis que le *front end* en HTML et CSS. L'application en elle-même a été développée avec le *framework* Flask.

8.2.1 Back end : algorithmie

Le choix du langage de programmation et du framework

Puisque l'application a vocation à être développée de manière *standalone*, j'ai été particulièrement libre de choisir le langage dans lequel je souhaitais la programmer. Le choix du langage de programmation s'est arrêté sur Python pour plusieurs raisons. La première était le fait de rester dans une certaine continuité avec la mission que j'avais effectuée précédemment à savoir le développement de fonctionnalités de recherche et de visualisations de données. Ces dernières, développées dans des Jupyter Notebooks, ont été entièrement rédigées en Python¹. Par ailleurs, il y a la question de la maintenance de l'application. Afin que cette application puisse être utilisée de manière durable, elle doit être maintenue. La personne qui sera désignée pour cette tâche doit connaître le langage dans lequel l'application est programmée. Python est un langage largement connu. Il est également apprécié par de nombreux pédagogues qui considèrent Python comme un langage où la syntaxe permet une initiation aisée aux concepts de base de la programmation².

Un *framework*, aussi appelé environnement de développement en français, désigne un ensemble de composants logiciels structurels. Il sert à créer les grandes lignes ou les fondations d'un logiciel. Il guide ainsi l'architecture logicielle et conduit parfois le développeur à respecter des patrons de conception de part les bibliothèques dont il est constitué. Toutes ces technologies m'étaient connues avant le début de la mission et cela me permettait de me concentrer sur la logique de l'application et non l'apprentissage d'un nouveau langage ou d'un nouveau *framework*. Flask est un *framework open-source* pour le langage Python.

1. Voir chapitre 5.

2. Jeff Elkner, *How to Think Like a Computer Scientist*, URL : <https://www.greenteapress.com/thinkpython/thinkCSPy/html/preface.html> (visité le 10/08/2021), préface.

Ce *framework*, déjà connu puisqu'il est enseigné à l'École nationale des chartes, permet le développement d'application web. Sa documentation est particulièrement détaillée et contient de nombreux morceaux de code facilement adaptables à nos projets³. À ce sujet, Gilles Babinet⁴ informe qu'on dit souvent que les programmeurs actuels ne programment plus mais que leur métier consiste à coller des morceaux de programmes les uns avec les autres. Les développements sont sophistiqués et reposent sur des ensembles de codes écrits par d'autres. Il faut toutefois noter que cette dynamique s'insère dans la communauté de l'*open source*. Les programmeurs utilisent des portions de code disponibles en *open source* qu'ils retournent à la communauté en mettant à disposition leur code sur des plateformes comme GitHub. Les licences *open source* ont le point commun de mettre l'accent sur la liberté de l'individu d'aller modifier le code source et de se l'approprier pour ensuite, pouvoir créer de la valeur ajoutée qui retourne à l'ensemble de la communauté.

La structure de l'application

La structure du projet est relativement simple : l'application est contenue dans le dossier `app/`. En son sein, se trouve un dossier prêt à recevoir les données puis deux autres dossiers destinés au *front end* (`static/` pour les styles et `templates/` pour les fichiers HTML). Les fonctions logiques sont stockées dans deux fichiers `.py` différents : les fonctions génériques contenues dans `generic.py` puis la fonction qui permet de corrélérer les données est stockée dans `matching.py`. Les routes sont les fonctions qui font le lien entre le *front end* et le *back end*, c'est-à-dire qu'elles vont chercher dans le *back end* les données afin de les mettre à disposition de l'utilisateur dans le *front end*. Ces fonctions sont enregistrées dans le fichier `routes.py`. Toute la configuration de l'application est répartie dans `app.py` et `constantes.py`.

Quand l'utilisateur installe l'application sur son serveur local, aucune donnée n'est encore stockée en son sein. Seul un dossier est dispo-

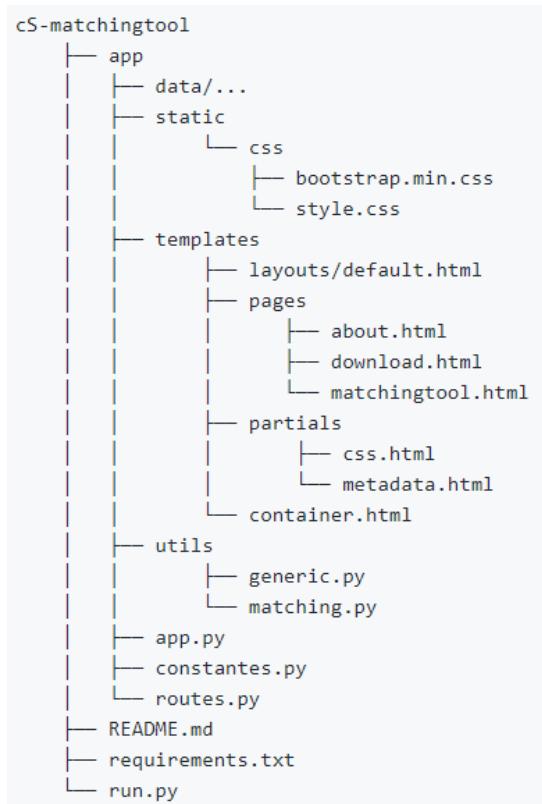


FIGURE 8.1 – La structure de l'application

3. La documentation est disponible sur flask.palletsprojects.com

4. Gilles Babinet, « 12. Open source », *Hors collection* (, 2020), Bibliographie_available : 0 Cairn-domain : www.cairn.info Cite Par_available : 0 ISBN : 9782100820764 Publisher : Dunod, p. 107-114, URL : <https://www-cairn-info.proxy.chartes.psl.eu/refondre-les-politiques-publiques--9782100820764-page-107.htm> (visité le 09/08/2021)

nible et est prêt à recevoir les données qui devront être comparées. Puisque ces données vont se limiter à deux fichiers à savoir : un fichier pour les données du Kalliope-Verbundkatalog et puis un second pour celles issues de correspSearch, développer toute une base de données avec SQLAlchemy n'a pas semblé nécessaire.

La librairie `xml.etree` pour prendre en charge des fichiers XML

La librairie `xml.etree`⁵ permet de se déplacer dans l'arbre de fichiers XML en le parsant. En effet, XML est un format hiérarchique et la meilleure façon de se représenter ce type de fichier est sous la forme d'un arbre. Cette bibliothèque dispose de deux classes qui sont *ElementTree* qui représente l'ensemble du document XML sous forme d'un arbre et *Element* qui est la représentation d'un seul noeud au sein de cet arbre. Ainsi, dans notre algorithme, le `tree` représente toujours notre document. Toutefois, notre fonction compare deux fichiers XML différents. Par conséquent, deux arbres sont présents dans notre algorithme, un pour chaque document.

Afin de récupérer les données voulues au sein de chacun des documents XML, certaines fonctions de la librairie `xml.etree` sont intéressantes :

- `Element.findall()` trouve seulement les éléments enfants de l'élément actuel. Il ne faut pas la confondre avec la fonction `Element.find()` qui trouve uniquement le premier élément enfant de la balise indiquée en tant qu'argument de cette fonction.
- `Element.text` permet d'accéder au texte contenu dans la balise.
- `Element.attrib` permet de récupérer les attributs d'un élément.
- `Element.iter()` permet d'itérer de manières récursive sur les balises qui lui sont inférieures c'est-à-dire les éléments enfants mais également les éléments de ses enfants.

Logique algorithmique

Afin que l'algorithme se lance, la première étape est le dépôt d'un fichier CMIF par l'utilisateur contenant des lettres de la correspondance d'Alexander von Humboldt. Tout comme les données du Kalliope-Verbundkatalog qui seront requêtées sur l'API correspondante, le document de l'utilisateur est enregistré dans le dossier `data/` prévu à cet effet. Les données sont stockées dans leur format d'origine à savoir en XML-TEI et plus particulièrement en CMIF pour celles de l'utilisateur. Une condition a été implémentée afin de vérifier que le document déposé par l'utilisateur est un document `.xml`, si cela n'est pas le cas un message d'erreur est alors envoyé à l'utilisateur. Une fois que le fichier valide est déposé alors la fonction `matching()`⁶ permettant de corrélérer les données est appelée. À chaque nouvelle utilisation, l'application envoie deux requêtes à l'API du

5. Voir la documentation.

6. Voir la fonction disponible sur le GitHub du projet.

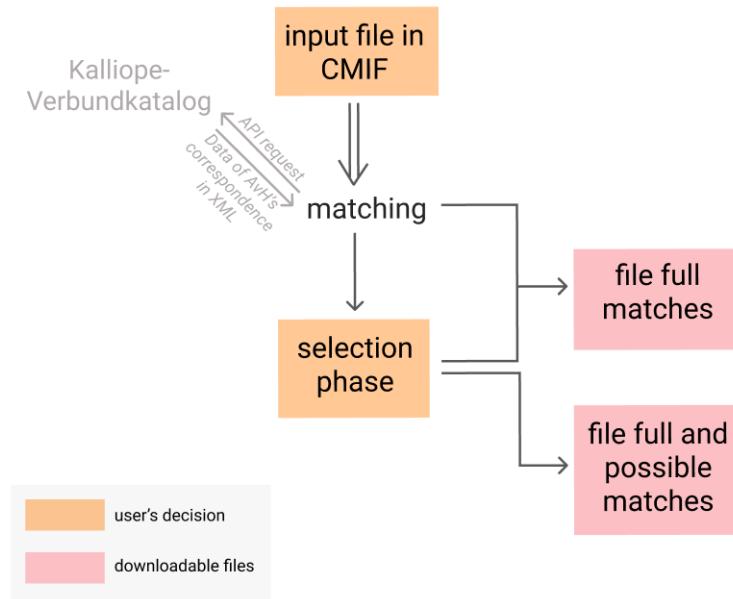


FIGURE 8.2 – Logique algorithmique de l’application schématisé

Kalliope-Verbundkatalog afin de récupérer les lettres reçues ainsi que les lettres envoyées par Alexander von Humboldt. Ces deux requêtes permettent de couvrir l’ensemble de la correspondance de Humboldt. Le format demandé lors de ces deux requêtes est MODS afin d’obtenir davantage d’informations sur les lettres comme notamment les URI vers les fichiers d’autorité des expéditeurs et destinataires de lettres⁷.

La fonction `matching()` se déplace à travers l’arbre du document contenant les données du Kalliope-Verbundkatalog et stocke dans diverses variables les éléments que nous souhaitons faire correspondre avec les données du document CMIF de l’utilisateur. Pour chacune des lettres du Kalliope-Verbundkatalog, l’identifiant de la lettre qui est l’URI vers le site du Kalliope-Verbundkatalog, le lieu d’envoi ainsi que l’URI de la notice d’autorité de l’expéditeur et du destinataire sont stockés dans des variables adaptées. Ces données stockées, la fonction passe ensuite au document contenant les données de l’utilisateur. Chaque élément que nous souhaitons corrélérer sont également stockés dans des variables. Les éléments issus du Kalliope-Verbundkatalog et ceux issus du fichier de l’utilisateur sont ensuite comparés c’est-à-dire l’expéditeur de la lettre du Kalliope-Verbundkatalog avec celui de `correspSearch`, le destinataire, le lieu d’envoi et la date de rédaction. Quand il s’agit d’un *full match* alors l’URI du Kalliope-Verbundkatalog est ajouté au sein du fichier CMIF dans l’attribut `@corresp` de la balise `correspDesc` de la lettre correspondante. Les *possible matches* sont, quant à eux, stockés dans une liste. La fonction renvoie cette liste et leur quantité. Ces matches seront mis à disposition de l’utilisateur qui pourra alors sélectionner ceux qu’il considère comme des lettres correspondantes. Les URI vers le

7. Voir l’annexe à ce sujet, présentant deux encodages de lettre, l’un en format MODS et le second en DC, l’annexe B.

Kalliope-Verbundkatalog des lettres sélectionnées par l'utilisateur sont par la suite ajoutées au fichier CMIF.

Lors de ce processus, l'utilisateur a la possibilité de télécharger à deux étapes différentes le fichier contenant les matches. En effet, il peut télécharger les *full matches* avant que les *possible matches* soient insérés au document. Une fois que l'étape de sélection des *possible matches* à intégrer au document est terminée, l'utilisateur peut télécharger le document complet c'est-à-dire contenant les *full matches* qui ont été ajoutés de manière automatique et ceux sélectionnés par l'utilisateur.

8.2.2 Front end : interface utilisateur

Les langages de programmation et le framework

Puisqu'on ne connaît pas le niveau technique de la personne qui utilisera cette application, il était important de développer une interface utilisateur. En programmation, cette interface, affichée dans le navigateur web de l'utilisateur est appelée *front end* et est programmée, dans notre projet, en HTML et CSS. Le HTML est un langage de balisage conçu pour les pages web. Ce langage peut maintenir du contenu, créer des formulaires de saisie, inclure toutes sortes de ressources multimédias. Il permet également d'écrire de l'hypertexte et de structurer de manière sémantique la page web. Il est aujourd'hui très utilisé conjointement au CSS. Le CSS est ce qu'on appelle également des feuilles de style en cascade. Il s'agit d'un langage informatique qui décrit la présentation des documents structurés en HTML ou XML. Dans le projet cS matching tool, ces documents sont stockés dans divers dossiers :

- `static/css/` qui contient tous les styles CSS pour les éléments HTML
- `templates/` contenant les fichiers HTML des modèles de pages

Le *framework* Bootstrap a également été largement utilisé afin de programmer le *front end* de l'application web. Bootstrap, gratuit et *open source*, est une collection d'outils pour développer le design de sites et d'applications web. Il contient des codes HTML et CSS, des formulaires, boutons, outils de navigations et éléments interactifs. Il est également livré avec plusieurs composants JavaScript sous la forme de plugins jQuery permettant de fournir des éléments supplémentaires comme des infobulles ou encore des carrousels. Chaque composant Bootstrap se compose d'une structure HTML, de déclarations CSS et, dans certains cas, de code JavaScript associé. Une fois utilisé dans un projet, Bootstrap fournit des définitions de style de base pour tous les éléments HTML et le développeur peut modifier ces styles ou en ajouter comme bon lui semble.

Le design de cS matching tool

Quand on explore les différents sites des projets initiés par la BBAW, on remarque plusieurs ressemblances. Le menu est toujours présenté sous la forme d'une barre latérale

en haut du site internet et contient le nom du projet en gras positionné à gauche suivit des sous-éléments du menu. La présentation des informations s'affiche sous la forme de trois colonnes, elles-mêmes constituées de cases pour chaque nouvelle information⁸. Les couleurs utilisées sont régulièrement les mêmes.

Ainsi, au sein du cS matching tool, les couleurs choisies sont semblables à celles utilisées pour le site de l'édition numérique des carnets de voyages d'Alexander von Humboldt. En effet, la couleur du fond du cS matching tool est le gris qui est régulièrement employé⁹ sur les sites de la BBAW tout comme l'ocre, particulièrement choisi pour la couleur des boutons ou des éléments sélectionnés dans la barre du menu. La structure de la BBAW d'une présentation divisée en trois colonnes a été mise en place pour la page **about** qui apporte des informations supplémentaires autour du projet à savoir : autour du développement, à propos de la logique algorithmique de l'application et au sujet du format CMIF.

Pour ce qui est de la présentation des fonctionnalités, celle-ci a pour vocation d'être la plus simple possible afin de faciliter l'utilisation de l'outil. L'utilisateur voit apparaître deux colonnes : celle de gauche permet de déposer le fichier CMIF qui a pour but d'être enrichi des liens vers le Kalliope-Verbundkatalog, celle de droite est dans l'attente de présenter les matches possibles. Une fois que le fichier est déposé et que l'utilisateur a cliqué sur le bouton **submit**, un indicateur de chargement apparaît afin d'indiquer le lancement de l'outil et de ne pas laisser l'utilisateur dans l'attente sans nouvelle information. Sans cet indicateur, il est difficile de savoir si l'outil est lancé ou si l'application rencontre un bug puisque l'application a besoin d'un certain temps avant d'afficher l'étape suivante : la sélection de matches par l'utilisateur.

Les matches à sélectionner se présentent sous forme de tableau. Pour chacun des

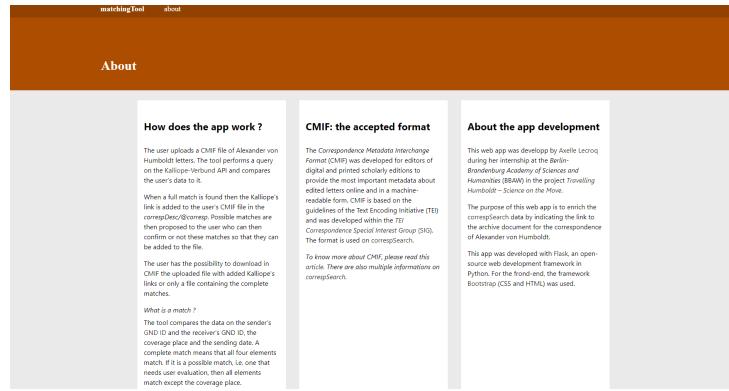


FIGURE 8.3 – Page **about** du cS matching tool

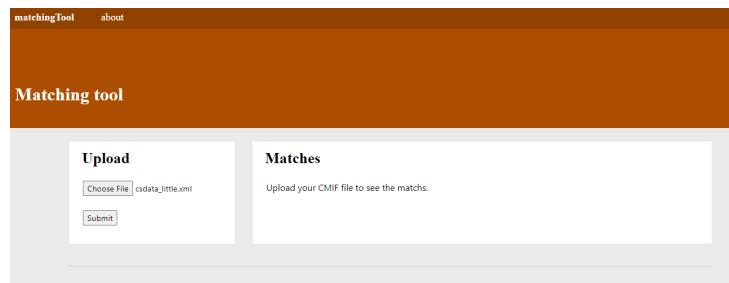


FIGURE 8.4 – Interface permettant à l'utilisateur de glisser son document CMIF

8. Voir l'annexe G.

9. Voir l'annexe G, (a), (b)

Matches [Download](#)

3 full matches were added. Thus, 26 letters have the @corresp attribute.
A full match means that sender, addressee, coverage place and date matched.

Possibles matches

The match is possible when sender, addresse and date matched. Here, only the coverage place doesn't match.
You can select the match if you consider it to be a match although the data doesn't match 100%. The selected matches will be added to the CMIF file once they are submitted.

| Element | Kalliope data | cS data |
|---|---|---|
| 0. <input type="checkbox"/> Sender | http://d-nb.info/gnd/118554700 | http://d-nb.info/gnd/118554700 |
| Addressee | http://d-nb.info/gnd/104260106 | http://d-nb.info/gnd/104260106 |
| Date | 1791-06-14 | 1791-06-14 |
| Coverage place [o.O.] | | Freiberg |
| Link | on Kalliope | |
| 1. <input type="checkbox"/> Sender | http://d-nb.info/gnd/118554700 | http://d-nb.info/gnd/118554700 |
| Addressee | http://d-nb.info/gnd/104260106 | http://d-nb.info/gnd/104260106 |
| Date | 1792-07-06 | 1792-07-06 |
| Coverage place Jena | | Jena |
| Link | on Kalliope | |
| 2. <input type="checkbox"/> Sender | http://d-nb.info/gnd/118554700 | http://d-nb.info/gnd/118554700 |
| Addressee | http://d-nb.info/gnd/104197102 | http://d-nb.info/gnd/104197102 |
| Date | 1789-08-05 | 1789-08-05 |
| Coverage place Ohne Ort | | Göttingen |
| Link | on Kalliope | |

[Create the CMIF file](#)

FIGURE 8.5 – Page *possibles matches* du cS matching tool

matches, les données du Kalliope-Verbundkatalog apparaissent sur la gauche du tableau et sont mises en comparaison avec les données du fichier de l'utilisateur, positionnées à droite du tableau. Avec cet affichage, l'utilisateur peut comparer l'URI vers la notice GND de l'expéditeur des deux fichiers ainsi que l'URI vers la notice GND du destinataire, le lieu d'envoi et la date d'envoi. Les URI du destinataire et de l'expéditeur sont cliquables. L'élément en rouge indique ce sur quoi le match ne s'est pas effectué. Le choix de cette couleur a pour but d'attirer l'attention de l'utilisateur. Un lien vers le document d'archive est également disponible c'est-à-dire vers le site du Kalliope-Verbundkatalog. L'utilisateur a ainsi directement accès au manuscrit et éventuellement à sa numérisation quand celle-ci est disponible. Il s'agit d'un élément supplémentaire afin d'aider l'utilisateur dans la corrélation et la reconnaissance d'une lettre. Il peut ensuite sélectionner les lettres qui ont trouvé leur bonne corrélation grâce à une petite case disponible à côté du numéro de ligne du tableau. Une fois tous les matches passés en revue, une soumission des sélections est requise afin d'ajouter les liens du Kalliope-Verbundkatalog des lettres correspondances au sein du fichier CMIF.

À cette étape, l'utilisateur a également la possibilité de télécharger le fichier CMIF dans lequel les *full matches* seulement ont déjà été ajoutés automatiquement par l'algorithme. Afin d'éclairer l'utilisateur, quelques lignes informatives sont jointes à cette étape et indiquent le nombre de lettres pour lesquelles un document d'archive était déjà rensei-

gné (le nombre de lettre ayant un attribut @corresp) ainsi que le nombre de corrélations qui ont été ajoutées de manière automatique. Une fois les matches sélectionnés par l'utilisateur ont été soumis, ils sont ajoutés au fichier CMIF et l'utilisateur a la possibilité de télécharger deux fichiers différents :

- un fichier CMIF contenant seulement les matches qu'ils soient complets ou bien qu'ils aient été ajoutés par l'utilisateur.
- le fichier CMIF déposé par l'utilisateur avec le lien vers le Kalliope-Verbundkatalog pour les lettres qui ont trouvé une corrélation. Ce document contient donc les lettres qui n'ont trouvé aucune correspondance et celles qui en ont trouvé une.

Ces fichiers sont téléchargeables et s'enregistrent directement sur l'ordinateur de l'utilisateur.

Chapitre 9

Apports et livrables

Ce chapitre a pour but de mettre en perspective ce projet, de pointer du doigt ce qui mériterait d'être amélioré mais aussi d'exposer les perspectives de cet outil. Il est également intéressant de savoir ce que l'outil a apporté jusqu'alors et de quelle manière il a pu enrichir les données de correspSearch. Son accessibilité sera également abordée ainsi que les utilisateurs cibles.

9.1 Résultats et perspectives de l'outil développé

9.1.1 Enrichissement des données de correspSearch : les résultats

Jusqu'alors, aucune lettre issue de la correspondance d'Alexander von Humboldt et stockée sur correspSearch ne possédait de lien vers son document d'archive. Grâce au cS matching tool, ces données ont pu être enrichies en ajoutant le lien correspondant vers le portail archivistique du Kalliope-Verbundkatalog.

Au sein de correspSearch sont stockées 6193 lettres issues de la correspondance d'Alexander von Humboldt. Grâce à plusieurs ajustements dans le code, le nombre de matches complets a pu augmenter de 199 au premier lancement à 440 au dernier lancement de l'outil. Le nombre de matches possibles, c'est-à-dire un match pour lequel le lieu d'envoi n'est pas correspondant, a augmenté de 601 à 714. Suite au regard d'un utilisateur et après sa sélection, le nombre de lettres enregistrées sur correspSearch et pour lesquelles il a été possible d'ajouter le lien vers leur document d'archive s'élève au nombre de 705. Cela représente 11,3% de la totalité des lettres enregistrées sur correspSearch issues de la correspondance d'Alexander von Humboldt. Est-ce beaucoup ou au contraire très peu ? Il est difficile de le dire et aucune estimation ou d'objectif à atteindre n'ont été évoqués avant le développement de cet outil. Il faut également préciser le fait que beaucoup de lettres n'ont pas des données complètes. Comme cela a été mentionné dans la définition

des objectifs de cet outil¹, beaucoup de lettres n'ont aucune indication concernant la date ou de lieu d'envoi. Quand un élément permettant corrélation entre les deux sets de données manque, il devient difficile de faire matcher deux lettres ensemble. Toutefois, l'efficacité de l'outil n'est pas seulement lié aux données et certaines parties de l'algorithme mériteraient d'être améliorées.

9.1.2 Améliorations et perspectives

Améliorations

Certaines lettres ont été rédigées à plusieurs mains, cela signifie que plusieurs expéditeurs sont enregistrés dans les données. Toutefois, l'outil ne prend pas encore en compte les expéditeurs enregistrés sous forme de liste. L'algorithme fait donc correspondre toujours l'expéditeur enregistré en première position du document CMIF et des données issues du Kalliope-Verbundkatalog. Si la liste des expéditeurs n'est pas rangée de la même manière ou tout du moins le premier expéditeur enregistré n'est pas le même, l'algorithme ne détecte aucune corrélation possible. La fonction `matching()`, après améliorations, pourrait prendre en charge les listes d'expéditeurs et destinataires. Cela permettrait d'accroître le nombre de corrélations de lettres et d'enrichir par conséquent les données de correspSearch.

D'autre part, l'outil compare les identifiants GND du destinataire et de l'expéditeur des lettres. Faire correspondre les lettres sur des identifiants de notice d'autorité permet d'assurer plus facilement un match. Il serait également possible de faire correspondre les données sur la chaîne de caractère des noms et prénoms des destinataires et expéditeurs cependant nous ne sommes pas à l'abri des erreurs de typage ainsi que des diverses orthographes utilisées pour un seul et même nom. Faire correspondre les données sur l'identifiant GND s'accompagne également de désavantages. Certaines personnes ne possèdent aucun identifiant pérenne et ce pour plusieurs raisons : soit il s'agit d'un manque d'informations pour les identifier soit ils ne sont tout simplement pas enregistrés sur une base de bibliothèque et ne possèdent pas (encore) de notice d'autorité. Les lettres pour lesquelles le destinataire et/ou l'expéditeur n'ont pas d'identifiant ne pourront trouver corrélation. Sur le site internet de correspSearch², il est indiqué que certaines personnes ne possède parfois pas d'identifiant GND enregistré et cela est complètement normal. La création d'une notice d'autorité prend du temps. Les bibliothèques n'ont pas beaucoup de temps à investir dans la création de nouvelles notices et une demande peut recevoir réponse plusieurs mois plus tard. Ainsi, si nous relançons l'outil dans quelques mois, certaines lettres qui n'ont pas trouvé corrélation actuellement trouveront peut-être à ce moment-ci.

1. Voir section 8.1

2. Voir la rubrique ueber sur le site de correspSearch.

Perspectives

CorrespSearch mentionne sur son site que les données issues des éditions numériques sont régulièrement actualisées de manière automatisée. Il serait judicieux d'insérer l'algorithme du cS matching tool lors de l'actualisation des données des éditions numériques vers correspSearch. Ainsi, les données de correspSearch seraient directement liées à leur manuscrit via le lien du Kalliope-Verbundkatalog lors de leur actualisation dans le portail des documents épistolaires sans avoir besoin de lancer l'outil développé. Cela rendrait la chaîne de mise à jour des données plus efficace car une étape, celle de l'enrichissement des données vers leur source archivistique, disparaîtrait. Les données seront ainsi mises à jour et par la même occasion enrichies. Il serait toutefois important de relancer de temps en temps le cS matching tool car les données du Kalliope-Verbundkatalog s'actualisent régulièrement et indépendamment de celles des éditions numériques.

D'autre part, le cS matching tool prend en charge pour l'instant seulement la correspondance d'Alexander von Humboldt. En effet, l'utilisateur n'a pas la main sur les requêtes envoyées à l'API du Kalliope-Verbundkatalog. Qu'importe si l'utilisateur dépose un fichier contenant des lettres de la correspondance d'une autre personne, la fonction enverra toujours deux requêtes pour récupérer les données de la correspondance de Humboldt. La fonction n'effectue d'ailleurs aucune vérification dans le document déposé par l'utilisateur afin de savoir si le fichier contient exclusivement des lettres issues de la correspondance d'Alexander von Humboldt. L'outil a été réalisé seulement pour répondre à un besoin très précis : l'enrichissement des données de correspSearch de la correspondance d'Alexander von Humboldt et non toutes les correspondances stockées sur le portail. Cela est dû au fait que mon stage s'est réalisé au sein du projet académique *Alexander von Humboldt auf Reisen* et non dans le pôle des humanités numériques, TELOTA, qui touche à tous les projets scientifiques de la BBAW. Toutefois et à l'avenir, il serait intéressant d'enrichir les capacités de l'outil et de développer l'enrichissement des données des échanges épistolaires d'autres personnages présents sur correspSearch. Pour cela, l'utilisateur pourrait par exemple indiquer l'identifiant GND de la personne pour laquelle il souhaiterait enrichir les données. Cet identifiant serait injecté au sein des requêtes envoyées à l'API du Kalliope-Verbundkatalog permettant ainsi de récupérer les données correspondantes.

9.2 Cible et accessibilité du projet

Le cS matching tool a été développé avec l'idée d'enrichir les données de correspSearch et par conséquent d'être utilisé par un technicien de la BBAW ou bien par un ou une stagiaire. Dans tous les cas, il n'a pas à vocation d'être grand public. Toutefois, l'expérience utilisateur est suffisamment simple afin que la personne qui doit réaliser cette tâche puisse la faire de manière aisée.

À présent et avec quelques améliorations du code et des fonctionnalités énoncées précédemment, l'outil pourrait être utile à tous les projets d'éditions numériques de documents épistolaire dont les données sont stockées en CMIF. Ces projets d'éditions numériques pourraient utiliser l'outil afin d'enrichir leurs propres données en ajoutant un lien vers le Kalliope-Verbundkatalog. Bien entendu, l'enrichissement a lieu seulement si l'outil trouve une corrélation et/ou que l'utilisateur en a sélectionné.

En attendant que ces fonctionnalités soient implémentées, le projet est entièrement disponible en *open source*. Il est hébergé sur la plateforme GitHub³ tout comme le projet qui a été mené précédemment à savoir l'exploration de la correspondance d'Alexander von Humboldt par des fonctions de recherche et des visualisations de données⁴. *Alexander von Humboldt auf Reisen* n'avait encore aucune plateforme permettant d'héberger les divers projets numériques qui étaient développés en son sein. L'outil chronotopographique⁵ est quant à lui non disponible en *open source* pour l'instant même si il a vocation de l'être dans le futur. Pour cela, l'équipe m'a demandé de créer une organisation GitHub qui héberge tous ces projets innovants en les centralisant en un seul et même endroit : edition-humboldt-collection. Cette organisation contient pour l'instant deux dépôts contenant les projets menés au cours de mon stage mais d'autres viendront s'ajouter à l'avenir.

3. Voir le projet sur le dépôt GitHub.

4. Voir la partie chapitre 6 sur les diverses accessibilités du projet.

5. Voir section 4.3.

Conclusion

Après un séjour à Paris, Alexander von Humboldt retourne le 6 décembre 1827 à Berlin et débute une série de lectures publiques sur la description physique de la Terre et de l'Univers. À la fin du mois de mars 1828, Humboldt a dispensé seize lectures⁶. En proposant ses lectures de manière publique, Humboldt se positionne dans une éducation ouverte à tous et non plus exclusive à élite savante. Il indique, par cette prise de position novatrice, les découvertes des sciences naturelles émergentes comme appartenant à tous. Lors de ces conférences particulièrement appréciées du public, un nombre inhabituel de femmes y assistent. Depuis 1800, il revendique le fait de rendre accessible ses connaissances et ses publications au grand public. Humboldt apparaît comme un précurseur de la démocratisation et la vulgarisation du savoir. Ces conférences publiques ont débouché sur la rédaction de son oeuvre majeure, *Cosmos*, publiée après 1845.

Le projet *Alexander von Humboldt auf Reisen* se situe dans cette continuité, il contribue à rendre accessibles l'oeuvre et l'héritage manuscrit du scientifique. Afin d'atteindre cet objectif, les données produites par l'édition sont publiées dans des formats standards interopérables et accessibles sur le site internet qui lui est dédié⁷. L'importance de l'*open data* et du *linked data* ont été soulignés au cours de ce mémoire, mais il est nécessaire de rappeler que ces deux concepts sont primordiaux pour des projets de recherche comme celui-ci. En effet, en plus d'utiliser des données externes et d'enrichir de cette manière son corpus, publier des données ouvertes et liées insère l'édition au sein d'un large réseau et lui apporte une plus grande visibilité. La question de l'accessibilité aux données a été un thème transversal de ce mémoire et est parfaitement lié à l'*open data*. Le souhait de Humboldt de rendre accessible à tous son héritage scientifique et manuscrit trouve écho dans le fait que les deux projets développés au cours de mon stage soient disponibles ouvertement sur la plateforme GitHub : edition-humboldt-collection.

Ce mémoire s'est attaché à mettre en perspective le travail réalisé au cours du stage de quatre mois que j'ai effectué au sein de l'équipe de ce projet. Les missions qui m'ont été confiées consistaient à l'élaboration de fonctions de recherche et de visualisations de la correspondance d'Alexander von Humboldt d'une part et au développement d'un outil de

6. Andreas W. Daum, « Resonance space - The cosmos Lectures », dans *Wilhem und Alexander von Humboldt, Berlin Cosmos*, dir. Paul Spies, Ute Tinteman et Jan Mende, Wienand, 2020, p. 149-151

7. Voir le site edition humboldt digital (ehd).

corrélation dans le but d'enrichir les données de la correspondance du scientifique issues de correspSearch d'autre part. Mes réalisations se sont attachées à fournir des solutions à chaque étape de la conception de ces deux missions qui ont évolué au fur et à mesure en s'adaptant aux attentes de l'équipe de recherche.

Maintenant que ces outils sont opérationnels, l'équipe de correspSearch disposent à présent d'un outil de corrélation permettant d'enrichir les données de la correspondance d'Alexander von Humboldt éditée et répertoriée dans sa base avec des données externes issues du Kalliope-Verbundkatalog. Ce projet, répondant à un besoin particulièrement précis, pourrait être élargi de façon à enrichir les données de toutes les correspondances de correspSearch. Plus largement et de manière plus ambitieuse, cet outil pourrait évoluer dans le but de proposer un enrichissement des données pour toutes les éditions numériques ayant des données encodées en CMIF. Les membres du projet *Alexander von Humboldt auf Reisen* disposent, quant à eux, d'outils opérationnels répondant à des attentes définies en amont. L'exploration et la découverte de la correspondance d'Alexander von Humboldt à travers des fonctions de recherche bénéficieront dans un premier temps à l'équipe du projet académique et aux personnes extérieures ayant suffisamment de compétences techniques pour installer de manière autonome ces fonctions sur leur propre ordinateur. Ce projet a permis de prendre connaissance de l'ampleur de la correspondance de Humboldt pour ma part et de redécouvrir cette correspondance grâce à de nouvelles technologies pour les chercheurs. Un site web serait intéressant à développer afin que toute personne puisse avoir accès à ces données sans condition de compétences techniques préalables. Cela ferait également résonance à la volonté d'Humboldt énoncée dans les premières lignes de cette introduction.

En fonction de la suite de ces outils, de l'envergure future à laquelle on les destine, du public cible et de la vocation de leur pérennisation, diverses évolutions et améliorations peuvent ainsi être envisagées et de nombreuses solutions aux problèmes soulevés pourront efficacement être mises en place.

Bibliographies thématiques

Sur les humanités numériques

ALBOUY (Ségolène), *Médiation des données de recherche. Élaboration d'une plateforme en ligne pour une base de tables astromiques anciennes*, Mémoire le Master Technologies numériques appliquées à l'histoire, 2019.

CHATELUS (Marjorie), *L'édition électronique d'un catalogue du XVIII^e siècle : le manuscrit 813 de la bibliothèque de Saint Omer*, Mémoire pour le diplôme "Technologie appliquées à l'histoire", 2014.

CLAVERT (Frédéric), « Vers de nouveaux modes de lecture des sources », dans *Le temps des humanités digitales*, dir. Olivier Le Deuff, Fyp Editions, 2014, URL : <https://orbi.lu/bitstream/10993/34980/1/CLAVERT%20-%20Sur%20le%20cas%20de%20l%27histoire.pdf> (visité le 14/07/2021).

COLLECTIF, « L'historien programmeur ? : Proposé par : Frédéric Clavert, Aurélien Berra, Franziska Heimburger », dans *THATCamp Paris 2012 : Non-actes de la non-conférence des humanités numériques*, Code : THATCamp Paris 2012 : Non-actes de la non-conférence des humanités numériques, Paris, 2012 (La Non-Collection), URL : <http://books.openedition.org/editionsmsmsh/305> (visité le 23/07/2021).

DUMONT (Stefan), « Briefe kommentieren im Semantic Web » (), p. 25.

DUMONT (Stefan), BÖRNER (Ingo), LEIPOLD (Dominik), MÜLLER-LAACKMAN (Jonas), SCHNEIDER (Gerlinde), DUMONT (Stefan), HAAF (Susanne) et SEIFERT (Sabine), « Correspondence Metadata Interchange Format », dans *Encoding Correspondence. A Manual for TEI-XML-based Encoding of Letters and Postcards in TEI-XML and DTABf*, 1^{re} éd., Berlin, 2019.

DUMONT (Stefan) et FECHNER (Martin), « Bridging the Gap : Greater Usability for TEI encoding », *Journal of the Text Encoding Initiative* (, Issue 8[2014]), Number : Issue 8 Publisher : Text Encoding Initiative Consortium, DOI : 10.4000/jtei.1242.

ELKNER (Jeff), *How to Think Like a Computer Scientist*, URL : <https://www.greenteapress.com/thinkpython/thinkCSpy/html/preface.html> (visité le 10/08/2021).

Förderkriterien für wissenschaftliche Edition in der Literaturwissenschaft, Publication à partir du colloque Sprachwissenschaften der Deutschen Forschungsgemeinschaft (DFG), Bonn, 2015, URL : https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf.

- « *Framework* », *Wikipédia* (, 22 févr. 2021), URL : <https://fr.wikipedia.org/w/index.php?title=Framework&oldid=180178641> (visité le 10/08/2021).
- GENET (Jean-Philippe), « Histoire, Informatique, Mesure », *Histoire & Mesure*, 1–1 (1986), p. 7-18, DOI : 10.3406/hism.1986.904.
- GRAHAM (Shawn), MILLIGAN (Ian) et WEINGART (Scott), *Exploring Big Historical Data : the Historian's Macroscopic*, Imperial College Press, London, 2016.
- Interopérabilité : définition et synonyme*, URL : <https://www.journaldunet.fr/web-tech/dictionnaire-de-l-iot/1208123-interoperabilite-une-capacite-essentielle-pour-l-iot/> (visité le 20/08/2021).
- JÄNICKE (S), FRANZINI (G), CHEEMA (M F) et SCHEUERMANN (G), « On Close and Distant Reading in Digital Humanities : A Survey and Future Challenges » (, 2015), p. 21, URL : <https://www.informatik.uni-leipzig.de/~stjaenicke/Survey.pdf>.
- Kalliope / Historie*, URL : <https://kalliope-verbund.info/de/ueber-kalliope/historie.html> (visité le 26/07/2021).
- KALLIOPE OPAC*, 26 oct. 2014, URL : <https://web.archive.org/web/20141026122746/http://kalliope.staatsbibliothek-berlin.de/> (visité le 26/07/2021).
- « Kalliope-Verbund », *Wikipedia* (, 15 janv. 2021), URL : <https://de.wikipedia.org/wiki/Kalliope-Verbund> (visité le 26/07/2021).
- LE ROY LADURIE (Emmanuel), « La fin des érudits », *Le Nouvel Observateur*, 8 (1968), p. 38-39.
- LEMY (Stefan), « *L'historien de demain sera programmeur... : Emmanuel Le Roy Ladurie et les défis de la science* », L'Histoire à la BnF, URL : <https://hypotheses.org/1505> (visité le 14/07/2021).
- MEYER (Till), « Datenmodellierung für digitale Editionen – Stand und Perspektiven zwischen XML/TEI, Linked Open Data und Ontologien », *Master Thesis* (, 2019), p. 70.
- MKADMI (Abderrazak), « XML et travail collaboratif : vers un Web sémantique », *La revue maghrébine de documentation et d'information*, 16 (2006), URL : <https://hal.archives-ouvertes.fr/hal-01340492> (visité le 25/07/2021).
- MORETTI (Franco), *Graphs, Maps, Trees. Abstract Models for a Literary History*, Verso, London, New York, 2005.
- PÉLISSIER (Chrysta), « Accompagner le chercheur en SHS à l'ère des humanités numériques », *Les Cahiers du numérique*, Vol. 13–3 (2017), Bibliographie_available : 1 Cairndomain : www.cairn.info Cite Par_available : 0 Publisher : Lavoisier, p. 167-194, URL : <https://www-cairn-info.proxy.chartes.psl.eu/revue-les-cahiers-du-numerique-2017-3-page-167.htm> (visité le 08/07/2021).
- « Société américaine de philosophie », *Wikipédia* (, 25 avr. 2020), URL : https://fr.wikipedia.org/wiki/Soci%C3%A9t%C3%A9_am%C3%A9ricaine_de_philosophie (visité le 28/07/2021).

- TEI-Correspondence-SIG/CMIF*, original-date : 2015-06-18T09 :51 :52Z, 19 nov. 2020,
URL : <https://github.com/TEI-Correspondence-SIG/CMIF> (visité le 07/07/2021).
- TOMIC (Yves), « De l'usage des API », *Documentaliste-Sciences de l'Information*, Vol. 51–3 (25 sept. 2014), p. 17-18, URL : <https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2014-3-page-17.htm> (visité le 08/07/2021).
- Use the APS Library*, American Philosophical Society, URL : <https://www.amphilsoc.org/library> (visité le 28/07/2021).
- VERLAET (Lise) et DILLAERTS (Hans), « L'enjeu du web de données pour l'édition scientifique », *I2D - Information, donnees documents*, Volume 53–2 (5 juil. 2016), Bibliographie_available : 0 Cairndomain : www.cairn.info Cite Par_available : 1 Publisher : A.D.B.S., p. 49-49, URL : <https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-49.htm> (visité le 24/07/2021).
- VITALI-ROSATI (Marcello) et E. SINATRA (Michael), *Pratiques de l'édition numérique*, Les Presses de l'Université de Montréal, Montréal, 2014 (Parcours Numériques), URL : <http://www.parcoursnumeriques-pum.ca/>.
- WENZ (Romain), « Hypertextualisation », *Revue de la BNF*, n° 42–3 (2012), p. 36-41, URL : <https://www.cairn.info/revue-de-la-bibliotheque-nationale-de-france-2012-3-page-36.htm> (visité le 26/07/2021).
- *L'open data, un levier pour l'évolution des catalogues*, 2016, URL : <https://www.cairn.info/vers-de-nouveaux-catalogues--9782765415138--page-13.htm> (visité le 07/07/2021).

Sur les données et visualisations de données

BABINET (Gilles), « 12. Open source », *Hors collection* (, 2020), Bibliographie_available : 0 Cairndomain : www.cairn.info Cite Par_available : 0 ISBN : 9782100820764 Publisher : Dunod, p. 107-114, URL : <https://www-cairn-info.proxy.chartes.psl.eu/refondre-les-politiques-publiques--9782100820764-page-107.htm> (visité le 09/08/2021).

BERNERS-LEE (Tim), HENDLER (James) et LASSILA (Ora), « The Semantic Web : a new form of Web content that is meaningful to computers wil unleash a revolution of new possibillities », *Scientific American Magazine* (, 17 mai 2001), URL : <http://web.archive.org/web/20081114135540/http://www.sciam.com/article.cfm?id=the-semantic-web&print=true> (visité le 25/07/2021).

CASTETS-RENARD (Céline) et GANDON (Nathalie), « Open data des données de la recherche publique : entre réformes législatives et retour d'expérience sur un guide pratique à destination des chercheurs », *LEGICOM*, N° 56–1 (8 mars 2016), p. 67-75, URL : <https://www.cairn.info/revue-legicom-2016-1-page-67.htm> (visité le 08/07/2021).

FABRY (Cécilia), ROUSSEL (Clotilde), COLLIGNON (Alain), PARMENTIER (François), MOREAU (Elise) et THOUVENIN (Nicolas), « Publier des données liées et ouvertes en sept étapes », *I2D - Information, donnees documents*, Volume 54–1 (1^{er} avr. 2017), Bibliographie_available : 0 Cairndomain : www.cairn.info Cite Par_available : 0 Publisher : A.D.B.S., p. 12-14, URL : <https://www-cairn-info.proxy.chartes.psl.eu/revue-i2d-information-donnees-et-documents-2017-1-page-12.htm> (visité le 09/08/2021).

FRANCONY (Jean-Marc), « L'éditorialisation des données aux bornes des API : enjeux et perspectives pour une analyse empirique », *Les Enjeux de l'information et de la communication*, N° 19/2–2 (2018), p. 69-79, URL : <https://www.cairn.info/revue-les-enjeux-de-l-information-et-de-la-communication-2018-2-page-69.htm> (visité le 09/07/2021).

GRAHAM (Shawn), MILLIGAN (Ian) et WEINGART (Scott), *Exploring Big Historical Data : the Historian's Macroscope*, Imperial College Press, London, 2016.

- MAIGNIEN (Yannick), « Chapitre 5. Les enjeux du web sémantique », dans *Pratiques de l'édition numérique*, dir. Marcello Vitali-Rosati et Michael E. Sinatra, Code : Pratiques de l'édition numérique, Montréal, 2014 (Parcours numérique), p. 77-93, URL : <http://books.openedition.org/pum/320> (visité le 25/07/2021).
- MANNING (Patrick), *Big Data in History*, New York, 2013.
- « Metadata Object Description Schema », *Wikipédia* (, 13 déc. 2020), URL : https://fr.wikipedia.org/wiki/Metadata_Object_Description_Schema (visité le 27/07/2021).
- MKADMI (Abderrazak), « XML et travail collaboratif : vers un Web sémantique », *La revue maghrébine de documentation et d'information*, 16 (2006), URL : <https://hal.archives-ouvertes.fr/hal-01340492> (visité le 25/07/2021).
- « Semantic Web and data model », *Data BnF* (), URL : <https://data.bnf.fr/en/semanticweb> (visité le 21/08/2021).
- « Traitement de données », *Wikipédia* (, 8 juin 2021), URL : https://fr.wikipedia.org/wiki/Traitement_de_donn%C3%A9es (visité le 30/07/2021).
- VERLAET (Lise) et DILLAERTS (Hans), « L'enjeu du web de données pour l'édition scientifique », *I2D - Information, donnees documents*, Volume 53–2 (5 juil. 2016), Bibliographie_available : 0 Cairndomain : www.cairn.info Cite Par_available : 1 Publisher : A.D.B.S., p. 49-49, URL : <https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-49.htm> (visité le 24/07/2021).
- « Web des données », *Wikipédia* (, 16 juin 2021), URL : https://fr.wikipedia.org/wiki/Web_des_donn%C3%A9es (visité le 14/07/2021).
- « Web sémantique », *Wikipédia* (, 13 juil. 2021), URL : https://fr.wikipedia.org/wiki/Web_s%C3%A9mantique (visité le 14/07/2021).

Sur les projets de la BBAW

At the intersection of sciences, humanities and technologies – A review of the edition humboldt digital – RIDE, URL : <https://ride.i-d-e.de/issues/issue-13/ehd/> (visité le 07/07/2021).

avhumboldt.de, avhumboldt.de, URL : <http://www.avhumboldt.de/> (visité le 07/07/2021).
correspSearch - Correspondence Metadata Interchange-Format (CMI), URL : <https://correspsearch.net/de/dokumentation.html> (visité le 07/07/2021).

DUMONT (Stefan), « correspSearch – Connecting Scholarly Editions of Letters », *Journal of the Text Encoding Initiative* (, Issue 10[2016]), Number : Issue 10 Publisher : Text Encoding Initiative Consortium, DOI : 10.4000/jtei.1742.

Ediarum. A toolbox for editors and developers – RIDE, URL : <https://ride.i-d-e.de/issues/issue-11/ediarum/> (visité le 07/07/2021).

KRAFT (Tobias) et DUMONT (Stefan), « The Humboldt Code : On creating a hybrid digital scholarly edition of a 19th century globetrotter », *Wiener Digitale Revue*–1 (1^{er} oct. 2020), Number : 1, DOI : 10.25365/wdr-01-03-02.

SCHWARZ (Ingo), « Zur Geschichte der Alexander-von-Humboldt-Forschung und der Berlin-Brandenburgischen Akademie der Wissenschaft », dans *Die Berliner und Brandenburgische Lateinamerikaforschung in Geschichte und Gegenwart. Personen und Institutionen*. Dir. Gregor Wolff, Wissenschaftlicher Verlag Berlin, Berlin, 2001, p. 107-127.

Sur Alexander von Humboldt

- « Alexander von Humboldt », *Wikipedia* (, 27 juil. 2021), URL : https://fr.wikipedia.org/wiki/Alexander_von_Humboldt (visité le 28/07/2021).
- BIERMANN (Kurt-R.), « Der Zugang an Briefen Alexander von Humboldts hält an », *Spektrum. Mitteilungsblatt für die Mitarbeiter der Deutschen Akademie der Wissenschaft zu Berlin*, 11, 2 (1965), p. 55-58.
- « Die Alexander-von-Humboldt-Forschung an der Akademie der Wissenschaften der D.D.R. - Ergebnisse und Ziele », dans *Boston Studies in the Philosophy of Science*, 1974 (15), p. 295-305.
- « Alexander von Humboldt als Initiator und Organisator internationaler Zusammenarbeit auf geophysikalischem Gebiet », dans *Mischellanea Humboldtiana*, Akademie Verlag Berlin, Berlin, 1990, p. 107-116.
- BIERMANN (Kurt-R.) et LANGE (Fritz G.), « Die Alexander-von-Humboldt-Briefausgabe », *Forschungen und Fortscritte*, 36, 8 (1962), p. 225-230.
- BOURGUET (Marie-Noëlle), *Le monde dans un carnet : Alexander von Humboldt en Italie (1805)*, Édition du félin, Paris, 2017.
- DAUM (Andreas W.), « Popularisierung des Wissens », dans *Alexander von Humboldt, Handbuch : Leben-Werk-Wirkung*, dir. Ottmar Ette, Springer-Verlag, Stuttgart, 2018, p. 200-204.
- « Resonance space - The cosmos Lectures », dans *Wilhem und Alexander von Humboldt, Berlin Cosmos*, dir. Paul Spies, Ute Tinteman et Jan Mende, Wienand, 2020, p. 149-151.
- DR. KRAFT (Tobias) et THOMAS (Christian), *Alexander von Humboldt auf Reisen - [Editions-]Wissenschaft in Bewegung*, 12 févr. 2020, URL : <https://cutt.ly/SmVemqj>.
- SCHUCHARDT (Gregor), *Fakt, Ideologie, System. Die Geschichte der ostdeutschen Alexander von Humboldt-Forschung*, Franz Steiner Verlag, Stuttgart, 2010.
- SCHWARZ (Ingo), « Die Korrespondenz », dans *Alexander von Humboldt, Handbuch : Leben-Werk-Wirkung*, dir. Ottmar Ette, Springer-Verlag, Stuttgart, 2018, p. 80-91.

Acronymes

- API** *Application Programming Interface.* 13, 18, 19, 22, 25, 27, 28, 31, 50, 57, 58, 66, 72, 80
- APS** *American philosophical Society* (en français : Société Américaine de Philosophie). 24–28, 95
- ARK** *Archival Resource Key.* 26, 28
- BBAW** *Berlin-Brandenburgische Akademie der Wissenschaft* (en français : Académie des sciences de Berlin-Brandebourg). 2–4, 8–10, 21, 38–40, 43, 45, 50, 57, 59, 61, 64, 68, 74, 75, 80
- BnF** Bibliothèque nationale de France. 14, 25, 26, 28, 30, 59, 67, 95
- CMIF** *Correspondence Metadata Interchange Format.* 20, 21, 64–69, 72–77, 79, 81, 83
- CSS** *Cascading Style Sheets.* 9, 70, 74
- CSV** *Comma-separated values.* 25, 29, 30
- DC** Dublin Core. 23–25, 28–30, 95
- DFG** *Deutsche Forschungsgemeinschaft.* 8
- DOM** *Document Object Model.* 29, 30
- DTABf** *Deutsches Textarchiv – Basisformat.* 8, 20
- EAD** *Encoded Archival Description.* 22, 25, 28–30, 65, 95
- ehd** *edition humboldt digital.* 3, 8, 19, 21
- ERNIE** *Encyclopedia of Romantic Nationalism in Europe.* 42
- GND** *Gemeinsame Normdatei.* 14–16, 19, 23, 24, 67, 76, 79, 80
- HTML** *Hypertext Markup Language.* 41, 46, 47, 70, 71, 74
- HTTP** *Hypertext Transfer Protocol.* 15, 26, 66
- JSON** *JavaScript Object Notation.* 29–31, 51

- MODS** Metadata Object Description Schema. 23, 73
- NDL** *New Diet Library*. 67
- ODbL** *Open Database Licence*. 18
- OFAJ** Organisation franco-allemande pour la jeunesse. v, 3
- RDA** République Démocratique Allemande. 37
- RDF** *Resource Description Framework*. 13, 25
- SIG** *Special Interest Group*. 64, 66
- SRU** *Search/Retrieve via URL*. 18, 22, 23, 26
- TEI** *Text Encoding Initiative*. 7–11, 14, 20, 41, 64–66, 72
- TELOTA** *The Electronic Life Of The Academy*. 3, 9, 40, 45, 64, 68, 80
- TNAH** Technologies numériques appliquées à l'histoire. 3
- URI** *Uniform Resource Identifier*. 15, 26, 67, 69, 73, 76
- URL** *Uniform Resource Locator*. 15
- VIAF** *Virtual International Authority File*. 14, 16, 67
- WYSIWYG** *what you see is what you get*. 9
- XML** *eXtensible Markup Language*. 8–14, 20, 22, 23, 25, 26, 28–30, 41, 65, 66, 72, 74, 95
- XSL** *eXtensible Stylesheet Language*. 9–11, 41
- ZKA** *Zentralkartei der Autograph* (en français : répertoire central des autographies). 22

Table des figures

| | | |
|-----|--|----|
| 1.1 | Extrait du carnet de voyage d'Italie (1805) dans l'environnement ediarum. | 9 |
| 1.2 | Les différents index directement accessibles dans ediarum | 10 |
| 1.3 | Les diverses approches textuelles accessibles sur le site de l'édition. | 11 |
| 1.4 | L'architecture du Web sémantique ou <i>semantic stack</i> . Source : Wikipedia, Semantic Stack | 13 |
| 2.1 | Gay-Lussac dans le registre des personnes. Capture d'écran de l'interface de l'édition numérique. | 19 |
| 2.2 | La ville de Cumana dans l'index des lieux de l'édition numérique | 20 |
| 2.3 | Le lien vers le Kalliope-Verbundkatalog dans le <sourceDesc> | 21 |
| 2.4 | Le lien vers le Kalliope-Verbundkatalog cliquable dans la version numérique de l'édition | 21 |
| 3.1 | Entrée au sein du le fichier XML-EAD des données de l'APS | 28 |
| 3.2 | Entrée au sein du le fichier DC issu du Kalliope-Verbundkatalog | 28 |
| 3.3 | Entrées du CSV composé des données du Catalogue général de la BnF . . | 28 |
| 4.1 | Exemple du tiroir contenant les cartes des noms commençant par la lettre L. | 38 |
| 4.2 | Aide à la recherche en version numérique | 39 |
| 4.3 | Copie d'écran générale de l'outil <i>Humboldt Chronotopographie</i> | 40 |
| 4.4 | Copies d'écran de <i>Humboldt Chronotopographie</i> | 41 |
| 4.5 | Copie d'écran du site ERNIE présentant l'itinéraire du voyage de Mme de Stael, itinéraire sélectionné dans le menu à droite. | 43 |
| 4.6 | Définir un itinéraire : l'exemple de Valtaï à Moscou | 43 |
| 4.7 | Superposer les cartes : exemple de la frontière kazakhe | 44 |
| 5.1 | Structure de données de la librairie pandas : dataframe de toutes les don- nées de la correspondance d'Alexander von Humboldt | 48 |
| 5.2 | Table pivot de pandas avec toutes les données de la correspondance d'Alexan- der von Humboldt | 48 |
| 5.3 | Exemple des dropdown menus apparaissant de manière dynamique par la fonction récursive. | 51 |

| | | |
|-----|---|-----|
| 5.4 | Exemple du dropdown menu pour les expéditeurs de lettres. | 52 |
| 5.5 | Échanges épistolaires entre Buschmann et Humboldt | 53 |
| 5.6 | Zoom sur l'Europe de la carte représentant l'ensemble de la correspondance de Humboldt. | 54 |
| 5.7 | Visualisation cartographique des résultats pour les lettres de l'année 1802 avec un zoom optimal. | 55 |
| 6.1 | Accès aux catalogues en ligne. | 58 |
| 6.2 | Exemple d'affichage : les 1134 lettres de Berlin | 58 |
| 6.3 | Exemple : les résultats pour la BnF | 59 |
| 7.1 | Fonctionnement de correspSearch. | 65 |
| 7.2 | Exemple d'une lettre en CMIF. | 67 |
| 8.1 | La structure de l'application | 71 |
| 8.2 | Logique algorithmique de l'application schématisé | 73 |
| 8.3 | Page <code>about</code> du cS matching tool | 75 |
| 8.4 | Interface permettant à l'utilisateur de glisser son document CMIF | 75 |
| 8.5 | Page <i>possible matches</i> du cS matching tool | 76 |
| A.1 | Le modèle de données en réseau de l' <i>edition humboldt digital</i> | 100 |
| B.1 | Lettre encodée en format DC | 101 |
| B.2 | Lettre encodée en format MODS | 102 |
| D.1 | Première page de l'index de lieux | 105 |
| D.2 | Détail des documents conservés pour les lieux de l'index | 106 |
| E.1 | Visualisation des données extraites de l'aide à la recherche sur l'outil chro-notopographique | 108 |
| E.2 | Les différents fonds de cartes disponibles | 109 |
| F.1 | Carte représentant toute la correspondance de Humboldt. La taille des points représentés est proportionnelle au nombre de lettres reçues ou envoyées. | 110 |
| F.2 | Carte représentant toute la correspondance de Humboldt représentée sous six périodes différentes. | 111 |

Liste des tableaux

| | | |
|-----|---|----|
| 2.1 | Quantité des documents et des institutions partenaires du Kalliope-Verbundkatalog | 22 |
| 4.1 | Tableau du nombre de documents d'Alexander von Humboldt conservés à la BBAW | 37 |

Annexes

Annexe A

Le modèle de données en réseau de l'*edition humboldt digital*



FIGURE A.1 – Le modèle de données en réseau de l'*edition humboldt digital*

Annexe B

Différence entre le format MODS et DC.

Exemple d'une lettre de Abich envoyée à Humboldt en 1852 ou 1853.

```
▼<srw:record>
  <srw:recordSchema>info:srw/schema/1/dc-v1.1</srw:recordSchema>
  <srw:recordPacking>xml</srw:recordPacking>
  ▼<srw:recordData>
    ▼<srw_dc:dc>
      <dc:identifier>DE-611-HS-1650160</dc:identifier>
      <dc:identifier>http://kalliope-verbund.info/DE-611-HS-1650160</dc:identifier>
      <dc:identifier>GA 72/84</dc:identifier>
      <dc:publisher>DE-611</dc:publisher>
      <dc:title>Brief von Hermann Abich an Alexander von Humboldt</dc:title>
      <dc:created>20100505</dc:created>
      <dc:modified>20100505</dc:modified>
      <dc:contributor>Universitätsbibliothek Freiburg</dc:contributor>
      <dc:language>ger</dc:language>
      <dc:type>item</dc:type>
      <dc:date>1852/1853</dc:date>
      <dc:coverage>Tiflis</dc:coverage>
      <dc:creator>Abich, Hermann (1806-1886)</dc:creator>
      <dc:subject>Humboldt, Alexander von (1769-1859)</dc:subject>
      <dc:format.extent>2 Br., 1 Br.abschr. + 3 Bl. Beil.</dc:format.extent>
    </srw_dc:dc>
  </srw:recordData>
</srw:record>
```

FIGURE B.1 – Lettre encodée en format DC

```

▼<srw:record>
  <srw:recordSchema>mods</srw:recordSchema>
  <srw:recordPacking>xml</srw:recordPacking>
  ▼<srw:recordData>
    ▼<mods xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.loc.gov/mods/v3" xsi:schemaLocation="http://www.w3.org/2001/XMLSchema-instance http://www.loc.gov/mods/v3">
      <identifier type="uri">http://kalliope-verbund.info/DE-611-HS-1650160</identifier>
      ▼<recordInfo>
        <recordIdentifier source="DE-611">DE-611-HS-1650160</recordIdentifier>
        <recordCreationDate encoding="marc">20100505</recordCreationDate>
        <recordChangeDate encoding="iso8601">20100505</recordChangeDate>
        <recordContentSource authority="marcorg">Universitätsbibliothek Freiburg</recordContentSource>
      ▶</recordInfo>
      ▼<relatedItem>
        <identifier type="localparentid">DE-611-BF-18213</identifier>
        <identifier type="uri">http://kalliope-verbund.info/DE-611-BF-18213</identifier>
        ▼<titleInfo>
          <title>Seibold, Ilse GA 72</title>
        ▶</titleInfo>
      ▶</relatedItem>
      <typeOfResource manuscript="yes">text</typeOfResource>
      ▼<name type="personal" authority="GND" valueURI="http://d-nb.info/gnd/116003383">
        <namePart>Abich, Hermann (1806-1886)</namePart>
        ▼<role>
          <roleTerm type="code" authority="marcrelator">aut</roleTerm>
          <roleTerm>author</roleTerm>
        ▶</role>
      ▶</name>
      ▼<name type="personal" authority="GND" valueURI="http://d-nb.info/gnd/118554700">
        <namePart>Humboldt, Alexander von (1769-1859)</namePart>
        ▼<role>
          <roleTerm type="code" authority="marcrelator">rcp</roleTerm>
          <roleTerm>addressee</roleTerm>
        ▶</role>
      ▶</name>
      ▼<titleInfo>
        <title>Brief von Hermann Abich an Alexander von Humboldt</title>
      ▶</titleInfo>
      <abstract type="content">2 Fotokopien, 1 Abschr. (Auszug); Beil.: Informationen dazu</abstract>
      ▼<originInfo>
        <dateCreated encoding="w3cdtf">1852/1853</dateCreated>
        ▼<place>
          <placeTerm type="text">Tiflis</placeTerm>
        ▶</place>
      ▶</originInfo>
      ▼<language>
        <languageTerm authority="iso639-2b">ger</languageTerm>
        <languageTerm type="text">Deutsch</languageTerm>
      ▶</language>
      ▼<location>
        <shelfLocator>GA 72/84</shelfLocator>
      ▶</location>
      ▼<physicalDescription>
        <extent>2 Br., 1 Br.abschr. + 3 Bl. Beil.</extent>
      ▶</physicalDescription>
    ▶</mods>
  ▶</srw:recordData>
</srw:record>

```

FIGURE B.2 – Lettre encodée en format MODS

Annexe C

Fonction en Python retournant les informations géographiques d'un lieu donné

```

1 def get_geolocalisation_place(place):
2 """
3     Give the geolocation, geoname ID and ehd ID of a giving place.
4     The function searches first in different local files :
5     - placw register of edition humboldt
6     - place register from GeoName
7     If the place isn't already stored in one of this file, then
8     a request is sent to the GeoName API.
9     :param place: str
10    :return: coverage_location
11    :rtype: dict
12 """
13 ortsregister = getJSON('data/edh_ortsregister.json')
14 or_geoname = getJSON('data/geoname_ortsregister.json')
15 coverage_location = {}
16 from_geoname = {}
17
18 try :
19     for o in ortsregister:
20         # If the giving place is stored in the ehd place's register
21         # then store infos in the dict coverage_location
22         if place == o['properties']['ContentHeader'] :
23             coverage_location["key"] = o['properties']['key']
24             coverage_location['geoname_id'] = o['properties']['geoname_id']
25             coverage_location['address'] = o['properties']['ContentHeader']
26             coverage_location['coordinates'] = o['geometry']['coordinates']
27
28         # If the giving place isn't in the ehd place's register then check
29         # if it's stored in the geoname_ortsregister
30         if bool(coverage_location)== False :
31             for o in or_geoname:
32                 for i in o:
33                     if place == o[i]["address"]:
34                         coverage_location["key"] = o[i]['key']
35                         coverage_location['geoname_id'] = o[i]['geoname_id']
36                         coverage_location['address'] = i
37                         coverage_location['coordinates'] = o[i]['coordinates']
38
39         # If the giving place is neither in both local place's registers
40         # then a request is sent to GeoName API
41         if bool(coverage_location)== False :
42             location = geocoder.geonames(place, key='USER_KEY', featureClass='P')
43             coverage_location['geoname_id'] = str(location.geonames_id)
44             coverage_location['address'] = location.address
45             coverage_location['coordinates'] = [location.lng, location.lat]
46
47             for ort in ortsregister:
48                 if ort['properties']['geoname_id'] == str(coverage_location['geoname_id']):
49                     coverage_location['key'] = ort['properties']['key']
50                     from_geoname[place] = coverage_location
51
52         # If the geolocation from geoname wasn't already stored in the local
53         # file, then they are added in it.
54         if bool(from_geoname) == True:
55             d = getJSON("data/geoname_ortsregister.json")
56             d.append(from_geoname)
57             writeJSON("data/geoname_ortsregister.json", d)
58     except :
59         print(place)
60
61 return coverage_location

```

Annexe D

Aide à la recherche dactylographiée

Exemples de pages tirées de l'aide à la recherche des archives Humboldt de la BBAW

| Ort | Briefe in den Kästen Nr. |
|--------------------|--|
| Aberdeen | 1 |
| Altenburg | 1 |
| Aachen | 1 |
| Amsterdam | 1 |
| Ann Arbor, Mich. | 1, 154 |
| Ann Arbor, MI | 1 → 15712 |
| Avignon | 1 |
| Bad Godesberg | vielte Briefe |
| Bad Homburg v.d.H. | 1 |
| Baltimore, MD | 1 |
| Bamberg | 1 |
| Barnaul | 2 |
| Basel | 2 |
| Bassano del Grappa | 153 |
| Bayeux | 2 |
| Bayville | → 15712 |
| Benesov | 2 |
| Berkeley | Berlino 16. 11. (Sel. Aus.) |
| Berlin, DDR | 2 bis 106, 108, 109, 144, 149 bis 152, 155, 158. |
| Berlin (West) | 37a bis 45 (vorm. in Marburg), 65-66 (vorm. in Tübingen), 107, 107a, 109, 128 bis 136, 141 bis 144, 154, 155, 157, 160, 163, 166, 167, 168, 172, 176, 182, 183 186 (Klein) |

FIGURE D.1 – Première page de l'index de lieux

| | von H. | an H. | sonst. |
|---|--------|-------|--------|
| Aberdeen (Univ.Libr.) | 1 | | |
| Altenburg/Thür/ ^g (LA) | 1 | | |
| Aachen | 1 | | |
| Amsterdam (Ak.Wet.) | 2 | 2 | |
| Ann Arbor (Clements Libr.) x) | | | 1 |
| Avignon (Bibl.Calvet) | 2 | | |
| Bad Homburg v.d.H. (Möbius) | 1 | | |
| Baltimore, Maryland (Mangled und Sac) | 35 | | |
| Bamberg (SA) | 53 | 12 | 1 |
| Baltimore, MD Pr.H. Library | | 1 | |
| x) Ann Arbor (Univ.of Michigan) <i>Side K 159</i> | | | |

FIGURE D.2 – Détail des documents conservés pour les lieux de l'index

Annexe E

L'outil *Humboldt Chronotopographie*

Captures d'écran de l'outil expérimental

FIGURE E.1 – Visualisation des données extraites de l'aide à la recherche sur l'outil chronotopographique

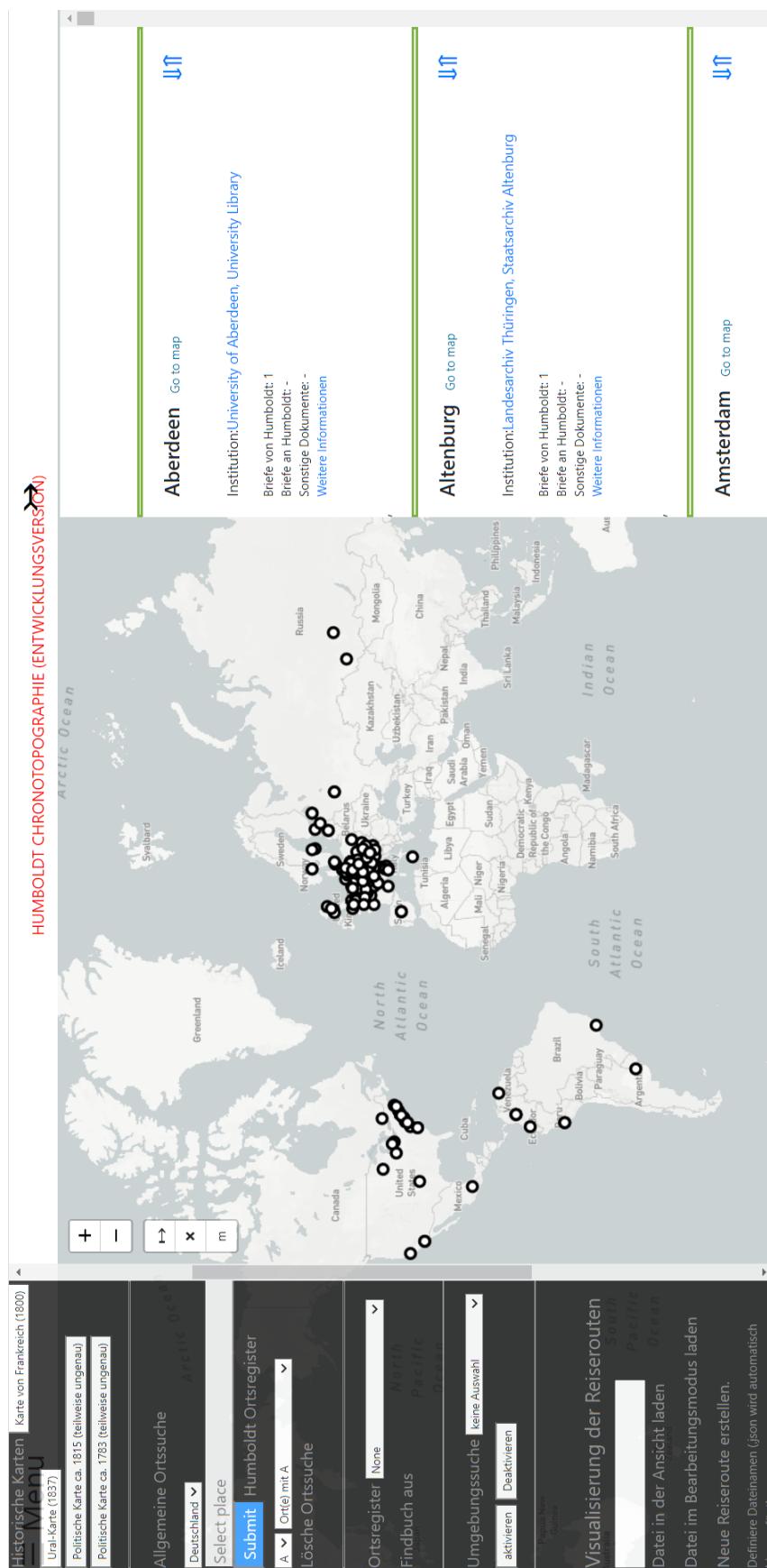
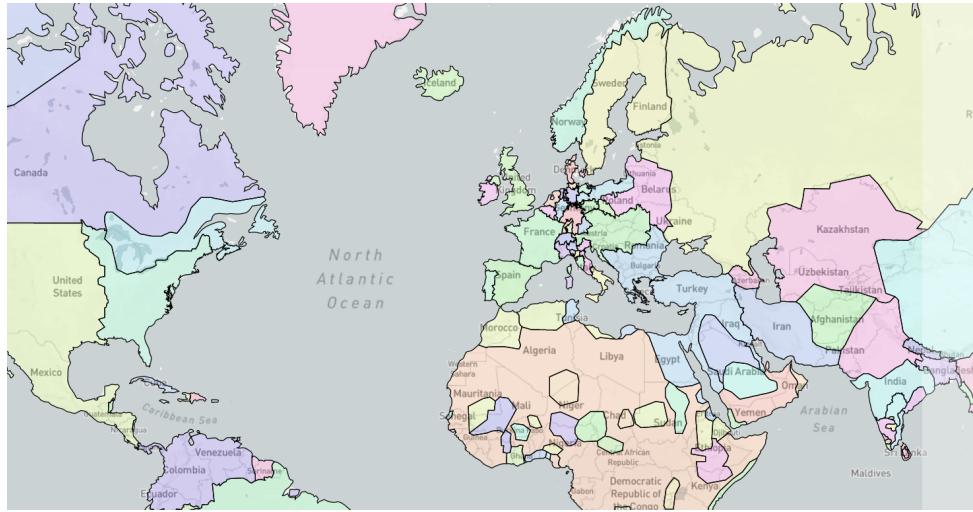


FIGURE E.2 – Les différents fonds de cartes disponibles



(a) Fond de carte mondiale des frontières politiques vers 1783



(b) Fond de carte mondiale des frontières politiques vers 1815



(c) Fond de carte de la France vers 1800.

Annexe F

Visualisations cartographiques

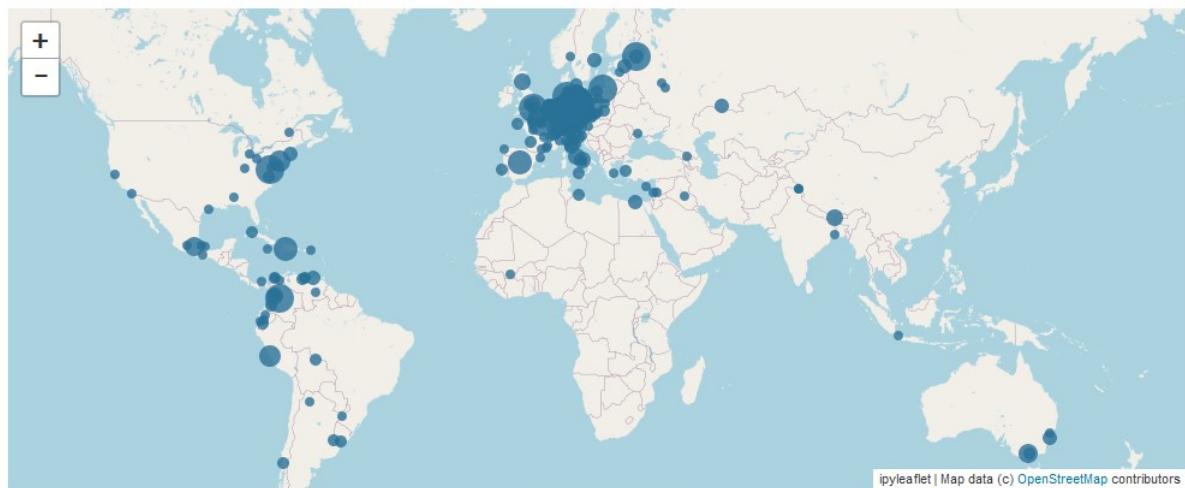


FIGURE F.1 – Carte représentant toute la correspondance de Humboldt. La taille des points représentés est proportionnelle au nombre de lettres reçues ou envoyées.



FIGURE F.2 – Carte représentant toute la correspondance de Humboldt représentée sous six périodes différentes.

Annexe G

Comparaison du design des sites web des projets de la BBAW

Über diese Edition

Das Editionsprojekt besteht aus zwei Teilen: aus der **Privatkorrespondenz Hirts** sowie aus dessen **Amtlichen Schriften**. Beide Teile ergänzen einander und sind durch die **Register** verbunden.

Die Edition dient in gleicher Weise der Quellenerschließung wie der kunst- und kulturhistorischen sowie wissenschaftsgeschichtlichen Grundlagenforschung und versteht sich als Beitrag zur Wiederentdeckung und Neubewertung eines einflussreichen Kunsteperten und Kulturpolitikers, eines Pioniers der Architekturgeschichte und Archäologie sowie ... [mehr]

Briefe

Das Briefkorpus besteht aus mehr als 250 privaten Briefen sowie ca. 250 verschlossenen Briefen aus den Jahren 1787 bis 1837, die erstmals im Gesamtzusammenhang ediert werden.

Amtliche Schriften

Das Korpus besteht aus mehr als 300 Einzeldokumenten (u.a. Gutachten, Voten, Promemorien, amtlichen Schreiben) aus dem Zeitraum 1796 bis 1837, die den jeweiligen Amtsbereichen Akademie der Wissenschaften, Akademie der Künste, Bauakademie sowie Museumskommission zugeordnet sind.

Briefwechsel aufschlagen

Suche

Suchbegriff hier eingeben

Suchen

[Erweiterte Suche]

(a) Extrait du site de l'édition numérique des écrits et de la correspondance d'Aloys Hirt

Briefe & Dokumente

Aufschlagen

Über diese Edition

Während seiner Direktionszeit von Dezember 1796 bis 1814 archivierte Iffland seine amtliche Korrespondenz sowie administrative und dramaturgische Dokumente. Ziel des Projekts ist es, die 34 von diesem Archiv überlieferten Foliobände mit ihren ca. 7.500 beschriebenen Blättern inhaltlich vollständig zu erschließen, um sie der Forschung und der interessierten Öffentlichkeit zugänglich zu machen

Suche

Suchbegriff hier eingeben

Briefe Suchen

[Erweiterte Suche]

Mehr zu dieser Edition

(b) Extrait du site de l'édition numérique des archives administratives et dramaturgiques d'August Iffland

New in Version 6 (October 2020)

Fragmente des Sibirischen Reise-Journals
A first glance at Humboldt's journal of his travels through Siberia and Central Asia in 1829 (beta edition).

Voyage d'Espagne aux Canaries et à Cumaná
Significant improvement in critical commentary and markup, including: correct conversion of the French Revolutionary Calendar; display of Humboldt's references to other sections of the diary.

Correspondence with the Schlagintweit brothers
Alexander von Humboldt's correspondence with the Schlagintweit brothers (complete edition).

The American Travel Journals
The journals of the American Journey (1799–1804) are the key source for understanding Humboldt's travel work and provide the basis for the reappraisal of his scientific heritage. The approximately 5,500 pages combine descriptions of the itinerary with measurement results, literary travel sketches, scientific essays, drawings, and sketches.

The Russian-Siberian Travel Journals
The journals of the Russian-Siberian Journey (1829) consist of three parts: Fragments of the Siberian Travel Journal 1829 and two journals with notes and measurements on geodesy and geomagnetism. The recordings of this journey provide the basis for Humboldt's three-volume work on Central Asia.

Humboldt's Papers
In addition to the travel journals, the project edits selected collections from Humboldt's personal papers: documents, letters, notes, and maps that are directly related to Humboldt's hemispheric journeys. First, we focused on the **topic life sciences**, in particular Humboldt's notes on phytogeography (editing period: 2015–2018). Since then, the focus has been on the **topic scientific journeys**.

Further information

Further Information

Discover Humboldt's Papers

(c) Extrait du site de l'édition numérique des carnets de voyages d'Alexander von Humboldt.

Moesia inferior, Thrace, Mysia, Troas

Coin of the Month

News

31 July 2021
Digital Classical Seminar
Berlin 2021/22

The call for papers is still open. We look forward to receiving your abstract by midnight on 31 July 2021

[Read More >](#)

A Patria Tradition in Tomis

This month, we present to you an enigmatic reverse type, which is attested only in the civic coinage of the Black Sea city of Tomis in Roman times. The coin image depicts a two wheeled chariot drawn by a bull. Seated in the chariot is a bearded man wearing a himation and extending his right hand, his gaze turned toward the direction of motion. In front of the bull, a clothed woman with an indistinguishable object in her right hand rushes to the left and looks back, stretching her left hand behind her.

The Coin of this Month is presented by Ulrike Peter

[Read More >](#)

[Show all news >](#)

(a) Extrait du portail web qui collecte et présente des données sur des monnaies provinciales romaines et grecques.

Briefe suchen

Mit correspSearch können Sie Verzeichnisse verschiedener digitaler und gedruckter Briefeditionen nach Absender, Empfänger, Schreiber und Datum durchsuchen.

Der Webservice aggregiert und wertet Dateien im "Correspondence Metadata Interchange"-Format aus, das auf der TEI-Erweiterung "correspDesc" der TEI Correspondence SIG basiert.

[Briefe suchen](#)

Mitmachen

Die bibliographischen Angaben der ausgewerteten Editionen finden Sie im Datenverzeichnis.

Der Bestand wird laufend erweitert. Jede gedruckte oder digitale Briefedition kann ihr Briefverzeichnis im "Correspondence Metadata Interchange"-Format beim Webdienst correspSearch registrieren.

[Mitmachen](#) [CMIF Creator](#)

Neu im Datenbestand

Johann Christoph Gottsched. Briefwechsel unter Einschluß des Briefwechsels von Luise Adelgunde Victorie Gottsched. Im Auftrage der Sächsischen Akademie der Wissenschaften zu Leipzig hrsg. v. Detlef Döring f.u. Manfred Rudersdorf. Band 14: November 1748–September 1749. Hrsg. und bearb. von Caroline Köhler, Franziska Menzel, Rüdiger Otto und Michael Schlotz. Berlin, Boston: de Gruyter, 2020. [Digitale Ausgabe, 2020. DOI: 10.1515/9783110679892]

[Anzeigen](#)

Die Korrespondenz der Constance de Salm (1767–1845). Inventar des Fonds Salm der Société des Amis du Vieux Toulon et de sa Région und des Bestands Constance de Salm im Archiv Schloss Dyck (Mitgliedsarchiv der Vereinigten Adelsarchiv im Rheinland e.V.). Elektronische Edition, DHF Paris 2020. <https://constance-de-salm.de>

[Anzeigen](#)

Aktuelles

correspSearch v2 online

Die nächste Generation des Webservices correspSearch ist online gegangen. Er bietet nun stark verbesserte Such- und Filtermöglichkeiten sowie eine eigene karten-basierte Suche, mit deren Hilfe man Briefe aus einem bestimmten Gebiet in einem bestimmten Zeitraum suchen kann. Die ganze Website ist nun responsive und kann somit auch auf mobilen Endgeräten benutzt werden. Schlussendlich hat sich einiges "unter der Motorhaube" getan: der komplette Harvesting- und Importprozess wurde verbessert.

Mehr Grußformeln

»Leben Sie wohl, bester Freund! – Ich möchte noch so manches mit Ihnen plaudern – aber für heute genug unveränderbar der Ihrige«

Alloys Hirt - Briefwechsel 1787–1837 An Karl August Böttiger. Berlin, den 11. April 1799. Donnerstag CC BY 4.0

Inspirierende Grußformeln für Ihre Korrespondenz erhalten Sie bei queteSalute!

Philip Jakob Spener, Briefe, hrsg. von

(b) Extrait du site de correspSearch

Über die digitale Edition

Jean Paul - Sämtliche Briefe digital präsentiert die Briefe von Jean Paul in der Fassung der von Eduard Berend herausgegebenen dritten Abteilung der historisch-kritischen Gesamtausgabe sowie die digital erstmals edierten Briefe aus Jean Pauls Umfeld und stellt sie der Forschung zur freien Verfügung.

Die Briefkorpora der digitalen Edition sind mit dem Gesamtregister der historisch-kritischen Gesamtausgabe verknüpft, welches zu diesem Zwecke digitalisiert, inhaltlich erweitert und mit Normdaten angereichert wurde. Somit sind die Korpora der Von- und Umfeldbriefe einheitlich erschlossen und gemeinsam durchsuchbar.

Zukünftig soll die digitale Edition um die Briefe an Jean Paul, die als vierte Abteilung der historisch-kritischen Gesamtausgabe erschienen sind, ergänzt werden.

Register

Das Gesamtregister umfasst insgesamt 11288 Einträge, darunter

- 5651 Personen,
- 1816 Orte,
- 474 Werke Jean Pauls und
- 2461 Werke.

Zwei eigens für die Edition der Umfeldbriefe entwickelte Register verzeichnen außerdem

- 40 Korrespondenzkreise und
- 137 Themen.

Daten & Schnittstellen

Die digitale Edition stellt die Daten aller veröffentlichten Briefe unter CC BY-SA-Lizenz bereit und bietet CMIF und BEACON-Schnittstellen.

Feedback

Die Erstellung der digitalen Edition und der Datenbestände ist ein fortlaufender Prozess. Hinweise und Ergänzungen werden dankbar entgegengenommen. Bitte schreiben Sie an bernauer@bbaw.de.

(c) Extrait du site de l'édition numérique de la correspondance de Jean Paul