::Savoirs LE FIL DES IDÉES

Mettre en ligne, annoter et explorer les fiches de lecture de Michel Foucault

Actes du colloque L'archive Foucault à l'ère du numérique. Fiches et fichiers dans la pratique philosophique, sous la dir. de Laurent Dartigues et Camille Noûs, 2022

Vincent Ventresque, Marie-Laure Massot et Richard Walter

Résumé

Le projet ANR collaboratif Foucault Fiches de Lecture visait la mise à disposition sur le Web des manuscrits de fiches de lecture de Foucault, conservés à la BnF. Cette mise en ligne supposait la numérisation des manuscrits, et un travail interdisciplinaire de description et annotation. Pour ce faire, nous avons développé une plate-forme originale, basée sur les technologies du Web sémantique, qui nous a permis de mutualiser les informations ajoutées par les chercheurs pour leurs travaux d'édition et d'analyse, mais aussi d'exploiter les données bio-bibliographiques de data.bnf.fr, et enfin de produire des visualisations interactives des parcours de lecture de Foucault. Ces développements nous ont également donné l'occasion d'expérimenter Transkribus, système expert basé sur l'intelligence artificielle, qui nous a servi à transcrire automatiquement l'écriture manuscrite de Foucault, moyennant une phase d'apprentissage. Les fiches de lecture numérisées ont enfin été publiées sur la plate-forme Eman, outil générique basé sur Omeka, que nous avons adapté par des plugins. Nous souhaitons ici dresser un bilan de nos travaux et rendre compte de nos présupposés méthodologiques. En particulier, il nous paraît important de distinguer cette exposition virtuelle, qui est un objet numérique évolutif, d'une édition numérique, ainsi que d'expliciter les principes de description que nous avons adoptés pour favoriser la collaboration interdisciplinaire.

Nous souhaitons dresser ici un bilan du projet ANR collaboratif Foucault Fiches de lecture (FFL, 2017-2020), inscrit dans l'axe « Révolution numérique », puisqu'à son terme nous sommes en mesure de présenter les outils réalisés et les données produites collectivement, mais aussi de rendre compte de nos présupposés et des choix méthodologiques qui ont guidé notre travail.

Objectifs et méthodologie

Commençons par expliciter notre principal objectif, qui était de mettre à disposition sur le Web les manuscrits des fiches de lecture de Foucault. En effet, il est important de lever d'emblée une confusion possible et de distinguer cette publication d'une édition numérique. Si la mise en ligne des images numérisées comporte bien un geste éditorial, il s'agit d'une *exposition virtuelle de documents d'archives*, permettant à tout·e internaute de consulter les manuscrits à distance, en bénéficiant de l'apport des données de description (indexation, classement) pour explorer le corpus des fiches. Notre travail visait en effet d'abord la production d'une base de données destinée à inventorier les manuscrits tout en repérant les auteurs et ouvrages lus ou cités par Foucault, de telle sorte qu'une recherche sur Freud par exemple, 1) renvoie les fiches qui mentionnent Freud ou citent ses œuvres, 2) situe ces fiches dans la structure des chemises établies par Foucault et réparties dans les boîtes d'archives, et enfin 3) donne accès au document numérisé. L'exposition virtuelle relève donc bien plus d'un inventaire que d'une édition, puisqu'elle vise d'abord à signaler, indexer et donner à voir le manuscrit, et non à en présenter un substitut, une représentation textuelle, comme c'est le cas d'une édition – ce qui ne nous a pas empêché·e·s de transcrire certaines parties du corpus, nous y reviendrons.

Il faut également rappeler que le projet FFL héritait de l'approche méthodologique de la Bibliothèque foucaldienne : traiter les fiches de lecture comme un ensemble de papiers de travail *qui ne constituent pas un texte, mais plutôt un fichier* [2], car les feuillets peuvent être déplacés pour être réutilisés dans plusieurs contextes – sachant que Foucault pouvait réunir un matériau bibliographique pour écrire un livre mais retravailler ce matériau pour préparer un cours ou une conférence. Comme le corpus donne à voir le travail de recherche du philosophe et contient une collection considérable de citations, souvent relevées sans autre ajout qu'un titre thématique, il peut être considéré comme le précurseur de la collection numérique Zotero qu'utilisent les chercheurs aujourd'hui. S'inspirant du concept d'hyperdocument, le projet FFL ambitionnait de construire une plate-forme numérique intégrant ces différentes dimensions : anthologie de citations, fichier multi-indexé, séries de dossiers organisés ou tâtonnement de la recherche.

Cet outil devait permettre aux chercheurs de reconstituer les « parcours de lecture » de Foucault et d'étudier les liens entre les textes : renvois internes d'une fiche à une autre ; résonances externes d'une fiche avec un brouillon de cours ou un chapitre de livre ; ou encore, renvois entre sources primaires et secondaires. L'analyse de ces parcours permet d'interroger le rapport entre un geste répétitif, un travail quotidien et artisanal de lecture, de copie et d'archivage, et la formulation d'une thèse inédite et originale de philosophie, et, par-là, de faire un pas décisif vers une « histoire matérielle des idées ». C'est pourquoi la plate-forme ne devait pas se contenter de donner à lire des documents manuscrits, mais surtout permettre de décrire, représenter et explorer les connexions entre les entités (concepts, auteurs, œuvres, contextes d'écriture, modalités d'énonciation) qui composent les fiches, sans imposer de parcours canonique ni de prescription interprétative.

En outre, la dimension collaborative du projet FFL était primordiale. Il était inenvisageable de mettre à disposition ce corpus très volumineux, qui n'avait encore jamais fait l'objet d'une exploration systématique, sans réunir une équipe composée de spécialistes à la fois de Foucault, des archives et des humanités numériques[3]. Plutôt que de traiter l'ensemble des fiches, évalué à environ 20 000 feuillets, il a paru plus pertinent de définir un programme de numérisation ciblé sur quelques boîtes formant une unité thématique et couvrant toutes les périodes du travail de Foucault [4] : M. Senellart a proposé de sélectionner les boîtes liées au thème de l'histoire de la sexualité, puisqu'il traverse toute l'œuvre foucaldienne et recoupe aussi bien l'analyse des institutions et la théorie politique que l'histoire des sciences. Le corpus ainsi délimité comptait environ 11 000 feuillets, et il n'a jamais été question de réaliser la mise en ligne de toutes les fiches. Toutefois, en cours de projet les conditions de la numérisation ont évolué et nous avons pu dépasser notre objectif initial : sans arriver à une complétude parfaite, nous avons pu reproduire plus de 18 000 feuillets.

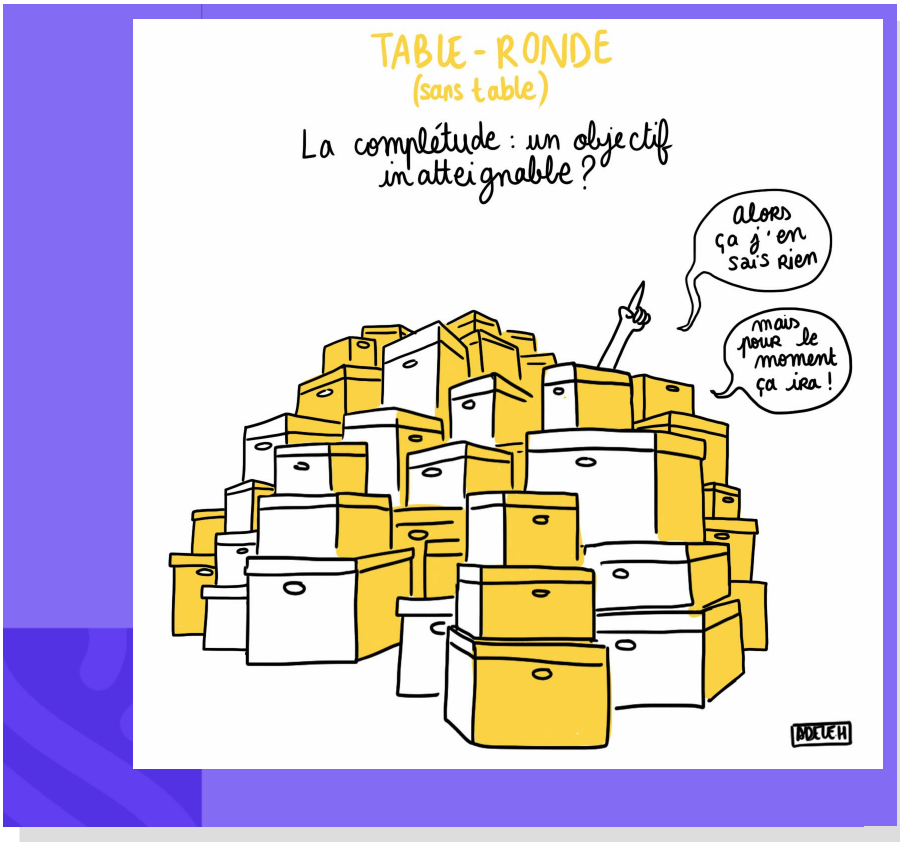

Figure 1. Figure 1 - Dessin tiré du reportage dessiné pour le colloque FFL 10-11 septembre 2020 à Caen.

Pour organiser ce travail collectif, nous avons proposé de réaliser un *traitement différencié* du corpus : 1) faire en sorte que l'ensemble des fiches numérisées soit suffisamment inventorié et indexé pour qu'un plan de classement et un moteur de recherche puissent faciliter la consultation ; 2) pour certaines chemises présentant un intérêt pour les travaux de recherche des membres du projet, proposer une description plus fine, ainsi qu'un système d'annotation. La plate-forme devait ainsi articuler plusieurs dimensions : à la fois une exposition virtuelle *et* un outil de travail, où les membres du projet pouvaient simultanément *mutualiser les données de description* et *baliser les documents pour les besoins de leur propre analyse*, constituant ainsi des sous-corpus spécifiques à une problématique de recherche. Deux grands types d'usage se sont donc rencontrés sur la plate-forme : le premier était celui des éditeurs, comme E. Basso et Ph. Sabot, qui avaient besoin de connaître les sources de Foucault pour réaliser l'édition critique des ouvrages sur Binswanger et sur la phénoménologie[5], en y précisant l'origine de certains concepts, en clarifiant certains passages où manquait la référence exacte. Cela nécessitait de sonder et décrire plusieurs boîtes (n^{os} 37, 38, 42 B), pour y retrouver les sources implicites et mesurer leur importance relative. Le second usage, plus thématique, impliquait une analyse détaillée de certaines chemises, comme celles de la boîte n^{o} 19 pour mettre en

perspective les recherches de Foucault sur le néolibéralisme ; ou encore, exhumer les pistes abandonnées autour du concept de biopolitique[6].

Le prototype FFL : description, annotation, exploration

Pour réunir en un seul espace les images, les données de description et les annotations produites au fur et à mesure par l'équipe, nous avons donc développé un prototype de plate-forme collaborative qui a permis à une trentaine d'utilisat·eur·rice·s – dont une dizaine de personnes très actives, chercheur·e·s et ingénieur·e·s – de travailler simultanément sur les manuscrits des fiches[7]. Ce prototype et le corpus ont donc constitué un *objet numérique en constante évolution*, à plusieurs niveaux : ajouts successifs de nouvelles images et informations, mais aussi développement par étapes de nouvelles fonctionnalités.

Au terme de ce travail collectif, nous sommes parvenus à produire la description de plus de 12 900 fiches, ce qui implique au moins un titre, et, si possible, une ou plusieurs mentions de personne et références bibliographiques (12 966 titres transcrits, 7486 références, 9168 mentions de personnes). Sur les 29 boîtes numérisées, 18 ont pu être décrites de manière complète (soit 11 436 feuillets répartis dans 491 chemises, chaque fiche ayant au moins un titre), 5 boîtes ont été décrites partiellement (entre 20 et 40 % de fiches décrites). Enfin, 6 boîtes ont été seulement dépouillées : nous en avons réalisé le plan de classement pour faciliter la consultation et donner un aperçu des auteurs et thèmes traités par Foucault. Nous avons également produit u n *index nominum* de plus de 2600 entités personnes et un *index bibliographicum* de plus de 2900 entités documents. En outre, cet index a été enrichi grâce aux informations biographiques et bibliographiques de la BnF : le projet comportait un axe d'expérimentation des technologies du web de données, et nous avons « aligné » les entités de nos index avec les données de data.bnf.fr (1255 personnes et 764 documents alignés).

Ces résultats n'ont été possibles qu'au prix d'une certaine liberté donnée aux membres de l'équipe : il s'agissait de proposer un outil comportant *un minimum de contraintes*, et où l'application de chaque règle garantisse à la fois un gain de temps et une simplicité d'usage. Par exemple, chacun·e devait utiliser une liste déroulante pour les entrées d'index qui présentait plusieurs avantages : 1) l'auto-complétion facilitait la saisie, 2) elle assurait l'homogénéité de la

description, 3) elle donnait la possibilité de corriger en une seule fois l'ensemble des occurrences du nom de personne ou de la référence d'un document. De même nous invitions les utilisat·eur·rice·s à enrichir l'index en associant les personnes et documents avec les notices de la BnF. D'une part ce travail « d'alignement » était facilité par un module de recherche pour interroger directement l'entrepôt de data.bnf.fr sans quitter la page de la fiche de lecture et supportant des requêtes sur des mots tronqués, ce qui est très pratique mais n'est pas possible dans le catalogue en ligne de la BnF. D'autre part, une fois l'alignement effectué, les informations de la BnF étaient automatiquement récupérées et consultables dans le prototype, ce qui est extrêmement utile par exemple lorsque Foucault cite par exemple Pinel sans préciser le prénom, et qu'il faut pouvoir distinguer entre Casimir, Philippe et Scipion. Ou encore, il était recommandé d'utiliser des crochets pour signaler l'intervention du scripteur dans le relevé des titres. Il existe évidemment des différences d'appréciation dans la manière de développer une abréviation ou d'indiquer un passage illisible, mais comme elles n'ont d'autre utilité que de donner accès au feuillet original, nous n'avons pas jugé nécessaire d'uniformiser *a posteriori* ces transcriptions de titres.

Nous avions fait d'emblée le pari de *l'économie de la contribution* (au sens de Bernard Stiegler) : en mutualisant les interventions individuelles, on prend certes le risque d'obtenir une description archivistique peu conforme aux normes en vigueur, mais les informations ajoutées, même imparfaites, aident chaque contribut·eur·rice à s'approprier le corpus et à trouver plus rapidement de nouveaux documents pertinents pour sa recherche – lorsque l'un·e ajoute par exemple des titres, ils sont immédiatement exploitables par le moteur de recherche. Le défi technique et architectural consistait alors à articuler le travail personnel de chacun·e, qui abordait le corpus avec ses propres objectifs de recherche, et la production collaborative de données partagées, communes. Plutôt qu'un contrôle *a priori*, qui aurait considérablement alourdi le processus collaboratif, nous avons préféré opérer certains traitements – notamment, développer des fonctionnalités pour détecter et fusionner les doublons, achever la description de certaines boîtes qui étaient presque entièrement traitées.

Et surtout, nous avons développé un modèle de données original, permettant d'apporter un *regard réflexif* sur les informations produites, et tirant partie des technologies du web de données. S'inspirant d'un mouvement de fond dans l'histoire de la bibliothéconomie et de l'archivistique, qui tend à diviser la *notice* de catalogue en ses composantes élémentaires pour en faire des informations indépendantes et réutilisables dans le web de données, notre système propose comme unité élémentaire *l'annotation*, et non une grille composée de champs d'annotation ou description. Pratiquement, plutôt que plusieurs personnes renseignent les champs d'une même

notice, chacun·e crée une annotation indépendante qui possède un identifiant unique et qui peut être associée (ou non) à d'autres. Chaque information ainsi ajoutée possède *ses propres métadonnées*, qu'il s'agisse d'un élément de description (titre, référence, mention de personne), d'une note éditoriale (signalement d'un problème, question à l'équipe…) ou enfin d'un élément d'analyse (mot-clé ajouté, mémo, commentaire personnel).

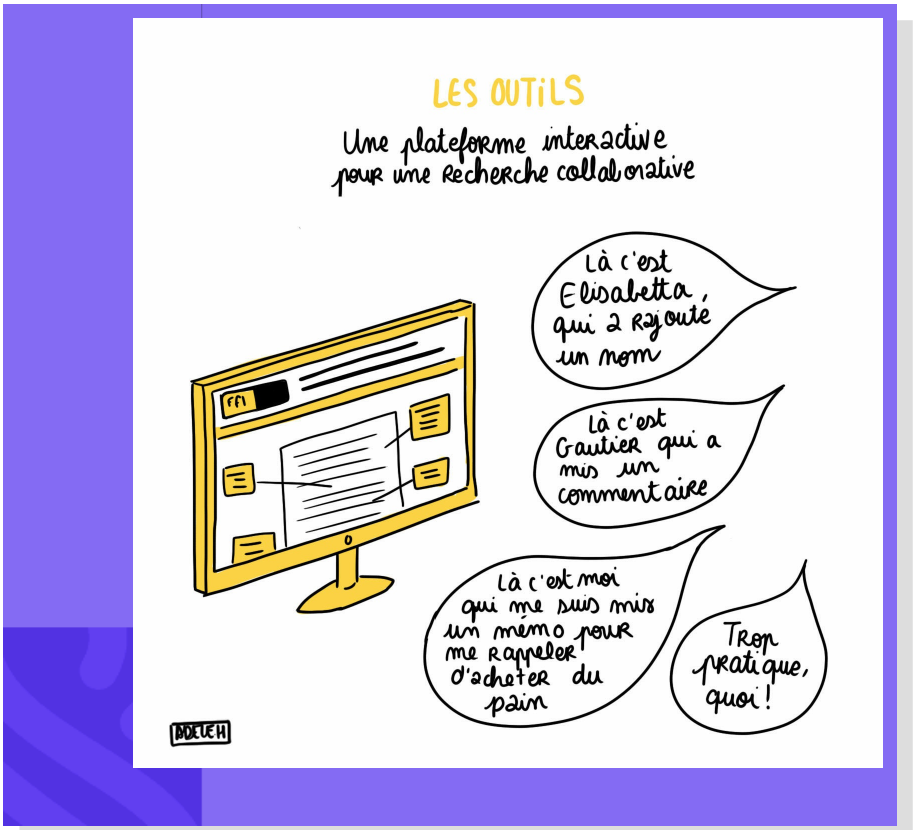

Figure 2. Figure 2 - Dessin tiré du reportage dessiné pour le colloque FFL 10-11 septembre 2020 à Caen.

Soient par exemple deux contribut·eur·rice·s A et B. Lorsque A enregistre une annotation « a1 : titre » pour une fiche F1, B peut simultanément enregistrer une annotation « b1 : commentaire personnel » ou « b2 : référence bibliographique ». A et B décident chacun·e du degré de visibilité de leur contribution : un titre est par défaut public, et un commentaire personnel privé, mais par exemple le commentaire pourrait très bien être partagé publiquement si B le choisit. À chaque annotation est ainsi affecté un ensemble de métadonnées qui en établissent la provenance (auteur, date), le degré de visibilité (public, restreint, privé), le type (titre, commentaire, mot-clé…) ; en outre, pour certaines, il est possible d'affecter un degré de fiabilité : ainsi, lorsque les indications données par Foucault dans la fiche ne permettent pas de déterminer avec certitude l'édition qu'il a consultée, la contribution peut être qualifiée de « référence approximative ».

L'indépendance des annotations produites par A et B pour la fiche F1, ainsi que l'existence des métadonnées associées à chaque annotation, permettent de modeler à volonté non seulement l'affichage des documents, mais aussi de personnaliser la recherche et l'exploitation des données. En effet, le premier bénéfice de ce système, qui offre une granularité supérieure à celui de la notice, est de restituer à chacun·e l'historique complet de ses interventions, et de séparer ses contributions de celles des autres contribut·eur·rice·s. De même, il est facile de restreindre la recherche de fiches au seul ensemble d'annotations ajoutées par telle et telle personne, ou après telle date. Ainsi se constituent plusieurs vues sur la même fiche, selon que l'on décide d'agréger ou non les informations produites par A avec les informations produites par B. Et puisqu'il n'existe pas de règle de catalogage imposant de remplir des champs préexistants et considérés comme indispensables, chacun·e décide d'utiliser plutôt des mots-clés, des commentaires personnels, des titres, selon les besoins de son propre travail. Enfin, il est possible à tout moment de mesurer précisément l'avancement du travail et la « couverture » du corpus, et de savoir, par exemple en utilisant le moteur de recherche pour explorer la liste des œuvres lues par Foucault, que les résultats renvoyés interrogent les 7486 références bibliographiques enregistrées pour 18 986 fiches numérisées, ou que la boîte n° 38 contient 699 références pour 929 feuillets, alors que la boîte n° 14 n'en contient aucune.

Cette réflexivité sur les données se traduit également par le recours à un *dispositif de visualisation cartographique*, qui nous a été particulièrement utile pour repérer des disparités de traitement, des anomalies résultant d'erreurs de saisie, même si nous l'avions d'abord conçu pour l'analyse scientifique du corpus. En effet, une idée centrale du projet FFL, qui rejoint d'ailleurs la conception foucaldienne du livre [8], était de représenter sous forme de réseaux les connexions entre les documents et de donner une vue synoptique du corpus, en dessinant les ensembles de fiches citant un même document ou mentionnant une même personne. La plate-forme donne la possibilité de générer et manipuler de manière interactive des graphes orientés construits sur le principe suivant : si les fiches F1 et F5 citent Freud, chacun des trois éléments étant représenté par un nœud, dessiner les arêtes « F1 → Freud », « F5 → Freud ». En appliquant ce principe à l'ensemble des annotations, on obtient des réseaux de co-occurrences, qui contextualisent les documents et permettent de répondre à la question : quels sont les auteurs et œuvres cités dans les autres fiches qui citent Freud et quelles sont leurs fréquences relatives ? Ou encore : quelle position cette fiche occupe-t-elle dans le réseau des auteurs, œuvres et concepts ; à savoir : quelles autres fiches mentionnent les mêmes auteurs ou œuvres, et que mentionnent-elles d'autre ?

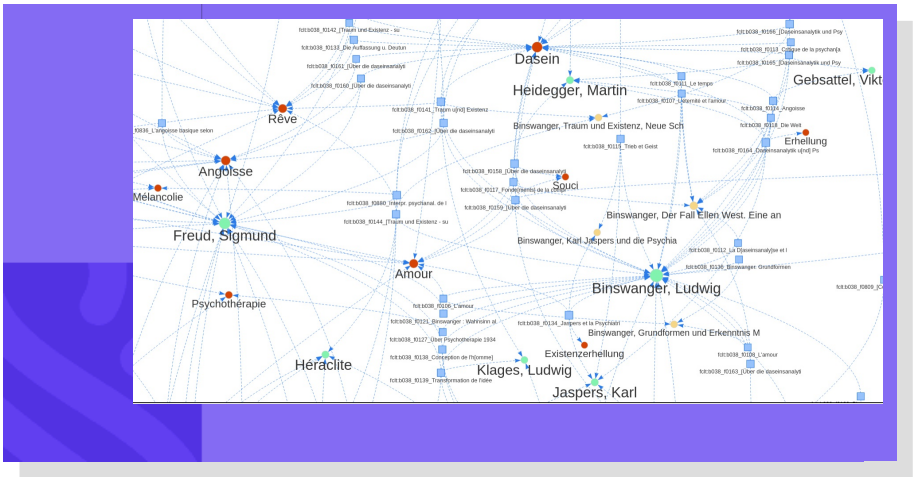

Figure 3. Figure 3 - Lectures autour de Binswanger et Freud dans la boîte n° 38.

Bien entendu les cartographies ainsi constituées dépendent elles aussi du point de vue qui a été adopté lors du travail de description et d'annotation et, de plus, leur interprétation requiert une certaine prudence pour ne pas se laisser illusionner par le caractère suggestif de ces représentations : certains rapprochements résultent des algorithmes de spatialisation, et un nœud peut se trouver au voisinage d'un autre sans présenter pourtant aucune connexion avec lui. Moyennant quelques précautions et un travail de retour aux documents sources – qui est facilité par les liens depuis chaque nœud-fiche vers la page de consultation –, les cartographies ont une fonction heuristique et exploratoire, en ce qu'elles font apparaître immédiatement des recoupements que les listes de résultats du moteur de recherche contiennent de manière seulement implicite ou potentielle. D'une part les cartographies ne sont rien d'autre qu'une mise en forme des résultats que l'on pourrait obtenir par plusieurs requêtes soumises à un moteur de recherche : elles organisent ces résultats en distribuant les fiches selon des rapports de co-présence entre concepts, auteurs et documents. Elles ne doivent pas être comprises comme des représentations figées et complètes, mais bien comme un outil de recherche dynamique, puisqu'il est possible de supprimer et ajouter de nouveaux nœuds sur un graphe, et de redessiner de nouveaux graphes en relançant la recherche autour de n'importe quel nœud. D'autre part cette fonctionnalité de visualisation formalise les repérages effectués pendant la consultation des fiches, et aide à tester des hypothèses : ainsi l'ajout de mots-clés aboutit au dessin d'une taxinomie qui répartit les fiches en régions thématiques, où l'on peut identifier des périodes correspondant aux sources travaillées par Foucault, ainsi que des auteurs servant de point de départ ou de commentaire pour l'exploration d'autres auteurs.

En effet, les graphes donnent à voir un contraste entre certains

ensembles de fiches centrés autour d'un auteur, et d'autres où les recroisements sont très nombreux parce que plusieurs auteurs sont abordés dans une même fiche : ce contraste est particulièrement intéressant pour l'étude du rôle de la littérature secondaire et de l'intertextualité. La cartographie de la boîte n° 38, par exemple, montre comment Foucault lit Binswanger en parallèle avec Freud et Heidegger, sur les thèmes du rêve et de l'angoisse existentielle. Ou bien, dans le cadre de la préparation du cours *Naissance de la biopolitique,* comme le montre la cartographie de la boîte n° 19, il s'intéresse à une série de concepts clés – comme État, intervention, planification – qui sont identifiés à partir de la lecture d'un livre de littérature secondaire de François Bilger, mais également travaillés par une lecture directe des auteurs (Wilhelm Röpke, Friedrich Hayek, Ludwig von Mises).

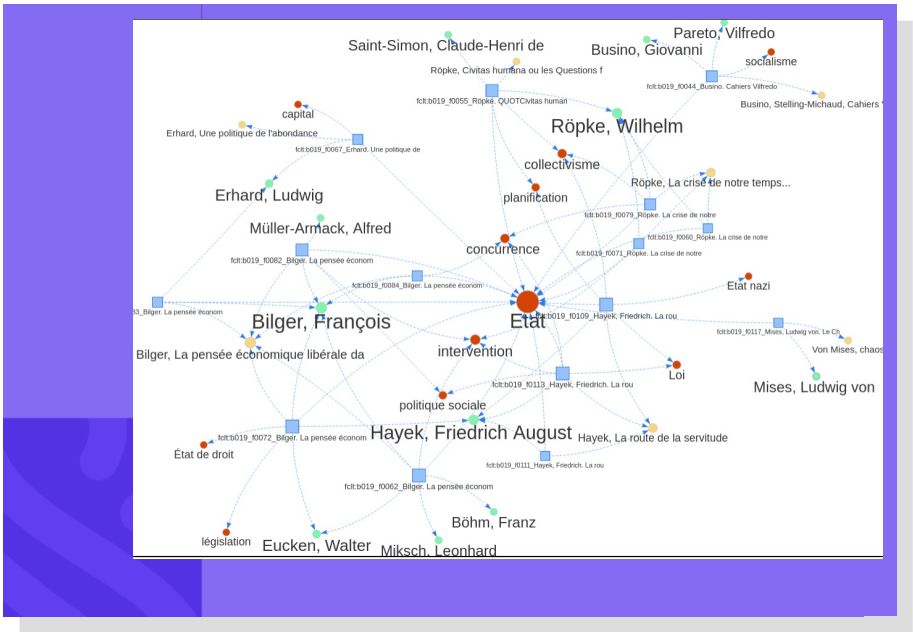

Figure 4. Figure 4 - Lectures autour des notions d'État et de néolibéralisme dans la boîte n° 19.

Ces développements originaux réalisés pour la plate-forme expérimentale ont donc été féconds pour l'appropriation du corpus, et ont grandement facilité la production collaborative des données de description, ainsi que le travail d'annotation des chercheur·e·s. Toutefois, les données ainsi produites et mutualisées résultent d'une *sélection* opérée par l'équipe, et ne donnent qu'un aperçu des contenus : titres, références, noms de personnes. Or le projet FFL souhaitait également donner les moyens d'explorer la dimension textuelle du fichier foucaldien, son lexique et la matérialité du travail du philosophe : sa manière de prélever des citations, de résumer ou catégoriser, d'abréger… C'est pourquoi nous avons expérimenté l'outil Transkribus, qui donne une autre perspective sur la prise de notes effectuée par Foucault, comme nous allons le voir.

Autre limitation importante de la plate-forme collaborative : elle n'a pas été conçue comme un outil pérenne, devant être accessible au grand public, mais comme un espace de travail et d'expérimentation limité aux membres du projet et destiné à l'analyse scientifique. De ce fait, nous avions prévu dès le départ de mettre en ligne le corpus sur deux autres supports plus pérennes et ouverts : la bibliothèque numérique Gallica[9] et le site FFL sur la plate-forme publique EMAN. Cette dernière, que nous allons également présenter, contient l'ensemble des données de description qui ont été mutualisées par les membres de l'équipe, et que nous avons isolées des annotations privées, puis exportées vers la base de données gérant le site FFL sur EMAN.

Transkribus : rendre lisible et indexer le texte des fiches

Développé dans le cadre du projet européen READ[10] (H2020), le système expert Transkribus promettait de révolutionner le travail sur les documents d'archives grâce à l'intelligence artificielle et à ses fonctionnalités de reconnaissance automatique d'écriture manuscrite (Handwritten Text Recognition, HTR). Une collaboration avec l'équipe du projet READ, entamée en 2017 nous a permis d'expérimenter Transkribus sur les fiches de lecture de Michel Foucault, pour voir dans quelle mesure ce logiciel pouvait nous aider à « lire » et à indexer les manuscrits du philosophe.

Figure 5. Figure 5 - Rendre lisible et indexer le texte des fiches.

Transkribus est une plateforme complète pour la reconnaissance, la transcription et la recherche automatisées d'archives manuscrites ou imprimées. Elle facilite la lecture, l'enrichissement et l'exploration des textes. Fondé sur les technologies du « *deep learning* », le moteur HTR doit être « entraîné » avec des données d'apprentissage, obtenues par la transcription manuelle d'une centaine de pages minimum, en établissant la correspondance ligne à ligne entre l'image du texte numérisé et sa transcription (segmentation). Comment faciliter la lecture et l'analyse d'un corpus aussi volumineux et fragmenté que le fichier foucaldien des notes de lecture, en particulier pour les lecteurs non habitués à l'écriture manuscrite du philosophe ?

Dans le cadre du projet, la pertinence de la transcription a parfois fait débat. Il n'était pas envisageable de transcrire manuellement la totalité des fiches de lecture et cela ne semblait pas utile à certains de passer autant de temps sur de simples notes de lecture où Foucault, pour l'essentiel, recopie des citations. En outre les ayants droit de Foucault ne souhaitaient pas que soient publiées des transcriptions imparfaites. Néanmoins, l'idée de collaborer à un projet européen innovant et de faire un test avec le corpus des fiches de lecture de

Foucault a été retenue. Cette collaboration était d'autant plus intéressante que le projet READ était encore dans une phase expérimentale : nous avons réalisé les premières transcriptions de manuscrits français modernes, et vu évoluer à la fois les fonctionnalités de l'outil et ses performances. Nous étions particulièrement intéressés par la recherche d'occurrences, d'auteurs, de concepts ou bien d'ouvrages, mais aussi par le travail que nous pourrions envisager sur le plein texte : analyses textométriques du lexique et repérage de styles de citations notamment. Ce dispositif promettait donc de faciliter l'exploration de manuscrits encore jamais étudiés pour les sonder automatiquement, mais aussi d'en faciliter la lecture, et d'accélérer considérablement le travail de transcription en fournissant un premier jet qu'il ne resterait plus qu'à corriger.

Nous avons créé les données d'apprentissage de l'écriture de Foucault en plusieurs étapes : dans un premier essai, l'équipe a formé un « modèle » de reconnaissance automatique de texte (HTR) en utilisant Transkribus pour segmenter et aligner 200 images[11] dont nous possédions une transcription (ce qui a considérablement réduit le temps de travail). À notre surprise, car il faut avouer que nous étions sceptiques vu la difficulté à déchiffrer l'écriture de Foucault, ce premier modèle s'est révélé capable de générer des transcriptions avec un taux d'erreur de seulement 15 % par caractère (85 % de caractères transcrits correctement par Transkribus). Nous avons donc ajouté 400 images supplémentaires, que nous avons transcrites manuellement, ce qui a permis une amélioration significative des résultats puisque nous avions alors un modèle HTR qui pouvait transcrire l'écriture de Foucault avec un taux d'erreur de caractères de seulement 8 % (c'est-à-dire 92 % des caractères transcrits correctement par Transkribus). Nous avons ensuite transcrit automatiquement de nouvelles images, et corrigé les transcriptions automatiques afin d'entraîner de nouveaux modèles. En outre les informaticiens de l'équipe de Transkribus ont amélioré progressivement leur modèle (HTR+). Aujourd'hui le taux d'erreur sur les caractères transcrits automatiquement par Transkribus est uniquement de 5 %[12].

La première et la plus importante des difficultés que nous avons rencontrées est bien sûr l'écriture du philosophe. Les fiches contiennent de nombreuses abréviations souvent difficiles à déchiffrer, qui peuvent signifier différents mots dans différents contextes. Foucault a également écrit plusieurs lettres de la même manière. Pour le test Transkribus, nous avons donc dû marquer par un système de balises les mots illisibles ou douteux, afin que le logiciel puisse les ignorer lors de l'entraînement d'un modèle. La transcription à plusieurs mains a nécessité par ailleurs de bien s'entendre sur la manière de préparer le corpus. Malgré ces difficultés, les résultats des tests Transkribus ont été très encourageants. Alors qu'il nous fallait

parfois plus de quarante minutes pour transcrire manuellement une seule fiche de lecture de Foucault, il suffit maintenant de cliquer sur un bouton et la transcription se fait automatiquement sur un lot choisi de fiches.

Il faut noter qu'il existe une certaine différence d'efficacité selon les différents types de manuscrits : l'écriture de Foucault change avec le temps, et le logiciel est sensible à la transparence du papier, au contraste de l'encre. Ainsi sur la figure 6 proposée ci-dessous, il s'agit d'une image avec une écriture peu lisible, très différente de l'écriture sur laquelle nous avons entraîné le modèle et, pourtant, les résultats sont très encourageants. Sur les fiches qui se rapprochent plus des fiches sur lesquelles le modèle a été entraîné, les résultats de transcription automatique sont excellents[13] (fig. 7)

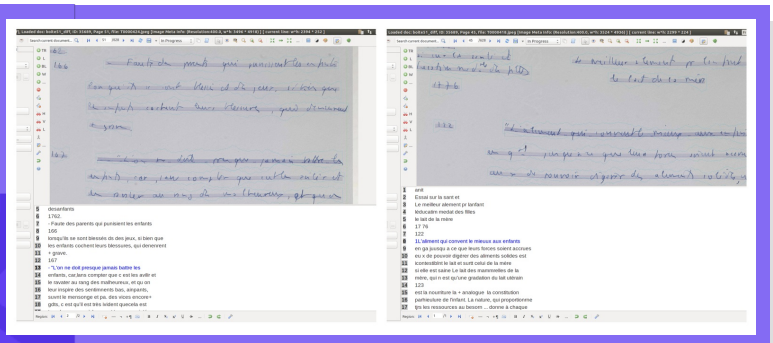

Figure 6. Figure 6 - Transcription automatique, boîte n° 51.
Capture d'écran de l'interface de Transkribus.

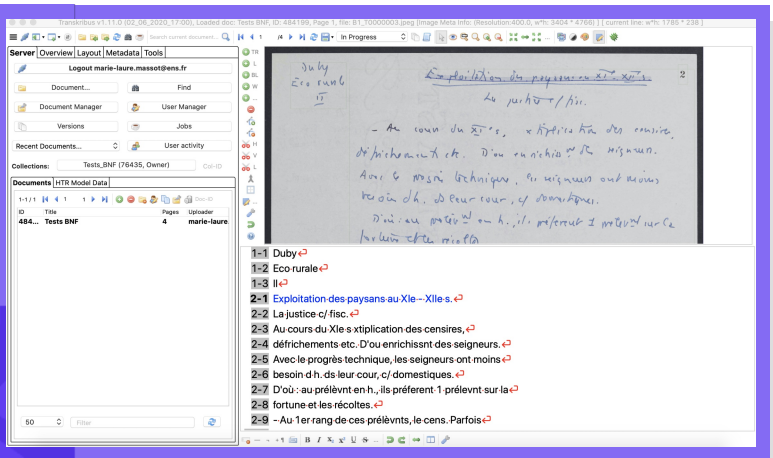

Figure 7. Figure 7 - Transcription automatique, boîte n° 1.
Capture d'écran de l'interface de Transkribus.

La fonction de repérage de mots-clés (*Keyword Spotting*, KWS) de Transkribus est un nouvel outil de recherche puissant pour rechercher des mots distincts dans une collection de documents. L'originalité de ce type de recherche repose sur le fait qu'elle est basée sur un système de comparaison des images et qu'elle est plus performante que les algorithmes classiques qui comparent les similarités entre un terme cherché et les termes de l'index (*fuzzy search*). Même si la transcription générée automatiquement contient des erreurs, KWS trouvera de manière fiable des mots, des phrases et même des parties de mots et des expressions régulières dans vos documents. De plus, il vous donnera un indice de fiabilité pour hiérarchiser les résultats. Sur la figure 3, nous avons recherché les occurrences des mots-clés « justice » et « seigneur ». Sur la fenêtre de résultats, nous voyons que même une transcription automatique fautive (« **c**eigneur », p. 457) est bien reconnue par KWS. Une image du mot est proposée en regard pour permettre une vérification.

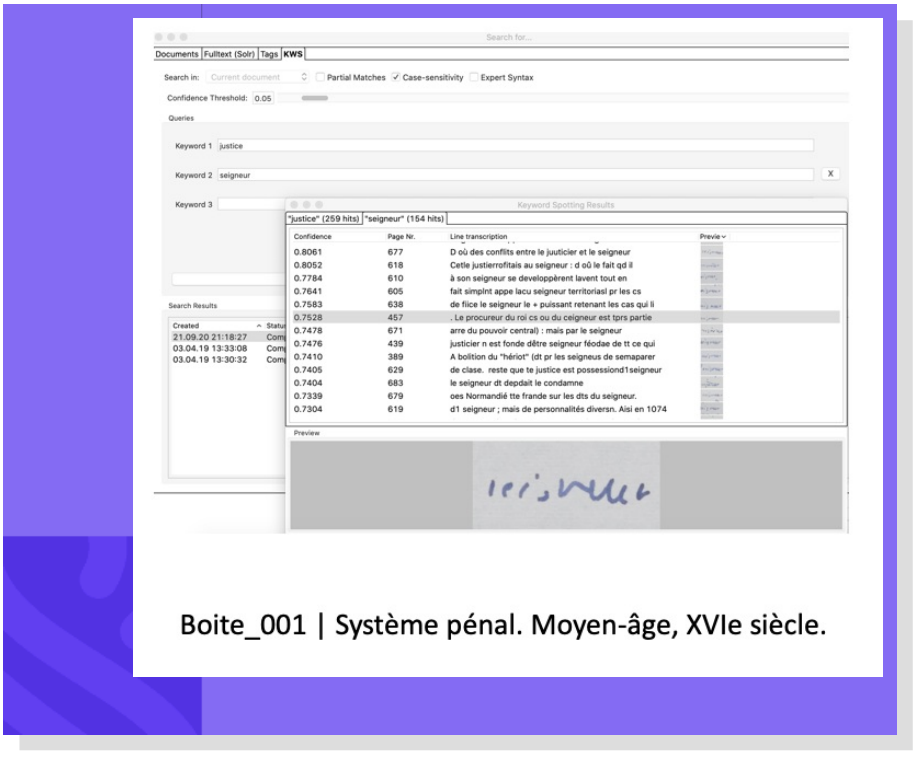

Figure 8. Figure 8 - Exemple de repérage de mots clés. Capture d'écran de l'interface de Transkribus.

Malgré l'imperfection des transcriptions, le dispositif s'avère extrêmement utile pour l'exploration d'un corpus. Comme nous le disions au sujet du traitement différencié pour la description des fiches et des cartographies, il faut garder en mémoire l'aspect heuristique et exploratoire des outils numériques. Ainsi l'absence de résultat sur une recherche d'occurrence n'est pas significative, puisque le terme peut avoir échappé à l'indexation ; en revanche, la présence d'une occurrence indique souvent une piste d'analyse et permet de faire apparaître des documents qu'on n'aurait pas songé à

consulter.

Non seulement la collaboration avec l'équipe READ et le test de Transkribus ont été une expérience très intéressante, mais ils ont totalement modifié notre point de vue sur le corpus et son exploitation. Si nous avions disposé de cet outil au début du projet, nous aurions pu partir du texte imparfait mais intégralement transcrit, nous contenter de corriger les titres de fiches et d'extraire les noms de personnes et les références de manière semi-automatique. En outre, il reste possible d'utiliser le modèle HTR pour de nouveaux projets d'édition ou des travaux d'analyse textométrique, ou encore de publier sur EMAN des transcriptions automatiques validées par les ayants droit. Notre expérimentation a également joué un rôle important pour d'autres projets d'humanités numériques hébergés sur EMAN.

EMAN : un outil modulaire pour la publication de collections numériques

EMAN[14] est une plate-forme permettant la mise en ligne et le travail collaboratif sur des corpus numérisés. Conçue et administrée par Richard Walter, elle compte actuellement plus d'une quarantaine de projets d'humanités numériques, dont le projet FFL. Elle offre une garantie de pérennité des données et de leur accessibilité : hébergée par Huma-Num[15], elle a été développée sur des bases open source et des formats interopérables (métadonnées Dublin Core et transcriptions XML/TEI). En outre, son architecture modulaire permet de personnaliser les fonctionnalités de consultation et de travail, ainsi que la navigation dans les corpus, puisqu'elle est basée sur le logiciel Omeka[16] auquel peuvent être ajoutés de nombreux plugins[17], dont une grande partie spécifiquement conçus pour la communauté EMAN.

Dans le cadre du projet FFL, nous avons choisi d'utiliser EMAN comme le support public final des images et données produites collaborativement dans le prototype. Sans entrer dans le détail technique des opérations, il faut commencer par rappeler ici que l'interopérabilité des outils est un enjeu majeur pour tout projet d'humanités numériques : exporter vers un autre support un corpus numérique aussi volumineux que celui des fiches de lecture serait difficile voire impossible sans la souplesse de systèmes comme Omeka et les technologies RDF. Comme indiqué plus haut, le modèle de données du prototype permettait d'isoler très facilement certains types d'information et de ne retenir que les contributions qui avaient une visibilité publique. Nous avons donc exporté les données suivantes : classement (chemise ou boîte d'origine du feuillet) ; titre du feuillet ; références bibliographiques et alignements avec data.bnf.fr ;

mentions de personnes ; remarques éditoriales. Omeka est construit sur un système arborescent classique de collections contenant des ressources numériques appelées « items » et décrites par des notices – sachant qu'une arborescence de sous-collections est facile à créer. Dans notre cas, un item est un feuillet recto-verso, soit deux images et une notice. Il a donc suffi d'agréger les données RDF de chaque feuillet pour recomposer une notice (abstraction faite des informations de provenance, à savoir la date et l'auteur·e des contributions). L'intégration finale des images et des notices sous formes d'items dans la plateforme publique EMAN a été facilitée par l'utilisation du plugin CSV Import, qui nous a d'ores et déjà permis de mettre en ligne 19 boîtes. Une limite importante de ce plugin est qu'il ne tient pas compte des sous-collections ; cependant la simplicité de la base de données d'Omeka nous a permis, grâce à une seule requête, de reclasser automatiquement les feuillets dans des sous-collections correspondant aux chemises.

La plus grande partie des fiches numérisées sont désormais en ligne et consultables par tout·e internaute, et la circulation dans le corpus est possible selon deux modalités : le plan de classement et les outils de recherche. L'arborescence du corpus est en effet exploitée à deux niveaux : une page qui permet de déplier le plan de classement et donc de développer chaque boîte en chemises, puis chaque chemise en feuillets ; et pour chaque boîte, une page « collection », qui renvoie à la fois à l'ensemble des feuillets et à la liste des chemises (sous-collections). En outre, une présentation a été rédigée pour présenter le contenu et l'intérêt scientifique de chaque boîte, et nous avons y ajouté quelques éléments chiffrés pour rendre compte de la couverture du corpus (nombre d'annotations et références) ainsi que des auteurs et œuvres les plus cités. Cette reconstitution du plan de classement a donc l'avantage d'offrir une vue d'ensemble du corpus et de respecter les chemises constituées par Foucault, et de faire apparaître ainsi les regroupements thématiques effectués par le philosophe ; par opposition aux outils de recherche, qui nécessitent de poser une question aux contenus indexés et donc de savoir déjà quoi chercher, cette vue d'ensemble donne la possibilité de découvrir le corpus en le feuilletant.

L'autre accès aux manuscrits des fiches est assuré par les dispositifs d'indexation, avec au premier chef un moteur de recherche SolR intégré au site, à la fois pour le plein texte, à savoir la recherche « simple » d'un terme dans tous les champs confondus, et pour l'exploitation des données structurées, en les interrogeant via une fonctionnalité de « facettes ». Ainsi, une recherche simple sur le mot « Freud » renvoie 333 résultats : soit le terme figure dans le titre du feuillet, soit dans celui de la chemise qui le contient, ou encore dans les mentions de personnes et les références bibliographiques. Les résultats peuvent être triés grâce à un système de filtres (facettes) en choisissant un ou plusieurs critères, comme la collection (boîte ou

chemise), les références ou les noms de personnes, par exemple pour extraire les sept fiches qui citent également Binswanger.

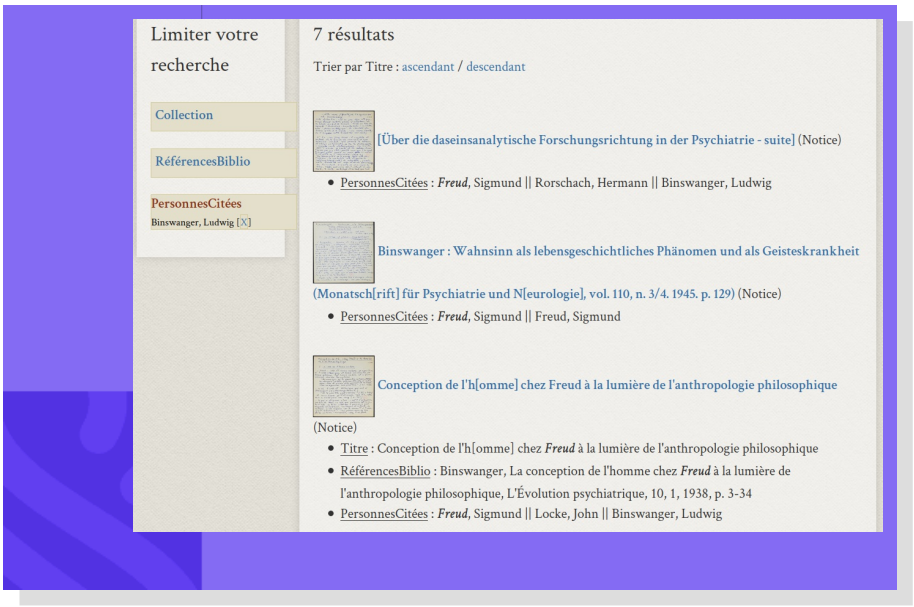

Figure 9. Figure 9 - Recherche à « facettes » : trouver les fiches qui mentionnent Binswanger parmi celles qui mentionnent Freud.

De plus, tout comme l'arborescence permet de feuilleter le corpus selon son classement, le plugin EMAN Index (page « index des valeurs ») donne à voir l'ensemble des références et personnes citées, en précisant pour chacune le nombre d'occurrences et les titres des fiches, qui sont accessibles par un simple clic. Ces informations sont parfois redondantes avec les titres donnés aux feuillets ou aux chemises, mais elles les complètent souvent : d'une part les titres n'indiquent pas toujours la source principale mise en fiche, d'autre part de nombreuses fiches citent plusieurs ouvrages ou auteurs non indiqués dans le titre. Il est à noter que pour des usages plus avancés, les internautes peuvent exporter les données en différents formats tabulaires ou XML, ou demander au comité éditorial l'ouverture d'un accès à l'interface privée d'EMAN, qui propose un moteur de recherche supportant des requêtes plus complexes, et donne la possibilité de participer à la correction ou à l'enrichissement des notices.

Enfin, le projet FFL a contribué à faire évoluer la plateforme EMAN. En effet, plusieurs plugins et fonctionnalités[18] ont été inspirés par les fonctionnalités du prototype, et développés avec les soutiens financier et technique du projet FFL. Ces plugins sont diffusés en open source et réutilisables par chacun des 35 projets de la communauté EMAN, et, plus largement, pour tout·e utilisa·eur·rice du logiciel Omeka.

Ainsi la réflexivité par rapport aux données et à leur production, qui était assurée par le modèle RDF du prototype, s'est traduite en deux plugins apportant une traçabilité des interventions des

utilisat·eur·rice·s. Le plugin *Historique* [19] permet de repérer les dernières modifications et de savoir qui en est l'auteur, ce qui facilite grandement les travaux de correction et d'harmonisation des notices, ainsi que l'organisation du travail en équipe. Le plugin *Bookmarks*, centré au contraire sur le travail personnel,permet de lister les contenus créés ou modifiés par l'utilisat·eur·rice, d'enregistrer des marque-pages et des notes personnelles, et de comparer des notices. De cette manière, toutes les informations liées au travail sur les collections numériques sont rassemblées au même endroit et accessibles en ligne.

Par ailleurs, la réflexion sur la réutilisation des données nous a inspiré deux autres plugins, le premier concernant les valeurs saisies dans la base de données du site FFL sur EMAN et le second exploitant les données en ligne de data.bnf.fr. *EMAN Index*, déjà mentionné ci-dessus pour le cas des références et des personnes, permet de générer en temps réel la liste complète des valeurs utilisées pour un champ de notice, et pour chaque valeur de renvoyer vers ses occurrences dans les collections, de les modifier au besoin. En outre, à cet index correspond une fonctionnalité d'auto-complétion pour faciliter la saisie d'informations dans les notices : cela évite l'effort inutile de retaper une information déjà enregistrée, ainsi que les fautes de frappe. Enfin, le plugin *BnfMashup* permet d'afficher automatiquement des données tirées de data.bnf.fr. Grâce aux alignements bibliographiques évoqués plus haut, qui consistent à associer à un document l'identifiant de sa notice BnF, les technologies du Web de données rendent possible d'incruster à côté de la fiche de lecture la reproduction numérique (stockée dans Gallica) du document lu par Foucault, et les informations bibliographiques (auteur, titre, date, tout ce qui est disponible dans la notice BnF du document, stockées dans l'entrepôt de data.bnf.fr).

En conclusion, l'objet numérique « corpus des fiches de lecture numérisées » reste en définitive un *work in progress*, avec ses imperfections et ses incomplétudes, et ne peut être comparé à une édition numérique classique : il s'agissait de donner la possibilité de consulter les manuscrits et de les analyser, tout en mutualisant les informations de description. Cependant, à la différence de nombreux outils numériques qui utilisent des données d'indexation pour faciliter l'exploration de ressources en ligne sans éclairer l'utilisat·eur·rice, notre travail a intégré dès le départ une exigence de réflexivité sur la production des informations, et il est possible d'en connaître précisément le degré de complétude, l'origine et la fiabilité.

Notes

1. Bachimont, 2007, p. 237-238.

2 . Bert, 2016. J.-F. Bert avait participé au projet LBF http://lbf-ehess.ens-lyon.fr .

3. Voir : https://eman-archives.org/Foucault-fiches/mapage/6

4 . La boîte n° 38 (voir https://eman-archives.org/Foucault-fiches/collections/show/261) contient des notes prises par Foucault lorsqu'il était normalien, probablement à partir de 1944 pour les plus anciennes. La boîte n° 28 contient les toutes dernières notes prises par Foucault, probablement autour de 1983-1984 (voir https://eman-archives.org/Foucault-fiches/collections/show/703).

5. Foucault, 2021.

6. Respectivement : C. Verlengia, thèse en cours, chemises de la boîte n° 9 ; L. Paltrinieri, analyse de la lecture de Malthus et chemise « Population » dans les boîtes nos 18 et 19. Signalons également les travaux de G. Dassonneville sur le jeune Foucault (boîtes nos 37-38) et de J. Guilhaumou sur la réception foucaldienne de la philosophie du langage (boîte n° 43), qui ont donné lieu à des expositions thématiques (voir http://eman-archives.org/Foucault-fiches/exhibits).

7. Prototype conçu et développé par V. Ventresque, entièrement basé sur le triple store Fuseki de la fondation Apache, et développé en PHP-JavaScript (Framework php léger Code Igniter. Librairie Vis.js pour les visualisations, et Jquery pour l'interactivité des pages web). Pour des précisions sur ses fonctionnalités et le modèle de données, évoqués au cours de cet article, voir ce billet sur le carnet du projet : https://ffl.hypotheses.org/2153. Le billet présente également Transkribus et EMAN.

8. Foucault, 2015, p. 23-24 : « C'est que les marges d'un livre ne sont jamais nettes ni rigoureusement tranchées : par-delà le titre, les premières lignes et le point final, par-delà sa configuration interne et la forme qui l'autonomise, il est pris dans un système de renvois à d'autres livres, d'autres textes, d'autres phrases : nœud dans un réseau. »

9 . Voir la présentation donnée par Laurence Le Bras : https://gallica.bnf.fr/blog/18112020/les-fiches-de-lecture-de-michel-foucault. Actuellement 21 boîtes sont en ligne ; pour les retrouver, indiquer la cote du fonds Foucault dans le moteur de recherche : NAF 28730.

10. Le projet READ (*Recognition and Enrichment of Archival Documents*) a été financé par la commission européenne dans le cadre du programme Horizon 2020 (2016-2019). À la suite de ce financement européen, l'équipe a créé une société coopérative européenne pour continuer de soutenir et de développer la plate-forme Transkribus : https://readcoop.eu/.

11. Notons aussi qu'il faut quelques heures pour apprendre à utiliser

efficacement Transkribus dans un projet. C'est pourquoi, il n'est conseillé d'utiliser le logiciel que si vous avez au moins une centaine d'images à transcrire.

12. Un compte rendu détaillé en français des tests a été mis en ligne sur l'archive ouverte HAL (Massot *et al.*, 2018) et un article en anglais a été publié dans le J *ournal of Data Mining and Digital Humanities* (Massot *et al.*, 2019).

13. Voir cette vidéo de démonstration : <http://ffl-public.huma-num.fr/doc/Transkribus_automatique.mov>.

14. EMAN (Édition de Manuscrits et d'Archives Numériques) : http://eman-archives.org/EMAN/.

15. Une très grande infrastructure de recherche consacrée au développement du numérique au sein des Sciences Humaines et Sociales et proposant plusieurs services ou outils aux acteurs des SHS en France. Site web : https://www.huma-num.fr/

16. Logiciel de gestion de bibliothèque numérique mis à disposition sous licence libre (GNU – General Public License) : https://omeka.org/.

17. En informatique, un plugin ou plug-in, aussi nommé module greffon ou plugiciel (ou extension dans les CMS ou Omeka), est un paquet structuré de codes informatiques qui complète un logiciel hôte pour lui apporter de nouvelles fonctionnalités.

18. Pour plus d'informations sur ces développements réalisés dans le cadre du projet ANR FFL voir le carnet de recherche : https://ffl.hypotheses.org/activites#eman. Les sources sont disponibles sur la plate-forme de diffusion de code GihHub : https://github.com/EMAN-Omeka.

19. Développé à partir du plugin pré-existant History Log : https://omeka.org/classic/plugins/HistoryLog/.

Références bibliographiques

- Bachimont, 2007 : Bruno Bachimont, *L'ingénierie des connaissances et des contenus*, Cachan, Lavoisier-Hermes science publications, p. 237-238.
- Bert, 2016 : Jean-François Bert, *Une histoire de la fiche érudite* , Lyon, Presses de l'enssib.
- Foucault, 2015 : Michel Foucault, *Œuvres*, tome III. *L'Archéologie du savoir*, Paris, Bibliothèque de la Pléiade, Gallimard.
- Foucault, 2021 : Michel Foucault, *Binswanger et l'analyse existentielle* , Paris, EHESS-Gallimard-Seuil.
- Foucault, à paraître : Michel Foucault, *Phénoménologie et psychologie* , projet de thèse, manuscrit conservé dans le fonds Michel Foucault, cote NAF 28730, boîte 46, dossier 2 ; édition établie par Ph. Sabot.
- Massot *et al.* , 2018 : Marie-Laure Massot, Arianna Sforzini et Vincent Ventresque, « Transcrire les fiches de lecture de Michel Foucault avec le logiciel Transkribus : compte rendu des tests », 〈hal-01794139v2〉.

- Massot *et al.* , 2019 : Marie-Laure Massot, Arianna Sforzini et Vincent Ventresque, « Transcribing Foucault's handwriting with Transkribus », *Journal of Data Mining and Digital Humanities*, Episciences.org, Atelier Digit_Hum, 〈hal-01913435v3〉.
- Ventresque, 2021 : Vincent Ventresque, « Les réalisations numériques du projet FFL : présentation et bilan », billet sur le carnet de recherche FFL, https://ffl.hypotheses.org/2153.
- Walter, 2020 : Richard Walter, « La plate-forme EMAN » (en ligne), site web EMAN (Édition de Manuscrits et d'Archives Numériques), http://eman-archives.org/EMAN/plateforme-eman.

Nos partenaires

Le projet *Savoirs* est soutenu par plusieurs institutions qui lui apportent des financements, des expertises techniques et des compétences professionnelles dans les domaines de l'édition, du développement informatique, de la bibliothéconomie et des sciences de la documentation. Ces partenaires contribuent à la réflexion stratégique sur l'évolution du projet et à sa construction. Merci à eux !

- CONCEPTION : ÉQUIPE *SAVOIRS*, PÔLE NUMÉRIQUE RECHERCHE ET PLATEFORME GÉOMATIQUE (EHESS).
- DÉVELOPPEMENT : DAMIEN RISTERUCCI, IMAGILE, MY SCIENCE WORK.
- DESIGN : WAHID MENDIL.