Archives Mathématiques d'Orsay (AMOr)

2020-2022

Emmylou Haffner & Elisabeth Kneller

I. Présentation générale du projet et des partenaires

Porté administrativement par l'Université Paris-Saclay et financé pendant deux ans dans le cadre de l'appel à projet de CollEx-Persée, le projet Archives Mathématiques d'Orsay (AMOr) était coordonné par Emmylou Haffner (à ce moment-là, chercheuse contractuelle à l'Institut de mathématique d'Orsay) et Elisabeth Kneller (responsable de la Bibliothèque mathématique Jacques Hadamard).

Le dépôt d'un projet CollEx-Persée était motivé en premier lieu par la volonté d'Emmylou Haffner et Elisabeth Kneller de prendre en charge les nombreuses archives de chercheur.ses du Laboratoire de Mathématiques d'Orsay déposées à la BJH, et qui n'avaient pu être traitées. Plus largement, nous souhaitions nous donner les moyens de développer une collaboration transversale entre bibliothèque et chercheur.ses. Notre démarche part également d'un constat plus large, appuyé sur nos expériences respectives d'historienne des mathématiques, de bibliothécaire et membres du Réseau National des Bibliothèques de Mathématiques : de nombreuses bibliothèques de mathématiques possèdent, dans leurs fonds documentaires, des fonds d'archives déposés par des scientifiques ou reçus par des legs, sans pour autant qu'une politique d'archivage ou de mise en valeur de ces archives n'existe. Plusieurs efforts ont déjà été entrepris, et l'importance d'une stratégie pour traiter les archives, les mettre en valeur, les conserver et pérenniser est plus qu'évidente. Ces efforts, dont nous avons pu être témoins et partie prenante, soulignent l'importance de mettre en place des actions collaboratives, transversales et interprofessionnelles pour développer une politique d'archivage pérenne et innovante. C'est l'un des objectifs principaux de ce projet. Pour aller plus loin et favoriser l'exploitation scientifique de ces fonds, il importe également de développer une stratégie d'éditorialisation scientifique et de patrimonialisation de ces archives. Celles-ci constituent en effet la mémoire de la communauté mathématique, et un objet de recherche précieux pour l'histoire. Soulignons, de plus, que le rapport de la communauté mathématique à ses archives et plus largement à la mémoire de la discipline est particulier, car les textes mathématiques possèdent une validité longue, et certains textes du passé restent des outils pour le développement des connaissances. Il s'agit d'un second objectif du projet, objectif ancré dans la collaboration entre chercheur.ses, bibliothécaires et archivistes. Cette collaboration, nous avons souhaité l'amener plus loin, en explorant également les possibilités de valorisation des fonds par les outils numériques.

Le projet AMOr s'articulait autour de deux axes. Le premier axe était à visée archivistique et concernait le recollement et le tri des archives du Laboratoire de Mathématiques d'Orsay, et a mené à la création d'un plan de classement et d'inventaires de la plupart des fonds. Nous nous sommes également penchées sur la question du signalement des archives dans des outils nationaux, dans le Répertoire de fonds en histoire et philosophie des sciences et des techniques (RHPST), et sur la valorisation et édition de certaines de ces archives. Les fonds étant, pour la plupart, très récents, un travail essentiel concernait l'étude du droit des archives en vue de leur diffusion. Cet axe a été mené par la Bibliothèque Jacques Hadamard, qui a pu embaucher une archiviste dans le cadre du projet AMOr. Nous avons, dans ce cadre, collaboré étroitement avec le Centre documentaire du Centre d'Archives en Philosophie, Histoire et Éditions des Sciences (CAPHÉS, ÉNS, Paris), ainsi qu'avec les membres de l'équipe Études sur les Sciences et les Techniques (EA 1610) de l'Université Paris-Saclay, spécialisée en histoire des sciences et de l'enseignement. Le second axe concernait l'édition de transcriptions de textes mathématiques, et le développement de modèles d'apprentissage de reconnaissance automatique des textes pour les formules mathématiques, avec comme objectif d'obtenir une transcription en LaTeX[1]. Le projet EMAN - Édition de Manuscrits et d'Archives Numériques, (Thalim, ENS-CNRS) s'est associé à AMOr pour développer l'édition

[1] LaTeX est un système de composition des documents très puissant pour les équations mathématiques, et largement utilisé dans la communauté scientifique.

des formules mathématiques dans un éditeur XML/TEI intégré à Omeka. La Direction des bibliothèques, de l'information et de la science ouverte (DiBISO) de l'Université Paris-Saclay a pris en charge le travail sur les modèles d'apprentissage, et a pu, grâce au projet AMOr, embaucher une ingénieure. Ce travail s'est fait en collaboration avec l'Institut des Hautes Études Scientifiques (IHÉS) qui a fourni le matériau principal pour tester les modèles avec le corpus des prépublications de l'IHES, tapées à la machine et, pour certaines, OCRisées auparavant.

Au cours de ces travaux, nous avons également grandement bénéficié d'échanges avec l'UAR Mathdoc (Université Grenoble-Alpes/CNRS).

II. Axe archivistique

Le but premier de cet axe était l'évaluation, le tri et le recollement, le traitement et l'inventaire des archives mathématiques conservées à la BJH. Pour cela, nous avons, grâce au financement CollEx-Persée, embauché une archiviste à temps plein pendant 9 mois. À cela s'est naturellement ajoutée une volonté de valoriser les fonds.

II. 1. Le corpus traité

Le département de mathématiques d'Orsay a été fondé en 1958 par les mathématiciens Hubert Delange et Jacques Dény. Il faisait alors partie du Département de mathématiques de Paris et d'Orsay, présidé alors par Henri Cartan. Jean-Pierre Kahane, l'un des premiers membres du département de mathématiques d'Orsay, fut à l'initiative de la création de la bibliothèque de mathématiques en 1962, initialement destinée uniquement à la recherche (aujourd'hui appelée Bibliothèque mathématiques Jacques Hadamard (BJH)). La création, en 1971, de l'Université Paris-Sud (ou Paris XI) a mené à une grande expansion du laboratoire. Celui-ci est devenu, en 1998, une unité mixte de recherche avec une double tutelle CNRS et Université Paris-Sud (aujourd'hui Université Paris-Saclay et le CNRS) sous le nom de laboratoire de mathématiques d'Orsay (LMO). À ce jour, la BJH est une unité mixte de service, initialement avec comme tutelles le CNRS, l'Université Paris-Saclay et l'IHÉS, et désormais sous la double tutelle de l'Université Paris-Saclay et du CNRS.

À la BJH, comme dans beaucoup de bibliothèques de recherche, ont été déposés plusieurs fonds d'archives sans mise en place de politique d'archivage. La plupart sont des cartons déposés par des chercheur.ses, au moment de leur départ du département, mais la provenance exacte de certaines de ces archives n'était pas connue. En effet, une partie provient d'une collecte organisée lors du déménagement du département de mathématiques de l'Université Paris-Saclay (alors Université Paris Sud) en 2017, au cours duquel un petit groupe de travail avait été constitué afin de s'assurer de la préservation des documents conservés dans les bureaux depuis la création du département.

Les fonds d'archives qui se trouvent à la BJH sont relativement récents. Le fonds le plus ancien est celui résultant du legs d'Albert Châtelet (1883-1960), qui n'a pas travaillé à Orsay mais était proche de certains membres du laboratoire. Il s'agit d'un fonds de 297 livres de mathématiques, physique et enseignement, datant des années 1850 à 1950. Ces livres (identifiables par un ex-libris) constituaient une partie du fonds de départ de la Bibliothèque Jacques Hadamard. Dans ce legs, toutefois, certains documents, trop anciens, hors format, ou plus personnels sont conservés dans les archives[2]. Les autres fonds appartenaient à des enseignant.es-chercheur.ses du LMO. Parmi les archives reçues se trouvent des documents pédagogiques, des notes de séminaire, des notes de travail, des tirés-à-part, des lettres personnelles, des dédicaces, des archives d'autres personnalités, ainsi que des archives créées dans des fonctions administratives. La plupart des archives, à quelques exceptions près provenant du fonds Châtelet, sont donc des archives contemporaines couvrant l'essentiel du XX^{e} siècle.

II. 2. Travail archivistique

Lors de son arrivée au LMO en 2019, Emmylou Haffner, a commencé à trier les documents et à évaluer leur valeur scientifique. Jusque-là, aucun inventaire et aucune référence de ces documents n'existaient. L'obtention d'un financement CollEx-Persée nous a permis de recruter une archiviste, Claire

[2] Voyez https://omekas.imo.universite-paris-saclay.fr/items/show/2141

Roulot, à partir de mai 2021[3] pendant 9 mois. Ce contrat a ensuite été prolongé de 6 mois supplémentaires grâce à un financement du LMO. Claire Roulot a été encadrée et conseillée par David Dénéchaud, chargé de ressources documentaires au CAPHÉS[4]. Ils ont amélioré ensemble la méthodologie de travail, en commençant par le fonds Illusie, en faisant un travail fin de reclassement, de foliation et cotation des dossiers et sous-dossiers en fonction d'un plan de classement discuté et établi en coordination avec les responsables du projet et des scientifiques du LMO – les mathématicien.nes ayant déposé leurs archives à la BJH se sont, pour la plupart, volontiers impliqué dans les discussions avec C. Roulot.

À ce jour, le fonds de la BJH se compose de sous-fonds ou de collections constituées soit selon le type des documents dont elles se composent (par exemple la collections des polycopiées de cours et d'exercices), soit selon le.la producteur.rice et/ou destinataire du fonds (identifié.e par des indices comme les dédicaces et ex-libris, ce qui est le cas des collections Châtelet, Mandelbrojt, Jean et Georges Cerf). Nous comptons 15 fonds et collections de taille diverses, classées dans 87 boîtes de conservation étiquetées par cotes :

- le fonds Luc Illusie (ILL, 22 boîtes)
- le fonds Anne-Marie Chollet (CHO, 8 boîtes)
- le fonds Myriam Dechamps (DEC, 7 boîtes)
- le fonds Sylvie Ruette (RUE, 11 boîtes)
- le fonds Marie-Claude David (DAV, 2 boîtes)
- la collection Albert Châtelet (CHA, 2 boîtes)
- la collection Szolem Mandelbrojt (MAN, 2 boîtes)
- la collection Élie et Henri Cartan (CAR, 1 boîte)
- la collection Georges et Jean Cerf (CER, 1 boîte)
- la collection Jacques Deny (DEN, 1 boîte)
- la collection Hubert Delange (DEL, 1 boîte),
- la collection des polycopiés (POLY)
- la collection des traductions manuscrites anonymes (TRA, 2 boîtes)
- la collection des publications scientifiques des années 60-80 (DOC, 2 boîtes)
- la collection des manuscrits et fascicules anciens sans marques d'appartenance (BJH Ms, 5 boîtes)

II. 3. Valorisation

En tant qu'administrateur du RHPST[5], David Dénéchaud a créé un compte au nom de la BJH et a formé Claire Roulot à l'outil OMEKA pour qu'elle puisse y signaler les fonds d'archives et collections de documents mathématiques traités au cours du projet. On peut consulter les fonds signalés ici : https://rhpst.huma-num.fr/collections/show/12. Ce signalement dans le RHPST a permis d'alimenter également le Catalogue collectif de France de la BnF, ce qui permet d'accroître la visibilité de ces fonds (voyez https://ccfr.bnf.fr/portailccfr/ark:/06871/0028638). Un signalement des fonds d'archives dans Calames est prévu, mais l'Université Paris-Saclay ne possédant pas encore de licence[6], cela n'a pas encore pu être fait.

Par ailleurs, nous avons créé une instance Omeka afin de développer une bibliothèque numérique pour l'Institut de Mathématique d'Orsay, que l'on pourra consulter ici : https://omekas.imo.universite-paris-saclay.fr/. Celle-ci présente l'ensemble des inventaires effectués par Claire Roulot, ainsi qu'une sélection de numérisations. Notre intention, en créant la bibliothèque numérique de l'IMO, n'est pas seulement de rendre facilement navigables les fonds conservés à la BJH, mais vise aussi à initier un travail à long terme de mise à disposition des archives. Les futurs inventaires y seront également mis à disposition. Il est également prévu une numérisation à la demande des documents qui le permettent. En effet, en 2021, grâce à un financement de l'INSMI (CNRS), la BJH s'est équipée d'un scanner professionnel. Au cours du recollement, nous avons commencé à numériser certaines archives pour lesquelles les questions de droit

[3] La crise sanitaire a considérablement retardé le recrutement.

[4] L'équipe du CAPHÉS a également fourni des conseils pour les commandes de matériel de conservation propre aux archives.

[5] Répertoire de fonds pour l'histoire et la philosophie des sciences et des techniques - https://rhpst.huma-num.fr/

[6] Une demande est en cours.

ne se posent pas – soit parce qu'elles sont rentrées dans le droit public, soit parce que nous disposons déjà des autorisations nécessaires. (Voyez également ci-dessous au sujet des droits.) Cette bibliothèque numérique fera, à terme, partie de la bibliothèque Numaclay de l'Université Paris-Saclay, dont la création est en cours.

II. 4. Traitement des questions de droit

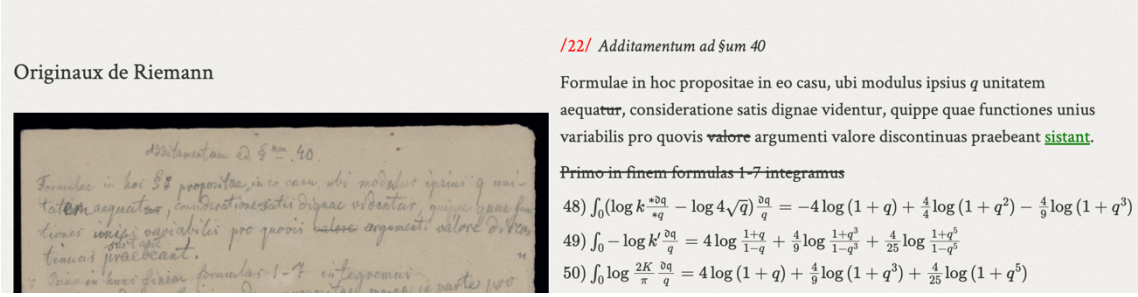
Les fonds conservés à la BJH étant pour la plupart assez (ou très) récents, les questions de droit ont été étudiées avec le plus grand soin. Notons qu'en théorie, les archives de la recherche sont des archives publiques et la communication de celles-ci est libre (sauf pour des documents de nature sensible tels qu'ils sont définis dans le code du patrimoine[7]). Toutefois, si a été établie une convention de don signée par la personne légataire, comme cela a été le cas pour certains de nos fonds, le souhait de celle-ci l'emporte. Lorsque cela n'était pas déjà renseigné dans la convention de don d'archives signée par les chercheur.ses eux-mêmes (comme c'est le cas, par exemple, pour les archives de Luc Illusie ou de Myriam Dechamps), Claire Roulot a contacté les chercheur.ses ou leurs ayant-droits pour demander l'autorisation de la diffusion des documents. Ainsi, pour certaines archives, des restrictions ont été déposées. Dans la plupart des cas, il s'agit d'archives qui contiennent documents sensibles, par exemple des rapports sur des dossiers de candidature, relevés de notes ou informations trop personnelles.

III. Axe numérique

Le deuxième axe du projet AMOr était consacré à l'étude de solutions numériques pour la transcription des textes mathématiques. Celle-ci s'est faite en deux temps : la mise en place d'un module de transcription des formules mathématiques en XML/TEI pour Omeka Classic, et des tests de logiciels d'apprentissages de reconnaissance automatique des textes.

III. 1. Transcription XML/TEI des formules mathématiques dans Omeka

Le projet EMAN, mené par Richard Walter (Thalim, CNRS/ÉNS), et dont Emmylou Haffner fait partie depuis 2017, s'est joint au projet AMOr pour travailler sur la question de la transcription en XML/TEI des formules mathématiques. Grâce au soutien d'AMOr, EMAN a développé la version 0.9 du plugin Transcript pour Omeka Classic, actuellement disponible sur GitLab[8]. Ce plugin permet à un.e utilisateur.rice non expert.e de saisir et annoter une transcription en proposant, en vis-à-vis de l'image, un bloc texte accompagné d'une barre d'outils pour l'encodage de la transcription avec des balises XML/TEI. Il utilise la TEI pour encoder les phénomènes éditoriaux ou annoter certains termes. Le module propose une page pour naviguer dans les items et les fichiers de l'instance Omeka, et d'en faire la transcription, avec des liens vers le fichier Omeka original, le fichier XML contenant la transcription, la notice du fichier Omeka et la notice de l'item Omeka qui contient le fichier courant.

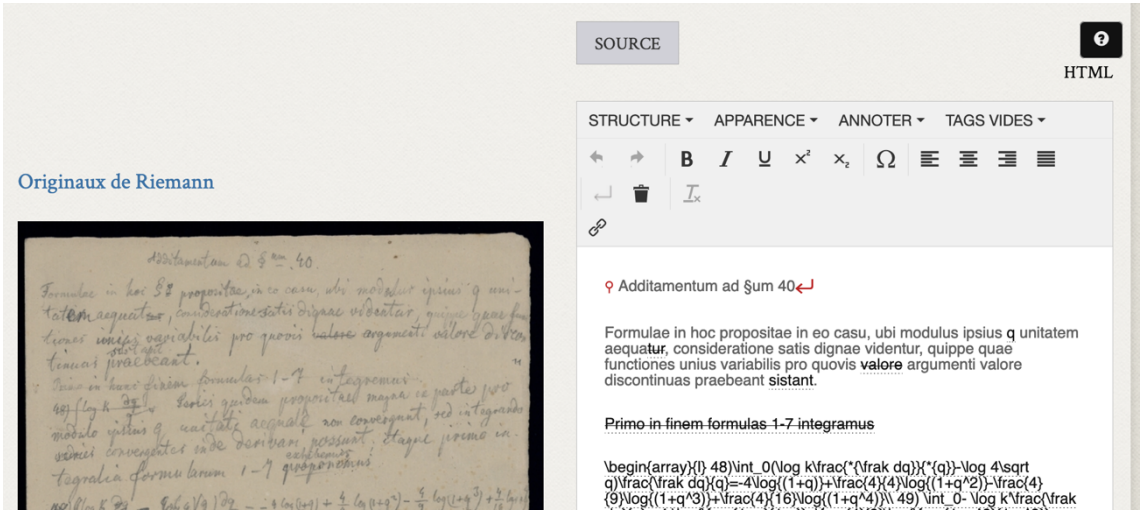


Capture d'écran d'un exemple tiré des recherches personnelles d'E. Haffner

Le plugin propose un mode « éditeur » en WYSIWIG avec l'utilisation d'un barre outils TinyMCE adapté aux tags TEI. L'utilisateur.rice n'a pas besoin de saisir directement les balises XML/TEI mais les sélectionne

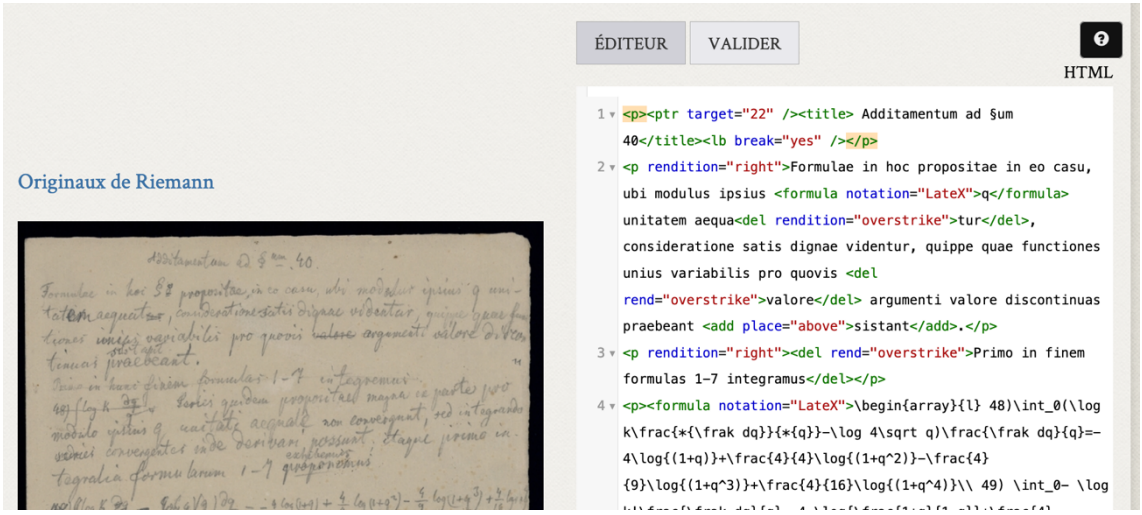
[7] Voyez l'Article L213-2 du Code du Patrimoine.
[8] https://github.com/EMAN-Omeka/Transcript

dans les menus de la barre outils. Des fenêtres pop-up facilitent la saisie des attributs des balises. Un bouton Valider permet de vérifier que le code saisi est conforme à la spécification TEI.



Capture d'écran d'un exemple tiré des recherches personnelles d'E. Haffner

Le plugin propose aussi un mode « source », donnant accès à une instance CodeMirror permettant de travailler directement en XML, avec des fonctions d'auto-complétion, de suggestion contextuelle de tags et de fermeture automatique. Le code est « *highlighted* », signalant les erreurs de syntaxe les plus évidentes et mettant en valeur les différents éléments (balises, attributs, texte, etc.). Le schéma avec les balises utilisables et les règles de cohérence est géré par un fichier XML (cm-tei-schema.xml) dont la structure est très simple.



Capture d'écran d'un exemple tiré des recherches personnelles d'E. Haffner

Le rendu des transcriptions en HTML n'est pas géré directement par Transcript. Sur EMAN, une instance TEI Publisher est utilisée ; l'URL de l'instance TEI Publisher est spécifiée à Transcript pour y envoyer les données à visualiser. Les transcriptions sont stockées dans un répertoire spécifique de l'instance Omeka et exportables en fichiers XML. Elles sont aussi stockées dans la base de données pour être intégrées au moteur de recherche SolR.

Le plugin offre également la possibilité de gérer un index avec la balise <term>. Un formulaire permet de décrire chaque terme avec des champs personnalisables et de faire des liens avec d'autres termes. La page de rendu du glossaire ajoute la liste des différentes occurrences du terme avec le contexte autour de l'occurrence. Le plugin ne permet pas encore de générer des index à partir des différentes balises TEI et d'autres fonctionnalités sont encore à imaginer.

Enfin, et toujours avec le soutien du projet AMOr, une version du plugin Transcript pour Omeka S a été développée. Cette version propose les mêmes fonctionnalités. Un travail spécifique a été fait pour pouvoir intégrer un encodage des équations mathématiques avec la balise <formula> et le langage LaTeX.

III.2. Études en vue de la réalisation d'un modèle HTR pour les formules mathématiques

Au sein du projet AMOr, la DiBISo était chargée d'explorer l'apport des outils IA dans la reconnaissance des formules mathématiques. Deux outils avaient été identifiés pour cela : la solution gratuite Watson d'IBM d'une part, la plateforme HTR Transkribus mise à disposition par READ. Ce dernier outil était lors du dépôt du projet une solution gratuite, dont le défaut principal était de ne pouvoir utiliser les résultats de ses travaux que dans le même environnement Transkribus. Au lancement effectif du projet, READ annonçait la transformation du modèle économique de Transkribus vers un modèle payant. Le retour des évaluations du projet ayant indiqué de privilégier des solutions ouvertes répondant aux principes FAIR, nous avons commencé à étudier des solutions alternatives, en particulier la solution Kraken/eScriptorium[9] portée par PSL, devenue plus mature et offrant plus de garanties sur le plan de la science ouverte.

Étant prévu que la partie IA serait abordée en fin de projet, nous avons profité des mois précédant la mise en place des travaux pour prototyper la chaîne de traitement permettant de nourrir la vérité terrain d'une part, et de tester le modèle d'apprentissage sur des documents numérisés d'autre part. Il est vite apparu que le travail sur un outil HTR serait déjà très conséquent et qu'il était plus raisonnable d'explorer la seule solution HTR fournie par eScriptorium.

Pour cette partie du projet, l'IHÉS a fourni des fichiers pdf, xml et txt.xml de prépublications de l'IHÉS qui avaient été préalablement numérisées et OCRisées. L'échantillon fournit comportait un grand nombre de formules mathématiques afin de créer un modèle pour tester les outils pressentis pour la reconnaissance des formules mathématiques. Ces travaux ont été menés par Luc Bellier, Silvia Silini et Emmanuelle Vietti, recrutée par le projet AMOr pour 4 mois.

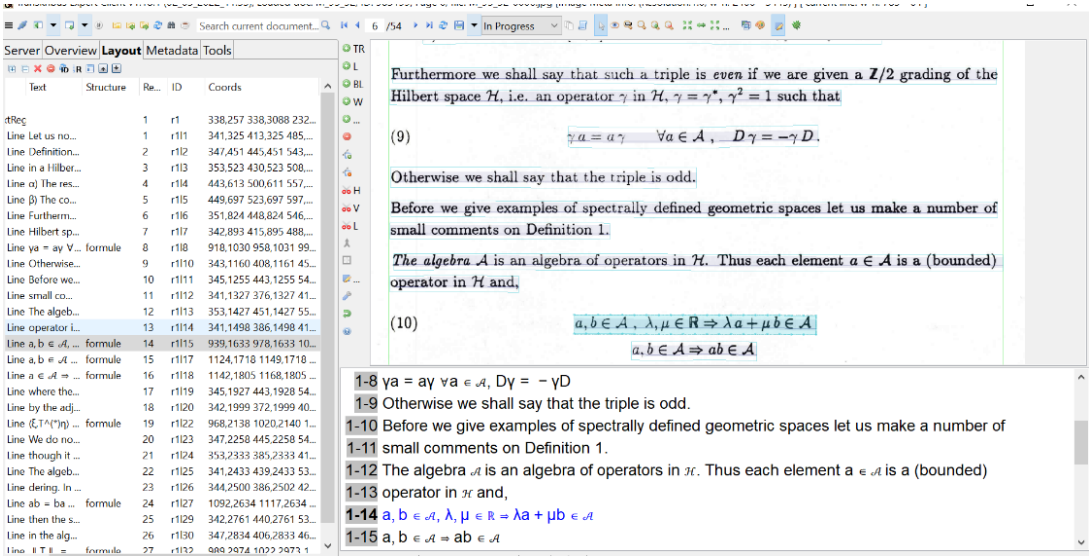
La première étape consistait à alimenter la vérité terrain, c'est-à-dire un ensemble d'informations validées par un agent humain faisant correspondre des images à du texte et des formules mathématiques. Cette vérité terrain sert de base d'apprentissage à l'IA. Pour rendre plus rapide cette étape pour les formules mathématiques, nous avons choisi d'utiliser Mathpix[10], un outil commercial qui permet de reconnaitre une formule après qu'elle a été découpée par un humain. Cette étape évite la saisie manuelle des formules, et nécessite la segmentation de la formule et la vérification et éventuellement la correction du résultat.

L'étape suivante concernait la segmentation des documents, c'est-à-dire l'étape primaire dans le traitement OCR ou HTR consistant à analyser une page dans son ensemble pour identifier les zones textuelles, et les zones non textuelles. Une fois les zones textuelles repérées, il s'agit de découper les différentes parties composant cette zone, du paragraphe aux caractères en passant par la ligne et les mots. Pour cette étape, qui n'est pas aussi aisée directement dans eScriptorium, nous avons pris le parti d'utiliser Transkribus qui le propose gratuitement avec un format d'export récupérable dans eScriptorium. Par ailleurs, cet outil permettant de taguer les zones, nous avons utilisé cette fonctionnalité pour taguer les formules mathématiques. Il est également possible à ce stade de corriger la segmentation. C'est une étape essentielle pour les formules, car les outils OCR classiques ne savent pas gérer les spécificités d'alignement de certaines formules, comme les fractions ou les indices et exposants qui sont écrits sur plusieurs lignes. Il est donc important de corriger la segmentation afin que les formules soient correctement identifiées à la segmentation. Transkribus permet enfin de traiter en OCR le document segmenté. Il faut souligner que les documents retenus, ici, sont pour l'essentiel des documents typographiés, avec d'éventuelles reprises manuscrites, particulièrement sur les formules. On procède ainsi à la reconnaissance des pages et on charge les informations correspondant aux formules mathématiques gérées en amont dans l'interface de saisie de Transkribus. La vérité terrain peut ainsi être produite sans une ressaisie complète des documents. On dispose alors avec le texte fourni par Transkribus d'une correspondance entre le texte et les images via des coordonnées spatiales. Ce sont ces informations que le HTR analyse.

---

[9] https://escriptorium.fr/. eScriptorium est une interface web développée dans le cadre du projet SCRIPTA (https://scripta.psl.eu/)

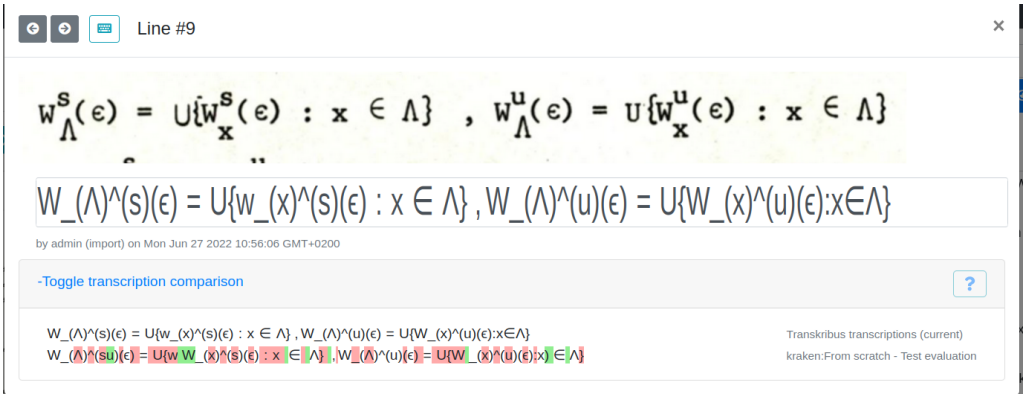
[10] https://mathpix.com



Ces données ont ensuite été exportées pour procéder à l'étape HTR dans eScriptorirum. L'export est fait au format XML page permis par Transkribus, qui permet de véhiculer les informations enrichies et complétées à cette étape. eScriptorium permet de réaliser les tâches de segmentation, transcription et d'entraînement des modèles de reconnaissance d'écriture. Une fois chargées les images avec les transcriptions et les segmentations, il est possible d'entraîner un modèle et de le tester sur un corpus d'images. Le prototype était prêt.

Le projet AMOr envisageait la possibilité de traiter directement les éventuelles numérisations d'archives du LMO, mais prévoyait qu'en cas de retard ou de manque de documents en quantité, il serait possible de s'appuyer sur les archives déjà numérisées de l'IHÉS, établissement membre associé de l'université Paris-Saclay et partenaire du projet. Ce sont ces archives et en particulier les pré-prints d'articles qui ont été utilisés. 302 pages ont été traitées (production des formules, correction, segmentation, correction et enrichissement). La vérité terrain produite après environ 4 mois de travail, il a été possible de commencer l'exploration de la mise en place d'un modèle d'apprentissage et de ses résultats.

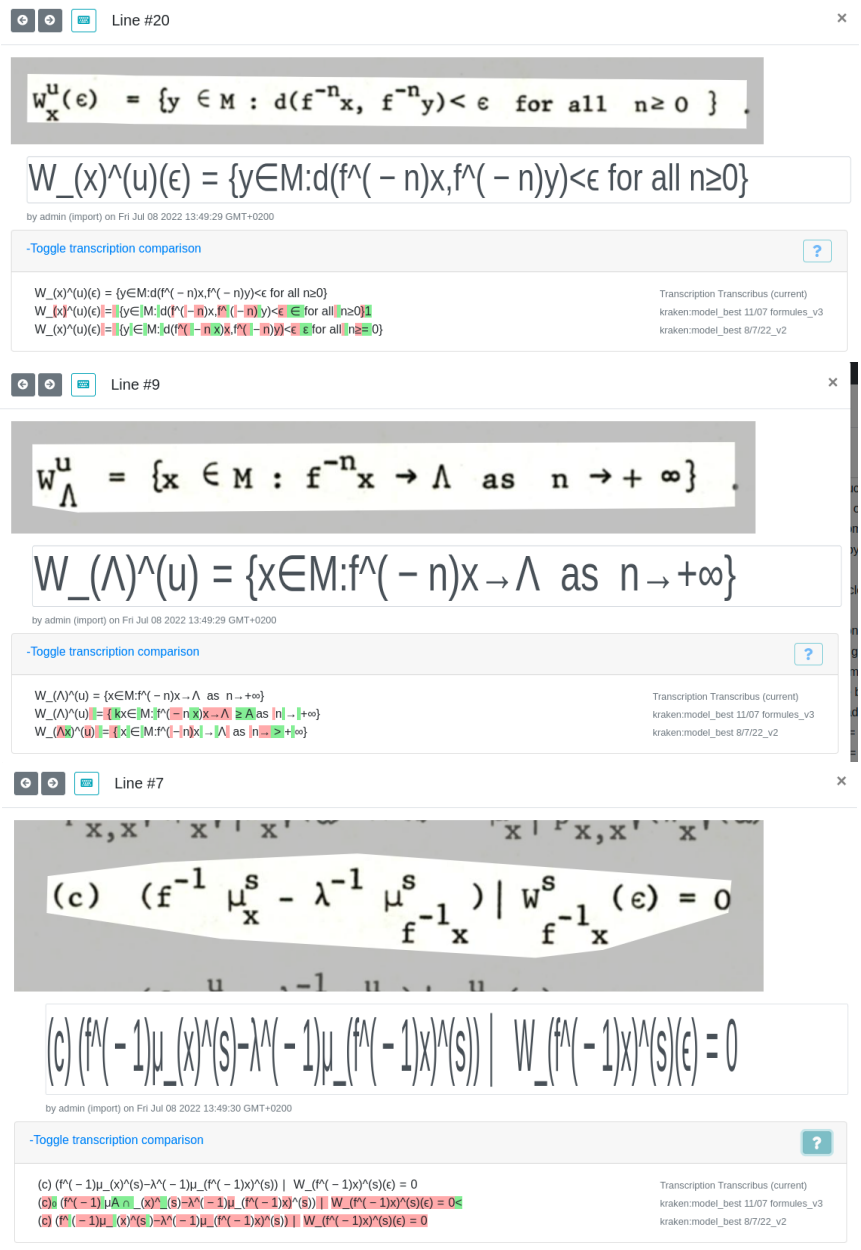
Pour les premiers essais, nous avons tenté la reconnaissance des pages dans leur ensemble (texte et formules). Avec cette approche, les résultats sur les formules n'étaient toutefois pas bons, même pour les formules les plus simples. Ci-dessous, un exemple de résultats du modèle : les couleurs représentent la différence avec la transcription Transkribus (ligne 1 dans la partie inférieure de l'écran). Les caractères en vert sont ceux ajoutés par le modèle et ceux en rouge ceux supprimés mais qui devaient être normalement présents.



Un autre essai a été réalisé en travaillant directement dans Kraken, le moteur d'apprentissage sur lequel s'appuie eScriptorium. Pour cela, nous avons entrainé un modèle avec 23 pages (document M_74_84) afin de savoir si l'entrainement allait se terminer sans difficulté ni erreurs. Vu le résultat positif de ce test, nous avons entrainé un modèle avec 170 pages (M79_321, M80_04, M80_06, M85_56, M95_52, M67_30). Ce modèle a été ensuite chargé sur eScriptorium et testé sur 143 pages. Beaucoup d'erreurs

étaient toujours présentes dans la transcription des formules mathématiques, et nous avons donc tenté d'entrainer un modèle avec une vérité terrain sélectionnée, en excluant les pages ne contenant que du texte de la vérité terrain. Ce modèle entrainé dans Kraken a ensuite été uploadé sur eScriptorium sous le nom « model_best 11/07 formules ».

Nous ne pouvons toutefois pas vraiment apprécier des progrès dans ces deux modèles. Les formules les plus compliquées restent toujours très mal reconnues. Pour les autres, comme on peut le voir ci-dessous, la reconnaissance reste très aléatoire.



Notre dernier test pour ce projet a été d'entraîner un modèle avec un document (P_78_243) constitué seulement par des transcriptions des formules mathématiques, dans lequel toutes les transcriptions du texte ont été supprimées. Cette stratégie a été choisie après avoir constaté que les modèles continuaient à bien reconnaître le texte mais pas les formules. Nous avons pu constater que ce modèle (model_best P_78_243 formules 8 /9/22) parvenait à reconnaitre des formules assez simples, mais continuait à faire beaucoup d'erreurs sur celles plus complexes, comme on peut voir sur ces captures d'écran.

Les 6 mois de travail prévus se sont malheureusement révélés insuffisants pour disposer d'un modèle fonctionnel de reconnaissance des formules. Ils ont cependant permis d'avancer dans la transcription (302 pages des transcriptions corrigés) et de pouvoir identifier de nouvelles pistes pour poursuivre le travail sur l'apprentissage automatique. Une première piste à explorer est celle de la modification de l'architecture du réseau de neurones de Kraken afin d'obtenir un modèle pouvant reconnaître tous types de formules. Une deuxième piste serait celle de segmenter les formules symbole par symbole de sorte que le modèle se base plutôt sur l'apprentissage des symboles que sur celui de formules en entier, mais cela impliquerait beaucoup de travail manuel puisque tout le processus de segmentation automatique serait à refaire.

IV. Ouvertures et retombées

Les archives du Laboratoire de Mathématiques d'Orsay ont grandement bénéficié du projet AMOr. D'inexploitables fin 2019, elles sont devenues un fonds ordonné, catalogué et référencé. Ce succès est précieux, d'autant que ces archives contiennent des documents inestimables pour l'histoire des mathématiques. Localement, nous espérons que ce travail saura être l'impulsion d'une conservation à long terme des archives de la recherche mathématique à Orsay. Depuis le début du projet, de nombreux chercheur.ses ont exprimé leur intérêt et le dépôt de nouveaux fonds au cours de ces deux dernières années porte la promesse d'une vraie pérennité pour notre travail.

Le succès de l'axe archivistique du projet AMOr, et l'intérêt marqué de la communauté des mathématicien.nes et des historien.nes des mathématiques a encouragé la reprise et le développement d'un groupe de travail « Archives » au sein du Réseau National des Bibliothèques Mathématiques, animé par Emmylou Haffner et Nayara Gil-Condé (Institut Henri Poincaré). Le groupe s'est rapproché du GDR Histoire des mathématiques et prévoit plusieurs actions, parmi lesquelles une journée au Séminaire d'histoire des mathématiques de l'Institut Henri Poincaré au mois de juin 2023, présentant les initiatives au sein du RNBM et engageant un dialogue avec les chercheurs. Nous envisageons également la publication d'articles sur le site *Images des mathématiques* – notamment présentant les archives du LMO. À plus long terme, nous souhaitons que la dynamique qui est en train de se mettre en place puisse mener à des actions (et demandes de financement) plus larges, peut-être nationales, pour la valorisation des archives de la recherche mathématique. La collaboration entre historien.nes des mathématiques et bibliothécaires s'étend également d'ores et déjà dans le cadre du projet ANR PatriMaths, porté en partie par l'EA EST.

Du point de vue des activités numériques du projet, la collaboration avec le projet EMAN a vocation à se poursuivre, en particulier, autour du développement du module Transcript – certains aspects de la visualisation et de l'édition critique des formules mathématiques demandant encore du travail. Nous souhaitons pouvoir développer plus avant les outils de transcription XML/TEI pour les mathématiques, notamment en collaboration avec des expert.es de la TEI.

La reconnaissance automatique des formules mathématiques est également un sujet sur lequel les membres du projet AMOr continueront à travailler. Pour cela, des collaborations sont engagées, notamment avec le projet ERC Philiumm (SPHERE, CNRS) porté par David Rabouin.

Avec le projet AMOr, nous souhaitions en premier lieu donner une vraie impulsion à des collaborations transversales pour l'étude et la valorisation des archives de la recherche mathématique. Il ne fait aucun doute que cette impulsion a été féconde, et nous avons le ferme espoir qu'il ne s'agit que de la première impulsion d'un plus long mouvement.