Projet CollEx-Persée Transcrire

(octobre 2020 – septembre 2022)

https://www.collexpersee.eu/projet/transcrire

Bilan scientifique

Rédigé par Fabrice Melka (IMAF), avec la collaboration d'Anaïs Wion (IMAF), Benjamin Suc Limonade&co), Véronique Ginouvès et Émilie Groshens (MMSH Aix-en-Provence), Michael Nauge (Laboratoire FoReLLIS), Nadine Dieudonné-Glad (Laboratoire HERMA), Gaëlle Hallair (UMR Géographie-cités).

Sommaire

Refonte et stabilisation de la plate-forme 2
Mise à jour technique de Omeka classique à Omeka S 3
Le module Scripto pour la transcription 3
Modèle de données 4
Import des images 4
Développements propres 5
Migration des données vers l'environnement Omeka S 5
Graphisme, ergonomie, nouvelles fonctionnalités 6
La recherche plein texte 12
Circulation des données 13
Collections et documents nouvellement déposés sur Transcrire 14
Travail de validation des transcriptions 15
Les ateliers de transcription collaborative 16
Utilisations de la plateforme Transcrire et exploitation des transcriptions 17
Sciences participatives 17
Meilleure accessibilité des collections et enrichissement des métadonnées 17
Articuler transcription manuelle et transcription automatisée 18
Valorisation autour du projet CollEx-Persée Transcrire (manifestations, articles, billets de blog, …) .. 19

La plate-forme Transcrire (https://transcrire.huma-num.fr) est un outil collaboratif qui propose à des chercheurs comme à des publics amateurs de contribuer à la transcription de documents manuscrits numérisés en mode image. L'un des objectifs est de faire découvrir des corpus de documents à un large public afin que celui-ci s'implique dans leur traitement (*crowdsourcing*) grâce à une nouvelle forme d'interaction et d'engagement : la transcription collaborative. Étant dédié à la valorisation du patrimoine scientifique, Transcrire constitue aussi le support à des projets de recherche en SHS en lien avec les collections de ses partenaires.

Son périmètre éditorial couvre les matériaux de terrains manuscrits provenant de fonds d'archives de chercheurs et chercheuses en sciences humaines et sociales (carnets de terrain, carnets de fouille, notes de voyage et d'enquête, correspondances scientifiques) et provenant de collections patrimoniales de plusieurs bibliothèques de recherche et centres d'archives partenaires. Initialement porté par le Consortium Archives des Ethnologues (TGIR Huma-Num) et ainsi constitué de fonds issus de travaux en anthropologie, Transcrire a suscité de nouveaux intérêts et accueille maintenant des archives et matériaux de terrains produits en ethnologie, géographie, archéologie et histoire.

Plusieurs institutions de conservation et équipes de recherche ont confié des collections à la plate-forme. Celles engagées dans le projet Collex sont les suivantes : Institut des mondes africains (IMAF), Musée du quai Branly, Grand équipement documentaire (GED) du Campus Condorcet, Centre de recherche bretonne et celtique (CBRC), Maison méditerranéenne des sciences de l'Homme (MMSH), SCD de l'Université de Poitiers.

La collaboration avec l'ensemble des équipes de ces institutions a été chaleureuse et constante durant ces deux années (octobre. 2020 - septembre 2022). Sur la poursuite et la continuité du projet, chacun a renouvelé sa confiance en la plate-forme et il y a lieu d'être optimiste quant à sa pérennité puisque le GED Campus Condorcet a ouvert un poste dédié à la problématique de la transcription et reprendra l'administration de l'outil pour les prochaines années. De même il s'est engagé lors du pré-AMI CollEx-Persée autour de la thématique de la transcription manuelle et automatique avec Transcrire comme brique possible d'un programme blanc.

De surcroît, d'autres institutions ou équipes de recherche ont confié des collections et/ou utilisent Transcrire pour des projets spécifiques : Archives nationales d'outre-mer, Bibliothèque nationale de France, Laboratoire HERMA (Université de Poitiers), Centre d'histoire sociale des mondes contemporains (CHS), Bibliothèque interuniversitaire de la Sorbonne, GéoCité.

Transcrire est développé à l'aide d'outils open source et privilégie les données en accès ouvert. Les bibliothèques et les projets de recherche partenaires prennent en charge la numérisation des documents manuscrits et renseignent les métadonnées puis ils confient l'ensemble aux gestionnaires de la plate-forme (Fabrice Melka de l'IMAF et, depuis septembre 2022 Alyx Taounza-Jeminet du GED Campus Condorcet) qui s'occupent de leur mise en ligne. La première version fut conçue en 2015 grâce aux financements du consortium Archives des ethnologues, consortium labellisé et soutenu par la TGIR Huma-Num. En 2020, la refonte technique de la plate-forme a été l'un des projets lauréats de l'appel CollEx-Persée.

Cette refonte a permis de nombreuses améliorations que nous allons développer dans ce qui suit.

Refonte et stabilisation de la plate-forme

La migration de l'ancien site web (sous Omeka classique), des fichiers images et des données existantes vers un environnement Omeka S a permis d'anticiper l'obsolescence technologique de la plate-forme et d'ancrer celle-ci dans une logique d'ouverture des données et de web sémantique.

Le chantier de refonte a été réalisée par la société Limonade&co.

Des réunions de travail régulières de l'instance de pilotage, composée de Benjamin Suc (Limonade & Co) et de Fabrice Melka (IE CNRS-IMAF), l'un et l'autre responsables de la méthode et de l'ensemble du suivi de la prestation pour les aspects scientifiques, techniques et fonctionnels du projet ont permis de valider le déroulement des différentes phases de la mission ainsi que les livrables finaux.

Mise à jour technique de Omeka classique à Omeka S

Ne sachant pas dans quelle mesure la version classique de Omeka serait encore maintenue dans les prochaines années, il nous a semblé nécessaire de faire cette importante mise à jour technique de la plate-forme.

De plus, il s'agissait de bénéficier des dernières avancées d'un module Scripto (gérant les fonctionnalités de transcription) sorti en octobre 2020 et ne fonctionnant qu'avec la version S de Omeka.

Enfin, celle-ci est clairement orientée web de données (format natif JSON-LD, URI, description des ressources à l'aide de vocabulaires RDF) et permet une interopérabilité plus large et variée avec d'autres systèmes, élément indispensable pour penser Transcrire en lien avec ses bibliothèques partenaires.

L'environnement de développement et de production sous le CMS Omeka S a ainsi été mis en place sur les infrastructures de la TGIR Huma Num, hébergeur du site.

- Installation d'Omeka S v3.0.1 et d'une BdD Sql (identifiants FTP, SSH et PhPmyadmin)

- Installation du module : SolR 7.2

- Installation d'une instance Mediawiki 1.34

- Installation et paramétrage des modules : Scripto, CSV Import, File Sideload, Search, OAI-PMH Repository, Simple Carousel, Block Plus, Clean Url, Contact us and report, Generic module

Le module Scripto pour la transcription

Après le paramétrage du back-end Omeka S, permettant de gérer les collections, les contenus, les fichiers images et leurs métadonnées dans la bibliothèque numérique, des développements du back-end Scripto ont été réalisés.

Le module Scripto permet la transcription collaborative des fichiers images stockés sur Omeka S, il s'appuie sur le logiciel Mediawiki (via son API) pour la mise à jour, le suivi et le contrôle des transcriptions et des comptes utilisateurs.

Sa nouvelle version permet de créer des projets de transcription différents pour chacun des partenaires. Ils existent ainsi indépendamment les uns des autres sur une seule installation Scripto. La mise en place d'un projet permet d'organiser les matériaux avec lesquels on veut travailler, en les sélectionnant depuis les collections créées dans Omeka, et de définir la manière de travailler avec ces contenus (gestionnaires, métadonnées, provenance).

Le module propose des interfaces de gestion, l'ajout de fonctionnalités (tri, filtre, …) et de nouveaux modes de suivi des transcriptions et de leur validation.

Une fois connecté, un partenaire dispose d'un tableau de bord qui comprend :

- Une vue de tous les projets dont il est propriétaire.

- Une vue de tous les projets qu'il révise.

- Une vue de ses contributions récentes.

- Une vue de la liste de médias qu'il a décidé de surveiller.

Modèle de données

Omeka S offre la possibilité d'utiliser des ontologies RDF combinables entre elles. Ainsi Scripto ajoute un nouveau vocabulaire à l'installation Omeka S, vocabulaire qui contient trois propriétés :

- Contenu : une représentation en texte brut du contenu d'une ressource.

- Transcription : Une transcription d'une ressource.

- Traduction : Une traduction d'une ressource.

Nous avons mis en place pour toutes les collections des partenaires un modèle de ressources relativement simple et reprenant les 15 champs Dublin Core auxquels est ajoutée la propriété Transcription de Scripto.

Cela couvre pour le moment l'ensemble des besoins en métadonnées des projets partenaires et il est tout à fait possible par la suite de créer des modèles spécifiques selon les besoins et les métadonnées fournies par les partenaires.

Import des images

Transcrire propose dorénavant deux modes d'import des médias images sur la plate-forme : import en masse sur le serveur (FTP et module File Sideload) et import en utilisant le protocole IIIF, augmentant sa capacité à échanger des données interopérables.

Deux collections récupèrent ainsi leurs données depuis des entrepôts distants grâce au IIIF. Le projet « Carnets scientifiques d'Antoine d'Abbadie », qui utilise l'API IIIF de récupération des images de Gallica, et le projet « Journaux et collectes d'Anatole Le Braz » dont les images ont été déposées sur Nakala (TGIR Huma-Num) par le Centre de recherche bretonne et celtique.

Deux guides d'import sont disponibles pour les partenaires. De même un tutoriel pour importer des contenus ayant pour origine un manifeste IIIF est disponible sur le site web de Limonade & Co :

https://www.limonadeandco.fr/omeka-s-et-le-protocole-iiif

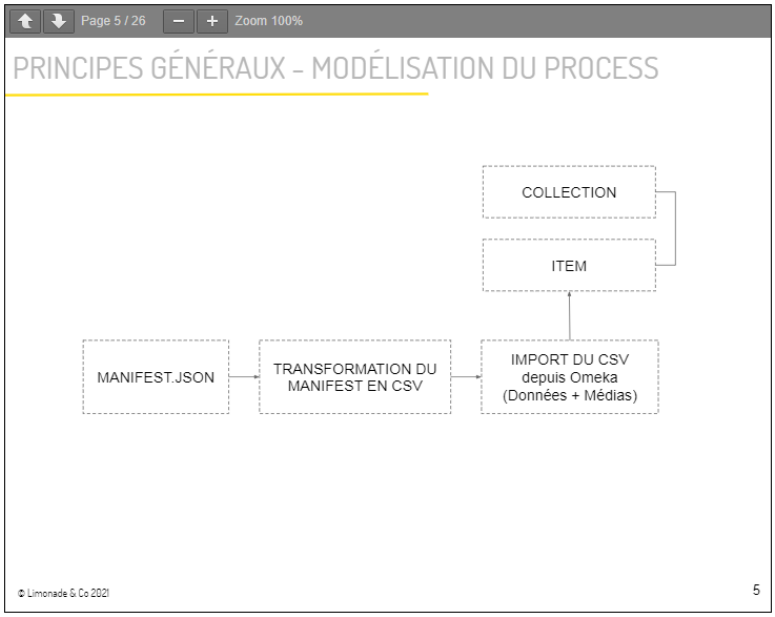

Fig. 1 : Importer des métadonnées et des médias depuis une API IIIF (Limonda&co, 2021).

Développements propres

Trois modules touchants à la fois au thème et au module de transcription Scripto ont été développés :

- Module *TranscrireSharedHeader* : Afin d'assurer une navigation fluide entre Omeka S et le module Scripto, le module TranscrireSharedHeader permet de partager un header commun. De la sorte, le site compte un seul et unique menu.

- Module *Scripto Search* : Le projet bénéficie d'une recherche avancée grâce à l'installation du moteur de recherche à facettes Solr. Sa configuration a été réalisée au moyen du module Solr pour la partie Omeka S (Collections et Items) et au moyen du module Scripto Search afin de gérer les transcriptions (Scripto).

- Module *Transcrire* : Ce module permet de gérer différentes fonctionnalités Scripto au sein du thème Omeka S.

L'ensemble des modules sont accessibles sur le Github de Limonade&co: https://github.com/Limonadeandcoscop

Migration des données vers l'environnement Omeka S

Une reprise/alignement et une migration des données de la version 1 de la plate-forme *Transcrire* (environnement Omeka classique) vers le nouvel environnement Omeka S était un des livrables demandés au prestataire.

Après un contrôle de la validité des données existantes, le périmètre des données à migrer comprenait :

- Toutes les collections, items, fichiers numériques et métadonnées.
- Toutes les transcriptions, les métadonnées liées aux transcriptions, les données utilisateurs (historique, comptes mediawiki/scripto …).

Cela totalisait 9 collections comptant plus de 12 000 fichiers image, 8000 transcriptions et près de 300 comptes utilisateurs.

Il n'existe pas de module permettant d'assurer la reprise des données Scripto Classique vers Omeka S. Cette opération a nécessité d'intervenir en particulier sur 4 domaines :

- Reprise des items
- Reprise des métadonnées
- Reprise des médias
- Reprise des utilisateurs et de leurs identifiants

Pour ce faire, le projet de reprise a été séquencé en plusieurs phases :

- Import des notices classiques (via le module Omeka2Importer)
- Création des Projets (la notion de Projet étant un nouveau concept dans la version S de Scripto)
- Synchronisation pour créer les items et média dans Scripto
- Création de la table de migration
- Export des identifiants des items Scripto Classique
- Génération des nouveaux identifiants des items Scripto S
- Migration des statuts des items

- Migration et mapping des métadonnées
- Migration des contributeurs
- Migration des comptes utilisateurs
- Gestion des mots de passes utilisateurs

Graphisme, ergonomie, nouvelles fonctionnalités

La refonte a également donné lieu à la réalisation d'une nouvelle charte graphique pour le site, les interfaces utilisateur ont été entièrement revues et un thème Omeka S (livrable) 'transcrire' a été spécifiquement créé. Un « Cahier des spécifications fonctionnelles » a permis l'intégration (HTML, CSS & PHP), la personnalisation de templates (interface graphique), la production des gabarits et des nouvelles fonctionnalités demandées. Enfin une version responsive du site a été élaborée afin d'adapter la mise en page du site à différent types d'écrans (desktop et mobile).

Transcrire propose ainsi un agencement nouveau des contenus et de nouvelles fonctionnalités qui améliorent l'affichage de l'information et la navigation dans les corpus d'images et les transcriptions.

La page d'accueil a été dynamisée grâce notamment au carrousel d'images qui permet d'éditorialiser les contenus en mettant en avant tel ou tel projet de transcription ou telle annonce d'événements.

Elle propose aussi, en plus d'une entrée par corpus ou par document, des navigations par discipline et par genre documentaire ainsi que la possibilité de s'atteler au déchiffrage de termes équivoques ou illisibles.

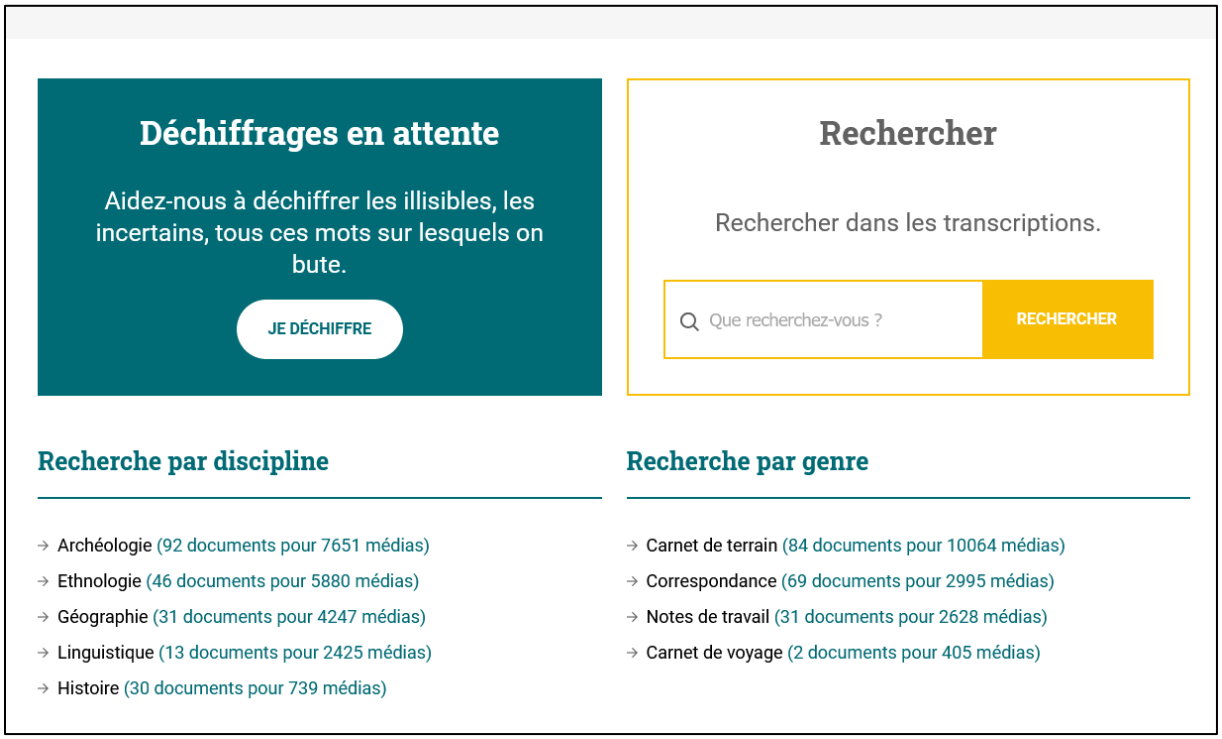

Fig. 2 : Sur la page d'accueil de Transcrire, l'accès aux « déchiffrages en attente ».

Chaque projet possède dorénavant une page d'accueil dédiée qui permet de valoriser le partenaire et qui comporte une courte présentation, des liens vers des ressources liées et vers une **page d'instructions** spécifiques. Ainsi, le projet des Carnets d'Antoine d'Abbadie propose 2 guides : un à

destination des utilisateurs qui débutent et la possibilité de télécharger un guide de type "mode expert".

https://transcrire.huma-num.fr/scripto/13/guidelines

Fig. 3 : Page d'accueil du projet Carnets d'Antoine d'Abbadie.

Le statut d'une page à transcrire - Nouveau, En cours, Terminé ou Approuvé – est constamment affiché et la mise en place d'une **Progress bar** signale le niveau d'avancement d'une transcription sur un carnet ou un lot de correspondances.

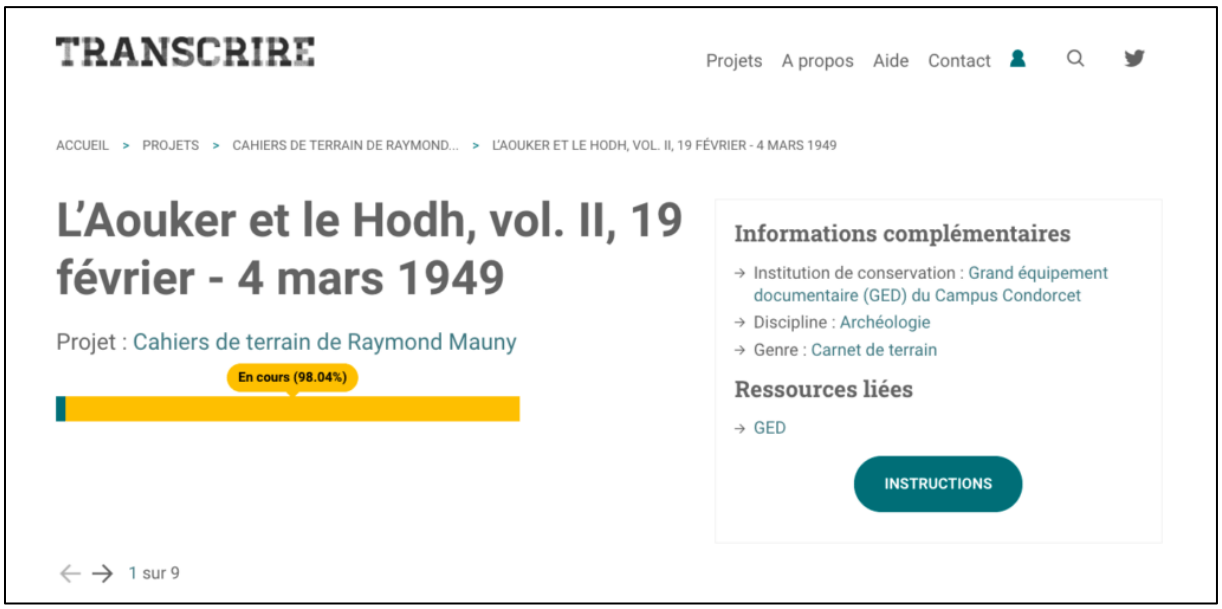

Fig. 4 : « Progress bar » permettant de mesurer l'état d'avancement de la transcription d'un item.

Une série d'**options de tris et de filtres d'affichage** des items et des médias permettent également aux utilisateurs de n'afficher que les pages nouvellement versées et qui sont donc entièrement à transcrire, ou, à l'inverse, de n'afficher que les pages avec des transcriptions approuvées.

Fig. 5a : Page regroupant l'ensemble des projets : fonctionnalité de tri et de visualisation.

Fig. 5b : Page d'un projet : fonctionnalité de tri par statut.

Le site a également pour objectif de créditer et de mettre toujours plus en valeur les contributeurs et les contributrices. La **liste des contributions** les plus récentes est désormais présente sur la page d'accueil mais aussi au niveau de chacun des projets et de chacun des items d'un projet. Elle permet de renseigner les informations suivantes : *Nom du contributeur* + *Titre média* + *Titre item* + *Titre Projet*.

Fig. 6a et 6b : Page item et page média : panneaux des contributions.

Chaque **compte utilisateur** dispose d'une page publique sur laquelle est rappelée l'ensemble de ses contributions.

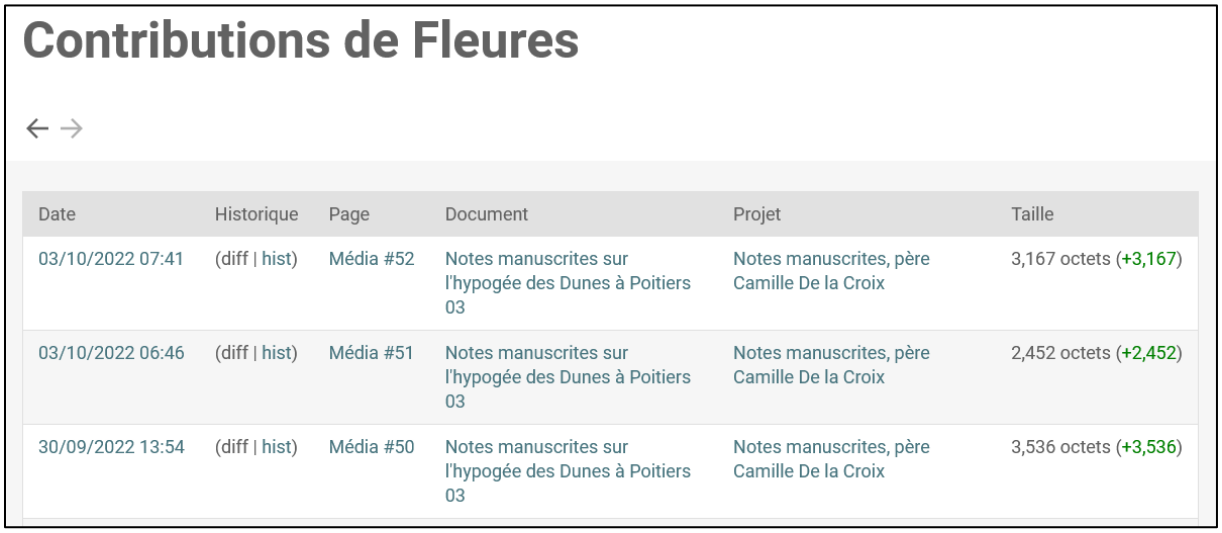

Fig. 7 : Page d'une contributrice

De même, les pages Révision permettent d'afficher tout l'**historique d'une transcription** et de pouvoir comparer les différentes versions de celle-ci dans le temps.

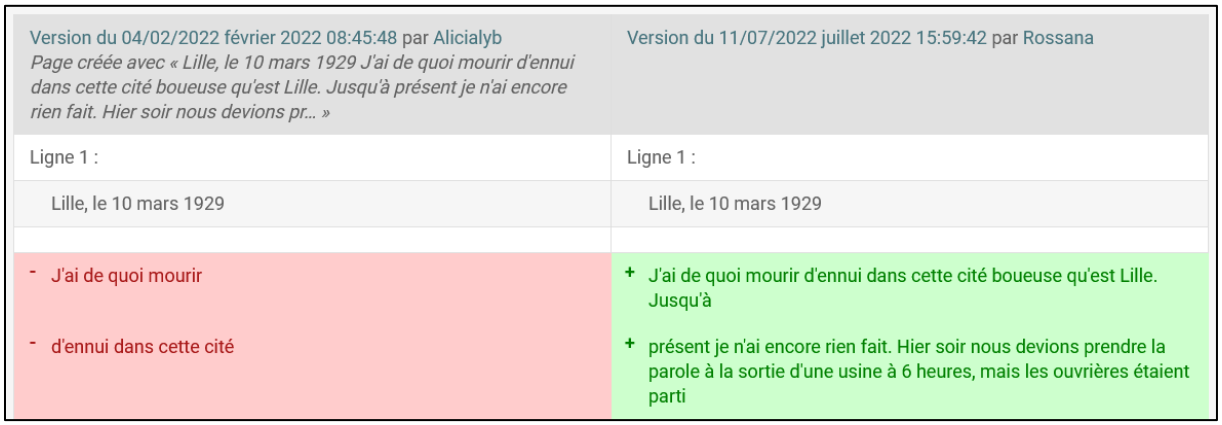

Fig 8a et 8b : Pages Révision.

Un **système de commentaires** (Notes) donne la possibilité à tous les utilisateurs de commenter la transcription d'un document. Celui-ci n'a pour le moment quasiment jamais été utilisé.

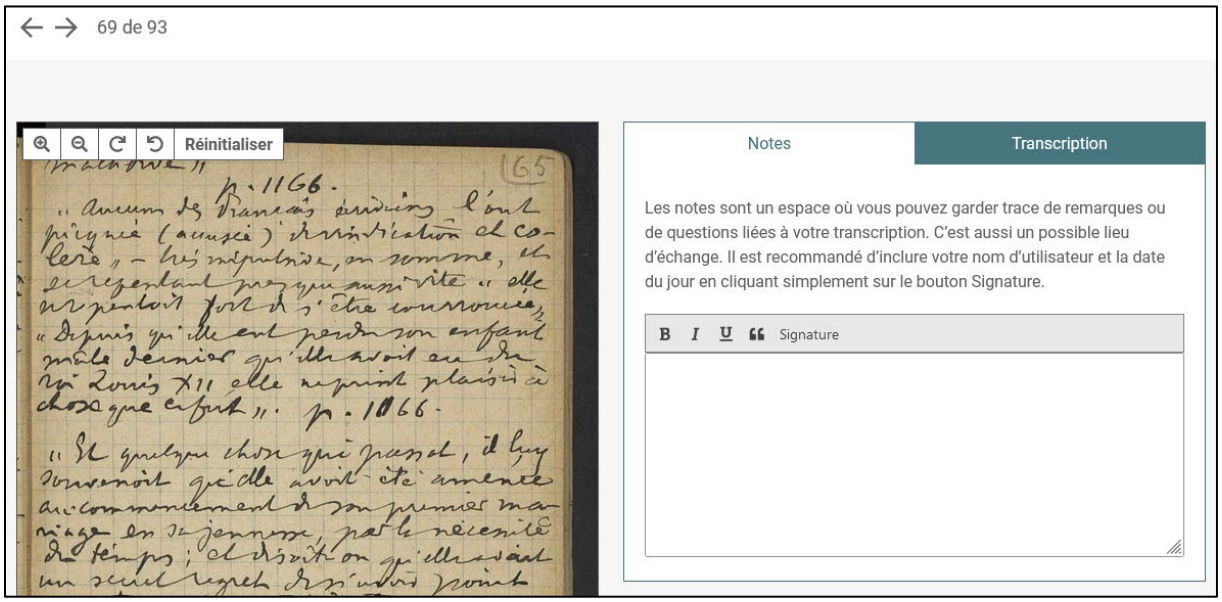

Fig. 9 : Page Notes

Enfin un contributeur peut à tout moment signaler (case à cocher) une **transcription terminée**, celle-ci est alors tout de suite visible pour les gestionnaires / valideurs d'un projet et peut ainsi être rapidement approuvée, cela permet alors de clore une transcription.

Au niveau d'un media (une page) **la visionneuse** unique permet à la fois de lire une image serveur comme une image IIIF, elle est accompagnée en vis à vis d'un **éditeur de texte**.

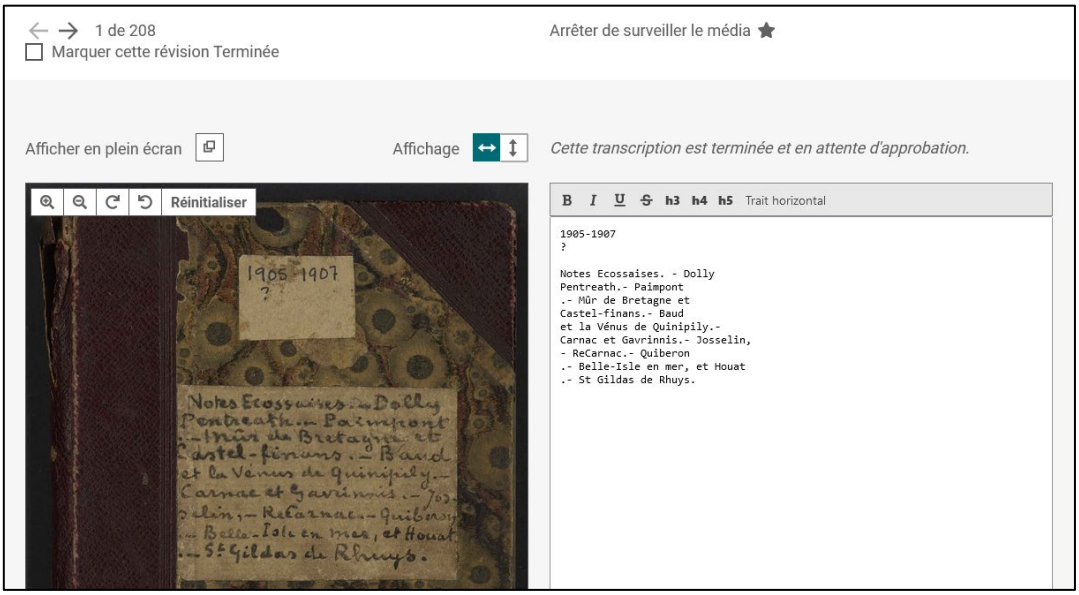

Fig. 10 : Page d'un media : révision terminée, mode plein écran, visionneuse et éditeur de texte.

L'affichage des métadonnées se fait lui aussi au niveau de chacun des items et de chacune des images numérisées. Cela contribue à tout moment à augmenter le niveau d'information sur le contexte de production de ces archives grâce à la présence des métadonnées Dublin Core qui, en plus d'éléments descriptifs, permettent de rebondir vers les inventaires (Calames) ou les catalogues des partenaires.

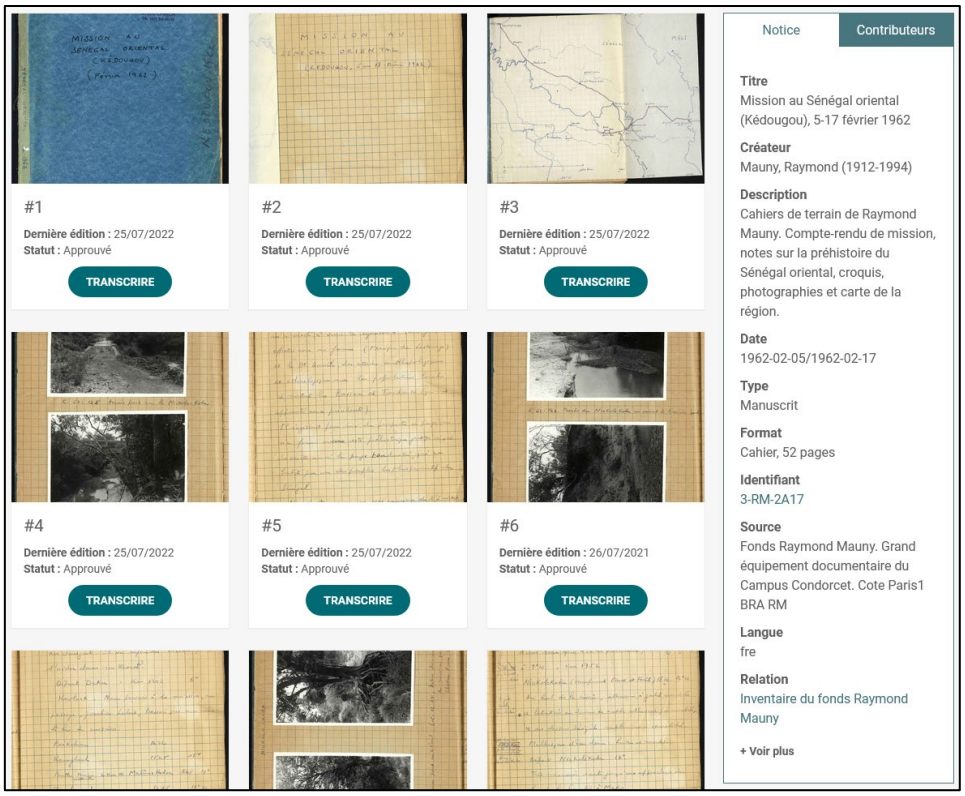

Fig 11 : Panneau 'Notice' qui affiche les métadonnées Dublin Core d'un item, ici un Cahier de terrain de Raymond Mauny avec un lien vers son emplacement dans l'inventaire Calames du fonds.

La recherche plein texte

Le site présente des fonctionnalités nouvelles permettant de faire des recherches parmi les contenus. Le module SolR a été modifié pour pouvoir répondre aux besoins de l'application : paramétrage de l'instance Solr, implémentation du module ScriptoSearch et indexation des contenus (métadonnées et transcriptions). La recherche à facettes donne aux utilisateurs les moyens de filtrer les résultats d'une requête en choisissant un ou plusieurs critères. Il est possible de naviguer au sein des différents contenus par statut, projet, document, genre, discipline et institution de conservation.

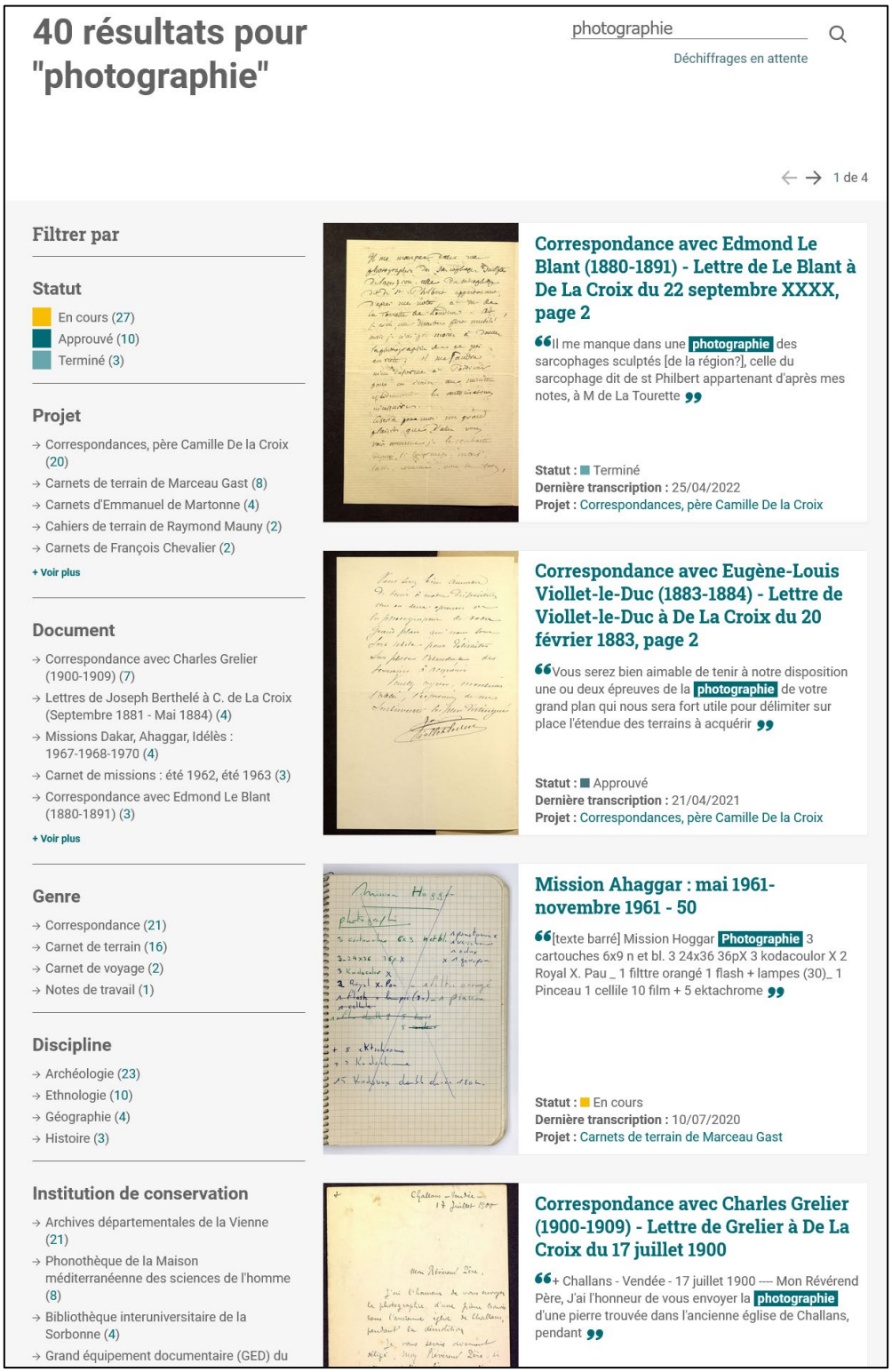

Fig. 12 : Système à facettes et résultats d'une recherche sur Transcrire.

Aide et documentation technique et fonctionnelle

Un Guide d'utilisation réalisé par le prestataire décrit l'architecture technique de l'application Transcrire sous Omeka-S. Il est utile en cas d'intervention technique sur l'applicatif ou la base de données (mises à jour de plugins ou d'Omeka par exemple).

Une présentation du travail de refonte est accessible ici : https://www.limonadeandco.fr/transcrire-avec-omeka-s-et-scripto/

Deux ateliers pratiques ont été compris dans la prestation : une formation a déjà eu lieue, la prochaine se déroulera à partir du 14 décembre. Elles se font sur 3 jours incluant une dizaine de personnes.

- Jour 1 : Découvrir Omeka S et juger de son potentiel

- Jour 2 : Atelier mécano - mettre en ligne ces premières données

- Jour 3 : Comment collaborer au projet Transcrire ?

Le but est d'aider les partenaires à monter en compétence et en autonomie, de les impliquer davantage en décentralisant la gestion des collections et la validation des transcriptions sur la plate-forme.

De même des livrets d'accompagnement pour le chargement des images (dépôt sur le serveur ou import via le protocole IIIF) ont été réalisés. Un est encore en cours de rédaction et concernera la chaine de validation des transcriptions avec Scripto.

Circulation des données

Ainsi que nous l'avons mentionné plus haut, l'import d'images et de métadonnées via le protocole IIIF depuis des entrepôts (Gallica, Nakala) est opérationnel sur Transcrire. Nous pouvons ainsi proposer cette procédure à nos partenaires, les encourageant à la fairisation de leurs données.

Partager les données produites sur Transcrire était aussi un des objectifs de ce projet CollEx-Persée. L'un des besoins exprimés par certains des partenaires était de pouvoir récupérer et réinjecter les données produites sur la plate-forme (**transcriptions, noms des contributeurs**) dans leurs propres systèmes d'information ou leurs bibliothèques numériques.

Plusieurs pistes ont été étudiées pour ce faire car il existe de nombreuses technologies pour ce type d'échanges. En premier lieu Omeka S dispose d'une API REST qui fournit une interface de programmation d'application (API). Celle-ci permet de visualiser, rechercher et faire circuler des informations en Json. Les transcriptions sont ainsi récupérables par des tiers. Néanmoins tous les partenaires ne disposent pas des compétences informatiques pour utiliser ce service web et effectuer ces opérations. Il nous a semblé important de simplifier l'accès aux données produites.

L'initiative fut portée par nos partenaires de Poitiers. Michael Nauge (Laboratoire FoReLLIS/MIMMOC, Université de Poitiers) a ainsi réalisé un **mécanisme d'export des transcriptions** :

https://gitlab.huma-num.fr/mnauge/omekas-transcrire_export

Les codes ont été réalisés lors d'un hackathon avec des étudiants dans le cadre de la formation CNAM Ingénieur Big Data et Intelligence Artificielle en convention avec l'Université de Poitiers.

https://gitlab.huma-num.fr/mnauge/cnam-athon-transcrire

Le dispositif consiste en un développement en Python (cf codes) et une interface graphique minimaliste utilisant un notebook Jupyter "déployé" en application web avec le système voilà.

Le partenaire récupère un dossier .ZIP nommé grâce au "title H1" du document (item d'un projet) et de la date-heure du lancement de l'export. Ce dossier contient :

- un fichier TXT par page du document. Chaque fichier TXT est nommé en utilisant l'identifiant présent dans la notice de chaque page.
    - exemple : Identifiant = FRAD86_16J3_38_138_10621.jpg pour la page https://transcrire.huma-num.fr/scripto/8/108/10582 va produire un fichier TXT nommé FRAD86_16J3_38_138_10621.txt
- un tableur XLSX nommé transcriptionsMerged.xlsx agrégeant l'intégralité des transcriptions. Ce tableur contient deux colonnes :
    - idFile : l'identifiant de page
    - transcription : la transcription de la page identifiée
- un tableur XLSX nommé contributorHistory.xlsx agrègent l'historique des interactions sur la transcription de chaque page. Ce tableur contient quatre colonnes :
    - identifier : l'identifiant de page
    - date : date d'interaction sur la transcription de la page
    - etat : valeurs possibles (en cours, approuvé)
    - auteur : le pseudo Transcrire de l'auteur de l'interaction
    - idPage : l'url de la page

Cette solution a permis à trois projets (Carnets Antoine d'Abbadie, Archives du Père De la Croix et Cahiers de Raymond Mauny) de récupérer les textes transcrits. Il est prévu de proposer, pour certains projets, le dépôt des fichiers textes de transcription sur l'entrepôt de données Nakala au côté des images (par exemple dans la collection Nakala des cahiers de terrain de Raymond Mauny), assurant ainsi leur partage et leur conservation à long terme.

Deux autres méthodes d'exposition des données sont en cours de test et seront opérationnelles d'ici la fin de l'année (grâce à un budget du GED Condorcet).

- Utilisation du protocole **OAI-PMH** grâce à des développements permettant de rendre moissonnables les métadonnées (transcriptions incluses) situées au niveau de l'item et du média.
- **Développement IIIF** : installation du plugin IIIF et récupération des métadonnées situées au niveau de l'item et du média pour les intégrer à des manifestes (json).

Collections et documents nouvellement déposés sur Transcrire

Avant le démarrage du projet CollEx-Persée en 2020, 10 collections (Projets) étaient présentes sur la plate-forme, comptant près de 12 000 fichiers image :

- 15 cahiers de terrain de Raymond Mauny sur l'Afrique de l'Ouest (archéologie)

- 5 carnets de terrain de Marceau Gast sur le Sahara (ethnologie)

- 7 carnets d'Annie-Hélène Dufour sur la Provence (ethnologie)

- 9 journaux de collectes d'Anatole Le Braz sur la Bretagne (histoire et ethnographie)

- 4 carnet d'Emmanuel de Martonne sur la France et la Roumanie (géographie)

- 2 carnets de François Chevalier sur le Mexique et l'Amérique centrale (histoire)

- 30 lots de correspondances scientifiques du père Camille De la Croix (archéologie)

- 31 lots de notes manuscrites du père Camille De la Croix (archéologie)

- 8 cahiers de terrain d'Edmond Bernus sur le Sahel nigérien (géographie)

- 9 lots de notes de terrain de Thérèse Rivière sur les Aurès en Algérie (ethnologie)

Depuis mars 2021, **trois nouveaux projets** ont été ajoutés, dont deux qui ouvrent le périmètre éditorial de la plate-forme à d'autres types de documents historiques (hors archives des sciences humaines et sociales) :

- Carnets d'Antoine d'Abbadie : comptant 13 carnets scientifique (2951 images) centrés sur la Corne de l'Afrique de 1837 à 1848. Les carnets sont conservés à la Bibliothèque nationale de France et disponibles sur Gallica. Ce projet fait partie du plan quadriennal de la recherche 2020-2023 de la BnF et est mené en collaboration avec l'Institut des mondes africains (IMAF).

- Correspondances militantes : comptant 23 lots de correspondances (262 images) utiles à la découverte de l'histoire ouvrière et sociale. Le GED Campus Condorcet et le Centre d'histoire sociale ont déposé ces archives et participent à leur transcription.

- Autour des correspondances féminines (1700-1840) : comptant pour le moment 5 lots de correspondances (80 images) et représentant une source importante pour faire une histoire de la langue française et des écritures féminines. La collaboration s'effectue avec le SCD et le laboratoire FoReLLIS de l'Université de Poitiers.

De même, de **nouvelles archives** à transcrire ont été ajoutées aux projets déjà présents :

- Notes de terrain de Thérèse Rivière : 7 lots de notes de terrain (316 images), Musée du quai Branly.

- Correspondances du père Camille De la Croix : 10 lots de correspondances (573 images), SCD Poitiers.

- Carnets de terrain de Marceau Gast : 3 carnets (343), MMSH Aix-en-Provence.

- Cahiers de terrain de Raymond Mauny : 5 cahiers (557 images), GED campus Condorcet.

- Carnets d'Emmanuel de Martonne : 5 carnets (174 images), Bibliothèque Inter-universitaire de la Sorbonne.

En septembre 2022, la plate-forme accueille ainsi **13 collections** et près de **17 000 images** au total.

A partir de la mise à disposition de la nouvelle version de Transcrire, en mars 2021, les contributrices et contributeurs du site (près de 50 nouveaux comptes actifs ouverts) ont poursuivi les tâches de déchiffrement des textes manuscrits. Actuellement **12 860 documents ont été complétement transcrits** ou sont en cours de transcription.

Travail de validation des transcriptions

Étant donné l'importance quantitative des textes produits par la transcription, la question de la vérification de ce travail devenait délicate.

Un des chantiers importants, porté par le musée du quai Branly - Jacques Chirac, concernait donc la validation de ces transcriptions. En 2022, deux personnes ont été recrutées spécifiquement afin de vérifier, corriger et **approuver plus de 7 500 pages transcrites**. Ce travail a permis de clore le traitement d'une importante partie des documents confiés qui sont dorénavant prêts à être réutilisés par les partenaires (voir quelques exemples plus bas dans la partie concernant les réutilisations).

Rappelons qu'une des nouvelles fonctionnalités de Transcrire donne la possibilité aux contributrices et contributeurs de participer aussi au processus de révision en signalant aux gestionnaires de collections les pages complètement transcrites.

Ce travail a aussi permis de clarifier la chaine Scripto de suivi validation des transcriptions dans le backoffice. Il s'agit ainsi tout d'abord des droits et des rôles à allouer aux personnes impliquées dans les projets, tels que contributeur, chercheur, éditeur, relecteur, superviseur. Ensuite les différents statuts d'un document ont été définis, nouveau, en cours, terminé, approuvé. Enfin, l'utilisation des fonctionnalités facilitant le traitement en lot des validations a été précisée. L'ensemble de ces éléments donneront lieu à de petits guides de prise en main personnalisés pour une plus grande autonomie des partenaires dans la gestion de leurs projets numériques.

Les ateliers de transcription collaborative

Afin d'animer les communautés de contributeurs, de réunir ceux et celles déjà engagées mais aussi d'attirer de nouveaux publics, nous avons durant ces deux années organisé plusieurs ateliers participatifs. Ces transcrithons se sont déroulés au sein des institutions de conservation et ont permis à ce que chercheurs, amateurs, archivistes et bibliothécaires se rencontrent.

Ils ont donné lieu à de mini-conférences au cours desquelles des chercheurs et des spécialistes présentaient au public les collections, les documents les composants et les scientifiques les ayant produits. C'était aussi parfois l'occasion d'exposer quelques-unes de ces archives afin de faire découvrir ce patrimoine dans toute sa matérialité.

En moyenne une quinzaine de personnes étaient chaque fois réunies lors de ces événements.

Une forme virtuelle de ces ateliers de transcription collaborative a également été mis en place sur le projet des Carnets d'Antoine d'Abbadie, à la demande des quelques personnes les plus engagées dans l'activité de transcription, quelques mois après le début du travail collectif. Nous avons mis en place un rendez-vous hebdomadaire, le mardi de 14h à 17h, pendant lequel une session de visio-conférence était ouverte et à laquelle se joignait qui voulait. Dénommé les "#Mardidabbadie", ces ateliers ont eu lieu sans interruption de janvier à fin juin 2022. Le travail de découverte et de transcription du texte est ainsi resté très collectif, suscitant des discussions précieuses sur les contenus et sur leurs usages futurs.

Listes des transcrithons :

- 7 juillet 2021 : Atelier participatif « Carnets Antoine d'Abbadie », Campus Condorcet, Aubervilliers. https://bnf.hypotheses.org/10423
- 30 septembre et 1 octobre 2021 : Deux ateliers organisés pendant les « Épopées » du Campus Condorcet à l'occasion de la fête de la science, GED Campus Condorcet, Aubervilliers. https://www.fetedelascience.fr/transcription-collaborative
- 3 novembre 2021 : Atelier participatif « Carnets Antoine d'Abbadie » à l'Ecole thématique Digital Areal réunissant des spécialistes de langues en écritures non latines, Fréjus.
- 13 décembre 2021 : Atelier participatif « Carnets Antoine d'Abbadie », DataLab de la Bibliothèque nationale de France. https://bnf.hypotheses.org/10944
- 18 mai 2022 : Atelier participatif autour des archives de l'anthropologue Marceau Gast (1927-2010), MMSH Aix-en-Provence à l'occasion du « Mois de l'engagement des archivistes ». https://phonotheque.hypotheses.org/36001
- 13 octobre 2022 : Transcrithon autour des notes de terrain de Thérèse Rivière : mission ethnographique dans l'Aurès (Algérie, 1934-1936), salon Jacques Kerschache, musée du quai Branly – Jacques Chirac.
- 7 novembre 2022 : Transcrithon durant l'Open Access Week organisée par le GED du campus Condorcet à Aubervilliers du 7 au 11 novembre.

Un petit guide de mise en place d'atelier de transcription collaborative, à destination des bibliothèques partenaires, est en cours de finition, réalisé en collaboration avec la MMSH d'Aix-en-Provence.

Utilisations de la plateforme Transcrire et exploitation des transcriptions

Sciences participatives

Qu'a permis ce projet ? En premier lieu d'étoffer un outil de sciences participatives, qui sont encore assez rares en France. De tels dispositifs donnent à d'autres personnes que les chercheurs la possibilité d'un engagement actif avec des documents d'archives. Ils participent ainsi au mouvement de la Science ouverte.

Au-delà d'une présentation virtuelle du contenu de leurs collections, les institutions scientifiques et culturels offrent une expérience de médiation originale impliquant de manière plus profonde et plus intense les usagers.

Ce n'est ainsi pas seulement un cadre productif pour améliorer un travail sur des collections, c'est aussi la possibilité d'utiliser et d'accroitre la créativité, l'intelligence et le savoir-faire d'un grand nombre de personnes, d'intéresser ceux-ci à la recherche, à la source primaire, au-delà de la synthèse historique.

Le processus de transcription associe en effet pleinement les contributeurs à la démarche savante. Devenant expert d'une écriture, ils participent au nécessaire travail préparatoire de traitement des sources. Grâce à une lecture précise et fine des documents, ils acquièrent une nouvelle compréhension de ces derniers et renouvellent les connaissances à leurs égards.

Meilleure accessibilité des collections et enrichissement des métadonnées

Nous l'avons vu la plate-forme Transcrire offre la possibilité d'extraire de l'information à partir d'images numériques et fournit des textes interrogeables au moyen du moteur de recherche. Cette mise à disposition du plein texte permet une grande variété d'exploitations, les objectifs scientifiques et documentaires étant différents pour chaque projet : nouvelles navigations au sein de corpus d'archives, au-delà des simples métadonnées, possibilité de croiser les documents, accessibilité plus grande aux données.

Qu'en est-il pour le travail des bibliothèques ? Voici quelques exemples de redocumentarisation à travers l'enrichissement de bibliothèques numériques, de catalogues ou de projets des partenaires.

- Pour la bibliothèque numérique du **Fonds d'archives de l'archéologue Camille de La Croix**[1], l'équipe récupère, grâce à l'outil d'extraction développé lors de notre projet, les transcriptions des notes et des correspondances de l'archéologue puis les réinjecte dans leur instance Omeka.

http://fondspdlc.edel.univ-poitiers.fr/

Sur Transcrire les contributeurs ont produit la transcription suivante :
https://transcrire.huma-num.fr/scripto/8/96/9878

… qui par la suite est réinjectée dans la bibliothèque numérique du Fonds :
http://fondspdlc.edel.univ-poitiers.fr/items/show/5918

On peut aussi voir, parmi les métadonnées, le nom de la transcriptrice.

[1] Projet porté par l'équipe de recherche HERMA (Université de Poitiers), le Service Commun de Documentation et la MSHS de l'université de Poitiers, les Archives Départementales de la Vienne.

- Dans le cadre du projet **Carnets d'Antoine d'Abbadie**, parallèlement et au fur et à mesure de la transcription des textes, les manuscrits ont été décrits dans le catalogue Archives et Manuscrits de la BnF. Cet enrichissement de la description de la collection de la BnF est considérable puisque jusqu'à présent la description de ces carnets reposait sur une reprise des tables des matières parfois proposées par Antoine d'Abbadie lui-même dans les carnets, qui étaient au mieux partielles, et souvent inexistantes.

Enrichissement de la notice du carnet 270, dans le catalogue de la BnF
https://archivesetmanuscrits.bnf.fr/ark:/12148/cc109208d
En retour les métadonnées du carnet sur Transcrire pointent vers une notice plus complète (champ Dublin Core Relation)
https://transcrire.huma-num.fr/scripto/13/14131/media

- Concernant le projet sur les **carnets de l'ethnologue Marceau Gast**, l'information transcrite complète les métadonnées des enregistrements sonores présents sur Ganoub, la base de données de la phonothèque de la MMSH d'Aix-en-Provence. Les données extraites des carnets offrent une nouvelle porte d'entrée dans le contexte de l'enquête, plus détaillée, permettant de retrouver des éléments biographiques sur des informateurs, des dates et d'autres informations précieuses pour décrire les matériaux audiovisuels. De la même manière cette information sert à redocumenter de lots de photographies par un meilleur repérage de lieux ou d'itinéraires.
https://medihal.archives-ouvertes.fr/search/index/q/*/authFullName_s/Marceau+Gast
Maryasha Barbé, « Croiser et connecter les archives : le fonds Marceau Gast », https://phonotheque.hypotheses.org/32130, 2020.

- Pour le projet sur les **cahiers d'Annie-Hélène Dufour**, l'inventaire du fonds disponible sur Calames mentionne et pointe vers les archives disponibles sur Transcrire : http://www.calamles.abes.fr/pub/ms/FileId-1760

- Autre exemple enfin, les **carnets de terrain d'Emmanuel de Martonne** sur la plateforme de recherche *Navigae* permettant de consulter des données issues de travaux en géographie.
https://www.navigae.fr/fieldnotes/?i=demartonne-c4

Une carte interactive y retrace l'itinéraire d'une excursion du géographe en Roumanie en 1921. Des photographies prises par de Martonne, les scans des carnets et la transcription des pages accompagnent ce dispositif.

Articuler transcription manuelle et transcription automatisée

Deux études ont porté sur l'articulation de la plate-forme avec des solutions de reconnaissance automatique d'écriture manuscrite (HTR). En effet le déchiffrement collaboratif de manuscrits numérisés permet aussi de produire des données d'apprentissage afin d'alimenter et d'entrainer un algorithme de reconnaissance d'écriture manuscrite.

Fin 2020, Quentin Mercurol, qui effectuait son stage professionnel au Grand Équipement Documentaire du Campus Condorcet lors sa formation à l'Enssib, avait pour mission de réfléchir à la mise en œuvre d'un service de transcription des textes manuscrits à destination des publics chercheurs. C'est dans ce cadre qu'il a réalisé un travail de comparaison entre les outils Transkribus et eScriptorium. En collaboration avec Fabrice Melka (IMAF) et Jean-Baptiste Pressac (Centre de recherches bretonnes et celtiques), il a préparé et lancé plusieurs tests à partir des transcriptions des journaux d'Anatole Le Braz produites sur Transcrire. L'idée était, à partir des mêmes données d'entrainements, de bien comprendre l'utilisation de ces outils, de comparer leurs fonctionnalités et leurs performances ainsi que les possibilités de circulation des données entre plates-formes.

L'ensemble de ce travail conséquent a été publié sur le carnet *DLIS, Digital Libraries & Information Science*. Une partie y est consacrée à la transcription automatique :

Quentin Mercurol, « Transcrire en bibliothèques ? Légitimité, enjeux et méthodes pour services aux chercheurs », 29/06/2021. https://dlis.hypotheses.org/5614

L'autre travail impliquant les technologies d'HTR est actuellement en cours. Il concerne le projet « Carnets d'Abbadie », assez spécifique de par ses documents à la mise en page complexe et ses problématiques d'écritures manuscrites non-latines. Là aussi Transcrire a permis de créer des jeux de données d'entrainement. Les transcriptions produites couvraient ainsi volontairement tous les cas possibles de mises en page et d'écriture (latine, signes diacritiques, alphasyllabaire éthiopien, arabe, hébreu, grec, sténographie) contenus dans ces carnets. Des essais avec les outils Transkribus et E-Scriptorium ont d'abord été réalisés entre décembre 2020 et juin 2021, avec un succès mitigé. Néanmoins, le texte de l'un des carnets (ms. 266) acquis par Transkribus fut reversé en ligne dans Transcrire. Il s'est avéré que l'une des transcriptrices a préféré corriger les résultats de cette HTR plutôt que de transcrire intégralement les pages manuscrites. De plus, même si relativement fautive, cette transcription issue de l'HTR permettait une fouille de texte et c'est ainsi que certaines pages au contenu identifié ont pu être transcrites en priorité en fonction des thèmes de recherche des participant·es.

Pour poursuivre le travail d'HTR il a été fait appel à un prestataire, Calfa, spécialiste de l'HTR pour les langues peu couvertes. Le travail d'alignement des textes et des images, la création et l'amélioration des modèles d'apprentissage, leur application et l'examen des performances de ces derniers, nécessitent du temps et une expertise qui explique le recours à un prestataire de service spécialisé et l'abandon des outils dédiés à un usage par les équipes du projet.

Actuellement (septembre 2022) la totalité des textes du corpus (17 cahiers) est en cours d'extraction par le prestataire Calfa, grâce des co-financements de l'EHESS, de la BNF et de l'Institut des Mondes Africains (IMAF). Les livrables sont attendus pour la mi-octobre 2022.

Valorisation autour du projet CollEx-Persée Transcrire (manifestations, articles, billets de blog, …)

Cette liste de productions et événements scientifiques est présentée par ordre chronologique.

- « Programmes de recherche - MSS-ABBADIE : Les carnets scientifiques d'Antoine d'Abbadie en Éthiopie (1840-1852) : numérisation, transcription et édition » http://actions-recherche.bnf.fr/BnF/anirw3.nsf/IX01/A2019000024_mss-abbadie-les-carnets-scientifiques-d-antoine-d-abbadie-en-ethiopie-1840-1852-numerisation-transcription-et-edition

- Maryasha Yael Barbé, « In the footsteps of the ethnologist: the missions of Marceau Gast in the southern Sahara through his field notebooks », sur le croisement des sources et l'utilité de la transcription pour de nouvelles métadonnées. Blog *Ethnologia*, https://ethnologia.hypotheses.org/1457, 2 novembre 2020.

- Vanessa Desclaux, Anaïs Wion, Mathilde Alain, « Les manuscrits d'Abbadie et la plateforme Transcrire », communication lors de l'atelier Transcription des écritures manuscrites organisé par le Data Lab de la Bibliothèque nationale de France. Video de l'atelier disponible sur https://bnf.hypotheses.org/10240, 26 janvier 2021.

- Mathilde Alain, « Transcrire les carnets d'Antoine d'Abbadie (1837-1848) », *Carnet de la recherche à la Bibliothèque nationale de France*, https://bnf.hypotheses.org/10253, 29 mars 2021.

- Anaïs Wion, « Transcrire les carnets d'Antoine d'Abbadie #1 », *Carnet d'Afriques, revue d'histoire*, https://afriques.hypotheses.org/1194, 19 avril 2021.

- Mathilde Alain, « Transcrire les carnets d'Antoine d'Abbadie #2 - Les trois journaux de voyage d'Antoine d'Abbadie : la première strate d'une longue activité scientifique », *Carnet d'Afriques, revue d'histoire*, https://afriques.hypotheses.org/1478, 27 mai 2021.

- Mathilde Alain, Vanessa Desclaux, Anaïs Wion, « Les carnets d'Éthiopie d'Antoine d'Abbadie », *Chroniques de la Bibliothèque Nationale de France*, vol. 92, 2021, p. 42–43.

- Quentin Mercurol, « Transcrire en bibliothèques ? Légitimité, enjeux et méthodes pour des services aux chercheurs », Blog *DLIS, Digital Libraries & Information Science,* https://dlis.hypotheses.org/5614, 29 juin 2021.

- Anaïs Wion, « TEI-XML edition of the notebooks of Antoine d'Abbadie », https://medihal.archives-ouvertes.fr/hal-03339615, 7 juillet 2021.

- Anaïs Wion, Mathilde Alain, « Transcribing Antoine d'Abbadie's notebooks : tutorial for using the transcription tool », https://medihal.archives-ouvertes.fr/hal-03339609, 7 juillet 2021.

- Anaïs Wion, Mathilde Alain, « Transcrire Antoine d'Abbadie : tutoriel d'utilisation de l'outil de transcription », https://medihal.archives-ouvertes.fr/hal-033395997, 7 juillet 2021.

- Yaël Caux « Transcrire»: la plateforme qui ressuscite les journaux de bord des ethnologues, RFI, https://www.rfi.fr/fr/culture/20210731-transcrire-la-plateforme-qui-ressuscite-les-journaux-de-bord-des-ethnologues, 31 juillet 2021.

- Nadine Dieudonné-Glad, « Les vestiges archéologiques découverts à Poitiers par le père de la Croix : du croquis numérisé et des notes, au plan et à la carte géoréférencée, construction d'un outil de connaissance et de gestion du patrimoine », Atelier *Numériser et mettre en ligne des fonds patrimoniaux anciens*, université de Poitiers, 4 octobre 2021.

- Anaïs Wion, « Transcrire de façon participative une source complexe : enjeux et difficultés du projet d'édition des carnets d'Antoine d'Abbadie », *Un œil sur la corne / An eye on the Horn*, https://cfee.hypotheses.org/7739, 1er novembre 2021.

- Véronique Ginouvès, « La transcription d'images de carnets de recherche sur la plateforme Transcrire", Atelier professionnel du Master Métiers des archives et des bibliothèques, médiations de l'histoire et humanités numériques, MMSH, https://masterabd.hypotheses.org/8256, 2 novembre 2021.

- Mathilde Alain, « Noter, classer, utiliser : les carnets de voyage d'Antoine d'Abbadie en Éthiopie ». *Sources. Materials & Fieldwork in African Studies*, n° 3 : 137-188. https://www.sources-journal.org/560, 2021.

- Gaëlle Hallair, Présentation de la plateforme Transcrire lors inauguration de l'exposition bilingue franco-roumaine intitulée « E. de Martonne : 1921-2021. Centenaire des grandes excursions en Roumanie », Bibliothèque de l'Institut de géographie de Paris, https://miralta-edito.wixsite.com/centenaireexposition, 13 décembre 2021.

- Les carnets E. de Martonne sont valorisés dans le cadre du projet de recherche intitulé « Corpus Emmanuel de Martonne » et coordonné par Gaëlle Hallair (CNRS) à l'UMR Géographie-cités. Ce projet est le fruit d'un partenariat avec la Bibliothèque Inter-universitaire de la Sorbonne, l'UMR PRODIG, le consortium IMAGEO, la TGIR Huma-num et des collègues étrangers.

- Véronique Ginouvès, « Avent 2021 – Annie-Hélène Dufour, l'anthropologue, l'enseignante et ses archives », blog *Archives de la recherche & Phonothèque*, https://phonotheque.hypotheses.org/34147, 8 décembre 2021.

- Maryasha Barbé, « Avent 2021 – Les archives de Marceau Gast, l'anthropologue pédagogue », blog *Archives de la recherche & Phonothèque*, https://phonotheque.hypotheses.org/33989, 18 décembre 2021.

- Dominique Harre, « Transcrire les carnets d'Antoine d'Abbadie #3 - Antoine d'Abbadie ethnographe, les caravanes et les marchands d'Éthiopie », *Carnet d'Afriques, revue d'histoire*, https://afriques.hypotheses.org/1478, 8 mars 2022.

- Atelier sur les archives de Marceau Gast, MMSH Aix-en-Provence, Mois de l'engagement des archivistes, un atelier collectif permettant à chaque participant de contribuer à l'enquête sur le contexte de production des documents, indispensable pour reconstituer le travail de recherche du chercheur, et à leur mise à disposition du public académique ou grand public via les sites de science ouverte ou d'archives ouvertes (le site Transcrire de transcription collaborative, le portail MediHal de diffusion d'images), 18 mai 2022.

- Claire Cialone-Grégoire, Véronique Ginouvès et Emilie Groshens, « Plan de gestion des données rétrospectif du fonds Annie-Hélène Dufour », le PGD précise l'archivage des transcriptions des carnets. Les textes seront exportés pour l'archivage pérenne au CINES avec les fonds, https://nakala.fr/10.34847/nkl.eef91t84, 1er septembre 2022.

- Véronique Ginouvès, « Consulter les archives d'Annie-Hélène Dufour, demain », blog *Archives de la recherche & Phonothèque*, https://phonotheque.hypotheses.org/37299, 1er septembre 2022.

- Fabrice Melka, Participation à la journée d'études du projet CollEx-Persée FamiliLettres, table-ronde *Les enjeux d'une transcription : pourquoi, comment et pour qui transcrit-on ?*, 7 octobre 2022.

- Julien Bondaz, « Learned excursions and scientific tourism on the occasion of the first International Conferences of West Africanists (1945-1949) », tiré du travail de transcription d'un des cahiers de Raymond Mauny présent sur Transcrire. Blog *The Scientific Conference: a Social, Cultural, and Political History*, https://sciconf.nu/blog, octobre 2022.