Transcript, un module EMAN pour Omeka classic : de la bibliothèque à l'édition numérique

Julie Giovacchini, Céline Bohnert

Transcript, un module EMAN pour Omeka classic : de la bibliothèque à l'édition numérique*

Julie Giovacchini, Centre Jean Pépin (UMR 8230) CNRS-ENS-PSL
Céline Bohnert, CRIMEL, Université de Reims Champagne-Ardennes / IUF

L'objet de cette contribution sera de présenter dans un premier temps la méthodologie et le contexte de conception du module Transcript, développé pour la communauté EMAN comme plugin intégré au sein d'un environnement de bibliothèque Omeka classic, puis de mettre en évidence les questionnements épistémologiques que soulève cet outil original d'édition au sein de cet environnement. Nous appuierons notre présentation sur des exemples issus de trois projets EMAN utilisant le module Transcript : le projet Mythologia, le projet Marcianus et le projet Epicurei.

La communauté EMAN, collectif qui réunit de nombreuses personnes du monde de la recherche et de l'enseignement supérieur, ingénieurs, enseignants chercheurs, agents des bibliothèques, collabore depuis plusieurs années à l'élaboration d'une plateforme destinée à la publication de bibliothèques numériques axées sur les publications d'archives et de fonds imprimés et manuscrits. Elle s'est progressivement ouverte aux problématiques d'édition numérique, en proposant un environnement d'abord de transcription puis d'encodage pour certains de ses objets. Cette démarche a permis une ouverture encore plus importante à de nouveaux types de textes, et un souci croissant de favoriser la bibliodiversité des contenus mis en valeur par la plateforme[1].

EMAN se compose :

- d'un comité de pilotage avec des représentants de chaque projet ;

- d'un bureau interne qui se réunit chaque mois pour discuter des développements ;

- de deux développeurs ;

- de groupes de travail dédiés à certaines fonctionnalités (transcription, cartographie, Zotero, etc)[2].

Le modèle de développement d'EMAN peut être qualifié de mutualiste : chaque développement financé profite à plusieurs projets et est testé pour déploiement potentiel sur l'ensemble de la plateforme. La plateforme regroupe actuellement plusieurs dizaines de projets et 28 sites publics centrés sur des corpus extrêmement divers, anciens ou contemporains. Les sites EMAN sont principiellement des bibliothèque numériques construites à partir du logiciel Omeka Classic.

Le manifeste EMAN insiste en particulier sur la centralité du document dans la démarche de ses membres :

> Le format numérique du document constitue le point de départ d'un travail ordonné de description, d'enrichissement, d'édition, d'exploitation ; différents traitements peuvent être appliqués au document

* Ce texte reprend et développe les éléments présentés le 27 juin 2023 lors de la conférence Humanistica 2023 à Genève. Le travail ici décrit a été rendu possible par le soutien financier de l'EUR Translitterae (programme Investissements d'avenir ANR-10-IDEX-0001-02 PSL* et ANR-17-EURE-0025) , du laboratoire Thalim, du Centre Jean Pépin, du CRIMEL, de l'URCA et de l'IUF.

1 Sur les risques liés à la standardisation croissante des contenus produits par et pour les humanités numériques, voir Pawlicka-Deger 2021, p. 9.

2 Maria Laura Cucciniello, « Organisation de la plate-forme », site "EMAN (Édition de Manuscrits et d'Archives Numériques)" Consulté le 22/08/2023 sur la plateforme EMAN, https://eman-archives.org/EMAN/organisation.

selon les besoins. Ces traitements génèrent des métadonnées, produisent de nouveaux objets et intègrent le document dans le cycle de vie des données.[3]

Omeka crée un dispositif de consultation qui imite l'exploration en bibliothèque : des collections abritent des documents qui peuvent être consultés en tant que totalité signifiante (notice d'item ou de contenus) mais également feuilletés page après page. Le document, instancié comme fichier image dans le logiciel, est ainsi à la fois le point de départ et le centre de gravité d'un site Omeka.

Comme les précédents modules développés par EMAN (mis à disposition sur Gitlab[4]), le module Transcript est le fruit d'un dialogue entre les besoins exprimés par les porteurs de projet et les possibilités informatiques offertes par Omeka classic assoupli et retravaillé par la plate-forme. Son développement obéit à une méthode ouverte et itérative qui procède par ajustements successifs, de manière à équilibrer la dimension générique de l'outil (gage d'une facilité d'installation et de maintenance qui vont dans le sens de la pérennité) et ses possibilités de personnalisation pour chaque projet[5].

Pouvoir transcrire les documents est la condition pour aller du traitement d'image au traitement de texte. Ce besoin a émergé dans la communauté EMAN du fait de problématiques croisées :

- arrivée de projets avec des corpus complexes pour lesquels la description par métadonnées, si riche soit-elle, ne suffit pas ;

- projets de recherche impliquant un accès direct au contenu des documents (fouille de texte, création de glossaires, édition critique) ;

- volonté d'aller encore plus loin dans les possibilités de navigation aussi bien à l'intérieur d'un site que vers d'autres sites et d'autres ressources numériques.

Une première version du module Transcript est née à la faveur d'un projet spécifique, la publication du corpus des Notes de cours de l'ENS[6]. Cette première tentative s'appuyait sur un cahier des charges particulier :

- une volonté de transcription diplomatique ;

- une technologie légère : un éditeur XML en format wyswyg intégré à l'environnement Omeka (TinyMCE).

Très vite les acteurs du projet se heurtent à deux écueils : d'une part la transcription diplomatique exhaustive, et sa traduction en icônes, créent un objet textuel peu lisible ; d'autre part l'outil est limité et peu adaptable à d'autres projets de transcription ; chaque projet doit en effet construire son propre schéma d'encodage et les méthodes de transcription ne sont donc pas reproductibles d'un site à l'autre.

L'équipe de développement, soutenue par le bureau et certains porteurs de projets, décide donc de développer le module différemment.

Le projet Transcript s'oriente alors vers une solution plus complexe pour intégrer de façon fluide l'édition du texte à son environnement de publication. Cette démarche soulève des questionnements importants, aussi bien techniques (quel standard d'encodage choisir[7], quel outil de visualisation de l'édition ?) que méthodologiques (comment harmoniser ce module avec l'environnement Omeka, comment coordonner le

3 Cucciniello et al. 2022, « Manifeste Eman » article 3b.
4 https://gitlab.com/eman8
5 Walter 2022.
6 https://eman-archives.org/coursENS/
7 Le choix du standard TEI a ainsi été déterminant, en essayant de tirer le meilleur parti d'un standard robuste et éprouvé, tout en évitant l'écueil de l'hypersingularisation qui irait contre la volonté de mutualisation des outils d'EMAN mais aussi contre les impératifs scientifiques d'interopérabilité (voir Schmidt 2014.)

travail des métadonnées propre à la bibliothèque numérique et le travail d'enrichissement sémantique propre à l'édition encodée ?) et épistémiques (quel objet créons-nous, s'agit-il toujours d'une bibliothèque, comment le dispositif ainsi conçu modifie-t-il les conditions de consultation et de lecture pour le visiteur du site ?).

Le besoin de transcription et sa mise en oeuvre diffèrent notamment selon qu'on se propose de travailler sur un document isolé (une page ou un ensemble limité de page constituant un seul item) ou sur un corpus (plusieurs centaines voire milliers de pages réparties dans différents items et différentes collections.)

Ces deux cas de figure induisent des développements spécifiques et des méthodes de travail avec des contraintes parfois opposées. L'enjeu du développement du module Transcript devient alors de proposer un outil compatible avec ces deux approches distinctes de la transcription. La particularité la plus évidente et la plus riche de possibilités de Transcript est qu'il ne s'agit en effet pas d'un module isolé, mais d'une brique supplémentaire connectée à un véritable écosystème de modules implémentés sur la plateforme, dont certains directement développés au sein d'EMAN. Cet écosystème s'organise autour de la notion de corpus, et construit des niveaux de lecture, de consultation, de circulation qui relient les documents et leurs métadonnées de façon variée et parfois extrêmement complexe, ce qui donne toute sa profondeur à la traversée des ensembles documentaires ainsi structurés. Cette particularité structurelle offre également des possibilités d'exploitation numérique et d'investigation scientifique qui lui sont propres. Un enjeu de l'édition numérique au sein de cet écosystème est de ne pas perdre cette fluidité de navigation, et cette vision synthétique globale des corpus, tout en permettant des focus sur des documents isolés qui auront été transcrits et édités, l'éditorisaliation jouant de cette manière sur plusieurs échelles[8].

Nous allons présenter plusieurs exemples qui permettent de montrer comment Transcript dans son état actuel apporte une réponse technique à ces deux types de projet de transcription. Notre premier exemple est celui du projet Mythologia[9], qui consiste en l'« édition numérique de quatre états d'un texte en mutation (Venise 1567, Francfort 1581, Lyon 1612 et Paris 1627) : plus qu'une œuvre, la *Mythologie [de Natale Conti]* constitue un corpus foisonnant auquel collaborèrent éditeurs, correcteurs, traducteurs et graveurs[10]. »

La caractéristique principale de ce projet est sa très grande complexité, et le souci permanent doit alors être de guider la navigation pour que l'utilisateur puisse pleinement profiter des transcriptions intégrées dans leur environnement. Une force de la bibliothèque Omeka est de ce fait de proposer une navigation qui puisse « guider l'utilisateur à travers le volume de données qui lui est présenté[11] ».

L'insertion de la transcription au sein de la bibliothèque rend possible une double approche, analytique et synthétique, du corpus, avec deux entrées de lecture proposées à l'usager : une entrée directe dans la transcription d'un passage via la liste des transcriptions[12], une entrée indirecte par l'exploration du corpus via la notice d'un contenu[13].

Le projet *Mythologia* a fait le choix initial d'éditer une partie du corpus (texte de Paris, 1627) ce qui crée un centre de gravité pour le corpus, qui structure l'ensemble du site et propose conjointement :

- une édition semi-diplomatique, dans une perspective d'histoire du livre : on entre dans la fabrique de l'objet via la graphie, les abréviations, les coquilles ;

---

8 Nous considérons donc que EMAN est une de ces bibliothèques multimodales souhaitées dans l'article fondateur de 2009 de Gregory Crane et al. :"We need fourth-generation collections that can seamlessly integrate image-books, accurate transcriptions, and machine actionable knowledge in various formats."

9 https://eman-archives.org/Mythologia/.

10 Texte de présentation du site, https://eman-archives.org/Mythologia/.

11 Leblanc, Elina. "Review of 'Omeka Classic. Un environnement de recherche pour les éditions scientifiques numériques'." RIDE 11 (2020). doi: 10.18716/ride.a.11.3. Accessed: 14.06.2023.

12 https://eman-archives.org/Mythologia/transcriptions

13 https://eman-archives.org/Mythologia/items/show/1124

- une édition savante avec en perspective la possibilité de comparaison avec les versions antérieures présentes sur le site ;

- une annotation sémantique (à venir) par localisation dans le texte des métadonnées personnalisées (entités historiques et mythologiques, attributs, toponymes, etc.) ;

- une annotation qui prépare une exploitation quantitative du corpus, pour les citations (en cours, actuellement sont signalées la nature des citations, prose / vers)[14].

Cette double approche du corpus est encore approfondie par la possibilité de créer des index thématiques à partir des transcriptions. L'indexation crée alors des rebonds au sein du site Omeka et offre un nouveau chemin possible de déambulation. Ce faisant, elle éloigne de la lecture linéaire et propose une forme d'expérimentation de lecture distante du corpus.

La pluralité des index et de leurs présentations reflète des choix scientifiques différents au sein des projets. Ces choix sont dictés par les besoins de recherche et par la nature même des objets documentaires traités dans les différents sites. Comparons ainsi les index proposés par le site du projet Mythologia et le site du projet Epicurei[15]. Le projet Mythologia propose à la fois des index thématiques[16], liés aux métadonnées des contenus présentés dans chaque collection, et un glossaire[17] des termes remarquables de l'édition transcrite de 1627. Le projet Epicurei propose quant à lui un glossaire thématique[18] permettant de regrouper des termes conceptuellement proches dans différentes collections du site.

Il s'agit là d'une différence prenant racine dans l'origine même de la question de recherche derrière chacun de ces projets. Le projet Epicurei, initié en 2020, avait en effet pour objectif premier la constitution d'un glossaire philosophique épicurien avec possibilités de relier de façon dynamique les entrées de ce glossaire et les textes dans lesquels les termes pertinents sont employés dans le corpus épicurien grec canonique (notamment les *Lettres et Maximes* d'Epicure transmises par Diogène Laërce au livre X des *Vies des philosophes illustres*), de façon à donner un contexte aux entrées du glossaire et de permettre des rebonds du texte au concept et du concept à ses occurrences. La possibilité de transcrire le corpus est alors essentielle car elle rend matériellement possible le rebond. Du côté du projet Mythologia, l'indexation double est plutôt un enrichissement des données qui ne structure pas la navigation mais qui propose d'autres chemins de navigation, transverses, soit par l'étude de la langue et des graphies (via la transcription), soit par l'étude des objets thématiques (noms, personnages, lieux, titres…) sur lesquels un complément de recherche et d'information est proposé.

Dans les deux cas, le point commun essentiel est que la lecture de ces deux sites suppose une approche globale d'un corpus. Cette approche est facilitée par les options de navigation, elle autorise des focales plus ou moins larges, mais l'objet publié reste un corpus et l'usage de Transcript est dicté par cette particularité : tout n'est pas transcrit, tout n'est pas à transcrire, et l'éditorialisation proposée n'interprète pas tant le texte qu'elle l'enrichit et élargit le point de vue à son sujet.

Le projet Marcianus quant à lui propose une approche très différente de la transcription et de ses objectifs. Nous avons affaire ici non pas à un corpus complexe mais à un seul objet, un manuscrit d'une grosse centaine de folios, datant du XIVe s., et contenant plusieurs collections disparates de textes littéraires en grec classique, accompagnés ou non de scholies grammaticales, et probablement destiné à l'enseignement du grec

14 Voir par exemple ici : https://eman-archives.org/Mythologia/transcript/browse?fileid=236
15 https://eman-archives.org/Epicurei/
16 https://eman-archives.org/Mythologia/emanindexpage
17 https://eman-archives.org/Mythologia/transcript/glossaire
18 https://eman-archives.org/Epicurei/transcript/glossaire

à des étudiants byzantins. Les objectifs de sa transcriptions sont spécifiques et liés aux particularités de l'objet :

- le document est illisible en l'état pour un non spécialiste du fait de ses particularités paléographiques ;

- son contenu est d'un intérêt scientifique considérable et susceptible de toucher un public beaucoup plus large que les quelques spécialistes capables de le consulter ;

- c'est un document à l'histoire complexe, plusieurs fois restauré, contenant des compilations imbriquées[19], ce qui induit un format et une organisation difficiles à rendre explicites.

Le projet Marcianus a été l'occasion d'une réflexion approfondie sur la nature exacte de la transcription possible. Une transcription diplomatique a été très rapidement évacuée : elle n'est pas indispensable pour un document pour lequel une image est fournie, et les accidents du document ne sont pas tous significatifs ni inédits (les abréviations sont classiques, la couche palimpseste est constituée d'un texte dont il existe des centaines de version, les ajouts et ratures ne manifestent pas l'originalité d'un auteur mais le plus souvent la maladresse d'un copiste). De plus, la nature composite du document rend certains phénomènes codicologiques d'une part extrêmement complexes à encoder et d'autre part susceptibles de gêner la compréhension des contenus réunis et de leur organisation. La transcription a donc pour objectif principal non pas de décrire le texte dans son état original, mais d'en proposer une version lisible, susceptible d'être acccompagnée d'enrichissements et surtout de notes critiques permettant de le contextualiser et de questionner sa lettre quand il existe d'autres versions des textes contenus par le manuscrit.

Le module Transcript est dans ce cas adapté à un travail éditorial proprement exégétique, impliquant donc des décisions quant à ce qui convient ou pas d'être mis en valeur ou explicité au cours de l'encodage ; il construit une version du texte, produite sous la responsabilité de chercheurs, qui n'est pas une description diplomatique mais bien une édition à caractère critique du document.

Ce dispositif induit un nouveau rapport texte/image, aussi bien sur le plan de ses soubassements technologiques que du nouvel acte d'édition et de lecture qu'il induit. L'objet donné à lire à l'usager du site est un objet proprement nouveau, inédit, qui n'est ni seulement une source numérisée ni seulement une édition, mais un compromis entre ces deux perspectives sur le texte, intégré au dispositif de la bibliothèque numérique. Ce dispositif a des conséquences sur l'acte de lecture. En effet, l'image et la transcription se commentent réciproquement et placent l'usager dans une position surplombante. Faisant dialoguer les possibilités offertes par l'image d'archive éditorialisée et les usages liés à la lecture d'édition critique, l'outil ouvre un tiers lieu[20] qui démultiplie la liberté critique de l'usager.

L'outil Transcript est encore actuellement en cours de développement, mais au stade où il est parvenu, nous avons déjà à notre disposition un outil qui répond à la plupart des besoins des projets qui en font usage. Les choix techniques des dernières années ont été faits dans le but d'améliorer son intégration dans l'environnement Omeka de départ et sa souplesse d'utilisation :

- du point de vue de l'intégration, l'accès aux transcriptions se fait par la bibliothèque numérique, aussi bien pour le lecteur que pour l'utilisateur ; les transcriptions sont accompagnées des métadonnées Dublin Core de chaque contenus, et bénéficient ainsi de l'ensemble de l'écosystème de la plateforme (plugin Zotero, liens entre contenus, métadonnées enrichies etc) ;

- le module utilisable soit par import direct de fichiers xml encodés, soit par une interface wyswyg plus élaborée que l'interface de départ et accompagnée d'un valideur TEI intégré ;

19 Voir l'arborescence de ce site : https://eman-archives.org/Marcianus/arbre-collections
20 Céline Bohnert et al. 2022.

- les transcriptions peuvent être faites isolément, feuillet par feuillet, ou regroupées grâce à une balise <ptr> ce qui rend possible un déroulement continu de l'image et de la transcription pour la totalité d'un contenu.

Transcript devient donc progressivement un outil global, capable de gérer toutes les étapes de l'édition numérique depuis l'import d'image jusqu'à la publication ; ceci résout ainsi le conflit habituel lié à la dissociation des étapes d'encodage et de publication web dans les projets d'édition numérique.

Le format choisi pour la transcription est celui de la TEI, et il convient de s'arrêter brièvement sur ce choix engageant. La TEI est un format d'encodage souple et de ce fait permet de couvrir un très large éventail de besoins et de spécificités de corpus ; ce format, de plus, est un format interopérable et robuste, et le choix d'un nombre considérable de projets d'éditions numériques au niveau international. Mais l'encodage en TEI a ouvert surtout pour nous la possibilité d'utiliser pour les visualisations html l'outil TEI Publisher, puissant et soutenu par une communauté internationale représentative. Chaque site faisant usage de Transcript est ainsi synchronisé à l'instance TEI Publisher d'EMAN, et les visualisations produites ainsi sont incrustées dans la page de la transcription de façon efficace et invisible pour l'usager.

Pour que l'utilisation du format TEI reste gérable et pour ne pas retomber dans les difficultés liées au premier Transcript, EMAN met en place un usage strictement encadré, avec l'établissement d'un schéma général commun pour la plateforme EMAN, issu du travail de réflexion collective des usagers désireux d'utiliser Transcript, réunis dans un groupe de travail se réunissant régulièrement. Les balises TEI utilisables sont en nombre limité, de même que les comportements possibles de ces balises au niveau de la visualisation.

Nous avons présenté dans ces quelques pages l'état actuel du module Transcript pour Omeka Classic, et montré comment Transcript cherche à établir un compromis satisfaisant entre transcription et édition pour répondre au cahier des charges d'un véritable outil d'édition numérique. Transcript a des atouts solides : il est intuitif, tout-en-un, co-développé et maintenu par une communauté d'usagers apte à en assurer la viabilité. Mais il a encore des faiblesses : commun à une multitude de projets différents, son usage est de ce fait contraignant et oblige les projets à adapter leurs besoins à ce que le schéma d'encodage propose, avec des possibilités de personnalisation limitées ; sa viabilité dépend de sa capacité à demeurer léger pour être maintenable dans les limites humaines et économiques de la communauté EMAN.

BIBLIOGRAPHIE

Bohnert, Céline. « L'édition numérique des *Mythologiae libri decem* de Natale Conti sur la plate-forme EMAN : un aperçu ». *Anabases - Traditions et réceptions de l'Antiquité* 34 (2021): 263. https://hal.science/hal-03463764.

Bohnert, Céline, Charlotte Dessaint, Marie Dupont, Jean-Sébastien Macke, Anne Réach-Ngô, et Richard Walter. « La plate-forme collaborative EMAN, un Fab Lab pour les Humanités ? » *Le Verger* Bouquet XXIII (30 avril 2022). https://shs.hal.science/halshs-03815700.

Crane, Gregory, Alison Babeu, David Bamman, Thomas Breuel, Lisa Cerrato, Daniel Deckers, Anke Lüdeling, et al. « Classics in the Million Book Library ». *Digital Humanities Quarterly* 003, n^{o} 1 (février 2009).

Cucciniello, Maria Laura. « Manifeste Eman ». https://eman-archives.org/EMAN, 25 janvier 2023. https://eman-archives.org/EMAN/manifeste_eman.

Giovacchini, Julie, et Stéphane Marchand. *Epicurei : index épistémologique et transcriptions enrichies de textes épicuriens*, 2022. https://hal.science/hal-03554873.

Leblanc, Elina. « Review of 'Omeka Classic. Un environnement de recherche pour les éditions scientifiques numériques'. » *RIDE* 11 (2020). doi: 10.18716/ride.a.11.3. Accessed: 14.06.2023.

Pawlicka-Deger, Urszula. « Infrastructuring digital humanities: On relational infrastructure and global reconfiguration of the field ». *Digital Scholarship in the Humanities*, septembre 2021. https://doi.org/10.1093/llc/fqab086.

Schmidt, Desmond. « Towards an Interoperable Digital Scholarly Edition ». *Journal of the Text Encoding Initiative*, n° Issue 7 (12 novembre 2014). https://doi.org/10.4000/jtei.979.

Walter, Richard. « EMAN & transcription : le module Transcript », 2022. https://hal.science/hal-03850288.