UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
Instituto Universitario de Análisis
y Aplicaciones Textuales

Deuxième atelier de travail international

Pratiques culturelles de l'édition numérique - Prácticas culturales de la edición digital

« Les Pratiques de transcription et d'annotation »

« Comment transcrire aujourd'hui sur des plateformes d'édition numérique ? »

1er au 5 avril 2019

Argumentaire

Dans le cadre du projet « Artlas » soutenu par le labex TransferS (ENS/PSL), nous organisons un atelier de travail pour faire dialoguer les projets de l'IIATEXT et ceux de la plateforme numérique EMAN.

Notre atelier vise à étudier les différentes pratiques de transcription et d'annotation de projets d'édition numérique de contenus patrimoniaux. L'édition numérique varie en fonction des contextes culturels, techniques et institutionnels où elle est produite. Il s'agit d'envisager les différentes contraintes et approches de l'objet à éditer numériquement, puis de voir comment ces différents contextes agissent sur les opérations d'édition numérique, en particulier sur les tâches spécifiques et centrales de la transcription et de l'annotation, thématique de cette année.

La reconnaissance des contenus porte de plus en plus sur des objets complexes, alliant images et textes dans une mise en page élaborée ; des logiciels OCR (Optical Entities Recognition) ou pour l'écriture manuscrite (comme Transkribus) commencent à donner des résultats impressionnants. Mais ce processus de reconnaissance puis de correction doit s'adapter à des contextes de production et des zones culturelles divers. Qui plus est, ces logiciels spécifiques, la plupart du temps professionnels, n'ont pas encore de dispositifs permettant de les adapter à des plateformes génériques, de grande masse et pour un public non spécialiste. Enfin, la transcription n'est utile que si elle est complétée par l'annotation qui donnera les clés de compréhension du texte transcrit.

De nombreux projets éditoriaux veulent maintenant aller jusqu'à donner la possibilité de consulter la transcription du texte des documents publiés sous forme d'images, notamment ceux qui sont les plus difficiles à déchiffrer. Mais les perspectives changent ; les problématiques habituelles de la transcription – proposer l'exhaustivité de l'édition diplomatique ? Rendre compte de tous les phénomènes vus (aperçus ?) sur le document – ont une autre importance avec Omeka ou avec les différents modèles de plate-forme d'édition numérique. Omeka donne accès à l'image du document et l'utilisateur a la possibilité de visualiser la transcription à côté de cette image. Ce n'est pas une obligation du logiciel mais, quand c'est le cas, l'objectif principal de la transcription doit changer : on

n'est plus dans la nécessité de faire du *reprint* mais plutôt de donner accès à une masse de textes pour pouvoir chercher et exploiter des hypothèses de travail.

L'automatisation de la transcription et de l'encodage pourraient faire gagner énormément de temps à de nombreux projets de la communauté scientifique utilisant Omeka ou d'autres plateformes génériques. Il est bien entendu que cette automatisation ne remplacera jamais la lecture humaine ; mais elle doit permettre d'accélérer le travail. Des initiatives et des outils existent pour faciliter une transcription et une publication des textes transcrits et annotés.

Le choix des fonds et des documents à transcrire et à annoter, des normes et des descripteurs varie en fonction des projets, des objectifs, mais aussi des expériences acquises et du contexte institutionnel et culturel au sein duquel intervient l'éditeur numérique. Les théories de l'édition numérique, les approches et les finalités peuvent être considérablement différentes (pédagogiques, patrimoniales, touristiques, scientifiques, etc.), et elles sont souvent imbriquées.

Il importe donc de définir des usages propres de certaines procédures numériques tel que l'OCR, la reconnaissance et l'annotation automatique d'images et de structures dans celles-ci, la transcription des résultats dans un encodage élaboré (essentiellement TEI, encodage de textes), et pour aller plus loin, la recherche de motifs partagés dans des bases documentaires d'images. Les expérimentations et l'outillage qui en résultera permettront de mieux équiper les projets des différentes structures, en particulier pour le traitement en grande masse de documents venant d'horizons divers.

La mise à disposition, au sein de la communauté Omeka, de cet outillage permettant une automatisation de certaines procédures aura beaucoup d'impact. Les utilisateurs d'Omeka disposeront alors d'une suite d'édition électronique complète, de l'acquisition à l'export de la transcription. Notre atelier veut précisément donner un espace de réflexions, d'analyses et d'expérimentation pour développer des méthodologies et des modules permettant de connecter ces nouveaux outils à Omeka ou à d'autres plateformes utilisant des standards.

L'atelier est résolument interdisciplinaire (de nombreux types de corpus sont pris en compte avec des approches scientifiques différentes) et international (les problématiques de transcription sont les mêmes malgré la différence linguistique). Nous allons alors faire dialoguer et expérimenter autour de la plateforme EMAN différences expériences venant de l'IATEXT, de l'équipe ELAN (laboratoire Litt&Art de Grenoble), du projet FFL (Fiches de Lecture de Michel Foucault, BNF-ENS Lyon-ENS Paris). L'atelier sera composé de tables rondes, de travaux pratiques sur document et de retours d'expérience sur les différentes étapes et protocoles de transcription et d'annotation des données d'un projet d'édition numérique : de la conception du corpus à la mise en place d'une chaîne de traitement des données et des métadonnées, les problématiques théoriques et techniques sont nombreuses.

De là doivent naître discussions et controverses sur les pratiques numériques actuelles, notamment sur l'idée (relativement courante) qu'il existerait une pratique *lambda* de la transcription et de l'annotation dans l'édition numérique s'adaptant à la diversité des usages. Or selon nous, il ne peut y avoir un modèle générique de transcription et d'annotation. À travers cet atelier, nous voulons au contraire montrer et analyser la diversité des pratiques de numérisation et d'édition aujourd'hui, en la reliant aux spécificités institutionnelles et culturelles au sein desquelles les projets s'inscrivent. Les modèles

éditoriaux comme les outils d'exploitation sont tributaires de ces spécificités, de même que les contextes scientifiques et culturels conditionnent la définition des objectifs de ces projets. Mais au-delà d'une telle diversité, il existe également des invariants, notamment dans le choix des métadonnées et du modèle d'édition critique, qu'il conviendra de déterminer et d'analyser.

L'atelier vise donc quatre objectifs principaux. Le premier est de faire dialoguer et de créer des synergies entre les projets des différents partenaires. Le second, à destination d'une communauté plus large, est de proposer une grille d'analyse et de comparaison pour les descripteurs et les modes opératoires de transcription et d'édition d'un contenu scientifique et/ou patrimonial, dans un contexte particulier. Enfin, pour chacun des partenaires, il permettra de mieux modéliser des schémas éditoriaux pour les projets développés au sein de chaque institution. Les confrontations de ces différentes expériences dans des contextes linguistiques et éditoriaux différents ne manqueront pas de lancer une dynamique pour l'organisation de futures discussions.

Organisation

L'atelier est organisé par l'IATEXT de l'Université Las Palmas (Veronica Trujillo) et l'ITEM, laboratoire CNRS-ENS (Richard Walter).

Les matinées seront consacrées à des réunions internes EMAN permettant de faire le point sur les pratiques d'édition numérique, spécialement de transcription et d'annotation dans les différents projets, pour rédiger ensuite des protocoles permettant des bonnes pratiques pour la plateforme EMAN mais plus généralement pour l'édition numérique de corpus élaborés. Un des livrables de cet atelier sera la publication d'un guide résumant les différentes méthodologies possibles pour transcrire et annoter dans différents contextes.

Pour les après-midi, les séances seront consacrées à des présentations théoriques et méthodologiques croisées, des séances de formation et de retours d'expériences, ainsi des tables rondes sur les problématiques de la transcription et des outils pouvant aider celle-ci (reconnaissance automatique, aide de référentiels et de listes d'autorité). Des séances spécifiques donneront une formation de base sur les pratiques de transcription et d'annotation, sur les usages de référentiels et de listes d'autorité et donc sur le web sémantique. Chaque présentation sera suivie d'un temps d'échanges. Les expérimentations des outils se feront sur des données test propres à chaque participant. Nous pourrons ainsi mener une étude comparative des différentes expériences de transcription sur le même document d'après-midi en après-midi.

Les langues de travail seront le français, l'anglais et l'espagnol.

Programme des matinées (9h-13h)

Mardi 2 avril

Retours d'expériences & état des lieux

Présentation des différents projets présents, utilisant ou non la transcription.

> Organisation & objectifs du workshop : Richard Walter (ITEM, CNRS-ENS) & Veronica C. Trujillo-González Trujillo (IATEXT, Universidad de Las Palmas de Gran Canaria)
>
> Présentation des projets présents

Mercredi 3 avril

Les limites d'EMAN et des outils génériques

Les outils génériques veulent prendre en charge la totalité d'un processus : constituer les données, les exploiter, les diffuser et les valoriser. Quels outils à disposition pour traiter ces différentes étapes ? EMAN peut-il tout faire ? EMAN, au départ bibliothèque numérique, est-il destiné à l'exploitation scientifique informatisée ? *In fine* quelles sont les limites d'utiliser un outil générique comme Omeka ?

> Les problèmes liés à la circulation des textes dans un document – retours d'expérience : Anne Reach-Ngo (Université de Haute Alsace & Institut universitaire de France, Mulhouse) & Céline Bohnert (CRIMEL, Université de Reims, Reims)
>
> Comment traiter le besoin d'encodage des symboles mathématiques – retours d'expérience : Charlotte Dessaint (Bibliothèque des lettres de l'École normale supérieure, Paris) & Emmylou Haffner (SPHERE, Paris-Sorbonne, Paris)
>
> Comment traiter le multilingue : Veronica C. Trujillo-González Trujillo (IATEXT, Universidad de Las Palmas de Gran Canaria)
>
> De l'outil générique à l'édition critique : Richard Walter (ITEM, CNRS-ENS)
>
> Les usages pédagogiques possibles : Richard Walter (ITEM, CNRS-ENS)
>
> Faire des outils *adhoc* : avantages et inconvénients – Retours d'expérience d'ELAN : Anne Garcia-Fernandez (Litt&Art, CNRS-Université Alpes, Grenoble) & Elisabeth Greslou (Litt&Art, CNRS-Université Alpes, Grenoble)

Jeudi 4 avril

Exploitation des données

Ce module permettra de réfléchir à l'exploitation scientifique des bases de données construites sur EMAN dans une perspective éditoriale. Une bibliothèque numérique donne un accès et une interrogation très réduits de l'ensemble des données accumulées dans la base de données. Ce qui pose la question des interactions entre EMAN et d'autres outils (comme le prototype du projet FFL ou la base de données du projet des Thresors de la Renaissance). Le débat est ainsi entre publication et exploitation : quelle distribution des outils selon les tâches et nos besoins ? Il y a de toute façon la nécessité de développer des passerelles entre les outils pour permettre une meilleure circulation des données. Cela servira aussi à la question : où arrêter EMAN ou autre outil générique pour la transcription? Quelles possibilités de se diriger vers d'autres outils ou de travailler en complémentarité ? Un outil de départ, un outil d'arrivée, un outil d'accompagnement de la recherche au jour le jour…

> Le prototype FFL : Vincent Ventresque (École normale supérieure, Lyon)
>
> Exploitation désirées sur les données d'une bibliothèque numérique : Anne Reach-Ngo (Université de Haute Alsace & Institut universitaire de France, Mulhouse)
>
> Comment faire « une cartographie des sentiments » : Jean-Sébastien Macke (ITEM, CNRS-ENS, Paris)
>
> Faire une transcription et éditer/valoriser : des outils différents ? Retours d'expérience d'ELAN : Anne Garcia-Fernandez (Litt&Art, CNRS-Université Alpes, Grenoble) & Elisabeth Greslou (Litt&Art, CNRS-Université Alpes, Grenoble)

Vendredi 5 avril

Valorisation

Comment assurer le passage de l'édition / publication (simple mise à disposition) à la valorisation / exploitation en environnement numérique. Là aussi il faut assurer la circulation entre des démarches de mises à disposition et d'exploitations scientifiques. Deux axes guideront les travaux de cet atelier : a/ Comment assurer la circulation des métadonnées en particulier avec le moissonnage OAI-PMH : être moissonné mais aussi moissonner et alors comment intégrer les métadonnées extérieures à nos propres métadonnées. b/ Comment utiliser des nouveaux dispositifs peu en usage dans la communauté scientifique, comme les expositions virtuelles, et ce que cela implique en terme de patrimonialisation.

> Présentation du fonctionnement de l'OAI-PMH : Richard Walter (ITEM, CNRS-ENS, Paris)

Retour d’expérience d’expositions virtuelles : Charlotte Dessaint (Bibliothèque des lettres de l’École normale supérieure, Paris) & Anne Reach-Ngo (Université de Haute Alsace & Institut universitaire de France, Mulhouse)

Présentation de la méthodologie et des actions de valorisation d’ELAN : Anne Garcia-Fernandez (Litt&Art, CNRS-Université Alpes, Grenoble) & Elisabeth Greslou (Litt&Art, CNRS-Université Alpes, Grenoble)

Les perspectives – Retour d’expérience sur les institutions culturelles et la patrimonialisation des données : Emmanuelle Bousquet (Université de Nantes)

Synthèse de l’enquête et des travaux du groupe Projet scientifique EMAN : Marie Dupond (UDPN & Association Guizot, Paris)

Tour de table final des projets présents

Programme des après-midis

Lundi 2 avril 17h-19h

Introduction : présentation de l'IATEXT et d'EMAN

Verónica C. Trujillo-González (IATEXT, Universidad de Las Palmas de Gran Canaria) et Richard Walter (ITEM, CNRS-ENS)

Usages du web sémantique

Richard Walter (ITEM, CNRS-ENS, Paris), Anne Garcia-Fernandez (Litt&Art, CNRS-Université Alpes, Grenoble)

Ce module proposera une présentation du web sémantique et dans le cadre du « Linked Open Data » de voir quels référentiels existent ? Quels dés/avantages ? Comment les utiliser d'un point de vue technique (comment les mettre en œuvre) et scientifique ?

Mardi 2 avril 15h-18h

Référentiels & listes autorité

Présentation des référentiels dans le monde de l'édition et dans le monde des bibliothèques, ainsi que ceux spécifiques à la géolocalisation. Le module devra déterminer quels liens avoir avec les listes d'autorité et les référentiels pour annoter, transcrire et exploiter les données. Comment les utiliser ?

Un retour sur l'usage de Data.bnf (https://data.bnf.fr) sera proposé en guise d'exemple.

Elisabeth Greslou (Litt&Art, CNRS-Université Alpes, Grenoble), Charlotte Dessaint (Bibliothèque des lettres de l'École normale supérieure, Paris), Vincent Ventresque (École normale supérieure, Lyon)

Mercredi 3 avril, 16h-19h

Principes de transcription

Ce module de formation concernera l'établissement des principes de transcription et donnera des pistes de réponse aux questions de base : Pourquoi transcrire, comment transcrire ? Quels modes de transcription (diplomatique, semi-diplomatique) ? Une ou plusieurs transcriptions ? Utilisation ou pas de logiciels de reconnaissance de caractère ?

Marie Dupond (UDPN & Association Guizot, Paris), Anne Reach-Ngo (Université de Haute Alsace & Institut universitaire de France, Mulhouse)

La plateforme TACT

Comment transcription avec cette plate-forme (https://tact.demarre-shs.fr) : exercice pratique sur des textes de Benoîte Groult.

Exemple d'atelier « Transcripthon » : https://elan-transcript.sciencesconf.org/

Anne Garcia-Fernandez (Litt&Art, CNRS-Université Alpes, Grenoble) & Elisabeth Greslou (Litt&Art, CNRS-Université Alpes, Grenoble)

Retour d'expérience sur Transkribus

Retours d'expérience sur la transcription automatique de manuscrits avec l'outil Transkribus (https://transkribus.eu).

Marie-Laure Massot (Caphes, CNRS l'École normale supérieure, Paris)

Jeudi 4 avril 16h-19h

L'outil Transcript

L'étape de la transcription avec le plugin Transcript développé par EMAN pour Omeka (https://eman.hypotheses.org/1067). Exercice sur un bac à sable avec document amené par les utilisateurs

Richard Walter (ITEM, CNRS-ENS, Paris)

Initiation à Scan Tailor & Abby FineReader

Atelier d'initiation aux programmes d'amélioration de documents Scan Tailor et de reconnaissance de documents numérisés Abby FineReader.

Zaida Bartolomé-Díaz (Universidad de Las Palmas de Gran Canaria)

Table ronde finale & perspectives sur les usages de la transcription