Consortium OLIO
Outils Libres Interopérables et Ouverts pour la recherche en Humanités

3.9.21

Document de travail

Contexte

La pratique du numérique dans le domaine des Humanités s'est, comme dans toutes les disciplines SHS, largement développée ces dernières décennies. Grâce à des efforts soutenus institutionnellement (InSHS, TGIR Huma-Num, Collex Persée entre autres), la quantité de données numériques, structurées et utilisant des standards, explose. Le consortium Cahier, le consortium international TEI, les associations en Humanités numériques ont permis de disposer de guides de bonnes pratiques, de retours d'expériences, de tutoriels, etc. pour permettre la création de données de qualité, qui suivent les principes FAIR et ceux de la Science ouverte.

Nous partons du constat que la communauté autour des corpus d'auteur a beaucoup travaillé sur les *données* et qu'un travail de même nature sur les outils et les pratiques est maintenant nécessaire. Des outils pour l'édition, la description ou l'exploitation de corpus sont utilisés de longue date et des formations spécifiques se sont développés à destination de divers publics. Toutefois, il existe peu de passerelles entre ces outils et le versant "exploitation scientifique outillée" des données publiées mérite d'être développé, y compris dans ses présupposés et ses implications scientifiques.

En parallèle, la production de corpus d'auteur s'élargit avec l'apparition de nombreux acteurs dans les mondes académiques et culturels, comme les nombreux projets de numérisation d'archives de bibliothèques ou les appels d'offres Collex. Mais cette masse inédite de contenus, qui répondent à des enjeux nouveaux de valorisation, de collaboration interprofessionnelle et plus largement de redéfinition de l'économie de la culture à l'heure du numérique, demande un outillage sur lequel l'expérience acquise au sein du consortium Cahier serait une vraie "plus value".

Un retour sur les usages et en amont sur les moyens, technologiques et méthodologiques, constitue uune manière originale de penser les données suivant une perspective interprofessionnelle et interdisciplinaire. Comme l'ont montré les travaux relevant de la *distant reading* (Moretti) /*référence : Franco Moretti, *Distant Reading*, Verso, Londres – New York, 2013 (https://www.sudoc.fr/177536683).*/, de la *machinic reading* (K. Heyles) ou des enjeux de la datavisualisation, il convient de considérer les données ou *data* plutôt que des « établis » (Latour) ou « *capta* » (Drucker).

Forte de son expérience d'une décennie de traitement de corpus, *la communauté constituée par* le consortium Cahier est désormais à même de proposer des scénarios de recherche, des parcours documentés avec des langages outils désormais bien connus, tout en offrant un espace d'innovation, du fait même de cette expérience diversifiée.

Le projet proposé par Olio en prolongation de Cahier réside dans ce parti-pris : engager un questionnement critique collectif sur la production, la manipulation et l'analyse des données à la lumière des démarches individuelles menées jusqu'à présent sur des corpus singuliers. Dans quelle mesure les écarts que l'on observe dans les pratiques traduisent-ils des démarches scientifiques différentes et dans quelle mesure les choix techniques ont-ils contribué à façonner l'objet scientifique étudié ? Quels sont les impacts méthodologiques distincts entre l'exploitation de données versées dans Omeka et de données structurées en XML/TEI ? Les développements ad-hoc s'opposent-ils aux chaînes de productions génériques ? Quelles contraintes institutionnelles pèsent sur le choix des uns ou des autres ? Quelles implications entre la transcription avec outils OCR/HTR et la transcription manuelle ? Faut-il fixer dès le début du projet la nature du résultat : une édition critique spécifique, une bibliothèque numérique, un corpus sémantique, etc. ? Le travail collectif sur ces étapes préalables à la

réalisation des projets scientifiques nous semble constituer un terrain de recherche à proprement parler qui ne peut pas se réduire pas à des préliminaires techniques.
Car la question des outils, au cœur du projet Olio, n'est pas seulement celle de la technicité de la recherche ou de l'appareillage des projets : elle soulève des enjeux épistémologiques fondamentaux. En humanités numériques, les outils sont des dispositifs intellectuels autant que des instruments techniques – ils sont d'ailleurs eux-mêmes les produits d'une recherche et d'une expérimentation. On peut ainsi envisager l'outil à la fois comme le résultat et comme l'origine potentielle de questions de recherche inédites, ou de nouvelles méthodologies et approches.

Les premiers résultats du groupe de travail Réutilisabilité, au sein duquel est née cette proposition pour l'avenir de Cahier, montrent bien l'importance de réfléchir à la mise en relation des corpus numériques existants au service de nouvelles problématiques scientifiques, ce qui demande l'appropriation, voire la création, de nouveaux outils. La proposition d'Olio est d'engager un travail collectif, auquel participeront chercheurs, ingénieurs, éditeurs numériques, documentalistes, en plaçant au cœur de la problématique des outils celle des usages – usages possibles, induits, contraints ou détournés, dans une tension entre normes et bonnes pratiques d'une part, inventivité et agilité de la recherche de l'autre.

La variété, tout comme les limites structurelles des outils, qui constituent des verrous autant que des occasions de retours réflexifs sur nos attentes et nos présupposés, rendent cette réflexion nécessaire. Certains outils sont propriétaires et payants.–D'autres sont libres et bénéficient d'une large communauté de développeurs. Il existe aussi un nombre impressionnant de "petits développements", d'utilisation confidentielle d'outils que l'on découvre au détour de colloques, journées d'étude ou autres. Assister à une conférence telle que DH peut donner le tournis tant la variété est importante. Bien sûr, il n'existe pas d'outil universel qui ferait tout, possédant toutes les options et fonctionnalités : chaque projet et chaque équipe a ses propres besoins, ses propres compétences. Pourtant, l'on aimerait parfois utiliser plusieurs outils en fonction des besoins et des objectifs. Là où l'outil était très performant pour acquérir des données, il n'est plus adapté pour leur diffusion ; là où le logiciel a permis la constitution d'un corpus accompagné de métadonnées riches, il n'est plus capable d'accompagner les analyses dont le projet de recherche a besoin, etc.
Passer d'un outil à l'autre n'est pas toujours aisé. Si nos données sont stockées dans une base de données, et qu'on veuille les intégrer à un entrepôt, il faudra bien souvent passer par plusieurs étapes, parfois très techniques. Les logiciels ne disposent pas toujours d'API compatibles et l'usage de cette technique n'est pas trivial. Un corpus en XML-TEI dont on souhaiterait faire une annotation morpho-syntaxique n'est pas immédiatement manipulable par les outils issus du TAL car les formats diffèrent. Les exemples d'interopérablité difficile ne manquent pas, malgré l'ouverture des corpus avec des formats standards et ouverts. Or, dans l'idéal, ne rêve-t-on pas de créer nos données avec un logiciel parce qu'il est efficace pour nos objectifs scientifiques, mais de pouvoir les transformer sur un autre, les diffuser sur un troisième, etc. ?

Corpus d'auteur

Le périmètre de notre projet est celui qui avait été défini par le consortium Cahier : « les corpus d'auteurs et plus généralement les corpus textuels constitués en référence à l'œuvre d'un auteur, d'une tradition éditoriale, d'une forme littéraire, d'un genre ». Ce périmètre permet de prendre en compte différentes disciplines et de travailler de façon transversale

(trans-séculaire, trans-disciplinaire) sur des sources de différentes natures, selon des objectifs variés : publication, édition brute ou éditorialisation, archivage, exploitation, valorisation.

Objectifs d'OLIO

L'objet principal du consortium que nous proposons est ainsi l'*outil pour la recherche* : qu'il soit logiciel, site web, plateforme, entrepôt, module/plugin, service..., qu'il soit outil d'analyse, de constitution, de visualisation, de publication, d'archivage... qu'il soit spécialisé ou générique, ouvert ou propriétaire (en privilégiant tout de même les logiciels libres), etc. L'objectif est, en partant des pratiques actuelles, d'accompagner des démarches de signalement et d'expérimentation des outils et des méthodologies, de réfléchir à des connexions possibles, de favoriser le développement de l'interopérabilité entre les outils et de contribuer à clarifier la nature des tournants qu'ils engagent.
Nous souhaitons intégrer autant que possible non seulement des usagers des outils, mais aussi des usagers des objets numériques qu'on fabrique grâce à ces outils. Nous imaginons cet idéal où, par exemple, un lot d'images entreposé sur Nakala, pourrait - *par le jeu des API et sans transfert des fichiers* - bénéficier d'une transcription au kilomètre par un outil HTR, passer *en un clic* par Grobid pour être structuré, puis *importé* directement sur un outil d'édition pour corriger et enrichir la transcription. Le parcours des transcriptions pourrait se poursuivre : intégration à EVT, TEI Publisher, Omeka ou autre ; dépôt sur l'entrepôt de données à côté des images initiales ; export dans des outils de TAL pour faire des analyses linguistiques, etc., etc.
L'objectif du consortium OLIO est de permettre de doter les projets d'outils adaptés aux objectifs de recherche, de travailler sur l'impact des outils sur la recherche produite et d'étudier les méthodes de recherche induites par ces outils. Cela implique la nécessité de connaître les outils (il ne s'agit pas seulement de former à leur utilisation, mais aussi d'expliciter les présupposés qui président à leur création et à leur développement), et, de là, de réfléchir collectivement aux possibilités d'utilisation et aux nouvelles formes de savoir qui peuvent en découler.

Car, à côté d'éditions (numériques ou pas) plus classiques, de plus en plus de projets créent ou utilisent des outils d'analyse, de visualisation et de fouille des données. Ces lieux d'exploration sont parfois déjà désignés sous les termes de « laboratoires numériques » ou « laboratoires de textes ». Ces objets encore mal définis, qui associent des corpus et de nouveaux outils, ouvrent de nouvelles voies/perspectives d'analyse et de recherche sur les corpus de textes.

Enfin, l'approche par l'outil et ses usages a l'avantage d'être structurellement interdisciplinaire. La plupart des outils sont employés très largement, par des communautés très diverses. L'analyse de leurs appropriations dans les différents champs disciplinaires nourrira ainsi une approche réflexive sur les Humanités (au-delà des seules Humanités numériques), en éclairant les points de rencontres, les lignes de forces et les divergences épistémologiques actuelles, sans préjuger de structurations a priori, dans la continuité des réflexions de Johanna Drucker notamment.

Un objectif complémentaire mais essentiel sera celui de la formation aux différents outils. Tout comme le consortium Cahier dont la réussite était basée sur des ateliers réguliers et fédérant une communauté, le travail du consortium serait animé par l'organisation régulière de retours d'expériences et de formations du niveau débutant au niveau expert. Des ateliers de

découverte et des ateliers de travail sur corpus prototypiques permettront de rassembler des spécialistes en corpus d'auteurs mais aussi les *nouveaux entrants*, en particulier les doctorants.

Il faut en effet donner une place aux *usagers* des corpus d'auteurs : lecteurs, public, chercheurs, enseignants qui utilisent ces éditions. Ainsi on fera appel à une communauté travaillant sur tel corpus d'auteur ou tels types de corpus pour avoir des retours sur les utilisations des objets numériques ainsi produits, et sur les besoins réels des éditeurs et des usagers.

Communauté OLIO

Le consortium veut fédérer celles et ceux qui ont besoin d'utiliser des outils pour les corpus d'auteur et celles et ceux qui les créent. Il ne s'agit pas de rassembler des structures mais des projets (et l'ensemble de leurs membres) ayant comme objet un corpus d'auteur ou un outil s'adaptant à celui-ci, personnes venant des institutions de la recherche en humanités (laboratoires, MSH, écoles doctorales) mais aussi du monde des bibliothèques ou de la culture, grands pourvoyeurs ou utilisateurs de corpus d'auteur (associations, bibliothèques, musées).

Le consortium peut aussi devenir un interlocuteur expert pour les autres opérateurs de numérisation et d'édition de corpus d'auteurs (Collex, campus d'excellence, EUR, etc.).

Loin de polariser les membres du consortium en opposant pourvoyeurs et consommateurs d'outils, il s'agit d'accompagner un dialogue structurellement nécessaire et fécond. L'engagement des développeurs dans une démarche « bottom-up », qui part du besoin exprimé, est naturel et reste pertinent ; mais, bien au-delà, le consortium se propose de devenir un lieu de réflexion sur le rôle et la place de l'outil dans les Humanités (au sens large). La manière dont l'instrumentation numérique de la recherche influence les Humanités doit faire l'objet d'une réflexion collective qui est encore loin d'être épuisée aujourd'hui, et dont le consortium OLIO souhaite s'emparer à travers le prisme des outils, aussi bien pour envisager leurs potentialités en terme de recherche que pour questionner leur apport.

Feuille de route

- Améliorer l'information sur les outils et leur offrir une visibilité à travers leur signalement ;
- Créer le cadre de réflexion et d'élaboration de chaînes de traitement à travers des retours d'expérience, des procédures spécifiques ou développées en commun, la rédaction de vademecums. Par exemple : comment favoriser l'interopérabilité de ses données en cours de projet ? Comment anticiper et élaborer une chaîne de traitement sans pour autant interdire un changement d'option, en conservant de la souplesse ?
- Signaler ou développer (les scripts, procédures et outils permettant le passage d'un format à l'autre, d'un environnement à l'autre ;
- Faciliter l'appropriation des outils par la communauté à travers la formation, des débats contradictoires basées sur de larges retours d'expérience et encourager les débats méthodologiques ;

- Nourrir la réflexion théorique et épistémologique : quel impact des outils sur la recherche ? quels principes, quels concepts, qu'ils soient explicites ou qu'ils relèvent de l'impensé, entrent en jeu dans la conception, le fonctionnement et l'utilisation des outils ?
- Regrouper et synthétiser l'actualité des outils pour les corpus d'auteur à travers un état de l'art annuel qui sera présenté lors de chaque AG annuelle et diffusé ensuite au sein de la communauté. Cet état de l'art annuel présentera les actualités autour des outils (nouvel outil, nouvelle version), les formations et les événements notables de l'année écoulée ; il pourra faire des focus sur des thématiques émergentes ou qui suscitent les controverses. Il n'aura pas vocation à être prescriptif mais il donnera toutes les ressources et tous les éléments de débats. Il sera conclu par un glossaire avec des définitions des notions faisant l'actualité.

Le consortium OLIO utilisera pour cela différents outils :

- Carnet de recherche avec liens vers différents réseaux sociaux : il comportera l'actualité des événements organisés par le consortium, les comptes-rendus des formations, des retours d'expériences, lieu de publication du consortium et de ses instances ;
- Agenda du consortium (avec module d'inscriptions et d'évaluation des formations organisées) et agenda sur l'actualité des outils pour les corpus d'auteur ;
- Espace collaboratif interne pour l'élaboration des signalements, recommandations à diffuser ;
- Publication de documents (vade-mécums, état de l'art annuel) sur supports numérique et papier.

Des outils

Une première liste d'outils ou de références :

- Pour (chaînes de) traitement des données : Metope, Textable, Dataiku, OpenRefine, TXM...
- Pour visualisation des données : Tableau, Gephi, ElasticSearch&Kibana...
- Pour travail sur les images : API IIIF, Mirador...
- Lemmatiseur et TAL : Pyrrha ; CLTK et NLTK (Python)...
- Fouille et analyse : Textable, iramuteq, TXM, hyperbase, voyant tools, Philologic ...
- Logiciels de reconnaissance de caractères (OCR, HTR) : EScriptorium, Tesseract (ocrmypdf), Kraken, Trankribus...
- Editeur ou outil XML/TEI : Oxygen, TEIPublisher...
- Plateformes de transcription : TACT, Transcrire, Transcript/EMAN...
- Bibliothèque numérique : Omeka Classic & S, Heurist...
- Description archivistique : atom…

Fonctionnement & statuts

Le consortium sera organisé en ateliers et groupes de travail (GT). Une assemblée générale annuelle du consortium fixera la thématique et l'agenda des ateliers et des groupes de travail (GT) qui présenteront leurs travaux à chaque AG.

Cette assemblée générale annuelle élira un comité de pilotage qui sera chargé de suivre l'avancement de ces travaux et une équipe de coordination tricéphale qui sera chargée de les mettre en œuvre.

Les statuts et le mode d'adhésion seront débattus lors de la première AG du consortium.

Gouvernance

L'équipe de coordination sera tricéphale : un/une coordinateur/trice principal(e) et deux adjoint(e)s qui devront être représentatifs des différents statuts et expériences rassemblés dans le consortium : producteur / utilisateur d'outil - chercheur, enseignant chercheur / ingénieur / bibliothécaire. Une variété disciplinaire, institutionnelle et géographique sera aussi nécessaire.

Structuration

La première Assemblée générale fixera le fonctionnement du consortium, les statuts, les modalités de participation à OLIO et d'élections de ses instances. L'Assemblée générale sera ensuite l'organe central du fonctionnement du consortium.

Première assemblée générale sur une journée complète :

- présentation du consortium et de ses objectifs,
- propositions d'organisations et de statuts du consortium (structure, mandats, adhésions, etc.),
- proposition à la volée de GT suivie de séance d'échanges en sous-groupe pour la définition de ceux-ci,
- création des premiers GT dont 1 personne coordinatrice et 1 responsable communication.

Fonctionnement régulier prévu

Au cours de l'année :

- les GT organisent régulièrement des séances ;
- l'équipe de coordination veille à ce que les GT soient actifs (relance, accompagnement, aide, si nécessaire), se tient au courant des avancées et s'assure que les GT ont des périmètres bien définis (pas de doublon) ;
- une équipe de communication (une personne responsable de la communication par GT et une au sein du trio de coordination) veille à ce que le carnet de recherche informe régulièrement la communauté des travaux du consortium ;
- préparation d'un atelier annuel de formation porté par un ou plusieurs GT, la thématique est fixée en AG;
- l'équipe de coordination accompagne la production de manuel, documentation ou tout autre support utile ; elle met en place le recueil d'informations pour l'état de l'art annuel.

Consortium OLIO
Outils Libres Interopérables et Ouverts pour la recherche en Humanités

3.9.21

Milieu d'année:

- atelier annuel :
- workshop de rédaction de l'état de l'art annuel (parties thématiques). Les deux événements peuvent être reliés.

AG de fin d'année

- retour des GT : difficultés/facilités, décision de poursuivre, arrêter, se transformer ;
- discussion sur la publication des livrables produits par les GT (manuels, vade-mecum, outils, etc.) : ils sont soumis à tous *avant* diffusion et leur diffusion est votée.
- retour sur l'atelier annuel et programmation du suivant ;
- désignation dans chaque groupe d'une personne responsable FAIR et d'une personne responsable de la communication ;
- discussion et finalisation de l'état de l'art annuel pour diffusion.

En début d'année X+1

- avec les responsables FAIR de chaque groupe, fairisation de tous les livrables du consortium : ceux de l'atelier (programme, supports de formation, comptes-rendus...), ceux éventuels des GT (documents, données test, exemples type, outils...), l'état de l'art annuel.