International Workshop on Computer Aided Processing of Intertextuality in Ancient Languages, Lyon, 2-4 juin 2014: bilan et perspectives.

Cet atelier international, coorganisé par les laboratoires français HiSoMA et LIRIS (Lyon) et le Centre allemand de Göttingen pour les Humanités Numériques (GCDH), constituait le colloque conclusif du financement de l'Agence française de la Recherche pour le projet Biblindex[1], dont l'objectif est la réalisation en ligne d'un index exhaustif des citations et allusions bibliques dans les textes de l'Antiquité et du Moyen Age. Cette sympathique réunion, la première de cette ampleur sur la thématique, a rassemblé des représentants d'une quinzaine de projets, européens et américains, dont le point commun était le traitement des phénomènes d'intertextualité dans des corpus en langues anciennes. Un public étudiant fourni complétait le groupe. Après une présentation générale des projets, quatre sessions, qui laissaient une large place à la discussion, ont permis aux chercheurs en sciences humaines et aux informaticiens présents de faire le point sur les concepts et techniques utilisés dans différentes étapes de la chaîne de traitement des citations. Ont été abordés les thèmes suivants :

- La complémentarité des approches statistiques et linguistiques dans le traitement automatique du langage appliqué aux *big data* ;
- Les spécificités linguistiques des approches ;
- La modélisation de la notion instable de citation et l'établissement de typologies ;
- Les choix d'encodage à mettre en œuvre pour sa caractérisation ;
- L'opportunité de la création d'un écosystème numérique dédié aux recherches d'intertextualité en langues anciennes.

On trouvera le détail du programme en ligne, les diaporamas des interventions ainsi que des hyperliens vers chacun des projets mentionnés, sur la page http://biblindex.hypotheses.org/1686. Le compte rendu ici proposé sera donc plutôt une tentative de synthèse des premiers résultats de cette rencontre.

Les types de projets

Une grande partie des projets visent à produire **une édition critique d'un texte ancien** à l'aide d'outils numériques (alignement multi-textes pour comparer les différentes leçons offertes par les citations ou les manuscrits, base de données de citations se référant au texte concerné). Deux approches de l'outil numérique au service des humanités se sont nettement distinguées :

1) une première catégorie de projets a pour objectif final la réalisation d'une édition imprimée ; les outils numériques ne sont alors conçus que comme un moyen de parvenir à une qualité supérieure d'édition critique ou à la reconstruction d'un texte perdu (*Vetus Latina Iohannes*[2], COMPaul[3], ECM[4]). Les outils numériques sont des

[1] http://www.biblindex.org.

[2] Edition synthétique de témoins de traductions latines de l'Evangile de Jean antérieures à la Vulgate : manuscrits bibliques et citations patristiques (ITSEE, Birmingham).

[3] Le projet COMPAUL étudie les commentaires grecs et latins les plus anciens des Epîtres de Paul comme sources pour le texte biblique lui-même (ITSEE, Birmingham).

[4] *Editio Critica Maior* du texte grec des Actes des Apôtres (INTF, Münster).

auxiliaires dans une démarche d'édition de texte qui reste celle des humanités classiques : un texte publié, avec son apparat critique, au format papier. Il faut toutefois signaler que les projets de ce type produisent aussi des données en ligne, puisqu'ils se proposent de laisser accessibles les outils numériques qu'ils auront développés (bases de données, catalogues de manuscrits numérisés et encodés, …).

2) une seconde catégorie vise à établir l'édition en ligne d'un texte donné, sans nécessairement de prolongement imprimé (LOFTS[5], Hyperdonat[6]). Le résultat escompté est donc un nouveau type d'édition des textes anciens : le texte unique et l'apparat sont remplacés par une présentation multitextuelle, avec annotation dynamique dans le corps du texte. Ces projets participent à l'alimentation et à la création de librairies numériques (comme Perseid[7] et SAWS[8]).

D'autres projets se concentrent sur **la création de bases de données interrogeables en ligne**, pour la recherche de citations, l'étude de l'intertextualité. La conception et les objectifs de ce genre de projets sont assez similaires, mais la méthodologie peut varier grandement suivant le type d'outils numériques mis en œuvre et le degré d'avancement des projets. Certains travaillent à partir de leur propre banque de citations (Biblindex, CASG[9], La Tradicion Literaria Griega[10], DGC[11]), d'autres s'appuient sur des librairies numériques (Tesserae[12] sur Perseus).

Les projets consacrés entièrement au **développement d'outils linguistiques en ligne** sont rares, mais beaucoup de projets produisent de tels outils parallèlement à leur objectif principal, outils susceptibles d'être réutilisés. Il s'agit essentiellement de dictionnaires, de lexiques bilingues (grec et langues orientales en majorité), de textes et de concordances lemmatisés (SHEBANQ[13], GREgORI[14]) ; de logiciels d'alignement de textes ou d'aide à la traduction en juxtalinéaire (GREgORI) ; de lemmatiseurs (Perseus, Biblindex, DataLift appliqué à l'arménien). Plusieurs projets cherchent à développer **des logiciels de détection**

[5] Le projet *Leipzig Open Fragmentary Texts Series* (Université de Leipzig) établit des éditions en ligne d'œuvres anciennes qui n'ont survécu qu'à travers des fragments ou des citations dans d'autres textes.

[6] Edition en ligne de commentaires de Donat (HiSoMA, Lyon).

[7] http://sites.tufts.edu/perseids/ : plateforme collaborative d'édition des textes sources de la littérature classique.

[8] Le projet *Sharing Ancient Wisdoms* (King's College, London) présente et analyse des citations grecques de textes sapientiaux et leurs traductions en arabe.

[9] Le projet *Corpus der arabischen und syrischen Gnomologien* (Université de Halle) rassemble des collections de sentences écrites en arabe et en syriaque.

[10] Repérage et classification des formes d'intertextualité dans un corpus de grammairiens grecs des 3e et 4e siècles.

[11] Le projet *Digital Greek Patristic Catena* (Université Aristote, Thessalonique) collecte les références bibliques faites dans l'édition de la *Patrologie Grecque* de J.-P. Migne.

[12] Ce projet (Université de Buffalo) étudie l'intertextualité dans la poésie latine.

[13] SHEBANQ (VU Amsterdam, Eep Talstra Centre for Bible and Computer) et les projets connexes comparent des textes bibliques en hébreu, syriaque, araméen.

[14] *Softwares, Linguistic Data and Corpus for Ancient GREek and ORIental Languages* (Université catholique de Louvain, Institut orientaliste).

automatisée des correspondances textuelles (eTRACES[15], SHEBANQ, Biblindex, QuotationFinder[16]).

On peut aussi classer parmi les outils linguistiques les projets de grammaires et d'aide à l'apprentissage des langues anciennes en ligne programmés par Perseus.

Enfin, quelques communications ont présenté **des outils numériques génériques**, susceptibles d'être utilisés par des projets en humanités numériques (DataLift[17], TAP[18]). Un outil plus spécifiquement dédié au travail sur l'intertextualité a été présenté : la plateforme TRACER. Développé pour des projets sur corpus limité (eTRACES, eTRAP), il a été pensé génériquement pour traiter les intertextualités en sept étapes : segmentation, préparation linguistique, découpage, sélection, mise en relation, évaluation de la pertinence, visualisation.

A quelques exceptions près, les projets n'en sont qu'à leur début – moins en terme de durée d'existence que de produits finis effectivement mis à la disposition des utilisateurs. On trouve essentiellement accessibles en ligne des textes numérisés, des bases de données – encore incomplètes – et des librairies numériques. Une grande partie des données produites pour les différents projets restent malheureusement cloisonnées et ne sont pas disponibles en dehors des équipes en charge. Outre les problèmes liés aux blocages conceptuels et à l'exploration des fausses pistes, deux causes principales, indissociables, expliquent ces difficultés à « produire » : le manque d'informaticiens travaillant de façon pérenne pour les projets ; le défaut de financement. A l'exception de Perseus, l'ensemble des projets fonctionnent sur des financements à court terme – deux à quatre ans en général – qui laissent peu de place à la planification sur le long terme dont ces projets ont besoin. L'absence de financement est due à la fois aux réticences des institutions devant ce champ relativement nouveau de la recherche et au manque de livrables qui pourrait démontrer auxdites institutions l'intérêt de ce type de projets : un cercle vicieux dont il semble difficile de sortir ! D'où la nécessité impérieuse de collaborations nouvelles entre les équipes : pour partager les expériences de chacun, et éviter les fausses pistes ; pour partager les outils générés (ontologies, système d'encodage, zones de stockage, données…) ; pour créer des dynamiques cohérentes.

Les outils numériques mis en œuvre

L'encodage des textes et le balisage des citations à l'intérieur des textes étaient au centre des journées. La discussion s'est articulée autour de l'usage ou non de normes communes pour l'encodage : la standardisation ralentit le travail et le conditionne parfois, mais l'absence de standardisation le ralentit encore plus et risque à terme de rendre tous les résultats caducs. En particulier, a été discutée l'opportunité d'utiliser la Text Encoding Initiative (TEI), qui semble être le meilleur dénominateur commun en matière d'encodage… ne serait-ce que

[15] Fédération de plusieurs projets de recherche de citations dans de grands corpus (coord. Université de Leipzig).

[16] Logiciel qui affine les résultats de recherches par mots-clefs dans de grandes bases de données textuelles, permettant de mieux distinguer citations et allusions.

[17] La plateforme DataLift se décompose en 5 étages : capture de jeux de données structurés mais hétérogènes ; convertisseurs/mappeurs qui appliquent des ontologies et produisent des jeux de données RDF ; stockage dans un triple store ; interconnexion de données ; exploitation des données élevées et interconnectées.

[18] Le Text Alignment Protocol (Dumbarton Oaks) est un outil génériquement pensé qui a pour finalité l'édition parallèle de textes multiples.

parce qu'elle est utilisée dans la librairie numérique de Perseids[19]. Il a été rappelé que la TEI n'est peut-être pas le meilleur système d'encodage d'un point de vue technique, mais qu'étant le plus partagé, le plus à jour, le plus collaboratif dans son fonctionnement, il paraît très adapté pour des projets qui visent à la synergie. Il est un moyen de discuter entre spécialistes pour arriver à un consensus, notamment en matière de terminologie. Par ailleurs, l'opportunité d'un vocabulaire commun pour encoder les citations et textes fragmentaires n'a pas fait l'unanimité au sein des participants. Il a été proposé de dépasser la TEI en adoptant un langage commun spécifique (TAP), mais il ne s'agit encore que d'un programme en développement. Ceci étant, il existe des outils permettant de convertir d'un langage d'encodage vers un autre (Datalift).

La plupart des projets font face aux mêmes questions théoriques, mais le passage à la pratique rend peu opérant la théorie commune. Un problème récurrent a été soulevé, celui du chevauchement des balises intégrées dans les documents XML (*embedded markup overlapping*). Certains éléments doivent être encodés au fil du texte dans le document lui-même, comme les éléments de structuration, mais les annotations devraient parfois être placées ailleurs pour ne pas rendre le document XML illisible, ce qui est parfois le cas des textes dans Perseus.

La question des **identifiants** a aussi été abordée. Deux outils ont été mentionnés : l'architecture CITE/ CTS (Homer Multitext project, SAW, LOFTS, TRACER, TAP), qui permet de créer des identifiants uniques pour les personnes et les objets (URN/URI), et ainsi d'aligner aisément des extraits de plusieurs textes par XML ou des graphes RDF. L'accent a été mis sur l'impératif de suivi des données mises en lignes (qui fait quoi) pour assurer leur qualité et leur possibilité d'évolution.

L'établissement d'une typologie commune permettant de classer les différents types de citations a été proposé. Plusieurs projets ont présenté leurs propres systèmes (Biblindex, La Tradicion Literaria Griega, e-pigramme[20], projet de narratologie de l'Université de Genève, ECM), mais il semblerait toutefois que ce type de classifications dépende nécessairement des corpus traités et de l'objectif de chaque projet et ne puisse donc être commun à tous. La TEI offre cependant déjà des possibilités de balisage souples et fines, incluant le typage des intertextualités. A ce stade, bien que tous les projets soient confrontés au problème de la classification des citations, peu d'entre eux ont élaboré une typologie. Trois ontologies, consultables en ligne, ont été évoquées durant l'atelier. La Prov-O ontology qui est utilisée par LOFTS et consultable en ligne ; l'ontologie développée pour le projet SAWS et également utilisée par le CASG (et consultable sur le site de SAWS) ; celle de CITO enfin, qu'aucun projet n'utilise à ce stade – à cause de problèmes spécifiques liés aux textes anciens qui ne précisent pas nécessairement le citant – mais sur laquelle il a été conseillé de s'appuyer.

Il a été suggéré de se mettre d'accord sur les catégories que tout le monde utilise, par exemple à partir des travaux des grammairiens anciens, puis de donner une définition de ces catégories en travaillant en TEI et en comparant les projets. Cette typologie provisoire serait proposée

[19] La question du choix TEI ou Epidoc n'en est pas une, puisqu'Epidoc n'est qu'une variété de TEI.
[20] Projet de numérisation et valorisation des inscriptions grecques du Musée du Louvre (HiSoMA, Lyon).

avec une liste non exclusive, à affiner en fonction des réactions. Le caractère diachronique de la question sera forcément à prendre en compte, car la définition des réutilisations de textes varie avec le temps. Ce que l'on nomme « citation » pour l'Antiquité s'appellerait bien souvent « plagiat » aujourd'hui !

Du point de vue plus théorique de la **modélisation**, il a été suggéré d'utiliser une visualisation en graphe plutôt qu'une simple arborescence. Sur cette question, il y a une tension entre simplicité pratique du modèle et recherche de l'exhaustivité. L'arborescence est la seule typologie immédiatement lisible pour tout le monde et un autre modèle risque de se révéler inefficace d'un point de vue pratique.

Une dernière question – et sans doute celle qui a le moins reçu de réponses pratiques ! – concerne **le stockage et le partage** des données produites par les différents projets. Les deux questions sont liées parce que le lieu de stockage des données conditionne en partie leur accessibilité. Comme les projets ne sont viables que sur les court et moyen termes, du fait de leurs modalités de financement, il convient de s'assurer que les résultats obtenus, même partiels, ne seront pas perdus. Faut-il créer une nouvelle bibliothèque en ligne ou utiliser Perseus ? L'idéal serait d'utiliser des connecteurs de données, dans des espaces comme Europeana (*hubs/repositories)*. L'essentiel n'est pas le lieu de stockage, mais la publication pour une exploitation commune des données.

La création d'un écosystème numérique pour les humanités numériques occupe actuellement plusieurs équipes, en particulier celles de Perseus et du GCDH. Datalift propose un espace de stockage des outils numériques créés pour les projets (ontologies, outils linguistiques…), mais pas pour les données elles-mêmes. L'importance d'avoir un espace de stockage est admise par tous, mais les dispositifs d'archivage numériques existants sont payants et à des prix exorbitants.

Le partage et l'accessibilité des données ont fait l'unanimité. En matière d'humanités numériques, c'est la mise en ligne en *open access* qui fait office de publication. C'est également une condition essentielle des développements de synergies. Malheureusement, les conditions propres à l'élaboration – et au financement – de chaque projet, et parfois la nature du corpus traité, peuvent conduire à l'impossibilité d'ouvrir les données produites par les différents projets (données sous licence, partenariat avec des organismes privés…).

Malgré toutes ces difficultés, le bilan de l'atelier a été très positif, de l'avis unanime des participants, car pour beaucoup il constituait une première occasion d'échanges. Espérons qu'il inaugurera des collaborations nouvelles, ne serait-ce dans un premier temps que par la diffusion d'informations sur les avancées des projets. A cet effet, le Google group « historical text re-use », déjà en place, constitue un premier outil pertinent. La préparation d'une *cheatsheet* commune consacrée à l'encodage des citations sur le wiki du site de la TEI est également en cours.

Auteurs du compte rendu : Clément Crosnier, Laurence Mellerin (HiSoMA, Institut des Sources Chrétiennes, Lyon)