Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien

Peter Anthony Stokes

EPHE
Sciences historiques et philologiques

Annuaire de l'École pratique des hautes études (EPHE), Section des sciences historiques et philologiques
Résumés des conférences et travaux

153 | 2022
2020-2021

Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien

Peter A. Stokes

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/ashp/5750
DOI : 10.4000/ashp.5750
ISSN : 1969-6310

Éditeur
Publications de l'École Pratique des Hautes Études

Édition imprimée
Date de publication : 1 septembre 2022
Pagination : 529-531
ISSN : 0766-0677

Référence électronique
Peter A. Stokes, « Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien », *Annuaire de l'École pratique des hautes études (EPHE), Section des sciences historiques et philologiques* [En ligne], 153 | 2022, mis en ligne le 13 juin 2022, consulté le 15 juin 2022. URL : http://journals.openedition.org/ashp/5750 ; DOI : https://doi.org/10.4000/ashp.5750

HUMANITÉS NUMÉRIQUES ET COMPUTATIONNELLES APPLIQUÉES À L'ÉTUDE DE L'ÉCRIT ANCIEN

Directeur d'études : M. Peter A. Stokes

Programme de l'année 2020-2021 : *Vers une paléographie transversale : méthodes numériques pour la description et l'analyse de l'écriture*.

Au cours du premier semestre, nous avons poursuivi notre travail sur la modélisation de l'écriture manuscrite. Nous avons consacré quelques séances pour revisiter le modèle Archétype et les questions qui se posent quant à son application à différentes écritures, y compris le cas particulier du polygraphisme[1]. Nous nous sommes ensuite attelé à introduire au sein du modèle, non seulement la représentation visuelle de l'écriture, mais également sa fonction. En nous appuyant sur la théorie sémiotique et notamment sur les travaux de Monella et de Klinkenberg et Polis, entre autres, nous avons pris en compte les différentes fonctions des signes écrits[2]. Klinkenberg et Polis distinguent deux types de fonctions. Il y a, d'une part, les fonctions autonomes : le signe y est indépendant de son contexte, du moins théoriquement. Il existe, d'autre part, des fonctions relationnelles : le signe est en relation avec d'autres signes qui l'entourent. Parmi les fonctions relationnelles, Klinkenberg et Polis distinguent les fonctions relationnelles syntagmatiques et les fonctions relationnelles paradigmatiques. Les fonctions relationnelles peuvent également être définies comme lexicales, morphologiques, syntaxiques et prosodiques. Par exemple, les signes ayant une fonction lexicale syntagmatique comprennent les déterminatifs des hiéroglyphes égyptiens ; ceux ayant une fonction morphologique paradigmatique comprennent les formes majuscules dans les systèmes d'écriture européens modernes ; les fonctions syntaxiques syntagmatiques sont fournies par l'espace et la ponctuation[3]. Ce schéma est complexe à mettre en œuvre de manière concrète dans un système numérique, mais le principe n'est pas éloigné de celui du standard Cidoc-CRM, et notamment de son extension à l'écriture connue sous le nom de CRMtex[4]. Le Cidoc-CRM est un modèle conceptuel de référence *(conceptual reference model)*, c'est-à-dire qu'il fournit des définitions abstraites de concepts et des relations qui les lient. Un exemple est la classe E37 Mark qui est définie comme de la manière suivante :

> This class comprises symbols, signs, signatures or short texts applied to instances of E24 Physical Human-Made Thing by arbitrary techniques, often in order to indicate such

1. P. A. Stokes, « Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien », *Annuaire. Résumés des conférences et travaux, 152e année, 2019-2020*, Paris, EPHE, PSL, SHP, 2021, p. 506-508.
2. P. Monella, « An Ontology for Digital Graphematics and Philology », Wuppertal, 2020, http://www1.unipa.it/paolo.monella/wuppertal2020/ ; J.-M. Klinkenberg et S. Polis, *Les fonctions de l'écriture : un modèle général*, Liège, Collège Belgique, 2019, https://orbi.uliege.be/handle/2268/241566.
3. J.-M. Klinkenberg et S. Polis, *Les fonctions*.
4. C. Bekiari *et al*. (éd.), *Definition of the CIDOC Conceptual Reference Model, Version 7.1.1*, Icom-Cidoc, 2021. F. Murano, A. Felicetti et M. Doerr, *Definition of the CRMtex: An Extension of CIDOC CRM to Model Ancient Textual Entities Version 1.0*, Icom, Cidoc, 2020.

> things as creator, owner, dedications, purpose or to communicate information generally. Instances of E37 Mark do not represent the actual image of a mark, but the abstract ideal (or archetype) as used for codification in reference documents forming cultural documentation[5].

Comme on peut le voir à travers cette définition, le modèle CIDOC-CRM fait également la distinction entre l'écriture physique (l'encre sur la page, l'incision dans la pierre…) et la forme idéale du signe que la marque physique représente visuellement. Il fait également la distinction entre le signe et l'objet linguistique indépendamment de son support physique particulier (écriture visuelle, son enregistré…). Ce modèle a donc des points communs avec ceux de Klinkenberg et Polis et de Monella, ainsi qu'avec ceux de Diehr et Kronemeyer sur l'écriture maya[6], bien que le CIDOC-CRM soit naturellement beaucoup moins développé en ce qui concerne l'écriture en soi.

Comme déjà indiqué, le CIDOC-CRM est un modèle de référence conceptuel, il n'est donc pas formellement défini que dans l'abstrait. Cependant, ce modèle a également été mis en œuvre dans des formats numériques appelés *ontologies*, qui peuvent ensuite être traités automatiquement par des ordinateurs. Parmi les exemples concrets, citons l'implémentation Erlangen du modèle FRBRoo (ce qui combine le CIDOC-CRM et un autre modèle, le FRBR), ainsi qu'une adaptation du Erlangen développée dans le cadre du projet BIBLISSIMA[7]. Nous avons consacré plusieurs séances à ces implémentations en utilisant l'outil de développement d'ontologie appelé Protégé. Nous nous sommes notamment interrogés sur la manière dont ces implémentations pourraient être adaptées à nos écritures et à nos questions de recherche spécifiques.

La deuxième partie de la conférence a été consacrée à l'intelligence artificielle, et plus particulièrement à l'apprentissage profond pour l'analyse automatique d'images de texte. Ce travail, largement pratique, s'est appuyé sur les logiciels Kraken et eScriptorium, tous deux développés par l'équipe d'Humanités numériques de l'EPHE. Nous avons d'abord discuté des principes de l'apprentissage automatique appliqués à deux problèmes. Le premier consistait à trouver des lignes de texte et d'autres régions dans des images de différents formats et contenant différentes écritures. Le second était la transcription manuelle et automatique de ces écritures. Nous avons appliqué les logiciels et les méthodes sur des échantillons issus de nos propres projets, notamment des manuscrits sur parchemin écrits en latin, en vieil anglais et en arabe, des inscriptions sur pierre en vieux khmer et en vieux cambodgien, ainsi que des manuscrits sur feuille de palmier écrits en vieux javanais.

Ce travail a soulevé des questions à la fois d'ordre théorique et pratique. Les humanités numériques exigent une connaissance toujours plus fine des problématiques du domaine étudié, dans notre cas la philologie, la paléographie, l'épigraphie et la codicologie. Nous avons donc pris le temps de discuter de manière transversale des cas que nous avons rencontrés, tels que l'identification des mains, la fabrication

5. C. Bekiari *et al.*, *Definition*, p. 83.
6. F. Diehr *et al.*, « Modellierung von Entzifferungshypothesen in einem digitalen Zeichenkatalog für die Maya-Schrift », dans A. Kuczera, T. Wübbena et T. Kollatz (éd.), *Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten*, Wolfenbüttel, Forschungsverbund Marbach Weimar, 2019, http://dx.doi.org/10.17175/sb004_002.
7. « Ontologie Biblissima », https://doc.biblissima.fr/ontologie-biblissima.

des supports de l'écriture et leurs implications pour les images numériques, et ainsi de suite. Nous avons entamé une discussion importante sur les normes de transcription et sur la manière de traiter des particularités des différentes écritures. Par exemple, l'écriture latine peut être fortement abrégée, mais il est souvent impossible pour le logiciel Kraken de rendre des telles abréviations automatiquement avec leur expansion (c'est-à-dire, le mot complet). Les textes en vieux khmer et en vieux javanais sont généralement publiés dans les éditions scientifiques sous forme de translittérations en alphabet latin avec des signes diacritiques. En principe, les translittérations ne posent pas de problème pour Kraken, tant qu'il existe une relation directe entre l'image d'un signe et sa transcription. Toutefois, l'entraînement d'un tel système nécessite de grandes quantités de données préparées à l'avance qui sont ensuite analysées par la machine pendant le processus d'apprentissage. Cependant, pour les langues rares, quand il existe des corpus numériques, leur volume est souvent peu important. Les méthodes standard d'intelligence artificielle utilisées par les grandes entreprises et de nombreux chercheurs ne fonctionnent pas dans ce contexte. Le partage et la réutilisation des données et des modèles sont un moyen de pallier ce problème. Toutefois, le partage des données exige que toutes les personnes impliquées utilisent les mêmes formats et les mêmes normes. En réalité, les différentes disciplines et les différentes écoles ont toutes leurs propres pratiques en matière de transcription, de définition des régions sur la page, de définition de l'écriture à partir d'une ligne de base, d'une ligne supérieure ou d'une colonne verticale, et ainsi de suite. Tout cela s'ajoute à la variété des normes techniques d'échange de données, comme les formats XML Alto et Page. Ces normes sont bien définies, mais dans la pratique, elles sont utilisées dans différentes versions et avec différents détails de mise en œuvre. De plus, ces normes sont souvent conçues pour les écritures occidentales modernes et intègrent des hypothèses qui ne s'appliquent pas à tous les systèmes d'écriture du monde. Comme toujours avec les humanités numériques, notre travail a donc impliqué à la fois des discussions théoriques et des applications pratiques, en s'efforçant de découvrir ces préjugés et ces hypothèses dans les outils et les méthodes existants, et en voyant si et comment ils peuvent être adaptés à la large gamme d'écritures et de langues qui sont traitées dans notre École.