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

151 | 2020
2018-2019

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/3997
DOI : 10.4000/ashp.3997
ISSN : 1969-6310

Éditeur
Publications de l'École Pratique des Hautes Études

Édition imprimée
Date de publication : 1 septembre 2020
Pagination : 437-439
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], 151 | 2020, mis en ligne le 09 juillet 2020, consulté le 06 juillet 2021. URL : http://journals.openedition.org/ashp/3997 ; DOI : https://doi.org/10.4000/ashp.3997

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

Directeur d'études : M. Peter A. Stokes

Programme de l'année 2018-2019 : I. *Introduction aux humanités numériques*. —II. *Approches numériques et computationnelles appliquées à l'étude de l'écrit ancien*.

En plus de la formation habituelle à l'« Introduction aux humanités numériques »[1], et une formation supplémentaire « Introduction à XSLT et XQuery », la conférence elle-même de cette année a porté sur la modélisation des écritures et sur les outils numériques permettant de les analyser plus efficacement. En particulier, la problématique était de savoir comment, dans quelle mesure et même si les méthodes existantes développées pour la paléographie latine pouvaient être étendues à d'autres systèmes d'écriture, notamment le chinois pré-impérial.

Cette discussion est sous-tendue par l'une des problématiques les plus importantes dans l'utilisation des outils numériques pour la recherche en sciences humaines, à savoir comment (et, encore une fois, si) on peut prendre la complexité illimitée de la production et de l'interaction culturelles humaines, et en représenter d'une certaine manière une partie dans un modèle qui le rende maniable pour l'ordinateur. Cela exige nécessairement de la simplification et de la représentation : en effet, toute recherche est nécessairement simplification d'une réalité complexe, et c'est le cas aussi pour le numérique. Le défi spécifique, cependant, est de savoir comment « les uns et les zéros » de l'ordinateur peuvent être utilisés pour représenter notre objet d'étude de telle manière que le modèle qui en résulte puisse nous aider à répondre à nos questions et à obtenir de nouvelles perspectives sur notre matériel, perspectives qui n'auraient peut-être pas été possibles autrement.

Pour commencer, nous avons examiné différents types de modèles qui sont généralement utilisés dans des contextes numériques. L'un des plus connus est le modèle relationnel, grâce auquel le contenu est représenté comme une série d'entités, chacune avec ses propriétés et des relations entre elles. Une entité peut donc être « écriture », qui pourrait avoir des propriétés telles que « régions où elle était utilisée », « dates auxquelles elle était utilisée », et ainsi de suite. Chaque entité peut être liée à d'autres entités telles que « Caractères » ou « Langues », qui à leur tour auraient leurs propres propriétés et d'autres relations. Ce modèle est le fondement des bases de données relationnelles « classiques » qui sont encore largement utilisées et qui ont une valeur pragmatique significative. En raison de l'importance de ce modèle et de la technologie qui lui est associée, un peu de temps a été consacré à son exploration et à la manière de le mettre en œuvre avec un logiciel libre, Open Office Base.

1. Voir P. A. Stokes, « Introduction aux humanités numériques », *Annuaire. Résumés des conférences et travaux, 150e année, 2017-2018*, Paris, EPHE, PSL, SHP, 2019, p. 495-496.

Un autre modèle est l'Entité-Relation, tel qu'il est exprimé dans le « Unified Modelling Language » (UML) ; il est similaire en principe au modèle relationnel, bien que différent dans le détail. Ces deux approches ont été utilisées pour créer le modèle Archetype qui a d'abord été développé pour le projet DigiPal et qui a depuis été appliqué avec succès à un large éventail d'écritures différentes, y compris le latin, le grec, l'hébreu, et expérimentalement le maya, l'arabe et le chinois moderne[2]. Après avoir examiné certaines de ces autres applications du modèle, nous avons cherché à voir comment il pourrait être adapté à l'écriture chinoise pré-impériale. Certains aspects clés semblaient en effet manquer dans le modèle existant, notamment ceux de la position relative des composants dans le caractère, ainsi que de la distinction entre les composants et les traits. La fonction différente des composants des caractères en chinois par rapport à ceux en latin semble également pertinente et doit être saisie, en particulier pour identifier ces composants qui donnent une valeur phonétique au caractère. Pour la position relative des composantes dans le caractère, nous avons envisagé un système de grille qui n'est pas loin de celui utilisé par le Wenlin « Character Description Language » qui fait actuellement partie du standard Unicode[3].

En plus de l'écriture, un autre défi de la modélisation d'objets historiques est celui des dates. Ceux qui créent les logiciels informatiques « standards » ont tendance à supposer que les événements se sont produits à un moment précis et unique et que ce moment de la production est connu avec un haut degré de confiance. Par exemple, de nombreux catalogues en ligne donnent les dates des manuscrits en années exactes, ou dans des fourchettes étroites d'années. Un livre daté de la « première moitié du XIe siècle » pourrait donc être stocké dans une base de données comme ayant été écrit entre les années 1000 et 1050. Cependant, on pourrait bien se demander si la première année du XIe siècle est 1000 (comme c'est presque toujours le cas dans les logiciels en ligne) ou 1001 (comme on peut s'y attendre plus strictement), ou d'autres dates selon le calendrier utilisé. Plus important encore, aucune de ces dates n'est une représentation correcte pour la datation des sources historiques. La « première moitié du XIe siècle » est nécessairement imprécise et peut inclure (par exemple) l'année 999, ou l'année 1051, ou même peut-être l'année 995. Dans la pratique, les dates peuvent souvent être plus complexes que cela, avec des possibilités telles que « première moitié du XIe siècle (vers 1020 ?) » ou « *c.* 1015 (ou vers 1065 ?) » [4]. Dans tous ces cas, la représentation du niveau d'incertitude dans un ordinateur est loin d'être simple, et nous avons donc envisagé quelques moyens possibles de régler ce problème.

En s'éloignant davantage de la question de la modélisation au sens pur, nous avons également considéré certains éléments pratiques qui prennent de plus en plus d'importance dans le travail numérique avec des écritures et des images. L'un d'entre eux est le « International Image Interoperability Framework » (IIIF : https://iiif.io/), qui

2. S. Brookes, P. A. Stokes, M. Watson et Débora Marques de Matos, « The DigiPal Project for European Scripts and Decorations », dans A. Conti, O. da Rold and P. Shaw (éd), *Writing Europe 500–1450: Texts and Contexts*, Cambridge, D. S. Brewer, 2015, p. 25-58.
3. Voir § 18.2 « Ideographic Description Characters », dans *Unicode Standard v. 12*, The Unicode Consortium, Mountain View (CA), 2019, p. 723-726.
4. P. A. Stokes, « The Problem of Digital Dating: A Model for Uncertainty in Medieval Documents », dans *Digital Humanities Book of Abstracts*, Sydney, The University of Sydney, 2015.

est un standard permettant aux logiciels de communiquer avec des dépôts d'images, ce qui signifie que l'on peut désormais travailler directement avec des images qui sont hébergées par différentes institutions à travers le monde. Cette innovation importante signifie non seulement que les images sont disponibles en ligne, mais aussi qu'elles ne sont plus verrouillées sur le site web d'une seule institution : au contraire, nous pouvons créer nos propres sites web ou d'autres logiciels qui peuvent automatiquement trouver et afficher des images provenant de nombreuses institutions à travers le monde. Cela permet d'effectuer des comparaisons et des analyses qui ne sont pas limitées par la localisation fortuite de documents historiques dans les bibliothèques d'aujourd'hui, mais qui nous permettent plutôt de « recréer » virtuellement des collections historiques comme elles l'étaient autrefois. Cela permet également la récolte automatique d'images de plusieurs milliers d'objets, ce qui permet alors des analyses automatiques et computationnelles. Un point de départ de cette analyse est souvent l'apprentissage automatique pour transcrire les manuscrits pour nous (OCR / HTR), et l'un des outils de pointe pour ce faire est Kraken (http://kraken.re/), qui a été développé par M. Ben Kiessling de l'EPHE. Une session de la conférence a donc été consacrée à un atelier donné par M. Kiessling au cours duquel nous avons fait des transcriptions de textes en pratique avec Kraken.