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

154 | 2023
Annuaire de l'EPHE, section des Sciences historiques et philologiques (2021-2022)

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/6630
DOI : 10.4000/ashp.6630
ISSN : 1969-6310

Éditeur
Publications de l'École Pratique des Hautes Études

Édition imprimée
Date de publication : 1 septembre 2023
Pagination : 517-524
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], 154 | 2023, mis en ligne le 22 juin 2023, consulté le 01 décembre 2023. URL : http://journals.openedition.org/ashp/6630 ; DOI : https://doi.org/10.4000/ashp.6630

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

Directeur d'études : M. Peter A. Stokes

Programme de l'année 2021-2022 : *Vers une paléographie transversale : le numérique et l'IA pour modéliser, comparer et analyser l'écriture ancienne*.

Au cours de cette année, nous avons mis en pratique nos réflexions théoriques de l'année dernière sur un projet pratique : créer un modèle conceptuel pouvant être appliqué à des objets inscrits dans n'importe quelle langue ou écriture, implémenter notre modèle avec une base de données, remplir cette base avec des données existantes et créer une interface web pour interagir avec les données[1]. En réalisant ces tâches nous avons rencontré et discuté d'un grand nombre de questions théoriques et pratiques sur la modélisation, l'échange de données, l'implémentation, la documentation et autres, qui sont au cœur des humanités numériques.

Le modèle conceptuel

La première étape de ce processus a été de développer un modèle conceptuel. Il s'agit d'une représentation plus ou moins formelle des concepts que nous souhaitons représenter. Nous avons essayé de définir ces concepts ainsi que les relations qu'ils entretiennent entre eux aussi précisément que possible. Cette étape a l'avantage de faciliter la communication et d'aider les autres à comprendre la base de données que nous allons créer. Elle est également précieuse pour clarifier notre propre pensée, car de nombreux concepts dans les sciences humaines sont ambigus ou possèdent des significations différentes en fonction du contexte dans lequel ils sont employés. Dans la pratique, l'élaboration d'un modèle conceptuel peut s'avérer très difficile, car elle nous oblige à reconnaître et à expliciter des représentations implicites, et à identifier des ambiguïtés, voire des contradictions, dans des concepts qui nous semblent clairement définis.

Afin de développer notre modèle, une première étape a consisté à examiner les différentes entités et à décider lesquelles inclure, puis à leur fournir des définitions claires. Il s'agit d'une tâche particulièrement ardue. En effet, il convient de trouver des termes suffisamment concrets pour être utiles, mais en même temps assez abstraits pour pouvoir être appliqués à l'ensemble des types d'écriture et des objets. Par exemple, un point de départ évident serait *Manuscrit*, qui fait référence à l'objet contenant le texte manuscrit. Cependant, dans la pratique, le terme « manuscrit » tend à designer un codex ou une autre forme de document portable, mais pas une inscription sur pierre, même si cette inscription est écrite à la main. En outre, les informations que l'on enregistre normalement pour une inscription sur pierre sont généralement différentes de

1. Pour le modèle conceptuel et les discussions théoriques de l'année dernière, voir P. A. Stokes, « Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien », *Annuaire. Résumés des conférences et travaux, 153e année, 2020-2021*, Paris, EPHE, PSL, SHP, 2022, p. 529-531.

celles des écritures sur des feuilles de parchemin, de papier, de feuilles de palmier ou d'autres supports portatifs. Cela suggère alors le concept de *Text-Bearing Object* comme classe, avec des cas spécifiques tels que *Codex* ou *Inscription* comme sous-classes. Cette approche est utile, mais il est difficile de décrire l'« archéologie » de l'objet, c'est-à-dire ses différents états à différentes époques[2]. Les livres, les inscriptions et les autres objets sont souvent très différents aujourd'hui de ce qu'ils étaient au moment de leur production : ils peuvent être divisés en unités qui circulent séparément, ou des unités qui étaient à l'origine séparées peuvent être jointes plus tard, du texte peut être ajouté ou supprimé, etc. Heureusement, ces problèmes de terminologie ont déjà été discutés par Andrist, Canart et Maniaci. Ces derniers distinguent l'*Unité de production (UniProd)* et l'*Unité de circulation (UniCirc)*[3].

Ensuite le défi est comment modéliser le texte en soi, ainsi que le contenu paléographique. Développé par la Fédération internationale des associations des bibliothécaires et des bibliothèques[4], le modèle FRBR semble être le plus approprié. Ce modèle distingue entre quatre entités principales : *Œuvre*, « une création intellectuelle ou artistique déterminée », *Expression*, « la réalisation intellectuelle ou artistique d'une œuvre », *Manifestation*, « la matérialisation de l'une des expressions d'une œuvre », et *Item*, « un exemplaire isolé d'une manifestation »[5]. Ils donnent en exemple l'*Œuvre* Œ1 *Playback* de Ronald Hayman, qui trouve sa réalisation dans l'*Expression* E1 le texte de l'auteur, mis en forme pour l'édition, qui se concrétise dans la *Manifestation* M1 le livre édité en 1973 chez Davis-Poynter, qui est représentée par *l'Item* I1 un exemplaire avec dédicace autographe de l'auteur. Dans notre cas, l'*Item* correspond directement au *Text Bearing Object*, c'est-à-dire la combinaison de l'*Unité de production* et de l'*Unité de circulation*. Le concept d'œuvre est aussi utile, parce qu'il nous donne la possibilité de grouper tous les objets qui contiennent une telle œuvre. En revanche, l'*Expression* nous semblait moins pertinente et nous ne l'avons pas incorporé dans notre modèle. Enfin, le concept de *Manifestation* est pertinent pour notre étude, mais son introduction génère des difficultés importantes. La définition d'une *Manifestation* est « la matérialisation de l'expression d'une œuvre », et elle « regroupe tous les objets matériels présentant les mêmes caractéristiques, tant du point de vue du contenu intellectuel que du point de vue de l'aspect matériel »[6]. Cette définition fonctionne assez bien pour les livres imprimés, mais elle est moins pertinente pour les objets crées à la main, où chacun a ses propres caractéristiques matérielles. Pour cette raison, quelques versions du standard FRBRoo, fondé sur celui de FRBR, ajoute le concept de *Manifestation Singleton* pour mieux prendre en compte les objets faits à la main. Toutefois, cette entité a été supprimée des versions

2. Pour l'« archéologie » du livre, voir F. Masai, « Paléographie et codicologie », *Scriptorium*, 4 (1950), p. 290 et 292-293, et A. Gruijs, « Codicology or the Archaeology of the Book? A False Dilemma », *Quaerendo*, 2 (1972), p. 87-108.
3. P. Andrist, P. Canart et M. Maniaci, *La Syntaxe du codex : essai de codicologie structurale*, Turnhout, Brepols, 2013, p. 59-61.
4. Groupe de travail IFLA sur les Fonctionnalités requises des notices bibliographiques, *Fonctionnalités requises des notices bibliographiques : rapport fina*l, 2e édition française, Paris, Bibliothèque nationale de France, 2012.
5. Groupe de travail IFLA, *Fonctionnalités*, p. 14.
6. Groupe de travail IFLA, *Fonctionnalités*, p. 19.

plus récentes[7]. Quoi qu'il en soit, la *Manifestation* est essentielle pour nos besoins et nous l'avons donc retenue pour notre modèle.

Les autres entités sont plus simples. Nous voulons inclure les lieux, par exemple le lieu de production du livre, le lieu où l'inscription se trouve actuellement, et ainsi de suite. Les personnes morales et physiques sont aussi essentielles, par exemple pour modéliser non seulement les scribes et les auteurs, mais aussi les institutions telles que les bibliothèques, les monastères et ainsi de suite. Nous avons aussi ajouté les évènements historiques comme entités, pour modéliser (par exemple) la production du livre, la donation d'une parcelle de terre, la casse d'un objet en deux parties, et ainsi de suite. Nous avons aussi décidé de modéliser les ordres religieux et les statuts ou les professions des personnes comme entités. Ces derniers pourraient être des attributs d'une personne (par exemple), mais dans ce cas il serait difficile, voire impossible, d'indiquer des informations telles que la période pendant laquelle la personne a exercé cette profession. Les images numériques constituent aussi une entité essentielle. Comment modéliser la paléographie est une question assez difficile, mais pour l'instant nous avons ajouté les mains comme entités, ainsi que la disposition des pages, même si ces deux concepts ne sont pas encore bien définis.

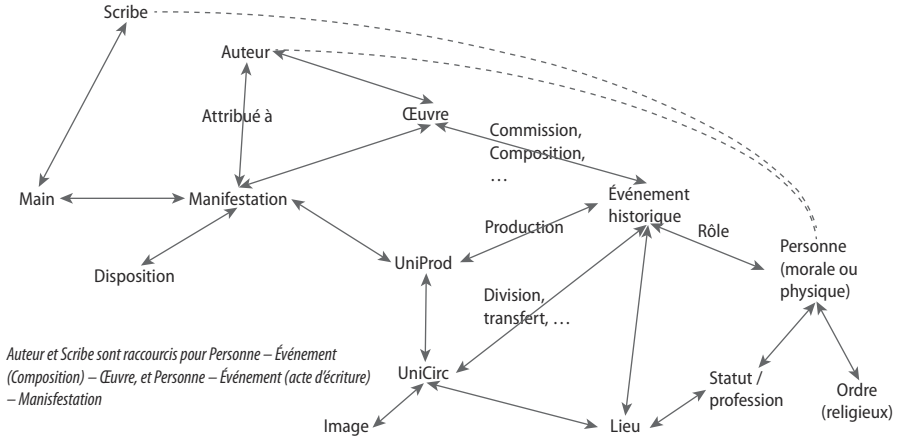
Une fois que les entités sont bien définies, il faut préciser les relations qui les unissent. La plupart des relations sont simples : par exemple, il y a une relation évidente entre une *Unité de circulation* et une *Unité de production*, ou entre une *Œuvre* et la *Personne* qui l'a écrite. Il y a également des relations moins claires. Nous avons, par exemple, ajouté une relation entre une *Personne* et la *Manifestation* pour indiquer l'auteur qui est désigné par le texte de cet objet n'est pas forcément le véritable auteur de l'œuvre. Un autre point de discussion était la relation entre l'*Unité de circulation* et la *Manifestation*. En principe, selon *La syntaxe du codex*, chaque ajout de contenu crée une nouvelle *Unité de production*, même sans addition du support matériel[8]. Dans ce cas, il semble clair que la *Manifestation* appartient à l'*Unité de production* et non pas à l'*Unité de circulation*. Cependant, ce principe implique qu'un manuscrit avec des gloses ajouté par une dizaine des personnes possèdera une dizaine d'*Unités de production* et donc une dizaine des *Manifestations*. Cela rendrait la base de données très difficile à gérer. Cette approche est aussi contraire au standard de FRBR, qui précise que :

> Des modifications survenues après l'achèvement du processus de production d'un exemplaire isolé […] ne sont pas réputées déboucher sur une nouvelle manifestation. L'exemplaire qui en résulte est simplement réputé être un exemplaire (ou item) présentant des caractéristiques déviantes par rapport aux autres exemplaires de la même manifestation[9].

Pour des raisons pratiques, nous avons aussi ajouté deux raccourcis : *Scribe* pour une *Personne* liée à un *Évènement Historique* type « acte d'écriture » et une *Main*, et *Auteur* pour une *Personne* liée à un *Évènement Historique* type « acte de composition » et une *Œuvre*.

7. Bekiari *et al.* (éd), *LRMoo (formerly FRBRoo) Object-Oriented Definition and Mapping from IFLA LRM, draft version 0.9*, 2022, p. 70.
8. Andrist *et al*., *La Syntaxe*, p. 64.
9. Groupe de travail IFLA, *Fonctionnalités*, p. 20.

Le modèle conceptuel peut donc être exprimé avec le diagramme suivant :



L'alignement du modèle conceptuel

Une fois le modèle conceptuel établi, il convient de l'aligner avec les modèles et les standards faisant autorité. Le principe est de préciser comment les entités de notre modèle correspondent à celles des autres standards : par exemple, notre *Manifestation* est l'équivalent exact de l'entité *frbr:F3 Manifestation* du standard FRBRoo ; nos entités *Unité de production* et *Unité de circulation* sont équivalentes à l'entité *frbr:F4 Item* de FRBRoo ; elles sont aussi des types (ou sous-classes) de *crm:E22 Human-Made Object*.

Notre modèle	Modèle FRBRoo	Type de Relation
Œuvre	*frbr:F1 Work*	Équivalence directe
Manifestation	*frbr:F3 Manifestation*	Équivalence directe
Text-Bearing Object, *UniCirc*, *UniProd*	*crm:E22 Human-Made Object*	Sous-classe
Événement Historique	*crm:E5 Event*	Sous-classe
Lieu	*crm:E53 Place*	Équivalence directe
Image (numérique)	*crm:E36 Visual Item*	Sous-classe
Organisation	*crm:E39 Group*	Sous-classe
Personne (physique)	*crm:E21 Person*	Équivalence directe
Personne (morale ou physique)	*crm:E39 Actor*	Sous-classe
Writing Sample	*tex:TX7 Written Text Segment*	Équivalence directe
Main (scribale)	*tex:TX7 Written Text Segment*	Sous-classe

Par ailleurs, dans notre modèle plusieurs relations correspondent aux entités ou aux autres structures de FRBR :

Notre modèle	FRBRoo	Type de relation
Rôle	crm:P14 *carried out by* crm:E39 Actor crm:P14.1 *in the role of* crm:E55 Type	Équivalence indirecte (raccourcie)
Période	crm:E52 Time Span	Équivalence directe
Événement + type	crm:E12 Production, crm:E11 Modification, crm:E6 Destruction, crm:E9 Move, crm:E8 Acquisition, crm:E10 Transfer of Custody	Équivalence indirecte (réification)
Auteur, *Scribe*	crm:E21 Person crm:P14.1 *in the role of* crm:E55 Type	Équivalence indirecte (raccourcie)

Implémentation avec Heurist et importation des données

Une fois que le modèle conceptuel est défini, il nous fallait l'implémenter dans une base de données. Un nombre énorme de logiciels sont disponibles pour créer et gérer des bases de données, mais nous avons choisi d'utiliser Heurist. Créé par une équipe de l'université de Sydney, Heurist est conçu pour les sciences humaines. Une instance de ce logiciel est disponible sur l'IR* HumaNum, l'infrastructure de recherche française pour les humanités numériques[10]. Un grand nombre de détails est déjà pris en compte par le logiciel, ce qui nous aide avec l'implémentation, et qui nous permet de créer la plupart de notre base de données sans besoin de programmation. Par exemple, plusieurs de nos entités sont déjà présentes dans Heurist, telles que *Personne*, *Institution*, *Lieu*, *Œuvre*, *Événement historique*, et *Image*. Nous avons facilement ajouté les autres entités, telles que l'*Unité de circulation* et l'*Unité de production*. D'autres entités sont présentes dans Heurist mais ne sont pas parfaitement adaptés à nos besoins, et donc nous les avons modifiés. Enfin, pour compléter le modèle et donc l'implémentation, il faut aussi décider quelles informations on souhaite enregistrer pour chaque entité. Par exemple, pour une *Unité de production*, nous avons choisi les informations suivantes :

— Identifiant et source d'identifiant (par ex. le numéro dans un catalogue publié) ;
— La date de création ;
— Le lieu de création (c'est-à-dire, une relation entre l'Unité de production et un Lieu) ;
— Le format de l'objet (codex, inscription murale…) ;
— Le matériel (parchemin, papier, pierre…) et une indication de sa qualité ;
— Une description de l'objet ;
— Un ou plusieurs liens vers (relations entre) les Manifestations ;
— Des références bibliographiques.

La figure 1 est une capture d'écran d'une Unité de production dans Heurist.

Un des avantages d'Heurist est que l'on peut s'appuyer sur les travaux existants et donc éviter la programmation. Par exemple, si l'on ajoute aux lieux leurs coordonnées en latitude et longitude, on peut sélectionner des lieux selon des critères précis et les afficher sur une carte (fig. 2).

10. L'installation Heurist de HumaNum est disponible à https://heurist.huma-num.fr/.

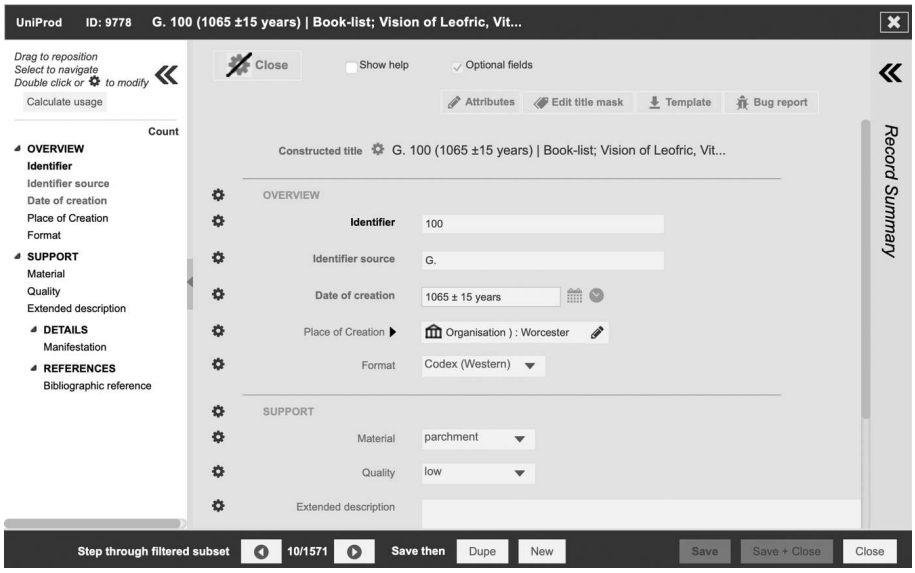


Fig. 1. — Structure d'une Unité de production dans notre base de données Heurist.

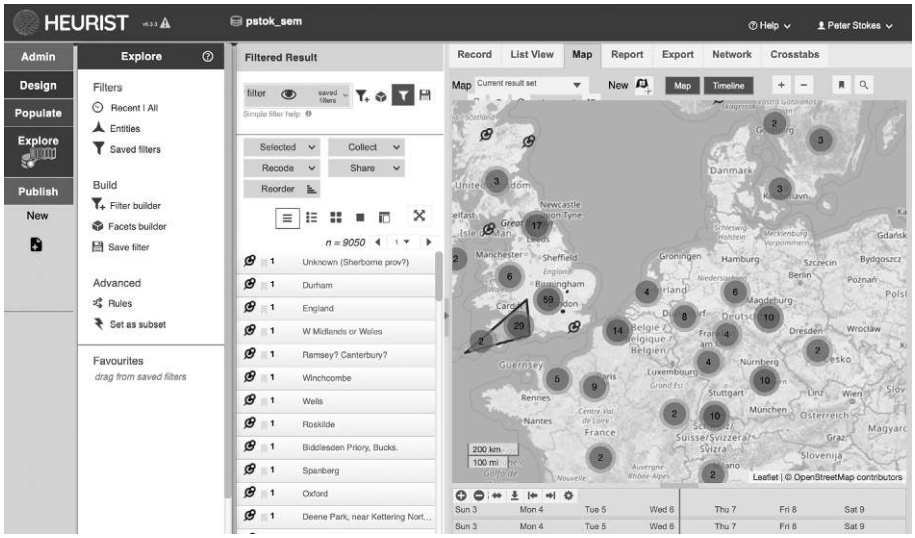


Fig. 2.

Importation de données depuis DigiPal et GeoNames

À ce stade, la structure de la base de données était en place : toutes les entités sont définies, ainsi que les champs qui les décrivent et les relations qui les unissent. L'étape suivante consistait à saisir les données. Cela pourrait être fait à la main, mais nous avons décidé d'importer les données d'un projet existant, DigiPal. Le projet DigiPal est le résultat d'une « Starting Grant » de l'European Research Council qui

s'est déroulée de 2010 à 2014 et qui comprenait la conception et la création d'une base de données et d'un site Web pour l'étude paléographique des écrits du XIe siècle en langue anglaise[11]. Au moment de la rédaction de cet article, la base de données DigiPal contient des entrées pour 1 675 manuscrits contenant 1 477 mains de scribes et plus de 60 000 images annotées de lettres écrites par ces scribes. La base de données DigiPal est donc clairement pertinente aussi pour notre base de données dans Heurist, et cela a soulevé la question de savoir si les anciennes données pouvaient être importées dans le nouveau site.

DigiPal est basé sur la plateforme Archetype (qui a également été développée par l'auteur et son équipe au King's College de Londres), et celui-ci possède une API Web qui permet à un logiciel externe de récolter le contenu de la base de données et de l'utiliser directement à d'autres fins[12]. Cependant, cette récolte n'est pas facile à faire dans la pratique, car le modèle de données de DigiPal est différent de celui que nous avons développé pour Heurist. Dans certains cas, les différentes entités sont suffisamment claires pour être liées directement : par exemple, un Historical Item de DigiPal correspond plus ou moins directement à une Unité de circulation de Heurist. En outre, le modèle DigiPal / Archetype a été aligné au modèle FRBRoo / CIDOC-CRM[13] que nous avons déjà évoqué. Cela signifie que FRBRoo peut être utilisé comme format intermédiaire pour la conversion entre DigiPal et Heurist. Cette approche a été relativement fructueuse, malgré quelques pertes de données résultant de légères différences entre les modèles. La difficulté la plus notable était peut-être celle des lieux. Le modèle DigiPal pour les lieux comprend un simple texte sans autres détails tels que les coordonnées sur une carte. Cela présente l'avantage considérable de permettre une grande souplesse qui est essentiel pour prendre en compte l'incertitude dans la représentation des lieux historiques, avec des exemples tels que « Worcester or York », « Abingdon (or Continent?) », et « France (prob. Brittany, or South-West France?) ». Cependant, l'absence de coordonnées rend les lieux impossibles à reporter sur une carte, ce qui est l'un des objectifs de notre base de données Heurist. À cette fin, nous avons produit un script simple et plutôt naïf pour la géolocalisation, c'est-à-dire pour identifier les coordonnées de chaque lieu et insérer cette information dans la base de données. Pour ce faire, nous avons utilisé GeoNames, qui est un répertoire géographique pour les lieux[14]. Il permet à une personne ou à un ordinateur de rechercher le nom d'un lieu et de trouver ses coordonnées, ainsi que le nom du lieu dans un grand nombre de langues et d'autres données. Il était très facile d'écrire un court programme informatique pour parcourir les listes dans DigiPal, les consulter sur GeoNames, puis de créer des fichiers XML avec les coordonées et les importer dans Heurist.

11. S. Brookes *et al*., « The DigiPal Project for European Scripts and Decorations », dans A. Conti, O. Da Rold, et P. A. Shaw (éd.), *Writing Europe 500–1450: Texts and Contexts*, Woodbridge, 2015, p. 25-58.
12. Pour le logiciel Archetype voir https://github.com/kcl-ddh/digipal, et G. Noël et P. A. Stokes, « The Web API Syntax », https://github.com/kcl-ddh/digipal/wiki/The-Web-API-Syntax.
13. P. Stokes, « Aligning Archetype: Towards a Formal Model for a Transversal Palaeography », *Comparative Oriental Manuscript Studies Bulletin*, à paraître ; et P. Stokes, « The Archetype Ontology », version 0.7.1 (2021), https://doi.org/10.5281/zenodo.5771601.
14. GeoNames, https://www.geonames.org.

Ce processus fonctionnait, mais il ne tenait pas compte de deux facteurs importants : premièrement, l'ambiguïté des noms et, deuxièmement, l'incertitude des lieux historiques. En ce qui concerne le premier facteur, de nombreux noms de lieux se retrouvent dans plusieurs endroits du monde, comme Canterbury en Angleterre et Canterbury en Nouvelle-Zélande, pour n'en citer qu'un. Ce problème a été réduit en donnant la priorité aux résultats en Europe lorsqu'ils étaient disponibles, mais cela n'a pas permis de résoudre toutes ces ambiguïtés. Deuxièmement, des noms de lieux tels que « Abingdon (or Continent?) » n'existent pas dans GeoNames, et bien que le code informatique ait pu être adapté pour rechercher simplement « Abingdon » dans ce cas, cela n'aurait toujours pas résolu le problème sous-jacent, c'est-à-dire comment afficher l'incertitude sur une carte. Enfin, d'autres particularités des données DigiPal ont empêché l'identification des lieux, comme l'utilisation de certaines abréviations idiosyncratiques (« WiOM » pour « Winchester, Old Minster » et ainsi de suite), ainsi que le nom « Unknown » que GeoNames a situé à l'extrême nord-ouest du Bangladesh. La seule solution faisable était d'examiner chaque nom, de décider de la meilleure façon de le représenter cas par cas, et puis de modifier les données en conséquence. Cette approche n'est pas idéale, mais est plus ou moins inévitable lorsqu'on essaie de combiner différents ensembles de données. Il s'agit d'un exemple utile pour montrer certains des défis qui découlent des sources de données ouvertes, en particulier dans les sciences humaines avec toute la richesse et la complexité de ce matériel.