Méthodologie et environnement pour le traitement de données appliquées aux Sciences Humaines et Sociales

Tiphaine VAN DE WEGHE

Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour
Avenue de l'Université
64000 PAU

t.van-de-weghe@univ-pau.fr

MOTS-CLÉS : *données, planification, modélisation, collecte, traitement, représentation*

KEYWORDS: *data,planning,modelization, collecting, processing, representing*

ENCADREMENT. *Philippe Roose (MCF) et Marie-Noëlle Bessagnet (MCF)*

1. Contexte

Les recherches génèrent de nouvelles informations. Si on prend le nombre moyen de projets par Université, multiplié par le nombre d'Universités en France, on est face à un grand volume de données. Imaginons donc ces éléments à l'échelle mondiale. En Sciences Humaines et Sociales (SHS), ces renseignements peuvent se multiplier très vite. Les recherches se font sur des supports retrouvés en archives, ou encore sur des études déjà effectuées, des événements étudiés etc. Les SHS regroupent des disciplines qui analysent les humains et les sociétés, qui ont existé et existent. Ces sciences possèdent une matière première diverse et complexe (images, vidéos, sons, textes non numérisés,etc.). Durant cette thèse, nous allons traiter plusieurs points, pendant lesquelles les données subirons des transformations. On dénombre parmi ces étapes : (i) la collecte de données, (ii) la modélisation, (iii) le stockage, (iv) le traitement/analyse de l'information (v) la valorisation. **L'objectif de cette thèse est de couvrir l'ensemble de ces phases et d'aider au mieux les chercheurs en SHS à valoriser leur données de la recherche.** Dans cet article, nous allons présenter l'intérêt d'une collaboration

entre les SHS et les Sciences Exactes, avec, pour fil conducteur l'intervention de l'informaticien pour la gestion des données, tout en faisant un bref état de l'art. Par la suite, nous pourrons présenter les problématiques autour des cinq points principaux (i à v) de cette thèse énumérés ci-dessus. Puis, nous déploierons les actions réalisées et futures, et nous conclurons.

2. État de l'art

L'OCDE[1] définit les données de la recherche comme " des enregistrements factuels (chiffres, textes, images et sons) utilisés comme sources principales pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires à la validation des résultats de recherche ". L'Université Humboldt de Berlin (Schöpfel *et al.*, 2017) distingue deux types de données : les sources (brutes) et les résultats (traitées). Elles suivent l'acheminement du cycle de vie de la donnée qu'offre le schéma de UK Data Service. (Reymonet *et al.*, 2018) montrent qu'il y a plusieurs intervenants (archivistes, documentalistes, informaticiens, chercheurs, etc.) dans le plan de gestion des données. L'implication de l'informaticien et sa mise en relation avec le chercheur en SHS est incontournable. En effet, le chercheur en SHS possède des ressources ou supports informatisés ou non. L'informaticien l'aide à les structurer dans des bases de données, ou encore, aider à réaliser de l'extraction d'information semi-automatique, comme l'ont stipulé (Kergosien *et al.*, 2016). Le type de données que le chercheur en SHS possède est : des textes, des images, des enregistrements sonores, des vidéos etc. Mais, les études sur les données de la recherche des Universités de Lille 3 et Rennes 2 (Prost *et al.*, 2015) et (Serres *et al.*, 2017), montrent que les principales ressources des chercheurs en SHS sont des textes. Bien entendu, pour que l'informaticien intervienne, il est nécessaire que le matériel des chercheurs en SHS soient numérisées. Malheureusement, les documents textuels ne le sont pas systématiquement.

3. Problématique

Quels sont les besoins des chercheurs en SHS ? Qu'ont-ils comme support (ressources) ? Au regard de notre expérience, deux profils sont déterminés : Dans un premier cas, le chercheur a des connaissances en système d'information, tel que l'archéologue qui use des systèmes d'information géographique (qui est le plus utilisé et qui a la plus grande communauté). Il se construit sa propre base de données, et parfois il étudie lui-même en cartographiant ses données. Celles-ci sont souvent analysées statistiquement par le gestionnaire de données. D'un autre côté, il y a des chercheurs qui nécessitent un accompagnement complet, c'est-à-dire de l'étude des besoins à la valorisation des données. Ils ont des problématiques, des supports, et un savoir, mais ne savent pas comment valoriser et utiliser ces informations, à l'aide de l'outil

1. Organisation de Coopération et de Développement Économiques

informatique, ni d'ailleurs ce qu'un tel outil peut lui apporter. Notre rôle sera alors de comprendre et de modéliser ses éléments, en définissant des méthodes et outils qui les automatisent. Après avoir analysé leurs demandes, il faut penser à l'organisation, au tri de leurs données. Il s'agit de les structurer, de les enregistrer. L'utilisation d'une base de données est nécessaire. De plus, à la fin des projets de recherche, certaines valeurs doivent être codées et exportées vers des institutions pour expositions et/ou pérennisation. A partir de cette collection, qu'est-il possible de faire avec ces données ?

4. Actions réalisées

En tant qu'ingénieur d'étude au laboratoire ITEM EA 3002[2], j'ai pu observer et analyser les pratiques des chercheurs en SHS. Ce laboratoire de recherche est composé de 25 enseignants chercheurs (anthropologie, archéologie, histoire, histoire de l'art, et étude hispaniques) et d'une cinquantaine de doctorants. Les principaux axes de recherches sont :

– territoires, mobilités

– identités, patrimoines

– méthodologie de la recherche : "archives et corpus"

Cela m'a permis de recenser les pratiques des anthropologues, des archéologues et des historiens, en terme de collecte de données et surtout les formats de données utilisés par ces chercheurs. Actuellement, le principal comportement des chercheurs est la consultation de bases de données bibliographiques spécialisées, sites Internet, archives, etc. Nous avons recensé, dans le laboratoire, différents types de support qui sont : des textes, des images, des vidéos, des enregistrements sonores, des valeurs quantitatives, ou encore des données qualitatives. Les documents textuels du laboratoire sont généralement des lettres de correspondance, des articles, etc. Nous modélisons de façon à être interopérable dans le futur. C'est pour cela que nous avons également étudié la norme Dublin Core avant de déterminer des variables plus appropriées aux recherches. Le Dublin Core, comme le définit la BnF (Bibliothèque Nationale de France) est une norme internationale de variables (titre, créateur, langue, contributeur, etc.) qui est utilisée par de nombreuses institutions, ce qui facilite leur exportation. Dans certains cas, il est nécessaire de gérer des données sources autrement (collecte et stockage), puis, de respecter la norme Dublin Core, avec des données résultats (sélection des informations pertinentes pour le chercheur). Pour des raisons de confidentialité, parfois un stockage suffit pour des éléments sources, seuls les résultats sont importants (éléments sensibles). Cependant, pour une insertion des valeurs, les chercheurs doivent pouvoir les stocker afin de pérenniser leur travail.

---

2. Identités, Territoires, Expressions, Mobilités Équipe d'Accueil

5. Actions futures

Après la collecte des données suivent les phases de traitement et d'analyse qui permettront de représenter les données, de manière compréhensive par tous. Nous extrairons les informations répondant aux problématiques des chercheurs. Ces analyses entraînent un traitement avec des méthodes statistiques et/ou informatiques. Quelle sera alors leur complétude ? En effet, l'informatique permet d'extraire les informations des corpus selon 3 dimensions : thématique, spatiale et temporelle. Nous appliquerons notamment des méthodes liées à la fouille de texte, ainsi que des démarches statistiques et de la cartographie. Notre question de recherche sera alors : comment combiner judicieusement ces types d'analyses ? Les supports sont souvent présentés sous la forme d'anciens écrits. Le chercheur en SHS va tenter de récupérer une majorité de ces documents. Pour des raisons politiques ou des droits d'utilisation, il rencontre des difficultés à se les procurer. Ses recherches sont donc basées sur un minimum de textes de taille plus ou moins grande. Alors, comment adapter les méthodes de fouille de texte sur de tels corpus ? Les documents en langues étrangères et anciennes apparaissent comme un autre défi. Quels sont les méthodes et outils informatiques pour aider les chercheurs en SHS dans l'analyse de tels textes ? Le chercheur en SHS demande également des outils et méthodes pour une meilleure visualisation des résultats. Parfois sur des corpus conséquents, la recherche d'information sera développée. Est-ce un type de valorisation ? En quelque sorte, cela met en avant les données. Il existe, tout de même, d'autres techniques qui seront définies. En général, des publications, des applications, etc. sont la suite logique de la valorisation des données. Mais aussi, l'exportation des données sources vers d'autres établissements où la fusion se fait avec d'autres données de recherche. Chaque institution choisit son mode de structuration de données, avant de procéder à leurs insertions. Il est important de se renseigner sur leurs moyens de travail (langage utilisé) afin de s'adapter. Aujourd'hui, un projet de recherche doit faire face à cette exportation, qui se fera en XML (langage de balisage extensible).

Cette thèse a démarré en janvier 2018. Elle reprend les points du cycle de vie des données, en développant des automatismes qui correspondent au mieux aux attentes des chercheurs en SHS, après avoir fait une analyse des besoins. L'enquête tend vers des projets de recherches en SHS comme Acronavarre[3] (actes royaux de Navarre), le patrimoine d'encre pyrénéen et TCVPYR[4] (Thermalisme, Culture, Villégiature dans les Pyrénées). Des modèles de données ont été créés pour la collecte. Ces programmes de recherches fournissent la matière appropriée à cette expérimentation. Ces derniers demandent un travail complet sur les données, c'est-à-dire, de la planification à leur valorisation.

3. projet ANR : https ://acronavarre.hypotheses.org/
4. projet européen FEDER : http ://tcvpyr.iutbayonne.univ-pau.fr/

6. Bibliographie

Kergosien E., Bessagnet M.-N., Sallaberry C., Le Parc-Lacayrelle A., Royer A., « Analyse géographique de séries de publications : application aux conférences EGC », *EGC'2016 (Extraction et Gestion des Connaissances)*, p. 371–382, 2016.

Prost H., Schöpfel J., Les données de la recherche en SHS. Une enquête à l'Université de Lille 3., Technical report, Lille 3, 2015.

Reymonet N., Moysan M., Cartier A., Délémontez R., « Réaliser un plan de gestion de données « FAIR » : guide de rédaction », 2018.

Schöpfel J., Kergosien E., Prost H., « « Pour commencer, pourriez-vous définir'données de la recherche' ? » Une tentative de réponse », *Atelier VADOR : Valorisation et Analyse des Données de la Recherche ; INFORSID 2017*, 2017.

Serres A., Malingre M.-L., Mignon M., Pierre C., Collet D., Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2, Technical report, Université Rennes 2, 2017.