Des données particulières : les données de la recherche en Sciences Humaines et Sociales

Tiphaine van de Weghe, Marie-Noelle Bessagnet, Philippe Roose

Des données particulières : les données de la recherche en Sciences Humaines et Sociales

Tiphaine VAN DE WEGHE
LIUPPA, Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour, et, ITEM, Identités, Territoires, Expressions et Mobilités
PAU, FRANCE
t.van-de-weghe@univ-pau.fr

Marie-Noëlle BESSAGNET
LIUPPA, Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour
PAU, FRANCE
marie-noelle.bessagnet@univ-pau.fr

Philippe ROOSE
LIUPPA, Laboratoire Informatique de l'Université de Pau et des Pays de l'Adour
PAU, FRANCE
Philippe.Roose@iutbayonne.univ-pau.fr

ABSTRACT

Les données de la recherche en Sciences Humaines et Sociales (SHS) sont au cœur de tous travaux des chercheurs. Gérer ces données tout au long du cycle de vie nécessite un travail commun entre chercheurs en SHS produisant et exploitant les données qui ont pour leur part été structurées par les chercheurs en Informatique. Afin de répondre aux problématiques posées par les traitements sur les données de la recherche en SHS depuis leur recueil jusqu'à leur valorisation, nous proposons dans cet article un cadre conceptuel et méthodologique, une chaîne de traitements ainsi que des outils de traitement. Nous aborderons deux types de données de la recherche : des documents textuels, le plus souvent anciens, que les chercheurs en SHS doivent retranscrire en premier lieu avant tout traitement informatique et/ou statistique et des données plus hétérogènes issues du Patrimoine Culturel Immatériel (PCI).

KEYWORDS

Données de la recherche, sémantique, statistique, TALN

ACM Reference Format:
Tiphaine VAN DE WEGHE, Marie-Noëlle BESSAGNET, and Philippe ROOSE. 2018. Des données particulières : les données de la recherche en Sciences Humaines et Sociales. . , 12 pages.

1 INTRODUCTION

Les Sciences Humaines et Sociales (SHS) sont confrontées aux enjeux de l'évolution des technologies. De plus, elles ont un besoin et une nécessité d'utiliser l'informatique, ne serait-ce que pour les projets de recherche qui ont besoin d'être valorisés. Les données de la recherche en SHS sont relativement particulières de par leur complexité et leur diversité. En effet, ces données peuvent, non seulement se trouver sur Internet, être collectées sur le terrain, ou encore être sur papier. Selon la discipline, les données peuvent être des textes, des images (schémas, photos), des vidéos ou encore des observations de l'être humain, de ses pratiques. Ces données, comme le détermine l'Université de Humboldt [13], sont des données sources. Par la suite, elles vont être analysées de différentes manières et donneront des données résultats. Aujourd'hui, les chercheurs en SHS effectuent des traitements manuels sur ces données, parfois répétitifs, qui prennent du temps, notre but est d'automatiser certaines tâches. Dans la gestion de ces données, nous sommes confrontés à plusieurs défis tout au long de leur cycle de vie, de l'énonciation de la problématique du chercheur en SHS jusqu'à leur valorisation. Tout au long de ce cycle de vie, des méthodes et outils doivent être mis en place. Nous tentons, quelque soit le domaine SHS traité, de concevoir un environnement générique couvrant l'ensemble des étapes du cycle de vie, de l'étude des besoins du chercheur en SHS jusqu'à la valorisation des données. Dans cet article, nous établirons, dans une deuxième partie, un constat sur les données de la recherche en SHS et leur cycle de vie. Nous présenterons également la solution nationale Huma-Num. Dans une troisième partie, nous présenterons notre cadre méthodologique et conceptuel en abordant la chaîne de traitement mise en place nous permettant d'appliquer des approches statistique et sémantique. Nous illustrerons nos propos par des expérimentations sur des corpus dans une quatrième partie. Dans une cinquième partie, nous discuterons de données particulières issues du Patrimoine Culturel Immatériel demandant un autre type de traitement. Enfin, nous conclurons et montrerons les perspectives.

2 DES CONSTATS

Bonvallot [3] décrit la pratique des chercheurs en SHS où cette synthèse détermine deux types de recherche : la notion de littérature primaire, document à étudier et à analyser (par exemple, un livre) et la notion de littérature secondaire où les documents primaires ont été analysés et résumés (par exemple, un article de revue). Ces documents sont principalement des textes. Lorsqu'un chercheur en SHS étudie ses documents, en général, il lit et interprète les corpus. Ensuite, il récupère les informations pour les traiter. Afin de pouvoir y faire quelques calculs et les représenter graphiquement le chercheur les intègre manuellement dans une table à une dimension. Dans cette table apparaissent les données liées à trois dimensions : spatiale, temporelle, et thématique. Dans un but d'automatisation de certaines tâches, les liens entre les SHS et l'informatique se structurent depuis quelques années. En effet, les SHS ont des ressources et des besoins que

BDA 2018, Oct. 2018, Bucarest, ROUMANIE

l'informatique peut aider à analyser et à exploiter. Cette collaboration permet de faciliter les traitements que le chercheur en SHS doit opérer sur ses données ainsi que la visualisation automatique des données traitées à des fins d'analyses. [19] explique que de grandes infrastructures ont été fondées afin d'offrir un service de gestion des ressources mais aussi des données pour les humanités numériques. Les auteurs de [13] ont évoqué que les données de la recherche sont diverses, il existe celles : d'observation, d'expérimentation, de simulation, dérivées, de référence . Ils mentionnent également le point sur les données sources et résultats, plus précisément les données non traitées et respectivement les données traitées. Quelle définition peut-on retenir pour les données de la recherche ? Quelles seront les problématiques posées par une exploitation automatique des données de la recherche tout au long de leur cycle de vie?

2.1 Données de la recherche et cycle de vie

Plusieurs définitions des données de la recherche peuvent être trouvées dans la littérature. Nous avons retenu celles de l'OCDE [1] : les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons - auxquelles on ajoute les vidéos), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche. ou encore La valeur des données réside dans leur exploitation. L'accès total et ouvert aux données scientifiques devrait devenir la norme internationale pour l'échange des données scientifiques issues de la recherche financée sur fonds publics . Ces deux définitions insistent sur la perspective d'ouverture et d'accès aux données de la recherche, financées sur fonds publics. Aussi, des chercheurs se sont penchés sur la mise en oeuvre d'un plan de gestion des données (DMP) [14] permettant de traiter des documents littéraires des chercheurs en SHS, définis comme des ressources ou encore des données sources. Un cycle de vie des données de la recherche a été défini et commenté notamment par the UK Data Service[1] (Fig. 1). Il s'agit de planifier la recherche, déterminer les types de ressources, pour ensuite réaliser un modèle qui permettra de collecter des données, puis de les traiter, de les publier, de les préserver mais également de les rendre réutilisables. Cet cycle convient aux données pour se préparer aux éventuels changements : éléments nouveaux, à modifier, à supprimer (possibilités d'erreurs), etc. Lors d'enquêtes [13], on peut lire que Les données de la recherche deviennent l'un des nouveaux défis de la gestion scientifique .

Deux études menées par les universités de Lille 3 [16] et de Rennes 2 [17] ont déterminé que les principales ressources des SHS sont des textes. Par ailleurs, ces études révèlent que ces textes ne sont pas ou peu numérisés (environ 10% des ressources). Notamment, l'un des problèmes soulevés concerne la retranscription manuelle de ces ressources.

[1] https://www.ukdataservice.ac.uk/manage-data/lifecycle

2.2 Une solution nationale : HUMA-NUM

Le concept d'humanité numérique prend de l'ampleur. [19] définit les humanités numériques comme "un cadre méthodologique et technologique qui opère sur des sources de données SHS et permet :

- la création, la numérisation et la structuration de toutes les sources de la connaissance;
- l'exploration, l'analyse et l'interprétation des informations numériques;
- la diffusion, le partage et la capitalisation des connaissances."

Nous retrouvons les éléments clés du cycle de vie des données, où d'un point de vue technique des infrastructures et des projets fédérés offrent des solutions.

Huma-Num[2] propose des outils (Nakala, Nakalona, ShareDocs, etc.) pour la gestion des données, afin que les chercheurs puissent structurer, stocker, partager et valoriser leurs ressources, tout en pensant à ce que ces données soient pérennisées. Ces services qu'offre Huma-Num (Fig. 2) sont dédiés uniquement aux chercheurs en SHS. La plupart des données sont structurées en accord avec la norme Dublin Core[3]. Le Dublin Core, comme le définit la BnF[4], "est un format descriptif à la fois simple et générique, comprenant 15 éléments différents, qui a été créé en 1995 à Dublin (Ohio) par OCLC [5] et le NCSA [6]." Les données ont vocation à être exportées et doivent être interopérables. La BnF [8] définit l'interopérabilité[7] comme "le fait de mettre en relation des données qui sont contenues à l'intérieur de bases de données distinctes, de les décloisonner pour offrir un espace commun de navigation et de recherche". De grandes institutions, comme l'Unesco, la BnF utilisent la norme Dublin Core, afin de faciliter cette interopérabilité.

Cette description sur Huma-Num montre l'intérêt de cette infrastructure pour les chercheurs en SHS. Toutefois, pour le moment, elle ne répond pas à toutes les problématiques des chercheurs. Par exemple, l'outil NAKALA[8] mis en place par la TGIR[9] Huma-Num facilite l'accès et le partage des données avec un dispositif destiné aux dépôts en grand nombre. Le chercheur procède à l'insertion média par média. Ce travail est relativement long. De plus, les enregistrer un par un, demande un travail minutieux pour éviter qu'il y ait une marge d'erreur. En effet, il n'offre pas la possibilité d'insérer un groupe de média possédant des métadonnées pour ce groupe. De plus, lorsque le chercheur est en possession de corpus numérique, il ne sait pas forcément utiliser les outils adéquats pour traiter les données, même si Huma-Num propose ces outils (par exemple ArcGis[10], Sphinx[11], outils de stockage, etc.). Les humanités numériques ont un avantage d'allier

[2] https://www.huma-num.fr/services-et-outils
[3] http://www.dublincore.org/documents/dces/
[4] Bibliothèque Nationale de France
[5] Online Computer Library Center
[6] National Center for Supercomputing Applications
[7] http://www.bnf.fr/fr/professionnels/anx_pro_videos/a.video_cnfpt_interoperabilite.html
[8] https://www.nakala.fr/
[9] Très Grande Infrastructure de Recherche
[10] https://www.arcgis.com/features/index.html
[11] http://www.lesphinx-developpement.fr/

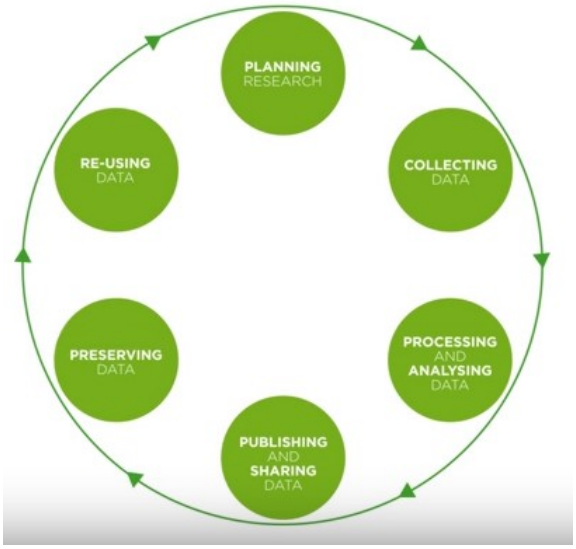

Figure 1: Cycle de vie des données

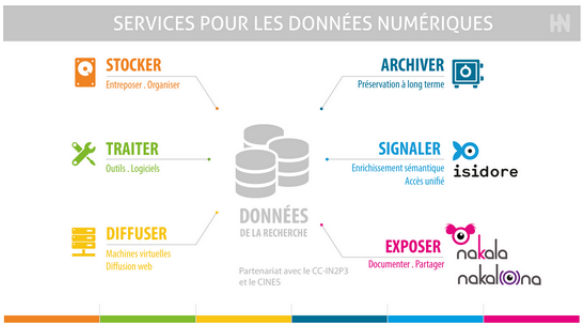

Figure 2: Services proposés par Huma-Num

les sciences exactes et les SHS, toutefois, il est primordial que les chercheurs en SHS s'intéressent au numérique [9]. Par ailleurs, la collaboration ouverte et valorisée des communautés peut apporter beaucoup de bénéfices intellectuels et financiers. Cependant, en général, les chercheurs en SHS ne peuvent seuls s'approprier ces environnements. Ils sont souvent démunis. Une réelle coopération entre chercheurs en SHS et chercheurs en informatique est nécessaire. Ainsi, la production de données par les chercheurs en SHS demande aujourd'hui qu'on s'y intéresse notablement. En effet, de nombreux problèmes autour de la gestion de ces données publiques se posent avec notamment l'avènement de politiques de type Open Data (les données doivent être libres, accessibles et réutilisables). Les chercheurs en SHS, contrairement aux chercheurs en informatique, ne possèdent pas encore cette culture de la sauvegarde numérique facilitant le partage, la communication, et l'accès aux données. Nous allons présenter notre approche pour tenter de répondre aux défis de la gestion des données de la recherche en SHS.

3 NOTRE CADRE MÉTHODOLOGIQUE ET CONCEPTUEL

Dans cette partie, nous traiterons de la nature des données en SHS pour montrer leur diversité ; nous définirons notre cycle de vie puis présenterons nos approches statistique et sémantique à appliquer sur ces données.

3.1 Nature des données

En accord avec Thierry Fournier dans Arabesque n°73[12], la nature et le périmètre des données de la recherche, sont très dépendants du cadre disciplinaire dans lequel s'effectue la recherche. En effet, les données en sociologie (enquêtes, séries statistiques etc.) sont très différentes des données en archéologie (rapports de fouilles etc.) ou encore en linguistique (corpus de textes etc.) ou encore en histoire, géographie, anthropologie sur les données du patrimoine, que ce soit le Patrimoine Bâti et Paysager ou le Patrimoine Culturel Immatériel (PCI). La nature des données dépend également de leur état : les données sont hétérogènes. Elles peuvent être textuelles, sonores, des images, des vidéos, etc. Cependant, les données de la recherche ont des points communs :

- elles sont numériques ou des moyens sont utilisés pour les rendre numériques. C'est une condition nécessaire pour envisager des traitements informatiques à opérer sur ces données et également pour leur diffusion
- elles sont massives ou au moins de plus en plus massives. Ceci implique des coûts de traitement et de stockage non négligeables.

[12] Dernier accès le 2/5/2018 http://www.abes.fr/Publications-Evenements/Arabesques/Arabesques-n-73

C'est la raison pour laquelle, nous devons réfléchir à un cadre conceptuel et méthodologique qui permettra de mettre en place des traitements informatiques et/ou statistiques similaires pour ces données, notamment textuelles. Nous travaillons sur l'extraction d'information selon trois dimensions : thématique, temporelle et spatiale [11]. Les patrons créés pour extraire de l'information temporelle et spatiale seront sûrement génériques sur nos corpus. Par contre, concernant la thématique, des ressources spécifiques devront être créées, tels que des thésaurus, des ontologies, des gazetiers. Dans ce cas, les traitements permettant de créer ces types de ressources devront être génériques.

3.2 Cycle de vie

Nous avons défini (Fig. 3) un ensemble de phases que l'on pourrait qualifier de cycle de vie sur les données de la recherche montrant la complexité des services de gestion des données à mettre en oeuvre tout au long de cette chaîne. Comme pour les données en entreprise [4], la gestion des données de la recherche pose de nombreux défis : le recueil ou la capture des données, le stockage, la production de nouvelles données, la structuration des données, l'extraction, l'intégration, l'analyse, la restitution mais également la valorisation de ces données. Cette complexité fait appel à de nombreux domaines de l'informatique tels que les bases de données, la recherche d'information, le Traitement automatique du langage naturel (TALN), la fouille de données, le web sémantique, les systèmes d'information géographiques mais également le domaine de la statistique.

Ce cycle de vie définit les différentes phases de transformation des données. Notre socle repose sur des données sources (données non traitées) recensées par le chercheur. Ce dernier nous communique ses interrogations, ses axes et tendances supposés. Puis, nous analysons, étudions la faisabilité de ces demandes. Pour résoudre les questions de recherche du chercheur en SHS, nous proposons en général un modèle structuré de présentation de ces données sources. Les résultats de ce scénario seront tout autant passés à la loupe du chercheur en SHS. À partir de ce point, les données subissent une première transformation. Cette base entérinée, validée, nous pourrons démarrer les premiers traitements. Nous appliquons simultanément des procédures d'analyses statistiques, informatiques, cartographiques, etc. Au final, nous présentons les résultats par des graphiques, des cartes, des frises temporelles, ou bien, sur tous supports facilitant la communication et la compréhension des nouvelles orientations. La synthèse des travaux de recherche en SHS appartient, évidemment, aux chercheurs en SHS, mais ils pourront s'appuyer sur nos valeurs de traitement pour argumenter leurs conclusions.

3.3 Approches sémantique et statistique

Les données de la recherche se présentent le plus souvent sous la forme de documents textuels non structurés. Sur ces données, nous appliquons une approche générique de traitements sémantique et statistique afin de répondre aux problématiques de recherche des chercheurs en SHS. Dans un premier temps, nous expliquerons ce qu'apporte ces deux approches distinctement, ensuite, dans un deuxième temps, nous verrons, l'avantage d'utiliser ces deux approches conjointement.

Côté sémantique, nous travaillons sur trois dimensions dans les documents : spatiale, temporelle et thématique [11]. Nous nous intéressons donc à l'extraction d'Entités Spatiales (ES), d'Entités Temporelles (ET) et d'Entités Thématiques (ETh) dans les documents. Les ES reposent sur le concept d'entité spatiale absolue caractérisant les informations propres à un lieu nommé (par exemple, la ville de Lescar) et le concept d'entité spatiale relative caractérisant des indications spatiales associées aux localisations (par exemple, près de Pau). Les ETh à annoter étant liées, dans notre cas, au domaine d'études, nous nous appuyons sur des ressources propres. Nous visons à terme à proposer une approche générique en donnant la possibilité d'intégrer aisément une nouvelle ressource sémantique de domaine. En ce qui concerne les ET, nous cherchons à marquer des entités calendaires, qui parfois demandent des traitements supplémentaires (cas du français ancien).

Côté statistique, le prétraitement est indispensable sur les textes, pour ne pas biaiser les calculs, comme pour un jeu de données, où les valeurs aberrantes faussent les résultats. A la suite du nettoyage et du prétraitement, il est possible de réaliser des traitements statistiques (Loi de Zipf, fréquence des termes). Plus précisément, il est possible de faire des analyses descriptives, c'est-à-dire, calculer par exemple la fréquence des mots, la répartition des mots dans différents textes, la corrélation des mots entre eux. Selon ces résultats, d'autres statistiques peuvent être pratiquées, comme l'analyse de données (ACP[13], AFC(M)[14], classification, etc.), ou encore, LDA[15], qui permet de faire des groupes de thèmes (topic model), des règles d'association, des classifications naïves Bayésiennes. Après les différentes analyses statistiques, nous pouvons représenter les données résultats par le biais de graphiques, schémas etc., [6].

Ainsi, nous nous sommes fixés pour objectif d'analyser du point de vue spatial, temporel, thématique et statistique l'ensemble des corpus sur lesquels nous travaillons. Aussi, nous décrivons notre approche par la Fig. 4. Indépendamment de tout corpus de textes, la chaîne de traitement sémantique, lors d'une première étape identifie et annote les données spatiales, temporelles et thématiques. Une deuxième étape concerne l'indexation de ces données dans un moteur de recherche afin de les exploiter dans des stratégies d'analyse et de recherche d'information combinant des critères spatiaux, temporels et thématiques. Enfin, une troisième étape définit des processus génériques d'analyse et de présentation de données, applicables à des corpus de textes. Nous utilisons l'environnement GATE [16] pour effectuer nombre de ces traitements sémantiques, notamment les traitements linguistiques. L'environnement GATE nous permet de définir

[13] Analyse des Composantes Principales
[14] Analyse Factorielle des Correspondances (Multiples)
[15] Latent Dirichlet Allocation
[16] https://gate.ac.uk/family/developer.html

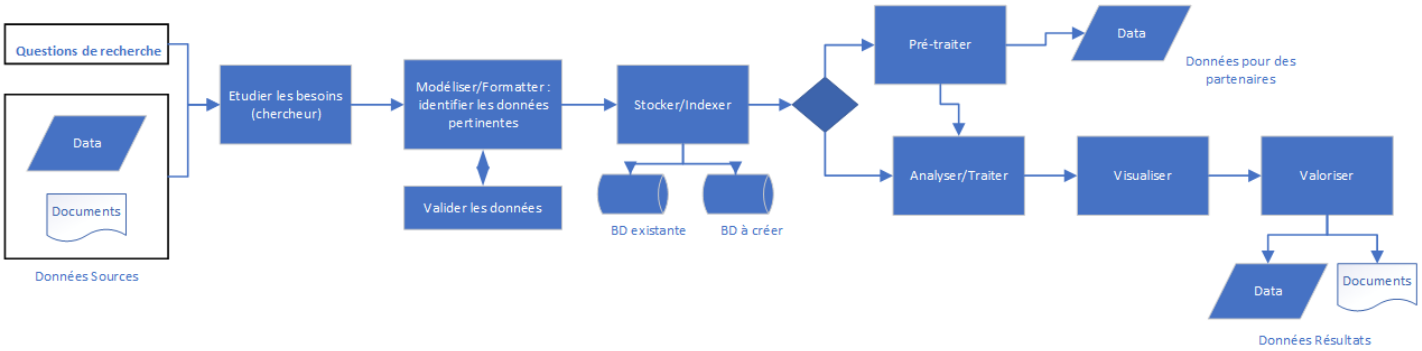

Figure 3: Cycle de vie

un pipeline générique où les modules de traitement seront généralement identiques quelques soient les textes, les adaptations portant sur les patrons à définir pour annoter les entités thématiques ainsi que sur la complétude des ressources sémantiques utilisées (gazetiers, ontologies, etc.). Concernant l'analyse statistique, il existe bon nombre d'outils pour traiter les données. En effet, nous avons la possibilité de travailler avec les langages Python[17], R[18], ou encore d'utiliser des logiciels propriétaires, comme SAS[19]. Nous avons choisi le langage R qui offre une multitude de bibliothèques pour analyser statistiquement les données et qui nous permet d'être le plus générique possible sur les données. A la suite des traitements opérés, il est primordial de représenter les données traitées selon des formes et schémas acceptés par les chercheurs en SHS : frise temporelle, carte, graphe de collaboration, graphiques divers et variés qui permettront aux chercheurs de faire une analyse plus approfondie de leurs données de la recherche.

L'étude sémantique permet d'extraire des informations pertinentes, comme le spatio-temporel, les thématiques. Ce qui donne l'accès à un document structuré qui nous permet de réaliser de nouvelles analyses statistiques. Les deux approches nous fournissent des éléments pertinents pour répondre aux questions des chercheurs, toutefois, allier celles-ci, nous permet d'ajouter de nouveaux résultats intéressants pour le chercheur. Dans la section suivante, nous observerons plus en détail les problématiques, ainsi que la complémentarité des deux approches.

[17] https://www.python.org/
[18] https://cran.r-project.org/
[19] https://www.sas.com/fr_fr/home.html

4 DES ANALYSES ET EXPÉRIMENTATIONS MENÉES SUR NOS CORPUS

Dans cette partie, nous croiserons les analyses et expérimentations menées sur nos corpus, en les définissant avec les questions posées, puis, en exposant les résultats des traitements statistiques et sémantiques. A la suite de ce travail, nous démontrerons la convergence des deux analyses pour aider au mieux le chercheur en SHS.

4.1 Les corpus et les questions posées

Nous avons travaillé avec des chercheurs en SHS sur des documents ayant trait aux domaines suivants :

(1) aux échanges entre artistes des avant-gardes qui ont joué un rôle considérable dans la littérature et le monde des arts au XXe siècle. Nous avons un corpus de 17 textes, la plupart étant des lettres, écrites par un artiste roumain du XXe siècle Gherasim Lucas à son ami Victor Brauner.
(2) des actes royaux du XVIe siècle. Nous avons un corpus de 33 textes actuellement retranscrits, sur 7000 actes à recenser (en vieux français, en espagnol),
(3) des lettres de correspondance en espagnol datant des XIXeme et XXeme siecles. Nous avons un corpus de 450 textes.

Tous ces documents ont été retranscrits numériquement par les chercheurs dans des corpus qui nous ont été fournis. Sur chacun de ces corpus, les chercheurs en SHS se posent des questions différentes mais liées aux trois dimensions : spatiale, temporelle et thématique.

Sur le cas d'étude numéro 1 (Correspondance Avant-Garde Roumaine), les questions posées sur les 3 dimensions concernent par exemple :

- Spatiale : Où la lettre a-t-elle été écrite ? Quelles sont les villes mentionnées dans les lettres
- Temporelle : A quelle date la lettre a-t-elle été écrite ? Quelles sont les entités temporelles décrites au sein de la lettre?

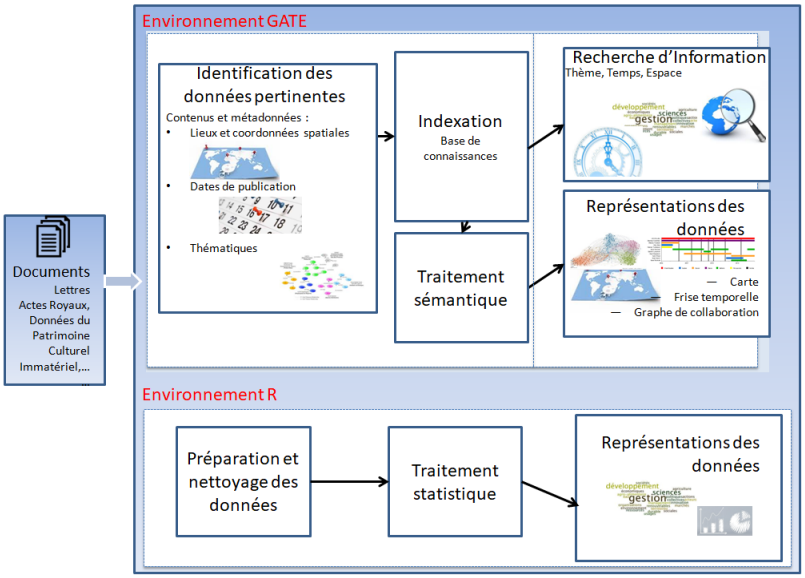

Figure 4: Chaînes de traitement génériques pour l'analyse sémantique et statistique de corpus textuels

- Thématique : Qui est cité dans la lettre ? Les personnes citées ont-elles des liens artistiques ? Une oeuvre artistique est-elle mentionnée ?

Sur le cas d'étude numéro 2 (acte royal), les questions posées sur les 3 dimensions concernent par exemple :

- Spatiale : Où ce texte a-t-il été écrit ?
- Temporelle : Quand ce texte a-t-il été écrit ?
- Thématique : Qui est cité dans ce document ? Les personnes citées ont-elles des liens familiaux ? Une transaction financière est-elle mentionnée ?

Sur le cas d'étude numéro 3 (correspondance en espagnol), les questions posées sur les 3 dimensions concernent par exemple :

- Spatiale : quels sont les lieux spécifiés dans les lettres (des villes, des pays ou des continents, mais aussi des lieux plus précis comme des manoirs, des quartiers ou des places)?
- Temporelle : quels sont les dates spécifiées dans les lettres (dates d'écriture des lettres le plus souvent)?
- Thématique : on s'intéresse aux *Personnes* spécifiées pour prendre connaissance de l'expéditeur et du destinataire par exemple ; au thème de la *Famille*, au thème du *Voyage* et enfin au thème de la *Littérature*.

Nous pouvons remarquer que pour cette approche, les questions se ressemblent quelque soit les cas d'étude. Ainsi, les chaînes de traitement (Fig.4) mises en place peuvent traiter tous ces cas de figure. Concernant l'approche sémantique, l'attention est à porter sur le vocabulaire qui varie d'un corpus à un autre, également sur la langue qui peut être différente, qui nécessitent, notamment, l'élaboration de gazetiers particuliers. Concernant l'approche statistique, la chaîne de traitement est générique.

En terme de statistiques sur les textes, les questions sont généralement les mêmes, où la fréquence et la pondération des mots jouent un rôle :

- Quels sont les mots qui ressortent le plus dans le texte ?
- Quelles sont les corrélations entre les mots les plus redondants et les autres ?
- Quelles sont les fréquences d'échanges entre les correspondants ?
- Quels sont les moments où il y a eu le plus d'échanges ?

En général, le chercheur en SHS est intéressé par les analyses descriptives : fréquence, moyenne, table de contingence. Il veut savoir quelle est la fréquence de ses données / des mots dans le texte, la présence et la corrélation des mots sélectionnés par le chercheur. En statistique, nous pouvons ajouter d'autres interrogations, comme :

- Quels sont les flux géographiques entre les correspondants ?
- Quels sont les échanges entre les correspondants ?
- Quelles sont les personnes qui écrivent le plus ?

Par ailleurs, afin de répondre à ces types de questions, nous devons procéder à un travail sémantique pour structurer les textes, pour pouvoir les analyser correctement.

4.2 Les traitements statistiques

Nous avons mené une analyse statistique sur les 17 lettres de correspondance i.e. le cas d'étude numéro 1 (Correspondance Avant-Garde Roumaine). En effet, il a été remarqué une hétérogénéité des lettres de faible volume. Ce qui peut entraîner des résultats biaisés et non significatifs. En statistique, nous prenons en compte chaque mot. Afin qu'ils soient tous traités de la même manière, nous devons passer par l'étape

de préparation et de nettoyage des données. Les lettres en majuscules ont été réduites en minuscules, les ponctuations ainsi que les chiffres, et les mots non significatifs (préposition par exemple) ont été supprimés. À la suite de ce nettoyage, la préparation des données consiste à rajouter des thèmes par lettres s'il en existe (c'est notre cas) et à transformer les textes en matrice document-terme qui est le croisement entre les mots et les documents. Nous avons le choix entre la présence ou non (choix binaire) du mot dans chaque document, ou encore la fréquence de chaque mot dans chaque document, etc. A partir de cette matrice, il est possible de réaliser des analyses statistiques. Tout d'abord, une analyse descriptive permet de visualiser les distributions des lettres, mais aussi les fréquences des mots par le biais de diagramme en barres, respectivement de nuage de mots (Fig. 5). Il a été possible de remarquer les corrélations entre des mots sélectionnés (ici, les plus fréquents) et d'autres qui possèdent un lien entre eux.

Figure 5: Nuage de mots

D'autres analyses plus poussées ont été réalisées comme une classification, LDA [20]. Ces études permettent de classer les mots puis, de les regrouper afin d'interpréter par le biais de ces groupes, des thèmes.

Concernant le cas d'étude numéro 3 (correspondance en espagnol), une étude préalable a été réalisée sur 107 textes. Ces 107 textes ont été dissociés dans un jeu de données, afin de distinguer les dates, les lieux, les expéditeurs, les destinataires. Nous sommes dans une situation de classification non supervisée. En effet, il n'y a rien à prédire, l'étude porte sur le passé à analyser. Tout d'abord, une analyse brute a été faite, afin de garder les mots les plus cohérents. Nous avons pu réaliser diverses statistiques comme une classification de mots. La classification ascendante hiérarchique met en évidence des groupes. Ces groupes sont formés selon leur ressemblance. La classification a été réalisée dans le but de constater les éventuelles thématiques, mais aussi de déterminer les lieux et l'aspect temporel. Des lieux (maison, Coruna, etc.), la temporalité (aujourd'hui, matin, etc.), des thématiques, comme la famille, la littérature, les voyages apparaissent dans le dendrogramme (Fig. 6). Il va être recherché une répartition des mots dans des classes. Le dendrogramme (Fig. 6), est découpé en 5 groupes. Il montre bien que les groupes sont hétérogènes. Pour établir ses conclusions, le chercheur s'appuiera sur ce travail afin de montrer les relations entre les mots. Il faut noter que amiga et bien sont démarqués, et se retrouvent seuls dans leur groupe. Alors que le groupe 5 est composé du maximum de mots. Actuellement, nous travaillons sur un

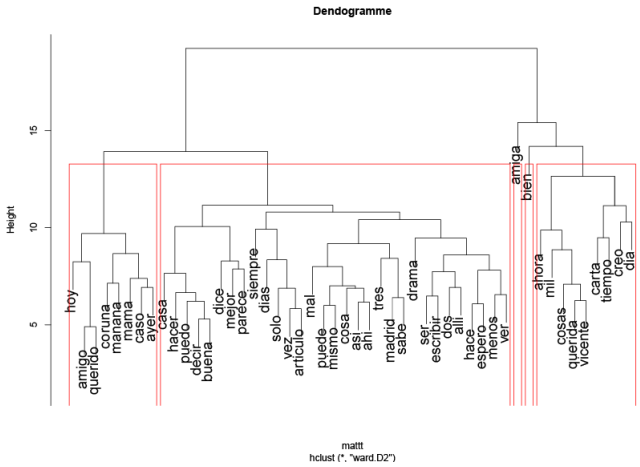

Figure 6: Classification des mots

corpus de 450 lettres. Concernant le cas d'étude numéro 2 (les actes royaux), 1000 actes (Fig. 7) sont en cours de retranscription par les chercheurs en SHS. Les chercheurs pourraient utiliser des plate-formes de transcription collaborative (ou crowdsourcing) des textes, accélérant ainsi le traitement.

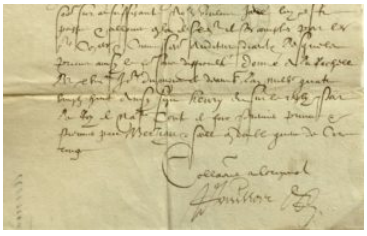

Figure 7: Acte royal

Ces corpus possèdent une structure où il sera possible de distinguer automatiquement un(e) lettre/acte d'un(e) autre ainsi que les éléments des lettres/actes différents du texte, comme la date, le lieu, l'expéditeur, le destinataire, la formule de politesse, les notes de bas de page. Par ailleurs, il faudra également préciser la langue pour l'ensemble des textes qui peut être différente, car le traitement ne sera pas le mêmes.

[20] Latent Dirichlet Allocation

4.3 Les traitements sémantiques

Sur l'ensemble des corpus, une chaîne de traitement a été mise en place dans l'environnement GATE. Cette chaîne contient un ensemble de modules (Fig.8), génériques pour la plupart, permettant d'annoter l'ensemble des entités nommées qu'elles soient spatiale, temporelle ou thématique. La chaîne de traitement a donc été mise en oeuvre sur la plateforme GATE [5] [2]. Elle intègre notamment l'analyseur morphosyntaxique Treetagger [15] et prend en charge la lemmatisation en langue française et en langue espagnole.

La reconnaissance d'entités nommées spatiales, temporelles et thématiques nécessitent parfois des traitements spécifiques : l'ajout de termes dans les gazetiers existants (noms de villes par exemple), la création de gazetiers particuliers (noms d'artistes, mots anciens, etc.) ainsi que l'écriture de patrons spécifiques. Prenons comme exemple la reconnaissance d'entités temporelles dans les actes royaux (Fig.9) pour lesquelles des patrons ont été écrits pour annoter les dates dites relatives (ici relatives à la date d'écriture de l'acte royal) et absolue (un jour précis): ceci est décrit dans la figure 9.

La figure 10 montre des exemples d'annotation mis en place sous GATE.

Après la phase d'annotation, on peut utiliser ces dernières à des fins de valorisation dans des interfaces de visualisation (Fig. 11) créées pour faciliter l'accès aux résultats par les chercheurs en SHS. D'autres types de valorisation existent tels que des cartes ou encore des interfaces pour faciliter la recherche d'information parfois combinée.

4.4 La convergence des deux analyses pour aider le chercheur en SHS

Auparavant, les chercheurs en SHS calculaient manuellement le nombre de mots dans leurs corpus, [10]. Dorénavant, il est possible de le faire automatiquement à condition que les documents soient dans un format numérique [7]. Dans une étude générale de fouille de texte en statistique, l'ensemble du contenu des corpus est pris en compte (chaque mot/terme): nous ne faisons pas la différence sur le sens du mot.

Nous savons que les deux approches, informatique et statistique, étudiées distinctement permettent de représenter, visionner les données en rapport à leur thématique, à leur fréquence, mais aussi au niveau spatio-temporel, selon les mots/phrases.

Dans l'exemple des lettres de correspondance et actes, nous nous intéressons à l'expéditeur, au destinataire, à la formule de politesse, à la date, au lieu et au texte (variables).

En sémantique, nous les repérons pour obtenir un document semi-structuré (Fig.12) composé des variables énumérées. À partir de ce nouveau document, nous pouvons réaliser davantage de statistiques. Par conséquent, si nous allions ces deux approches, nous avons la possibilité de représenter les échanges entre les correspondants (Fig.13), d'observer l'évolution des échanges, etc.

Pour être plus précis, lorsque le chercheur étudie des lettres de correspondance, il cherche à connaître les différents échanges qu'il y a pu avoir, grâce à l'extraction de notions tels que le spatial, l'expéditeur, le destinataire, la date. Ensuite, une étude statistique est réalisée. Nous avons pu constater qu'Emilia Pardo Bazan a envoyé 94 lettres. Nous avons cherché à représenter ces correspondances par le biais d'un graphique directionnel (Fig. 13).

Ces deux approches, statistique et informatique, sont totalement complémentaires. Il est important de noter que la constitution du corpus revêt un intérêt important puisque toute modification de celui-ci est susceptible d'altérer les résultats statistiques et sémantiques sur lesquels s'appuiera l'interprétation du chercheur. L'utilisation de cette approche duale sur un même corpus est un avantage, car elle permet de comparer les résultats obtenus, d'affiner l'interprétation, et au final de répondre aux attentes des chercheurs en SHS qui peuvent s'avérer extrêmement variées.

Notre cadre méthodologique s'applique correctement sur les corpus présentés (des textes plus ou moins structurés). Concernant d'autres données telles que les données du Patrimoine Culturel Immatériel, une autre démarche doit être mise en place nécessitant d'autres approches et traitements que nous allons expliquer ci-après.

5 DES DONNÉES DE LA RECHERCHE À TRAITER DANS UN CADRE PLUS AMBITIEUX : LE PATRIMOINE CULTUREL IMMATÉRIEL (PCI)

Dans le cadre d'un projet FEDER[21] pluridisciplinaire TCVPYR[22] [23] réunissant des chercheurs en géographie, en histoire, en anthropologie, en informatique mais également des chercheurs des Inventaires régionaux, nous menons un travail qui a pour objectif la valorisation du patrimoine culturel pyrénéen. Nous évoquerons la gestion du patrimoine culturel, puis le patrimoine culturel immatériel qui nécessite une attention particulière.

5.1 La gestion du patrimoine culturel

La gestion du patrimoine culturel est une tâche complexe qui implique l'engagement de différents types d'acteurs (institutions publiques ou privées, associations, entreprises et individus) que ce soit au niveau local, national voire international. Le projet TCVPYR comporte trois grands axes :

- (i) l'inventaire des données par des chercheurs de domaines disciplinaires différents (qui vont sur le terrain à cette fin),
- (ii) l'intégration et la structuration de ces données hétérogènes et géoréférencées dans un système d'information commun
- et (iii) la valorisation de ces données auprès du grand public afin de contribuer à la promotion du thermalisme et de la villégiature notamment.

[21] Fonds Européen de Développement Régional
[22] http://tcvpyr.iutbayonne.univ-pau.fr/
[23] Inventaire du patrimoine bâti et du patrimoine culturel immatériel de la villégiature et du thermalisme dans le massif pyrénéen français

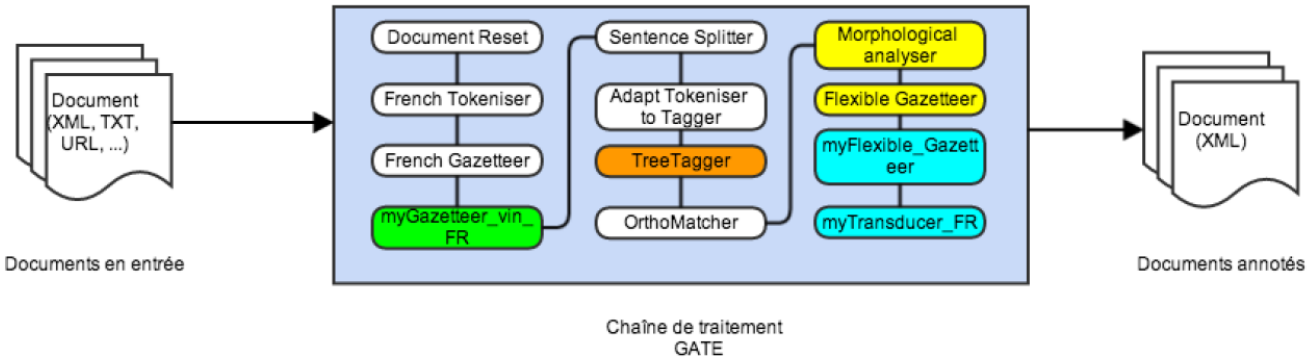

Figure 8: Chaîne de traitement GATE

[...] aux deux termes de Noël et de Saint-Jehan Baptiste [...] le XXe jour de juillet l'an mil cinq cens soixante-douze [...]

— Relative
— Absolue

Figure 9: Exemples d'entités temporelles

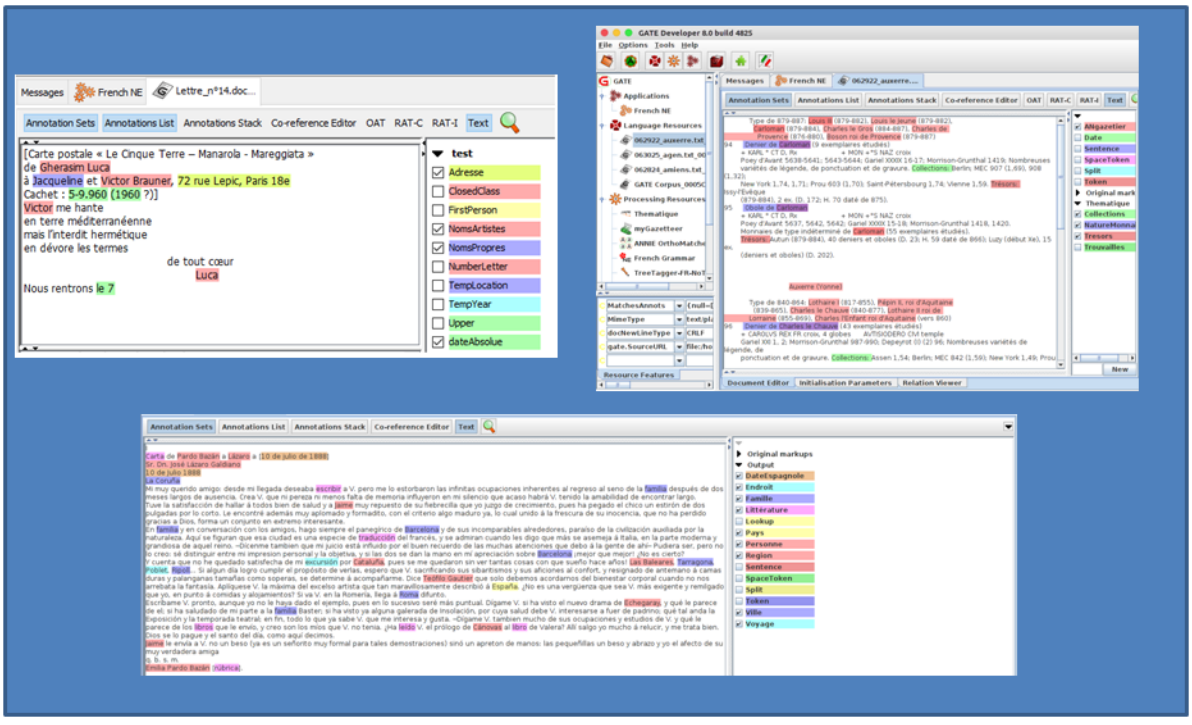

Figure 10: Exemples d'annotations sous GATE

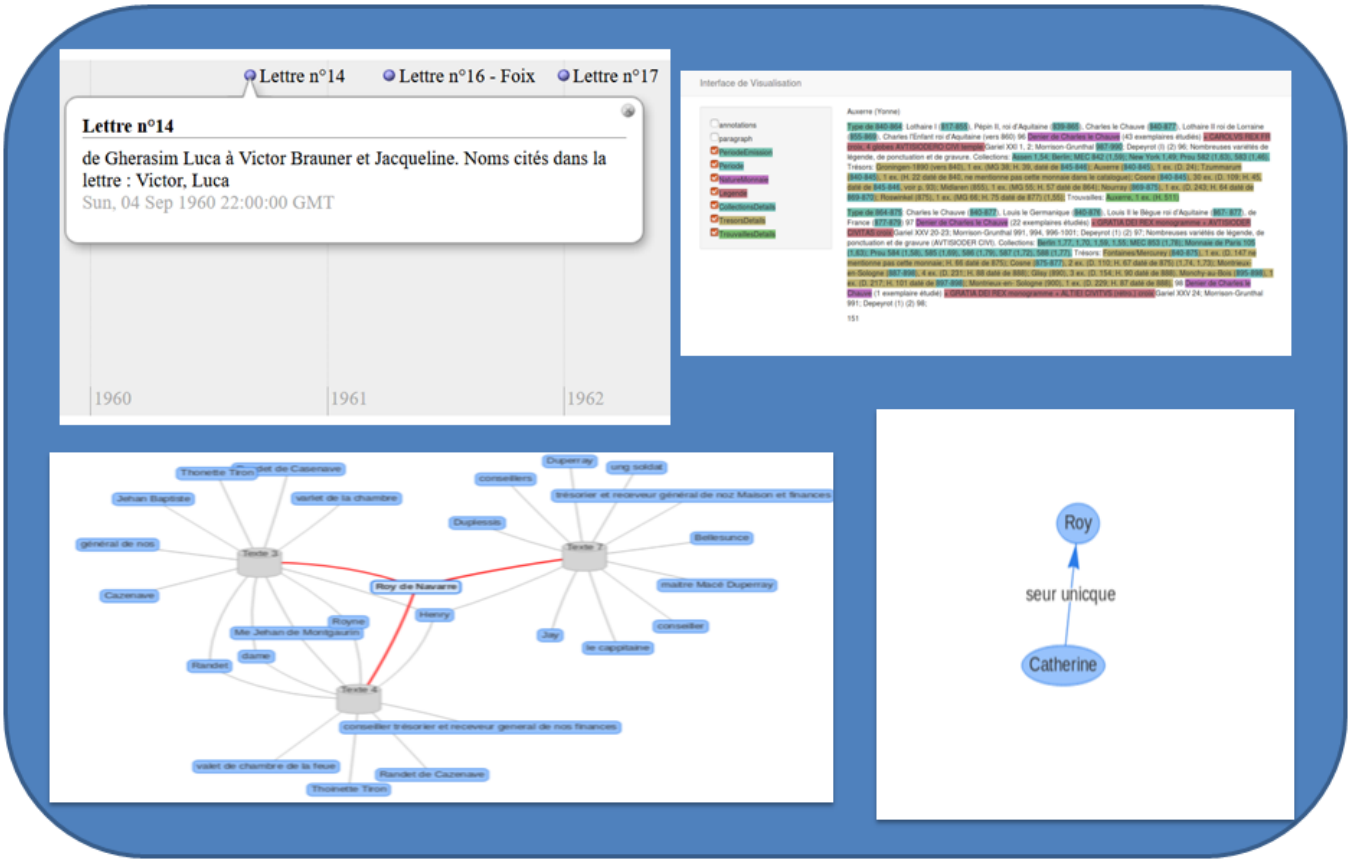

Figure 11: Frise temporelle et Interface Web de visualisation des résultats d'annotation, graphes des personnages et des relations familiales.

```
    {'title' : 'Lettre N°1',
'start' : '21 4 1946',
'description' : 'Lettre de Gherasim Lucas à Victor Brauner. Noms cités dans la lettre : Gherasim Luca, Nadine
```

Figure 12: Exemple de données extraites

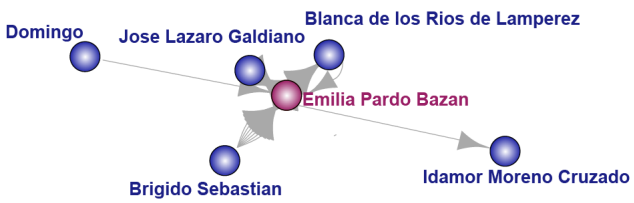

Figure 13: Echange entre les correspondants

Nous nous intéresserons dans cet article au premier axe (i). Cet axe concerne l'inventaire de données relatives à deux types de patrimoine : le patrimoine bâti et paysager (les bâtiments et le mobilier) et le patrimoine culturel immatériel (PCI), relatant une pratique qu'une communauté reconnaît comme partie de son patrimoine.

Les chercheurs en SHS vont sur le terrain pour collecter des ensembles d'information relative au patrimoine. Ces données sont ensuite stockées de façon hétérogène et sont difficilement accessibles par le grand public. Aussi, les chercheurs en SHS utilisent, pour le patrimoine bâti et paysager, deux applications Renabl[24] et Gertrude[25]. Ces applications sont utilisés pour Renabl dans la région Occitanie et pour Gertrude plutôt dans la région Nouvelle Aquitaine. Les formats de données utilisés sont différents. Quoiqu'il en soit, à partir de ces applications, l'ensemble des bases nationales Architecture et

[24] http://www2.culture.gouv.fr/culture/dp/inventaire/telechar/renabl/manuel_renabl.pdf
[25] http://www.inventaire.culture.gouv.fr/Chemin_annuaire1.htm

Patrimoine du Ministère de la Culture peuvent être renseignées. Ces bases nationales forment un ensemble cohérent, renseignant l'architecture (base Mérimée), les objets (base Palissy), les données iconographiques (base Mémoire). Ces données sont difficilement exploitables dans des applications de valorisation, raison pour laquelle nous mettons en œuvre dans le cadre du projet des méthodes et traitements permettant de mieux les structurer et de les valoriser.

5.2 Le patrimoine culturel immatériel (PCI)

Pour construire une connaissance partagée de ce domaine, il est primordial de définir précisément ce patrimoine, d'identifier l'ensemble des acteurs concernés possédant des données numériques liées au PCI, ainsi que leurs pratiques.

Donnons la définition du PCI par l'UNESCO[26] : "Le patrimoine culturel immatériel fait référence aux pratiques, représentations, expressions, connaissances et savoir-faire, transmis de génération en génération au sein d'une communauté, créés et transformés en permanence en fonction du milieu, de l'interaction avec la nature et de l'histoire". L'origine et l'approche du PCI sont mises en œuvre d'après de nombreux modèles d'analyse, dans le cadre d'observation des pratiques communautaires. Ensuite ces informations sont insérées dans une notice descriptive permettant d'isoler les facteurs organisationnels, comme le titre, la description, l'historique, le lieu, la date, etc. Lorsqu'il y a une enquête PCI, les chercheurs précisent si le patrimoine architectural ou mobilier est déjà répertorié dans leurs bases de données respectives (Mérimée pour l'architecture et Palissy pour le mobilier). Le chercheur réalise également des photographies, des vidéos, des enregistrements sonores, etc. annotés par des métadonnées qui sont spécifiées par le chercheur manuellement. Nous pouvons remarquer sur le schéma (Fig. 14) les données à prendre en compte dans le cadre d'une enquête PCI.

Concernant le PCI, il n'existe pas à notre connaissance une application commune permettant aux chercheurs en SHS de saisir, stocker et valoriser les données collectées. Aussi, un travail de plus long terme est à mener afin de définir d'une part une structure de données commune sous la forme d'une ontologie, et d'autre part de spécifier les diverses fonctionnalités requises dans une application pour qu'un chercheur en SHS puisse facilement saisir et valoriser les données collectées, sachant que ces données sont dans des formats hétérogènes : textes, images, vidéos, enregistrements sonores.

En 2012 les acteurs du PCI ont été recensés [27]. Les types d'acteurs sont principalement les associations, les chercheurs, ainsi que les institutions de recherche et de formation. Il est intéressant de cartographier le PCI à l'échelle nationale [12], afin de recenser les pratiques de chacun et de proposer un modèle commun, pour des données interopérables et réutilisables. L'ontologie résultante sera formalisée en XML

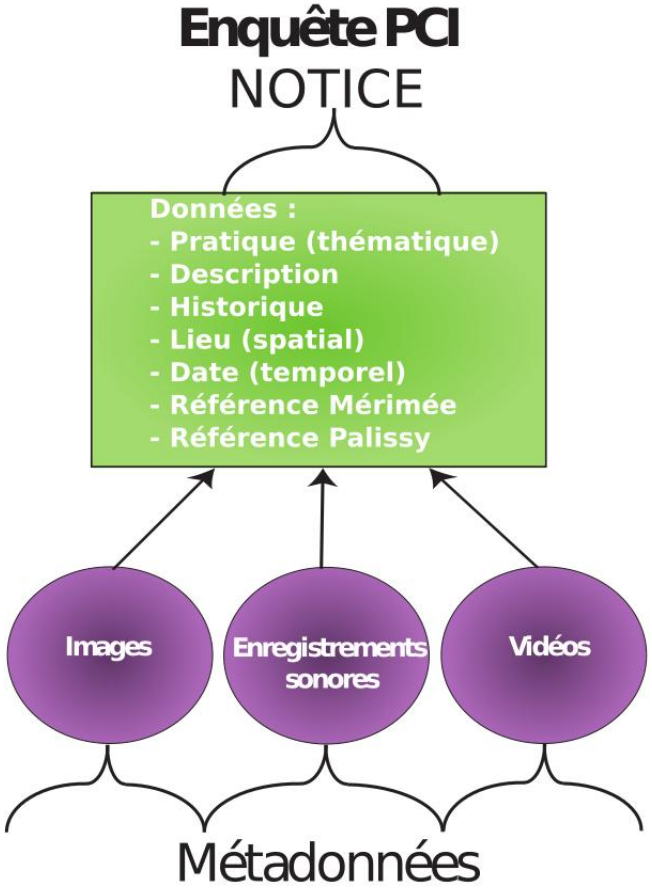

Figure 14: Déroulement de l'enquête PCI

CIDOC-CRM, modèle sémantique normé de référence pour la description du patrimoine [18]. Ce travail de structuration sera mené à partir également de deux documents fournis par le Ministère de la Culture [28] : un vade-mecum et une fiche-type de description d'éléments du patrimoine culturel immatériel en vue de leur inclusion à l'Inventaire national . Nous pourrons également exploiter le travail mené dans le cadre du projet PCI-Lab[29].

Afin que les chercheurs en SHS du laboratoire ITEM puissent démarrer leur travail de saisie des notices descriptives, nous avons élaboré sur la plate-forme Huma-Num un premier canevas de saisie en respectant d'une part la norme Dublin-Core et d'autre part la description des éléments du PCI émise par le Ministère de la Culture. Ces premiers traitements permettent de montrer aux chercheurs la nécessité de structurer les données, de les stocker avant de pouvoir les valoriser de manière simple en les exposant via Huma-Num ou

[26] Organisation des Nations unies pour l'éducation, la science et la culture
[27] http://www.culture.gouv.fr/Thematiques/Patrimoine-culturel-immateriel/Le-PCI-qu-est-ce-que-c-est/Les-acteurs-du-PCI
[28] http://www.culture.gouv.fr/Thematiques/Patrimoine-culturel-immateriel/L-inventaire-national/Fiche-type-et-vade-mecum
[29] https://www.pci-lab.fr/

de manière plus complexe via des applications de valorisation du territoire.

6 CONCLUSIONS ET PERSPECTIVES

Nous avons discuté dans cet article de la complexité de la gestion des données hétérogènes de la recherche en SHS. Cette gestion nécessite un travail commun entre chercheurs en SHS et chercheurs en informatique. Étudier la grande variété des besoins des chercheurs en SHS quant à la gestion de leurs données de la recherche afin d'y répondre demande la définition d'un cadre méthodologique et conceptuel que nous avons proposé et détaillé.

Les expérimentations que nous avons menées montrent :

- qu'il n'est pas simple de formaliser les besoins des chercheurs en SHS ;
- que la conception et le développement d'une chaîne de traitement générique sur des ensembles de données textuelles demandent encore un travail approfondi ;
- que la complémentarité des approches informatiques et statistiques permet de répondre aux attentes des chercheurs afin qu'ils puissent illustrer leurs conclusions ;
- qu'aider les chercheurs en SHS à valoriser leurs données de la recherche est primordial en proposant, par exemple, des modèles communs, des modalités de visualisation des données au moyen de statistiques, de représentations graphiques, de représentation calendaire ou spatiale, ou encore une recherche d'information combinée sur les dimensions spatiale, temporelle et thématique.

Diffuser ces données en Open data est également une phase essentielle quant au partage de ces données. Une réflexion est à mener avec les chercheurs en SHS sur la temporalité de diffusion des données en Open Data. En ce qui concerne le PCI, un premier travail a été effectué ciblant les types de données, les acteurs et les domaines concernés. Une enquête est en cours de réalisation pour analyser les méthodes et pratiques des acteurs du PCI, afin de créer un modèle unifié. Ce modèle ainsi que la réalisation d'une ontologie serviront à la création d'une application qui structurera les données mais aussi optimisera le travail des acteurs.

7 REMERCIEMENTS

Étude réalisée dans le cadre du programme de recherche européen TCV-PYR (2017-2020), financé par l'Union européenne (FEDER) en partenariat avec les régions Occitanie-Méditerranée et Nouvelle-Aquitaine.

REFERENCES

[1] [n. d.]. Convention de 2003 pour la sauvegarde. ([n. d.]), 46.
[2] Kalina Bontcheva, Valentin Tablan, Diana Maynard, and Hamish Cunningham. 2004. Evolving GATE to meet new challenges in language engineering. *Natural Language Engineering* 10, 3-4 (2004), 349–373.
[3] Valérie Bonvallot. 2007. *La pratique documentaire des chercheurs en SHS: la recherche d'information.* Report. Institut de l'Information Scientifique et Technique (INIST-CNRS).
[4] Christine Collet, Bernd Amann, Nicole Bidoit, Mohand Boughanem, Mokrane Bouzeghoub, Anne Doucet, David Gross-Amblard, Jean-Marc Petit, Mohand-Said Hacid, and Genoveva Vargas-Solar. 2013. De la gestion de bases de données à la gestion de grands espaces de données. *Revue des Sciences et Technologies de l'Information-Série ISI: Ingénierie des Systèmes d'Information* 18, 4 (2013), 11–31.
[5] Hamish Cunningham. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities* 36, 2 (2002), 223–254.
[6] Nicolas Dugué, Jean Charles Lamirel, and Pascal Cuxac. 2016. Visualisation pour la détection d'évolutions dans des corpus de publications scientifiques. *Les Cahiers du numérique* 12, 4 (Dec. 2016), 157–184. https://www.cairn.info/revue-les-cahiers-du-numerique-2016-4-p-157.htm
[7] Christian Fauré. [n. d.]. Introduction au Text-mining. ([n. d.]). http://www.christian-faure.net/2007/05/30/introduction-au-text-mining/
[8] Bibliothèque nationale de France. [n. d.]. BnF - L'interopérabilité : définition et enjeux pour les bibliothèques. ([n. d.]). http://www.bnf.fr/fr/professionnels/anx_pro_videos/a.video_cnfpt_interoperabilite.html
[9] Fabien Granjon and Christophe Magis. [n. d.]. Critique et humanités numériques. ([n. d.]). https://journals.openedition.org/variations/748
[10] Alain GUERREAU. 1989. POURQUOI (ET COMMENT) L'HISTORIEN DOIT-IL COMPTER LES MOTS? *Histoire & Mesure* 4, 1/2 (1989), 81–105. http://www.jstor.org/stable/24565515
[11] Eric Kergosien, Marie-Noelle Bessagnet, Christian Sallaberry, Annig Le Parc-Lacayrelle, and Albert Royer. 2016. Analyse géographique de séries de publications: application aux conférences EGC. In *EGC'2016 (Extraction et Gestion des Connaissances)*. 371–382.
[12] Eric Kergosien, Marta Severo, and Marie-Aimée Berthelot. 2018. Cartographier les acteurs d'un territoire : une approche appliquée au patrimoine industriel textile des Hauts-De-France. In *KARTHALA / colloque CIST 2016 "Demande(s) territoriale(s)*. pp 16. https://hal.archives-ouvertes.fr/hal-01708035
[13] Hélène Prost and Joachim Schöpfel. 2015. *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3.* Rapport. Lille 3.
[14] Nathalie Reymonet, Magalie Moysan, Aurore Cartier, and Renaud Délémontez. 2018. Réaliser un plan de gestion de données FAIR: guide de rédaction. (2018).
[15] H. Schmid. 1999. Improvements in Part-of-Speech Tagging with an Application to German. In *Natural Language Processing Using Very Large Corpora*, Nancy Ide, Jean Véronis, Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky (Eds.). Vol. 11. Springer Netherlands, Dordrecht, 13–25. https://doi.org/10.1007/978-94-017-2390-9_2
[16] Joachim Schöpfel, Eric Kergosien, and Hélène Prost. 2017. Pour commencer, pourriez-vous définir'données de la recherche'? Une tentative de réponse. In *Atelier VADOR: Valorisation et Analyse des Données de la Recherche; INFORSID 2017*.
[17] Alexandre Serres, Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, and Didier Collet. 2017. *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs: une enquête à l'Université Rennes 2.* Rapport. Université Rennes 2.
[18] Anne-Violaine Szabados and Rosemonde Letricot. 2012. L'ontologie CIDOC CRM appliquée aux objets du patrimoine antique. In *3e Journées d'Informatique et Archéologie de Paris-JIAP 2012*.
[19] Djamel Abdelkader Zighed. [n. d.]. Les Humanités Numériques en Sciences Humaines et Sociales. ([n. d.]).