Humanités numériques et archives orales : cartographies d'une mémoire collective sur les matériaux

Pierre Teissier, Matthieu Quantin, Benjamin Hervy

CAHIERS
FRANÇOIS VIÈTE

Série III – N° 4

2018

Actualité des recherches du Centre François Viète

sous la direction de
Jenny Boucard

Centre François Viète
Épistémologie, histoire des sciences et des techniques
Université de Nantes - Université de Bretagne Occidentale

Imprimerie Centrale de l'Université de Nantes
Mars 2018

Cahiers François Viète

La revue du *Centre François Viète*
Épistémologie, Histoire des Sciences et des Techniques
EA 1161, Université de Nantes - Université de Bretagne Occidentale
ISSN 1297-9112

cahiers-francois-viete@univ-nantes.fr
www.cfv.univ-nantes.fr

Depuis 1999, les *Cahiers François Viète* publient des articles originaux, en français ou en anglais, d'épistémologie et d'histoire des sciences et des techniques. Les *Cahiers François Viète* se sont dotés d'un comité de lecture international depuis 2016.

ISBN 978-2-86939-246-X

SOMMAIRE

Cahiers François Viète, série III, 4, 2018, p. 141-177

Humanités numériques et archives orales : cartographies d'une mémoire collective sur les matériaux

Pierre Teissier*, Matthieu Quantin† & Benjamin Hervy‡

Résumé

Les sciences humaines appréhendent classiquement les corpus de textes par des lectures qualitatives tandis que les « humanités numériques » les saisissent par des analyses quantitatives. Nous confrontons les deux approches en appliquant une méthode numérique originale (Haruspex) à l'étude d'un corpus textuel d'archives orales dédié à la recherche sur les matériaux. Pragmatique, notre démarche est aussi heuristique et réflexive. Heuristique car nous utilisons les artefacts numériques comme des outils pour renouveler les hypothèses de recherche en histoire. Réflexive car nos pratiques interdisciplinaires servent de base à une « philosophie de terrain » des humanités numériques.

Mots-clés : humanités numériques, extraction et chaînage de connaissances, analyse sémantique, archives orales, mémoire collective, recherche sur les matériaux, chimie du solide, épistémologie et histoire des sciences.

Abstract

Humanities classically perceive the corpus of texts through qualitative reading while "digital humanities" seize them through quantitative analyzes. We compare the two approaches by applying an original numerical method (Haruspex) to the study of a textual corpus of oral archives dedicated to research on materials. Pragmatic, our approach is also heuristic and reflexive. Heuristic because we use digital artefacts as tools to renew research hypotheses in history. Reflective because our interdisciplinary practices serve as a basis for a "field philosophy" of digital humanities.

Keywords: digital humanities, knowledge extraction and management, semantic analysis, oral archives, collective memory, materials research, solid-state chemistry, epistemology and history of science.

* Centre François Viète d'épistémologie et d'histoire des sciences et des techniques (EA 1161), Université de Nantes.

† École centrale de Nantes, Laboratoire des Sciences du Numérique de Nantes (LS2N UMR 6004) | orcid.org0000-0003-4315-7369.

‡ Université de Nantes, Laboratoire des Sciences du Numérique de Nantes (LS2N UMR 6004) | orcid.org0000-0002-5755-6478.

Les chercheurs en humanités produisent des corpus de textes, hétérogènes par leur forme et leur contenu, et spécifiques par leur terminologie et leur signification. Ces corpus sont classiquement analysés à travers des lectures exhaustives et des interprétations constamment remises sur le métier. Depuis les années 1950, des chercheurs des sciences numériques et humaines ont utilisé les possibilités de calcul des ordinateurs pour développer des méthodes quantitatives d'analyse de ces corpus textuels (Hockey, 2007). Ceci a fait émerger un champ d'institutions, de pratiques et d'outils, aujourd'hui identifié dans la sphère académique comme « humanités numériques » ou *digital humanities*. Un ouvrage collectif de synthèse (Schreibman, Siemens & Unsworth, 2004) marque la reconnaissance internationale du domaine.

Situé dans ce champ émergent, notre travail a pour objectif l'extraction de données quantitatives dans un corpus de textes et la confrontation de ces résultats numériques à une analyse qualitative du corpus. Le corpus, composé d'une quarantaine d'interviews retranscrites en français, est bien connu par l'un des trois auteurs (Pierre Teissier). Cette connaissance antérieure du corpus nous semble indispensable durant la phase de développement de la méthode numérique afin de s'appuyer sur une certaine fiabilité d'interprétation historienne des artefacts quantitatifs. L'outil numérique (Haruspex), qui génère automatiquement des liens entre les documents du corpus en fonction de co-occurrences de mots, a été construit par les deux autres auteurs (Matthieu Quantin et Benjamin Hervy). Haruspex calcule des graphes de proximité sémantique entre les entretiens du corpus sans *a priori* : ni apprentissage ni données extérieures. La sélection de différents critères de visualisation permet de générer différents graphes, qui semblent attendus, surprenants ou absurdes. Les allers et retours entre numériciens et historiens permettent de confronter les analyses numériques et historiennes afin de dégager une interprétation des graphes. Ils viennent ainsi confirmer certains savoirs antérieurs, dessiner des pistes fructueuses de réflexion ou conduire à des impasses qui semblent dénuées de sens historique. Des échanges réguliers nous ont permis de croiser nos compétences en histoire des sciences et en sciences numériques. Ce faisant, nous avons cherché à intégrer plutôt qu'à juxtaposer

deux domaines distincts par une démarche de terrain plus interdisciplinaire que pluridisciplinaire. Sans masquer les différences de perspective, les contradictions disciplinaires, voire les apories, notre démarche vise l'enrichissement mutuel de chaque domaine : création de connaissances historiques et de réflexions philosophiques et développement d'outils numériques adaptés aux sciences humaines et sociales. L'article est organisé en trois parties. Nous commençons par présenter l'objet d'étude — les archives orales — et l'outil d'analyse textuelle — la méthode numérique d'extraction de données Haruspex. Nous appliquons ensuite l'outil à l'étude de l'objet. Ceci permet de fabriquer trois types d'image numérique : une cartographie générale du corpus ; des zooms sur des territoires plus restreints du corpus ; des visualisations spécifiques du corpus par la sélection de critères particuliers. Ces artefacts numériques sont confrontés aux connaissances historiennes soit pour confirmer des acquis antérieurs, l'artefact numérique jouant alors le rôle d'indice supplémentaire pour le raisonnement historien, soit pour donner des représentations surprenantes, l'artefact numérique jouant alors un rôle heuristique. Cette étude de cas se termine par une réflexion épistémologique et réflexive sur les relations entre les sphères numériques et historiques. Une telle « philosophie de terrain » permet d'établir une typologie des inférences permises par l'interaction hommes/machines et un schéma de fonctionnement de la méthode numérique.

Contexte général de l'étude : corpus, outil et méthode

- *Présentation du corpus d'archives orales*

Histoire et mémoire de la recherche sur les matériaux au XX^e^ siècle

Notre corpus d'archives orales s'inscrit dans un programme de recherche plus large consacré à l'histoire et à la mémoire de la « recherche sur les matériaux » ou *Materials Research* en anglais (Bensaude-Vincent & Teissier, 2015). Ce programme collectif est matérialisé depuis 2011 par le site internet Sciences : Histoire Orale[1], simultanément « lieu de mémoire » et espace de réflexion épistémologique. Dans cet ensemble étendu et bigarré, nous avons délimité un corpus plus homogène selon deux critères. Premièrement, nous avons choisi les seuls entretiens que

[1] https://www.sho.espci.fr/?lang=fr

l'un d'entre nous connaissait suffisamment pour les avoir menés ou les avoir utilisés de manière approfondie. Ceci est nécessaire pour pouvoir discuter les résultats numériques à partir d'une connaissance qualitative préalable. Deuxièmement, nous n'avons retenu que les textes en français car l'analyse sémantique par mots-clés exige une unicité de langue.

Nous obtenons ainsi un corpus de 41 entretiens retranscrits, identifiés par le nom de la personne interviewée (il y a deux cas où deux personnes proches ont été interrogées ensemble). Les entretiens étaient semi-directifs, c'est-à-dire que l'interviewer interroge l'interviewé.e à partir d'une grille de questions préétablies. Les personnes interviewées ont ainsi été amenées à raconter leur carrière et leurs environnements professionnels, pour les plus anciens depuis les années 1940. Elles sont à la fois témoins et actrices de la recherche sur les matériaux du second XX^{e} siècle, que ce soit dans la sphère académique (chercheurs, enseignants-chercheurs, administratifs) ou industrielle (administrateurs, ingénieurs).

Qualitativement, le corpus d'entretiens n'est pas d'un seul tenant. Il est composé de trois sous-corpus présentés dans la table 1. Ces trois sous-ensembles se distinguent par la date de constitution, l'interviewer principal, la grille de questions et le thème de recherche. Détaillons-les pour expliciter leur teneur. Le sous-corpus 1 est le plus ancien. Il est composé de 7 entretiens, réalisés entre 2000 et 2003, pour la plupart par B. Bensaude Vincent. Il provient d'un programme d'histoire de la science et ingénierie des matériaux, parrainé par la *Sloan Fondation* et le *Dibner Fund* du MIT (Cambridge, MA)[2]. Le sous-corpus 2 est le plus étendu et le plus homogène. Il est composé de 26 entretiens, réalisés entre 2004 et 2007, pour la plupart par Pierre Teissier dans le cadre d'une thèse sur l'histoire de la chimie du solide en France (Teissier, 2007). Le sous-corpus 3 est le plus hétérogène. Il est composée de 8 entretiens, réalisés entre 2009 et 2016 par Pierre Teissier, en lien avec plusieurs objets d'étude historique : matériaux, batteries et piles à combustible, voitures électriques et à hydrogène.

[2] Les principaux chercheurs du programme, Hervé Arribart, Bernadette Bensaude Vincent et Arne Hessenbruch, ont ainsi collecté une trentaine d'entretiens de chercheurs en matériaux ayant fait leur carrière en Europe, aux États-Unis ou au Japon. Voir http://authors.library.caltech.edu/5456/1/hrst.mit.edu/hrs/materials/public/general.htm

Table 1 – Constitution du corpus d'entretiens : identité des interviewé.e.s et date d'interview

(a) Sous-corpus 1		(b) Sous-corpus 3	
Barboux, Philippe	2000	Beuzit, Pierre	2010
Boilot, Jean-Pierre	2000	Catonné, Jean-Claude	2012
Colomban, Philippe	2003	Fauvarque, Jean-François	2013
Friedel, Jacques	2001	Lisse, Jean-Pierre	2009
Griesemann, Jean-Claude	2001	Lucchese, Paul	2016
Livage, Jacques	2001	Poulain, Marcel et Michel	2016
Zarzycki, Jerzy	2001	Priester, Louisette	2011
		Vitet Sylvain	2009

(c) Sous-corpus 2			
Adam, Jean-Luc	2006	Lucas, Jacques	2005
Antic-Fidancev, Élisabeth	2006	Maestro, Patrick	2007
Bonino, Christiane	2006	Mondange, Hélène	2004
Caro, Paul	2005	Moradpour, Alexandre	2006
Dexpert, Hervé	2006	Ouvrard, Guy	2006
Étourneau, Jean	2005	Perez y Jorba, Monique	2004
Flahaut, Jean et Rivet, Jacques	2005	Pouchard, Michel	2004
Hérold, Albert	2006	Rousset, Abel	2006
Galy, Jean	2006	Serreau, Dannielle	2004
Hagenmuller, Paul	2004	Théry, Jeanine	2004
Jérôme, Denis	2006	Tournoux, Michel	2006
Kahn-Harari, Andrée	2004	Vivien, Daniel	2004
Lefrant, Serge	2006	Vitorge, Marie-Claude	2007

Enjeux de l'analyse numérique d'archives orales

Le corpus est intéressant pour l'analyse numérique dans la mesure où il offre plusieurs niveaux de granulométrie : il est suffisamment cohérent autour de la recherche sur les matériaux et suffisamment hétérogène par la présence de trois sous-corpus. Ceci permet de tester les possibilités offertes par Haruspex pour évaluer et représenter différents niveaux de proximité sémantique et thématique. Les caractéristiques quantitatives de base du corpus sont présentées dans la table 2.

En outre, une analyse rapide du corpus par un processus de *topic modelling* permet de s'assurer de l'homogénéité du contenu des documents constituant le corpus : il n'est pas trivial d'identifier des « thématiques » discriminantes au sein du corpus sur la seule base des mots employés. Cette homogénéité est nécessaire d'un point de vue méthodologique pour l'analyse des résultats produits par Haruspex.

Table 2 – Caractéristiques du corpus étudié

format de fichier	Open Document
structure interne (titre, sous-titre, etc.)	non
nombre de documents	41
nombre de mots	339k
nombre de mots après filtre	87 316
nombre de lemmes différents	9 884
moyenne du nb de mots par document	8268
écart-type du nb de mots par document	4837

Haruspex assimile chaque entretien à un nœud et calcule des intensités de lien entre ces nœuds. Le réseau de nœuds et de liens forme une cartographie numérique du corpus qu'il s'agit d'interpréter. Un raffinement est permis par l'attribution manuelle de métadonnées à un entretien. Ces métadonnées sont, pour notre corpus, des variables du type : date d'obtention d'une thèse (qui indique une génération), laboratoire d'exercice, statut social, genre, discipline, objets de recherche, etc. Elles ne sont pas nécessaires au processus d'extraction et de mise en relation mais sont cruciales pour l'interprétation des réseaux.

Un lien entre deux nœuds est fortement déterminé par le contenu sémantique des entretiens. Or, un entretien est un objet complexe, tissé d'informations et de significations : itinéraire professionnel de l'interviewé.e, institutions de rattachement, collègues plus ou moins proches, thèmes de recherche, laboratoires, communautés, théories et instruments, politiques scientifiques, stratégies commerciales et industrielles, etc. Il contient en outre une vision spécifique — celle de l'individu — du champ de recherche à travers l'espace social (institutionnel), géographique (villes et pays) et mémoriel (durée et précision des souvenirs variables en fonction des individus et des générations). Il entrecroise ainsi des questions techniques (recherche), des énoncés relationnels et affectifs (interpersonnels), des positionnements identitaires (discipline, génération, genre) et des jugements culturels. En résumé, un entretien articule une multitude

de niveaux de discours, de références plus ou moins explicites et d'interprétations des mondes humains et naturels. Il contient un nombre indéfini de significations. Au contraire, l'analyse numérique construit ses données comme des nombres et les relations entre les nœuds comme des opérations de calcul sur des nombres.

Il y a donc une irréductibilité des données entre sciences informatiques et sciences humaines. Ces données sont numériques et calculatoires pour les premières, sémantiques et herméneutiques pour les secondes. Cette rupture soulève quatre enjeux épistémiques de difficulté croissante :

- identifier différents niveaux de discours : scientifique, technique, institutionnel et politique certes, mais aussi épistémologique, parfois moral ou affectif ;
- saisir, à un niveau de discours donné, des structures : par exemple, au niveau institutionnel, repérer une école de recherche ou une communauté scientifique ;
- faire apparaître la dimension temporelle, qui focalise, sans doute, la principale interrogation des historiens : la structuration des discours pourrait par exemple faire apparaître des éléments propres à chaque génération de témoins ;
- appréhender un méta-discours comme celui des épistémologues, qui produisent des discours savants sur des discours savants : par exemple, l'analyse du travail du physicien qui décrit le comportement d'un nuage d'électrons.

• *Présentation de l'outil numérique*

Notre méthode d'extraction et de chaînage de connaissances a pour objectif de dépasser la linéarité de lecture d'un ensemble de textes en reliant ces textes en fonction de leur proximité sémantique. Cette cartographie numérique pourra avoir une valeur de validation (d'une interprétation préalable) ou d'heuristique (en suggérant de nouvelles hypothèses).

L'analyse de données textuelles en humanités numériques

Les humanités numériques se sont densifiées avec l'engouement des chercheurs et l'augmentation des financements depuis une vingtaine d'années. L'un des secteurs les plus stimulants est l'analyse de données textuelles à laquelle appartient Haruspex.

Plusieurs techniques d'analyse de données textuelles sont régulièrement exploitées pour des corpus en sciences humaines et sociales. Les techniques à base d'apprentissage supervisé notamment se servent de données annotées pour identifier des données brutes de même nature. Les solutions appartiennent à un ensemble fini, initialement pré-établi (par exemple : reconnaissance de caractère, à partir d'images de caractères typographiques associés à la lettre qu'ils représentent). À un niveau supérieur, l'apprentissage peut servir à reconnaître des entités (lieu, personne, date…). Ces données sont enregistrées avec des descripteurs standards (ontologies, métadonnées, TEI…), orientés vers le partage de données, la formalisation des connaissances ou leur valorisation. Les données sont parfois représentées sous forme graphique. Les analyses produites dans ce cadre sont limitées aux catégories pré-établies, principalement à destination des données massives.

Dans l'objectif d'étudier un corpus, ce type d'approche peut produire des analyses ou des traitements à l'échelle du corpus dans sa globalité. Il s'agit alors d'une « lecture en survol » (*distant reading*) du corpus (Moretti, 2005).

Notre approche oscille elle entre une « lecture attentive » (*close reading*)[3] et une « lecture en survol » (*distant reading*). Elle produit des représentations ou cartographies du corpus, à plusieurs échelles : de la lecture globale au lien très particulier. Ces cartographies constituent le point de départ des discussions entre numériciens et historiens.

Deux techniques sont alors source d'inspiration pour notre travail. Côté *distant reading*, le *topic modelling* (apprentissage non supervisé), consiste à proposer des classifications à partir de similarités dans les données brutes. Ces catégories ne sont ni interprétées ni présupposées (on cherche par exemple à regrouper des articles scientifiques à partir des mots qu'ils contiennent, sans présupposer de discipline…). Côté *close reading*, la lexicométrie, ensemble de mesures appliquées au texte, influence nos travaux dont l'approche est purement statistique.

Haruspex évite toute restriction du domaine d'étude : pas de descripteurs standard, pas de classes *a priori*, pas de textes annotés pour apprendre à étudier le corpus. En ce sens, et contrairement aux techniques

[3] Cette pratique consiste à étudier minutieusement un texte aussi bien qualitativement que quantitativement. Par exemple, dans *Ulysse Gramophone*, Jacques Derrida (1987) analyse la récurrence du mot « oui » dans *Ulysse* de James Joyce.

précédemment mentionnées d'analyse textuelle de lecture en survol, il permet une lecture attentive complémentaire, nécessaire pour répondre aux enjeux épistémiques décrits précédemment.

Le fonctionnement d'Haruspex

Haruspex fonctionne suivant quatre étapes principales : traitement du corpus, extraction d'expressions, enrichissement des expressions et création de liens. Quantin (2018) explique ces étapes en détail, nous les présentons ici succinctement.

1) Le traitement du corpus

Il consiste à traiter les fichiers en entrée (OpenDocument, LaTeX, PDF) : conversion, découpage, construction d'« unités documentaires » (documents ou parties de documents). Si le choix d'étude de corpus à taille humaine ($\approx 10^6$ mots) nous éloigne de certaines problématiques de *big data*, nous restons proches des techniques de *clustering* par l'analyse de relations inter- et intra-*clusters* (Jain, Murty & Flynn, 1999). Par la suite, nous appellerons *cluster* un groupe de documents partageant des caractéristiques quantitatives qui le démarquent des autres. Ces caractéristiques sont retranscrites visuellement dans les cartographies.

2) L'extraction d'expressions

Elle est réalisée grâce à un algorithme de traitement automatique du langage inspiré de l'*Automatic Natural Acquisition of a Terminology* (ANA) (Enguehard, 1993 ; Quantin et al., 2016). Cet algorithme extrait du corpus, sans entraînement préalable ni vocabulaire de référence, une liste d'expressions spécifiques. Les expressions discriminant des sous-ensembles du corpus seront privilégiées (Salton, 1983). Cela correspond également aux expressions plus longues, donc moins ambiguës (Finlayson & Kulkarni, 2011). Par exemple, l'expression « terres rares par cristallisation fractionnée », peu ambiguë, concerne une partie seulement du corpus. Notre analyse se base sur une approche statistique tirant parti de ces expressions en minimisant les choix en amont pour rendre l'extraction des expressions la plus indépendante possible.

3) L'enrichissement des expressions

Il consiste à associer les expressions extraites à des catégories grâces à des requêtes vers *Wikipedia* (Milne & Witten, 2008). Un taux de confiance dans ces catégories est calculé. D'autres indicateurs sont calculés, qui portent sur la rareté du mot dans la langue française, son ambiguïté, son pouvoir discriminant dans le corpus et la cohésion des termes (par exemple : « verres fluorés » versus « année de thèse »). Ils aident à modérer les résultats. Nous faisons l'hypothèse que seul un processus modéré permet d'obtenir des résultats de qualité sur des corpus non structurés[4] et sans données extérieures (apprentissage supervisé). Par ailleurs, la pratique historienne nous incite à éviter les effets de « boite noire » et à donner trop de confiance à la machine, l'historien ayant *in fine* la responsabilité du texte final.

4) La création de liens

Elle permet de construire des graphes formés par des liens pondérés entre les nœuds du corpus. La pondération est calculée en fonction de la distribution de l'expression au sein du corpus et au sein de chaque paire de documents concernés (un lien implique une paire de documents).

A : Au sein du corpus, une distribution entropique[5] (Shannon, 1948) de l'expression parmi les documents est favorisée (type IDF).

B : Au sein de la paire de documents concernés, la quantité et l'équipartition sont favorisées.

En A, une fonction à seuil permet de séparer les expressions génériques des expressions spécifiques à certaines parties du corpus (discriminantes). En B, nous utilisons une fonction logarithmique impliquant le nombre minimal d'occurrences de l'expression dans la paire. Notre analyse se base donc sur une approche statistique tirant parti du nombre d'occurrences d'expressions-clés complexes, révélatrices d'un sujet sans ambiguïté.

Nous faisons l'hypothèse générale que les représentations numériques du corpus, basées sur l'approche statistique, révèlent un contenu

[4] En informatique, un corpus est dit structuré s'il comporte des marqueurs explicites pour la machine, qui désignent des éléments de contenu comme des dates, des mots-clés, etc.

[5] L'entropie est définie par $H(X) = -\sum_i^n p_i \log_2(p_i)$ avec une source X comportant n symboles, un symbole i ayant une probabilité p_i d'apparaître.

informatif et sont susceptibles d'interprétations qui tantôt confirment, tantôt complètent la lecture linéaire du corpus. La deuxième partie de l'article met cette hypothèse générale à l'épreuve.

Application d'Haruspex à la cartographie d'une mémoire collective : le cas de la recherche sur les matériaux

- *L'analyse automatique du corpus par Haruspex*

L'originalité principale d'Haruspex tient à la deuxième étape de son fonctionnement : l'extraction automatique d'expressions spécifiques d'un corpus, c'est-à-dire sans spécification *a priori* de ce qui constitue une expression à retenir. La table 3 donne quelques exemples d'expressions extraites de notre corpus auxquelles elle associe deux caractéristiques quantitatives : leur nombre d'occurrences dans le corpus et le nombre de documents concernés. Elle intègre aussi la troisième étape d'enrichissement des expressions par la qualification thématique (manuelle) des expressions au moyen de requêtes vers *Wikipedia*.

Table 3 – Exemples d'expressions extraites, informations numériques associées et thématiques relatives

Forme extraite	Nb d'occ.	Nb. doc. concernés	Thématique
Bronzes de vanadium	26	5	Sciences
Chimie de coordination	22	8	Chimie
Thèse de troisième cycle	16	7	France, Éducation
Microscopie électronique à transmission à haute résolution	3	2	Physique
Four solaire d'Odeillo	4	3	Industrie, Énergie, Sciences

L'extraction automatique de données du corpus donne une liste brute de 2327 expressions. Nous avons traité cette liste afin d'éliminer les expressions trop générales (articles, pronoms, etc.) et de valider les propositions de fusion de l'algorithme (par exemple : l'expression « supraconductivité » est fusionnée avec « supraconducteur »). Une modération de trois heures environ a ainsi permis de réduire la liste à 1169 expressions pertinentes.

Ceci permet de passer à la quatrième étape : la création de liens. La présence d'une expression spécifique dans deux documents forme un lien entre les nœuds correspondants. La pondération du lien est fonction de la distribution de l'expression (voir p. 149). Au sein du corpus, l'entropie est favorisée. Par exemple une expression comme « laboratoire » est trop répandue dans notre corpus pour révéler des relations signifiantes. Elle est donc rejetée. À l'inverse, une expression comme « bronzes de vanadium », qui n'apparaît que dans cinq documents (cf. table 3), fournit des informations relationnelles. Le lien est donc validé. Au sein de la paire de documents concernés, la quantité et l'équipartition sont favorisées. Ainsi, pour « bronzes de vanadium », Michel Tournoux mentionne l'expression une seule fois alors que les quatre autres interviewés la mentionnent cinq fois ou plus : les 4 liens entre le nœud Tournoux et les quatre autres nœuds seront donc beaucoup moins forts que celui qui unit les quatre autres entre eux. Ceci évite de valoriser la simple évocation d'un sujet.

La pondération des liens est appelée « chaînage supervisé » par les numériciens, car elle consiste à enchaîner les nœuds du corpus les uns aux autres en supervisant la façon dont les liens sont fabriqués et leur intensité calculée. Le « chaînage supervisé » conduit, dans le cas de notre corpus, à créer plus de 127 000 liens spécifiques. Ce nombre de relations est trop important pour être complètement saisi par un cerveau humain. Des simplifications sont donc nécessaires. Des procédures de filtrage et de visualisation des données relationnelles permettent de produire des représentations numériques relationnelles ciblées qui servent de base, à une confrontation entre analyses quantitatives et qualitatives (p. 152 et suivantes).

- *Confrontation des analyses quantitatives et qualitatives*

Cartographie générale du corpus : chimie du solide, électrochimie et automobile

L'utilisation de filtres et de métadonnées permet de visualiser un réseau global de relations entre les entretiens du corpus. L'intensité des liens spécifiques a été additionnée pour donner un lien résultant entre chaque nœud et un seuil de visualisation a été imposé pour ne faire apparaître que les liens de plus forte intensité. Nous supposons que la figure résultante (figure 1) constitue une cartographie mémorielle de la recherche sur les matériaux. Cette hypothèse sera validée si l'interprétation qui en découle

est possible et convaincante par rapport à la connaissance historienne du corpus d'archives orales.

Transposition visuelle de valeurs quantitatives, la figure 1 montre des hétérogénéités dans la position, la taille et les relations entre nœuds, que nous appellerons *clusters* (voir p. 149) par la suite. La confrontation de cette visualisation à notre connaissance qualitative préalable du corpus conduit à formuler deux ensembles de commentaires.

1) L'analyse sémantique comme discriminant thématique

Le premier ensemble de commentaires concerne le rapport entre l'histoire de la constitution du corpus et sa représentation numérique. Le corpus a été formé par le regroupement de trois sous-corpus d'archives orales (cf. table 1) différenciés par la période de collecte (2000-2003, 2004-2007, 2009-2016), le principal interviewer (Bensaude-Vincent pour le premier, Teissier pour les deux suivants) et le thème de l'entretien. De ces trois différences *a priori*, la représentation numérique ne retient, au niveau de filtrage qui est le nôtre, que la dimension thématique. En effet, elle mêle les entretiens du sous-corpus 1 (science et ingénierie des matériaux) et du sous-corpus 2 (chimie du solide) sans qu'il soit possible de distinguer leur origine sans connaissance du corpus[6]. Les témoins correspondants appartiennent à la communauté de « recherche sur les matériaux ». Par rapport à cette partie centrale du corpus, cinq entretiens forment un *cluster* périphérique, isolé en bas à droite de la figure. Tous ont été interviewés sur le même thème spécifique : l'industrie des voitures électriques et à hydrogène en lien avec les réseaux électriques. Quatre d'entre eux (Beuzit, Lisse, Lucchese, Vitet) appartiennent au sous-corpus 3, le cinquième (Griesemann) au sous-corpus 1. En outre, les trois entretiens qui leurs sont le plus liés concernent la recherche et le développement des batteries et piles à combustible, deux d'entre eux issus du sous-corpus 3 (Catonné, Fauvarque), le troisième du sous-corpus 1 (Barboux).

Cette première confrontation entre analyses quantitative et qualitative montre deux atouts d'Haruspex. D'une part, l'outil est particulièrement adapté à l'analyse sémantique de textes puisqu'il parvient à isoler

[6] Dans une perspective réflexive, on peut noter ici que le second interviewer a réalisé une thèse sous la direction de la première interviewer, ce qui, à n'en pas douter, a généré des convergences de méthodes et de questionnements. Ils ont, en outre, réalisé certains entretiens ensemble.

des *clusters* centrés sur des thèmes périphériques par rapport au reste du corpus : industries automobile et électrique pour le *cluster* de cinq nœuds, isolé en bas à droite ; recherche et développement en électrochimie si on rajoute les trois nœuds les plus proches du *cluster* isolé. D'autre part, l'outil semble relativement indépendant des conditions extérieures de l'entretien (date, projet) et de la subjectivité de l'interviewer (formulation des questions, mode d'expression) comme le montre l'intégration des deux premiers sous-corpus réalisés par deux historiens différents ayant des projets différents à des dates différentes.

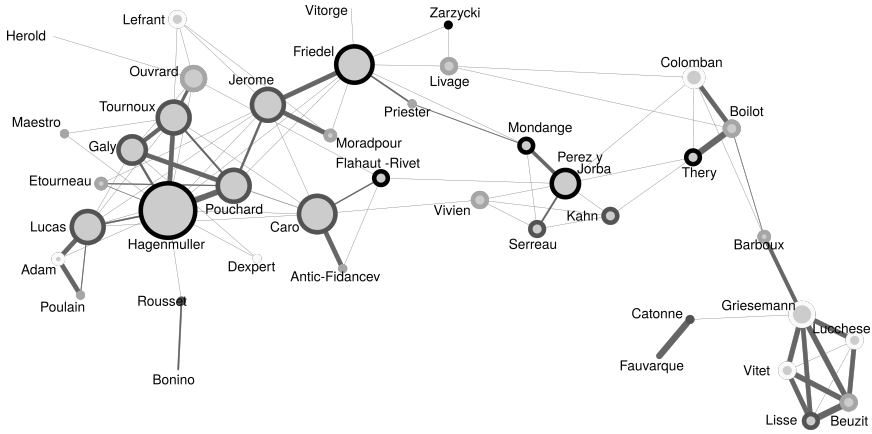

*Figure 1 – Chaque nœud représente un entretien, identifié par le nom du témoin interviewé. L'intensité des traits résulte de la somme de liens créés par partage d'une expression. Sous un seuil de pondération (*0.3*), les traits ne sont plus affichés. La taille d'un nœud indique son degré de connectivité : plus le témoignage a de connexions, plus il est volumineux. La couleur des nœuds renseigne sur la date de thèse de l'interviewé.e : plus les nœuds sont foncés, plus la décennie de soutenance est ancienne (noir dans les années 1950, blanc après 1980).*

2) Analyse de la communauté française de chimie du solide

Le deuxième ensemble de commentaires laisse de côté le *cluster* périphérique en bas à droite et analyse plus précisément la partie principale de la figure 1, composée de 33 entretiens, soit les trois quarts du corpus. Cet ensemble concerne la « recherche sur les matériaux » avec une focalisation sur la communauté française de « chimie du solide ». On y trouve le témoignage de 27 chimistes du solide, 4 physiciens du solide et 2 techniciens, travaillant tous en France.

La partie principale comprend cinq *clusters*. Une première lecture divise les parties droite et gauche de la figure 1 par rapport à l'axe vertical reliant Vivien et Livage. À droite, se trouvent les représentants de l'école de recherche initiée par Robert Collongues. L'analyse quantitative montre que cet ensemble de nœuds partage la mention de différents noms de chimistes (Jean « Talbot », Anne-Marie « Lejus », Noël « Baffier ») et de références épistémiques à des matériaux (« alumine » et « oxydes de fer ») et à des instruments (« fours à plasma », « fours à image »). Elle rencontre l'analyse historique car l'école Collongues (à laquelle sont associés les trois noms cités) s'est illustrée par l'étude de matériaux réfractaires (dont l'alumine et les oxydes de fer) et la mise au point de techniques de chauffage à haute température (dont les fours à plasma et à image). Il y a donc convergence des résultats quantitatifs et qualitatifs.

La partie droite de la figure peut, en outre, être subdivisée en deux *clusters*. Le premier, composé de cinq nœuds (Kahn, Mondange, Perez, Serreau, Vivien), recouvre le groupe historique de Collongues, qui commence à Vitry-sur-Seine puis migre à l'École de chimie de Paris. Le second *cluster*, composé de trois nœuds (Boilot, Colomban, Théry, quatre si l'on ajoute Barboux), correspond à peu près à un groupe de recherche héritier de Collongues, qui s'est établi à l'École polytechnique. L'analyse quantitative montre que le premier *cluster* se distingue du second par le partage de références institutionnelles au laboratoire de « Vitry » ou « CECM » (Centre de chimie métallurgique du CNRS).

La partie gauche de la figure 1 paraît plus dense, même si trois *clusters* peuvent encore être identifiés. Le premier d'entre eux est central dans le corpus. Il est composé, en son cœur, de cinq nœuds : Hagenmuller, qui constitue le nœud le plus connecté du corpus, et quatre de ses disciples directs (Etourneau, Galy, Pouchard, Tournoux). Des lignées peuvent ainsi être tracées d'une génération à la suivante : Hagenmuller - Tournoux - Ouvrard (Lefrant étant un collègue physicien d'Ouvrard à Nantes). On retrouve ici, comme dans le cas Collongues, une structuration en école de recherche, qui a essaimé depuis Bordeaux vers d'autres villes universitaires comme Nantes (Tournoux) ou Toulouse (Galy). Certains « anciens de Bordeaux » sont faiblement liés au *cluster* central, l'un (Dexpert) parce que ses choix thématiques l'ont éloigné de l'école, l'autre (Maestro) parce qu'il a fait carrière dans l'industrie des matériaux. Un « héritage indirect » est aussi repérable entre Hagenmuller et Lucas à Rennes, celui-ci étant

lui-même lié à son disciple (Adam) et plus faiblement à deux chercheurs de son laboratoire (Michel et Marcel Poulain)[7].

Les deux derniers *clusters* se trouvent au milieu de la figure. L'un, en bas, s'organise autour de Caro, lui aussi, figure centrale du réseau. Ses relations avec les autres sont peu marquées par des relations de filiation d'école : il est certes lié à son ancienne doctorante (Antic-Fidancev), qui reste dans son laboratoire, mais pas du tout à un autre chercheur de son laboratoire (Dexpert). Il est beaucoup plus lié à des pairs de la même génération que lui (Jérôme, Flahaut et Rivet, Pouchard) avec qui il partage des liens de type épistémique dont « four solaire », « isolant », « spectres » et « luminescence ». Contrairement à la structuration par écoles de recherche des trois premiers *clusters*, l'entourage de Caro s'apparente plus au *cluster* d'un chercheur influent (académicien) et original qui, sans faire école au sens institutionnel d'une filiation, a structuré le champ au niveau épistémique.

Le cinquième et dernier *cluster* identifié rassemble les physiciens du corpus (Jérôme, Lefrant, Priester) autour du physicien Friedel, qui jouerait le même rôle structurant que Hagenmuller si le corpus était centré sur la physique du solide plutôt que sur la chimie du solide. Il agrège aussi des chimistes qui n'étudiaient pas des cristaux minéraux, ce qui était rare à l'époque concernée : solides amorphes (Livage, Zarzycki) ou organiques (Moradpour, Vitorge). Mais ceci n'est pas une règle car d'autres chimistes « hétérodoxes » par leurs positionnements thématiques ou sociologiques (Boninon, Rousset et Hérold) sont quant à eux plus liés au *cluster* Hagenmuller qu'à celui des physiciens.

Ainsi, l'interprétation de la figure 1, notamment sa partie principale, confirme deux résultats principaux de la thèse sur l'histoire de la chimie du solide : une autonomie des communautés de physique et de chimie du solide en France ; une structuration de la communauté de chimistes du solide polarisée par la forte opposition entre deux mandarins tout puissants (Hagenmuller et Collongues) et clairsemée d'une multitude d'écoles de recherche moins puissantes (Caro, Flahaut, Hérold, Lucas, Rousset, etc.) (Teissier, 2014). Il est intéressant de remarquer que les liens entre les deux écoles de recherche les plus puissantes se font via les écoles de puissance intermédiaire (Caro, Flahaut) et par les physiciens, qui colla-

[7] Les témoignages de Hagenmuller et Lucas mentionnent l'influence du premier sur le second à Rennes même s'il n'y a pas eu direction officielle de thèse.

borent avec les chimistes depuis l'extérieur. L'analyse numérique révèle ainsi la mise en connexion des centres d'une communauté scientifique (chimie du solide) par les marges disciplinaires (physique du solide) et les outsiders de la communauté. La position dominante de Friedel parmi les physiciens interrogés permet d'extrapoler en suggérant une structuration mandarinale comparable en physique du solide. La position centrale du *cluster* Hagenmuller et la connectivité maximale du nœud Hagenmuller s'expliquent par trois éléments entremêlés. Une domination sociale d'une part : plus de 300 docteurs formés à Bordeaux durant la direction d'Hagenmuller ainsi que des laboratoires héritiers dans tout l'ouest de la France, notamment à Nantes et dans une moindre mesure Rennes. Une contingence historique ensuite : son grand rival, Collongues, ne figure pas sur le réseau car, décédé en 1998, il n'a pas pu être interrogé. Un « effet de sources » enfin : la collecte des témoignages entre 2004 et 2007 ayant été faite de proche en proche à partir du laboratoire Collongues puis Hangenmuller, ces deux écoles de recherche ont été favorisées dans la constitution du corpus.

Valider le connu et interpréter le surprenant pour des liens localisés

La représentation numérique globale du corpus a donc confirmé certaines conclusions historiques, notamment la structuration sociologique et épistémique de la communauté. Pour affiner cette approche globale, nous analysons trois liens localisés autour d'un nœud et entre deux nœuds. L'objectif est de montrer que la valeur numérique d'un lien est une réduction pratique mais trompeuse qui cache une grande variété de contenus et de significations.

1) Le nœud central : Paul Hagenmuller

Nous commençons par étudier un nœud très particulier : le plus gros du corpus, c'est-à-dire le plus connecté aux autres, Hagenmuller. Nous répertorions, dans la table 4, la liste des liens les plus forts de ce nœud et les expressions correspondantes. Ceci renseigne sur les thématiques fortes d'un témoin particulier et sur les témoignages les plus significativement liés à lui. Une telle analyse est indispensable pour comprendre le rôle et la place d'un acteur particulier dans le corpus. Le cas Hagenmuller donne deux types d'information.

Premièrement, ces liens forts sont souvent dus à des partages avec un ou deux autres témoins d'expressions ayant peu d'occurrences dans

le corpus. Par exemple, le lien le plus fort du corpus (poids de 0.606) concerne le « vanadium », un élément chimique que Hagenmuller partage avec Galy et Pouchard de manière quasi monopolistique : ces trois témoignages comptent pour 22 des 23 occurrences du mot dans le corpus. Ce monopole est encore plus frappant dans l'équipartition de « transition métal-isolant » avec Jérôme : 3 occurrences chacun, soient les 6 que compte le corpus. S'accaparer une expression-clé avec un ou deux autres témoins fabrique donc des liens de forte intensité. Deuxièmement, en extrapolant cette première tendance sur un grand nombre de liens, nous comprenons que ce qui fait le caractère central de P. Hagenmuller dans le corpus est son aptitude à savoir utiliser des expressions spécifiques qui ne sont partagées que par quelques spécialistes d'un domaine, de manière aussi précise qu'eux. Ceci lui permet de se lier fortement, et de manière privilégiée, avec beaucoup de témoins. Cette conclusion sémantique peut s'expliquer par l'organisation sociale de son laboratoire à Bordeaux : cet immense et riche institut était composé d'une douzaine d'équipes de recherche organisées autour de domaines spécialisés, dont le directeur avait une connaissance précise comme administrateur plutôt que spécialiste.

C'est un apport à la fois surprenant et fructueux d'Haruspex de suggérer un isomorphisme de l'espace social du laboratoire (connu antérieurement) et de l'espace sémantique de la communauté (suggéré par les représentations numériques). Dans cette perspective, un mandarin comme Hagenmuller n'est pas seulement celui dont tout le monde parle et qui forme le plus d'héritiers, ce qui était suggéré par l'analyse historique, mais encore celui qui parle le langage des spécialistes du corpus sans investir lui-même une spécialité.

2) Un chaînon manquant : Monique Pérez et Jeanine Théry

Le lien entre Monique Pérez y Jorba et Jeanine Théry a été choisi pour cette deuxième analyse locale en raison de son intensité étonnamment faible par rapport à notre présupposé historien. Son intensité atteint à peine le seuil d'affichage. Ceci est dû au fait que l'occurrence des mots-clés qu'elles partagent est faible par rapport au corpus. Ces deux chercheuses de la même génération ne partagent rien de spécifique alors qu'elles ont été les deux lieutenantes indéfectibles de Collongues pendant quatre décennies. Ainsi, leur position, côte à côte, sur la photographie des premières années du laboratoire Collongues à la fin des années 1950 (figure 2) tranche avec leur faible connexion sur la cartographie générale

Table 4 – Extrait (poids > 0.32) des mots-clés les plus significatifs dans leur mise en relation entre Paul Hagenmuller et les autres interviews. La colonne « Occ » indique le nombre total d'occurrences du mot-clé dans le corpus, les colonnes « in A » et « in B » le nombre d'occurrences dans l'interview Hagenmuller et l'interview associée.

Mot-clé	Poids	Occ.	in A	in B	Interview associée
vanadium	0.606	23	6	8	Jean Galy
vanadium	0.606	23	6	8	Michel Pouchard
Félix Trombe	0.533	66	4	45	Paul Caro
bronzes de tungstène	0.511	9	4	5	Michel Pouchard
John Goodenough	0.509	19	4	9	Michel Pouchard
Jacques Lucas	0.472	24	4	4	Jean-Luc Adam
fluor	0.441	12	3	6	J. Flahaut - J. Rivet
verres fluorés	0.421	55	2	23	MM. Poulain
Félix Trombe	0.413	66	4	4	J. Flahaut - J. Rivet
transition métal-isolant	0.409	6	3	3	Denis Jerome
verres fluorés	0.408	55	2	19	Jean-Luc Adam
bronzes de vanadium	0.395	22	2	15	Michel Pouchard
octaèdres	0.375	19	2	12	Michel Pouchard
théorie des bandes	0.330	17	4	4	Guy Ouvrard

du corpus (figure 1). Au contraire, chacune structure un *cluster* de l'école de recherche Collongues : Pérez au laboratoire historique ; Théry pour le laboratoire héritier de Polytechnique.

Comment interpréter une telle configuration ? L'analyse qualitative des témoignages suggérait déjà une rivalité profonde et durable entre les deux lieutenantes du professeur. Une telle rivalité a déterminé des stratégies différentes dans les choix épistémiques durant leur carrière, commencée dans les années 1950, et longtemps après, lors de l'interview (2004). Cette stratégie d'évitement diminue d'autant la probabilité de croisements thématiques, d'événements partagés, de rencontres interpersonnelles malgré l'appartenance à un même laboratoire durant toute une carrière. Lors de l'interview, chacune a ainsi peu de raison de parler des préoccupations de sa rivale. Cette interprétation qualitative est confirmée par la liste des expressions extraites communes aux deux interviews présentée dans la table 5. Les mots-clés communs les plus importants (alumine, réfractaires, Vitry-sur-Seine) sont des expressions courantes de l'école Collongues (voir p. 154 et suivantes). Ceci explique la faible intensité du lien entre les entretiens de Théry et Pérez.

Figure 2 – Photographie de Robert Collongues et ses six « maîtresses de recherche », novembre 1959 (Archives personnelles de Hélène Mondange)

3) Rapports de génération et de genre

Le troisième lien concerne les acteurs les plus anciens, repérés par un cerclage noir sur la figure 1, qui appartiennent à la génération qui a institutionnalisé la chimie du solide à partir des années 1950. Il montre une différence significative en termes de genre. Les hommes, qu'ils soient de puissants mandarins (Friedel, Hagenmuller) ou des professeurs moins influents (Flahaut, Hérold) sont éloignés les uns des autres dans le réseau. Leur ancienneté dans le champ social peut alors être vue comme cause de peuplement de leur environnement par des chercheurs plus jeunes, qu'ils ont notamment dirigés en thèse. Ils se trouvent ainsi éloignés les uns des autres. En revanche, les femmes restent proches les unes des autres qu'elles soient rivales (Pérez, Théry) ou pas (Mondange, Pérez). Cette fois-ci, l'ancienneté ne structure pas l'environnement social. Ceci est d'autant plus étonnant que si Mondange n'a pas occupé de position de pouvoir, Pérez et Théry furent chefs d'équipe et encadrèrent les thèses de doctorat du laboratoire Collongues. Ceci n'a, semble-t-il, pas suffi à garder

Table 5 – Liste exhaustive des mots-clés liant M. Pérez et J. Théry. La colonne « Occ » indique le nombre total d'occurrences d'un mot-clé dans le corpus, les colonnes « in A » et « in B » le nombre d'occurrences dans l'interview de Théry et Pérez.

Mot-clé	Poids	Occ	in A	in B
alumine	0.414	30	4	4
matériaux réfractaires	0.197	23	3	3
Daniel Vivien	0.093	28	2	2
Vitry-sur-Seine	0.062	50	3	6
Monique Pérez	0.052	11	1	1
ferrites	0.052	10	1	1

les jeunes générations dans une relation de dépendance sociologique ou épistémique vis-à-vis de ces chercheuses plus âgées.

En effet, dans les années 1950 et 1960, les femmes « n'étaient pas prises au sérieux » comme le rappelle Mondange dans son entretien. Les expressions « maîtresses de recherche » pour désigner, à l'époque, les collaboratrices de Collongues (titre de la figure 2) et « pères fondateurs » pour désigner, encore aujourd'hui, Hagenmuller et Collongues, montrent la position différenciée des hommes et des femmes dans le champ social des sciences de la deuxième moitié du vingtième siècle. Il y aurait donc là matière à approfondissement pour formaliser les articulations entre division du travail et rapports de génération et de genre dans le champ scientifique.

Approche heuristique des humanités numériques

Jusqu'à présent, nous avons commenté une représentation numérique du corpus (figure 1) grâce à des tables explicitant des listes d'expressions et des intensités de lien et à notre connaissance historique. L'artefact numérique donnait une perspective différente sur le corpus. Nous inversons ici la démarche en générant des représentations quantitatives à partir d'un questionnement *a posteriori* afin de tester les possibilités heuristiques de l'analyse numérique pour les historiens. Le travail historique a montré que la communauté française de chimie du solide s'est construite sur trois éléments forts : une rivalité identitaire, un apport (instrumental et théorique) de la physique, une structuration par les financements industriels (Teissier, 2014). Nous aborderons par la suite les deux premiers éléments de construction de la communauté, identitaire et disciplinaire,

en essayant d'évaluer l'apport de la méthode quantitative dans chacun des deux cas.

1) Interroger la structuration identitaire d'une communauté scientifique

L'émergence d'une communauté scientifique est induite par la construction d'une identité nouvelle, basée sur des organisations sociales et des perceptions psychologiques. Dans le cas de la chimie du solide en France, la mémoire collective est polarisée autour de l'opposition entre deux mandarins, Hagenmuller et Collongues, qui sont encore considérés aujourd'hui par de nombreux chimistes du solide comme « les deux pères fondateurs » de la discipline. Ce trait constitutif de la mémoire collective peut-il être visualisé ? Nous interrogeons le corpus en demandant à Haruspex de dénombrer, pour chaque entretien, les occurrences des expressions « Hagenmuller » et « Collongues » et nous visualisons la répartition obtenue sur la figure 3.

Celle-ci met en évidence un équilibre des populations entre les deux mandarins, ce qui confirme l'importance de cette polarité dans la mémoire collective. Le noyau dur de chacun des deux camps (interviews qui citent beaucoup plus un mandarin que l'autre) se situe près du mandarin, mais entre les deux : Kahn, Serreau pour Collongues ; Pouchard, Ouvrard pour Hagenmuller. Les nœuds proches d'un mandarin mais tournés vers l'extérieur (interviews qui ne citent qu'un mandarin) sont soit étrangers à la communauté (Fauvarque, Jérôme), soit marginaux dans la communauté par leur génération (Mondange) ou leur thèmes de recherche (Adam, Zarzycki). La ligne verticale centrale permet de visualiser les représentants des autres écoles de recherche (Caro, Flahaut, Rousset, Vitorge), ne prenant position ni pour l'un mais citant également les deux. Ainsi, la figure 3, issue d'un questionnement, affine l'analyse globale de la figure 1, confirme l'interprétation historique par des données quantitatives et ouvre des perspectives d'interprétation nouvelle sur les raisons de certains positionnements.

2) Sélectionner les registres sémantiques de la mémoire collective

Une communauté scientifique comme la chimie du solide se définit aussi par des thèmes de recherche (étude de la relation structure-propriété), des outils matériels (diffraction des rayons X) et théoriques (théorie des bandes), des objets d'étude (cristaux inorganiques). Nous construisons une nouvelle représentation numérique du corpus dans

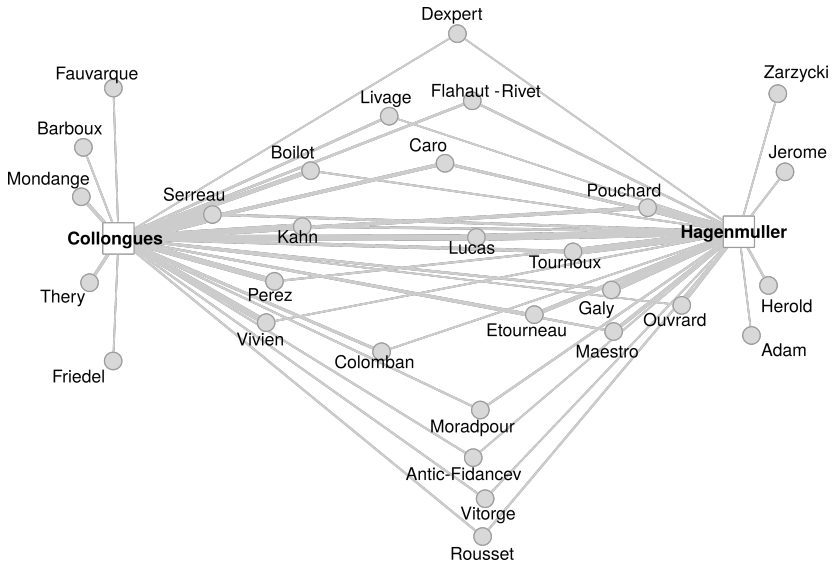

Figure 3 – Répartition des entretiens en fonction du nombre d'occurrences de Collongues ou Hagenmuller. Leur position, relative au nombre d'occurrences de Collongues et/ou Hagenmuller, est située par rapport aux deux pôles carrés. L'entretien de Hagenmuller a été retiré du corpus pour cette analyse.

sa globalité en sélectionnant uniquement les mots-clés liés au registre épistémique (disciplines, thèmes, outils, objets, etc.). Ceci nécessite une supervision manuelle des 1169 expressions-clés (étape d'enrichissement des expressions). Bien que laborieuse, cette étape ne peut probablement pas être automatisée à cause de potentielles erreurs d'indexation. La figure 4 fait ainsi apparaître une sous-structure épistémique du corpus global. Ont été écartées les relations entre nœuds des registres identitaires (noms propres notamment) et institutionnels (organisation des laboratoires, interactions industrielles). La cartographie épistémique rebat les cartes par rapport à la représentation globale du corpus en figure 1. Elle ne structure plus le corpus selon la polarisation gauche/droite des deux mandarins (Hagenmuller/Collongues) et dissout le *cluster* périphérique centré sur les industries automobile et électrique. En d'autres termes, en surprenant l'historien trop habitué à son corpus, elle le pousse à modifier ses habitudes de pensée.

La cartographie épistémique conduit à un réseau plus dense, plus régulier et plus éclaté dans la distribution des nœuds que la cartographie

globale. Son interprétation *a posteriori* permet de dégager trois ensembles d'observations. Premièrement, il est possible de retrouver des regroupements par école de recherche, quoique de manière moins saillante. Ainsi, en haut à droite, un *cluster* Collongues est regroupé par l'évocation de l'alumine, composé-phare du laboratoire des années 1960 et 1970 : Boilot, Colomban, Kahn, Perez, Thery, Vivien. Mais, trois membres de l'école ont été éloignés : D. Serreau, secrétaire du laboratoire, et Barboux, chercheur héritier, ont des intensités de lien trop faibles pour être liés au réseau (en bas au centre). De même, Mondange, qui a fait sa carrière à l'ombre de Collongues et de leur maître à tous deux, Georges Chaudron (chimie métallurgique), se trouve repositionnée dans un *cluster* métallurgie, entre physique et chimie, à gauche de la figure (Catonné, Friedel, Pouchard, Priester). Les relations de type école ou discipline sont amoindries devant les objets et thèmes de recherche : « acier », « dislocation » et « fer » pour le *cluster* métallurgie. Apparaissent aussi des thèmes marginaux de la chimie du solide. Lié au *cluster* métallurgie, on repère le *cluster* des solides faiblement organisés (grains, morphologie, sels, surfaces) par rapport à la norme dominante des solides cristallins (Bonino, Catonné, Fauvarque, Lefrant, Rousset, Zarzycki). L'exemple le plus emblématique, présentant les liens les plus forts en triangle, à droite (Adam, Lucas, MM. Poulain), est le laboratoire rennais d'étude des verres à base d'éléments autres que l'oxygène (« fluorés », « chalcogénures », « lanthanides »). Ce triangle est relié à plusieurs chimistes du solide étudiant les cristaux à base des mêmes éléments (Caro, Flahaut et Rivet, Ouvrard, Tournoux).

Deuxièmement, si les *clusters* par école de recherche sont démembrés ou affaiblis au profit des regroupements épistémiques (objets, thèmes), les modifications de relation entre sphères académiques et industrielles sont plus surprenantes encore : le *cluster* périphérique (automobile et énergie) est reconfiguré autour de deux « chaînes » distinctes, c'est-à-dire des enchaînements de trois ou quatre nœuds fortement liés par paires. La première chaîne est visible en bas à droite : Griesemann (pile à combustible chez Renault) est lié à Lucchese (énergie renouvelable au CEA), qui est lié à Maestro (développement de pigments chez Rhodia), lui-même associé à Caro (luminescence fondamentale au CNRS). La seconde chaîne part de Beuzit (manager chez Renault) puis Lisse (chimiste des piles à combustible chez Citroën) et se termine avec Pouchard (chimie théorique des matériaux électrochimiques à l'université). Ces deux chaînes parcourent toutes deux un chemin géographique de la périphérie vers le

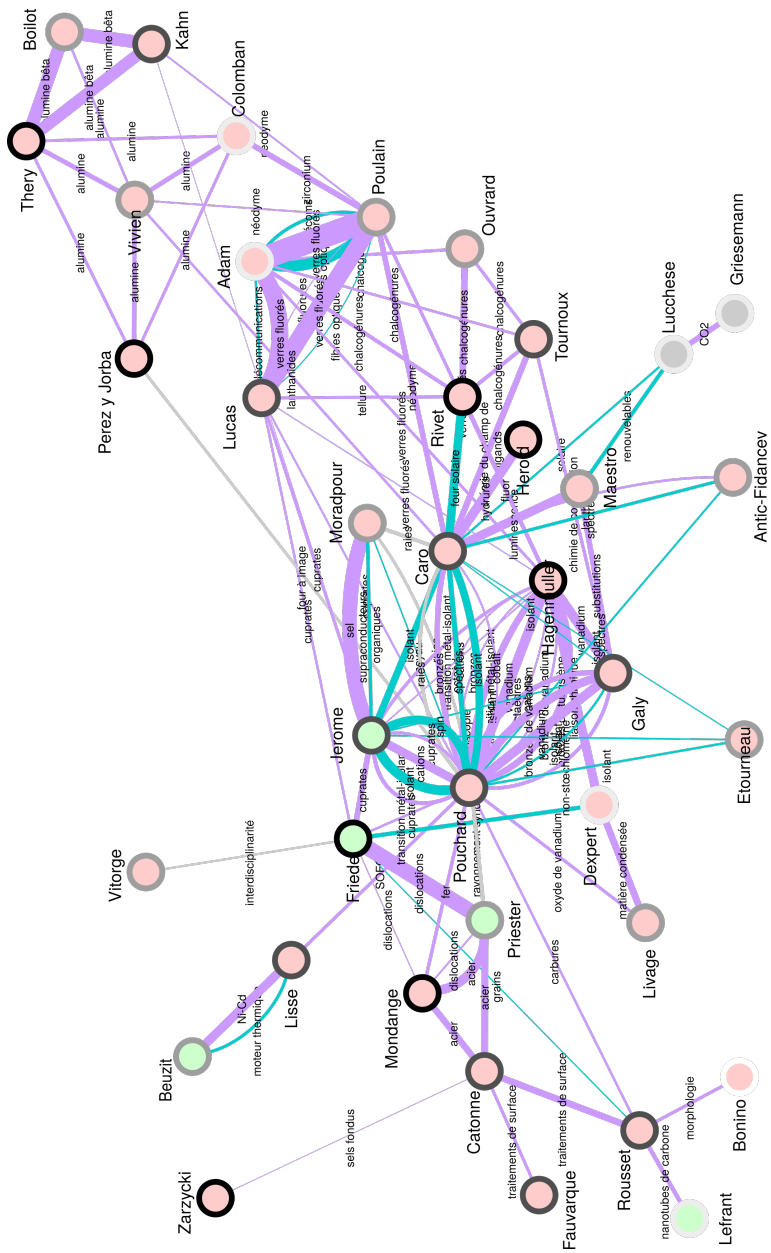

Figure 4 – Interviewés évoquant des sujets scientifiques : physique, chimie ou indifférencié (sciences). En dessous d'un seuil de pondération (0.32), les liens ne sont pas affichés. Les nœuds roses sont associés aux chimistes, les verts aux physiciens, les gris aux autres. Les liens roses sont de chimie, les bleus de physique, les jaunes sont indifférenciés. Les cerclages des nœuds sont fonction de la date de thèse : les plus anciens sont plus foncés.

cœur de la cartographie et un chemin professionnel du commercial et industriel vers le fondamental. La transformation des *clusters* en chaînes confirme la porosité des sphères industrielles et académiques dans la mise en place de systèmes d'innovation sur les matériaux. Cela suggère aussi des voies d'étude de ces interactions.

Abordons dans un troisième temps la partie la plus centrale de la figure 4, qui est aussi la plus épineuse à interpréter : un *cluster* dense formé par 5 nœuds au centre (Caro, Galy, Hagenmuller, Pouchard, Jérôme). Il ne s'agit plus d'un effet d'école comme pour la figure 1 puisque le nœud Hagenmuller est moins dominant qu'auparavant et qu'il est peu ou pas lié à certains de ses héritiers directs (Etourneau, Tournoux) ou indirects (Lucas, Ouvrard). De quoi s'agit-il alors ? Trois observations préliminaires avant de répondre : la présence notable d'un chimiste indépendant des deux écoles dominantes (Caro), lui-même lié à trois chimistes plus âgés que lui (Flahaut, Hérold, Tournoux) ; la présence non moins notable d'un physicien du solide (Jérôme de la même génération que Pouchard), lui-même lié à un second physicien (son mentor Friedel) et un chimiste de son laboratoire d'Orsay (Moradpour) ; enfin, des liens entre les nœuds basés sur des références à des structures cristallines (« bronzes », « cuprates »), des éléments chimiques (« vanadium »), des propriétés physiques (« conducteur », « isolant », « raies », « spectres ») et des concepts théoriques (« spin », « transition métal-isolant »). Le *cluster* central fait donc surgir des relations épistémiques fortes malgré les hétérogénéités sociales (physiciens *versus* chimistes, mandarins *versus* outsiders). Il rend compte d'un ensemble épistémique plus large qu'une école de recherche mais différent d'une discipline comme la chimie du solide. L'analyse numérique devrait ainsi permettre d'établir de nouvelles typologies de l'organisation sociale des sciences.

Au terme de cette analyse, la représentation épistémique de la figure 4 permet de dégager deux réflexions d'ensemble. Tout d'abord, au niveau réflexif, la surprise suscitée par une image inattendue et difficile à interpréter constitue un garde-fou méthodologique : elle refrène toute envie de généraliser ou de systématiser trop rapidement. Ensuite, la représentation épistémique augmente et relativise la représentation générale de la figure 1. En effet, elle stimule la réinterprétation en venant compléter, réfuter ou étendre les interprétations antérieures sans pour autant clore le débat. Dans notre cas, la figure épistémique amoindrit la place des écoles de recherche et des mandarins, qui structuraient profondément la figure

générale. La dissolution de la cellule de base de la communauté française de chimie du solide fait alors mieux apparaître des interactions hétérogènes (université/industrie, physique/chimie, outsiders/mandarins), qui ont pu être sous-estimées par la mémoire collective d'abord et par l'interprétation historienne ensuite. Sélectionner les mots-clés par le registre épistémique pourrait ainsi permettre de visualiser de multiples « communautés épistémiques », au-delà des formes les plus courantes : écoles de recherche, disciplines, institutions. Dans la perspective d'un élargissement du corpus au-delà de la chimie du solide, les cartographies épistémiques permettraient de s'affranchir des appartenances disciplinaires pour identifier et caractériser plus largement les aspects épistémiques de la recherche sur les matériaux dans le cas de la France (et ensuite au niveau international). Ceci affinerait la cartographie numérique de la mémoire collective sur les matériaux.

Épistémologie pratique entre les sphères numériques et historiques

Pour finir, nous proposons une réflexion épistémologique sur les humanités numériques basée sur le cas pratique développé dans l'article. Un premier temps est consacré à l'analyse du fonctionnement d'Haruspex. Un deuxième temps envisage, de manière plus large, le rôle que pourraient jouer les outils numériques dans les humanités, notamment l'histoire des sciences et des techniques.

- *Réflexions sur l'application d'Haruspex à l'étude des archives orales*

L'expérience menée permet de proposer une typologie des inférences hommes/machines pour Haruspex puis de schématiser son fonctionnement.

Typologie des interactions hommes/machines

Au terme de ce travail, l'utilisation d'Haruspex peut être définie selon huit étapes successives. Les quatre premières ont déjà été détaillées dans le processus linéaire théorique (p 149) : traitement du corpus, extraction d'expressions, enrichissement des expressions et création de liens. La figure 5 représente l'ensemble des étapes en indiquant les décisions humaines (flèches pleines) et les rétroactions machine (flèche pointillée) sous forme de boucles.

1) Les étapes 1 à 5

Les boucles de retour comptent pour beaucoup dans la robustesse et la finesse d'analyse d'Haruspex. Elles permettent le contrôle de résultats intermédiaires et la rectification du processus pour éviter les résultats aberrants selon une perspective humaine. La première partie de la figure 5 (étapes 1 à 5) comporte deux boucles imbriquées (2-3-3bis et 3-4-5) et une rétroaction (3bis vers 2). La rétroaction indique un apprentissage de la machine grâce à la modération de l'utilisateur humain qui influencera une extraction ultérieure. La première boucle concerne l'extraction d'expressions et la seconde la création de liens. La seconde partie de la figure est plus simple à comprendre car elle comporte une seule boucle (6-7-8).

Ce qui amorce une boucle est une décision humaine (symbolisée par un losange), motivée par la détection d'une *anomalie* dans les résultats produits. Qu'entend-on par là ? Une anomalie représente un *écart de valeur* entre le calcul algorithmique et l'évaluation humaine. Ceci requiert l'intervention d'un spécialiste du corpus. Le repérage d'un écart de valeur entre les sphères numérique et historique permet de *rectifier* les étapes antérieures avant de poursuivre.

Donnons quelques exemples d'anomalies. Au niveau de l'extraction de mots-clés (losange sous étape 3), la machine attribue une grande valeur (fort poids discriminant) à une expression comme « heure de cours » alors qu'elle a, selon le jugement historien, une faible valeur dans le corpus. La suppression d'un certain nombre d'expressions de ce type est suivie par une nouvelle extraction de mots-clés par Haruspex, qui donne une liste plus *signifiante* selon le jugement historien. Au niveau de la création de liens (losange sous étape 5), une anomalie peut être liée à une liaison locale, trop bien notée, à partir d'un mot-clé sans grande signification. Ainsi, « cycle de vie » nous semblait trop bien noté par la machine car l'idée n'était pas centrale dans le contexte de la « recherche sur les matériaux » et pouvait être exprimée de plusieurs façons. Cette seconde modération offre la possibilité de rectifier les causes d'erreur (mots-clés) plutôt que les effets absurdes (liens).

2) Les étapes 6 à 8

Au niveau de l'interprétation de graphes (losange sous étape 8), plutôt que d'anomalie, nous parlerons d'heuristique interprétative. En effet, l'interprétation de la figure 4 nous a conduit à fabriquer les figures 3

et 1. Cette boucle peut être parcourue un nombre indéfini de fois durant le processus de recherche. C'est aussi ce genre de considérations qui nous a permis d'écarter certaines interviews qui faussaient la lecture du corpus : lorsque nous disposions de deux entretiens pour un même témoin (Pouchard, Caro), nous n'en avons gardé qu'un pour éviter que les deux interviews d'une même personne forment une paire de nœuds trop fortement liés. Ce choix limitant le corpus pourrait être figuré par une boucle partant du troisième losange vers la première étape.

L'analyse typologique du fonctionnement d'Haruspex (figure 5) montre deux caractéristiques fondamentales de notre méthode : sa non-linéarité, ce qui rend indispensables des rencontres répétées et des discussions contradictoires entre informaticiens et historiens ; ses possibilités de *rectifications numériques* à partir de *jugements de valeurs historiques*. Ce dernier point montre clairement l'incommensurabilité des sphères numériques et historiques dans *l'évaluation de la valeur des relations*.

Schéma de fonctionnement d'Haruspex

Plutôt linéaire et chronologique sur la représentation précédente (figure 5), la typologie d'Haruspex peut être redéployée suivant deux autres variables : en abscisse, la part relative de travail humain et de travail machine ; en ordonnée, la part relative de fabrication de données et de savoirs (figure 6). Une telle représentation rend compte d'une conception continue et symétrique reliant données numériques et savoirs historiques[8].

1) Place de la machine dans la méthode

Le quart supérieur droit de la figure 6 est vierge. La conception d'Haruspex cantonne la machine aux inférences « bas-niveau » produisant de nouvelles données, éventuellement de nouvelles informations, mais ne livre pas de nouvelles connaissances ni n'établit de nouveaux savoirs. Les inférences « haut-niveau », de production de savoir, restent aux mains des humains. En donnant une préférence au *close reading*, Haruspex constitue une alternative aux approches *distant reading* de modélisation totale et *a priori* d'un domaine.

[8] Nous nous inscrivons dans une version continue du paradigme de la pyramide DIKW (*Data Information Knowledge Wisdom*, en français : données, informations, connaissances et savoirs) (Rowley, 2007).

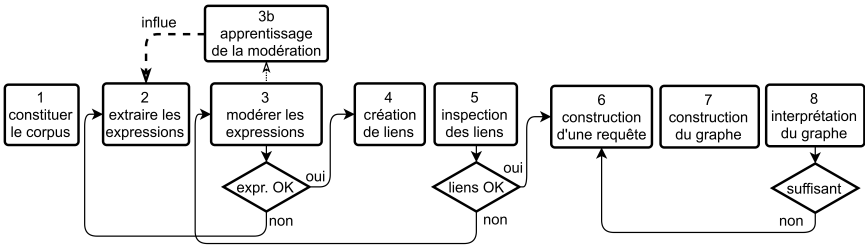

Figure 5 – Typologie des opérations d'Haruspex

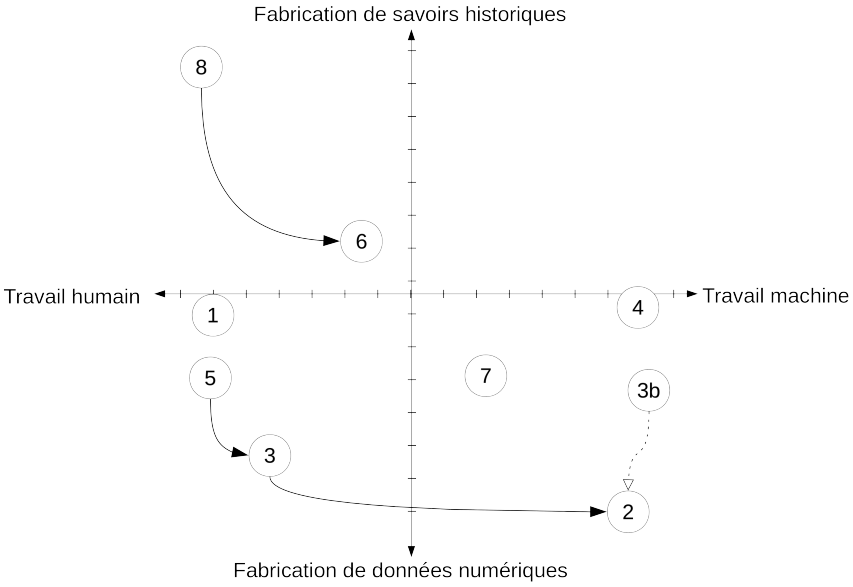

Figure 6 – Diagramme des opérations suivant le type de travail fourni et le type de production obtenue

2) Interactions

Les multiples allers-retours entre les parties droite et gauche de la figure 6 (de 1 à 2, puis de 2 à 3, puis 3 à 4, etc.) rendent compte de nombreuses interactions entre l'homme et la machine. Ces déplacements horizontaux sont couplés à des déplacements verticaux (de 1 à 2 ou 5 à 6 puis de 6 à 7 par exemple). Ainsi, les étapes de raisonnement alternent avec les étapes de récolement de données. Ces interactions s'opèrent

selon une tendance ascendante jusqu'à l'interprétation historienne de représentations quantitatives et graphiques du corpus (étape 8).

3) Rectifications et heuristique

Les trois boucles présentées précédemment ont la même forme : elles sont issues d'un travail humain et dirigées vers une étape précédente (indice inférieur), de travail davantage machine et fabrication de données. L'humain demande à la machine de rectifier les données avant de procéder à l'étape suivante de production de savoir.

- *Vers une approche interdisciplinaire des humanités et des sciences numériques*

Cette dernière section interroge le rôle épistémologique que pourrait avoir l'analyse numérique pour les sciences humaines, et plus particulièrement l'histoire des sciences et des techniques.

Expliquer et comprendre

Notre cas d'étude souligne le problème épistémologique fondamental de l'incommensurabilité des données numériques et des discours humains, marqués par l'incomplétude de leurs significations.On retrouve la différence introduite par Wilhelm Dilthey entre les sciences de la nature et les sciences de l'homme et de la société : « Nous expliquons la nature, nous comprenons la vie psychique ». L'explication des sciences naturelles tient surtout à la formulation d'un déterminisme de type causal, rendue possible par la quantification des objets naturels, notamment grâce aux outils mathématiques, numériques et expérimentaux. Les sciences humaines et sociales sont aussi orientées vers la recherche d'explications causales, sans possibilité d'expérimentation toutefois, ainsi que vers la formulation d'une compréhension des acteurs humains.

Cette intentionnalité humaine introduit des « effets de sens » dans les sciences humaines et sociales qui les rendent irréductibles aux seules explications causales. Les moments de rectification permis par Haruspex (p. 167) montrent cette incommensurabilité des sciences sciences numériques et humaines. L'appréhension des humanités par les sciences numériques ne doit donc pas être comprise comme la réduction du champ social au champ numérique mais plutôt comme l'introduction d'un ins-

trument supplémentaire (mais jamais suffisant) pour explorer le champ social.

Rôle instrumental des outils numériques

Pour l'analyse textuelle, Margaret Masterman présente l'outil numérique comme « un télescope pour l'esprit » (*a telescope for the mind*) (Masterman, 1962). Nous allons plus loin en défendant, avec Alfred N. Whitehead parlant des sciences instrumentales, que l'outil numérique stimulant l'imagination des chercheurs « montre les choses selon des combinaisons inhabituelles ». Plus qu'un grossissement télescopique, ceci provoque « une transformation » de l'objet d'étude[9]. Ainsi, chacune des représentations numériques a transformé notre perspective d'étude du corpus à la base de nouvelles interprétations.

La question est néanmoins de savoir si notre méthodologie fabrique une cartographie fiable ou une juxtaposition d'artefacts *ad hoc*. Trois critères de fiabilité ont guidé notre démarche : (1) l'*interdisciplinarité*, par laquelle la construction collective de représentations et d'interprétations fait écho à la singularité des regards individuels et disciplinaires ; (2) la *dualité* des productions numériques, qui sont simultanément signifiantes (image d'un objet d'étude) pour les historiens et signifiés (objet d'étude en soi) pour les numériciens ; et (3) la *sensibilité* d'Haruspex, dont les résultats évoluent à mesure que le corpus est modifié.

Au triple garde-fou méthodologique, il convient d'ajouter une souplesse conceptuelle et instrumentale. Nous suivons les leçons de sémiologie graphique quant au caractère nécessairement dynamique des représentations à produire.[10] Néanmoins, nous ne pensons pas qu'une cartographie sémantique, aussi fiable soit-elle, épuise la multiplicité des interprétations. L'instrument numérique multiplie les échelles d'observation et les artefacts signifiants mais n'épuise jamais le travail philologique, qui, sans cesse, peut proposer de nouvelles interprétations. Jusqu'à présent, et malgré la

[9] « The reason we are on a higher imaginative level is not because we have a finer imagination, but because we have better instruments. [...] a fresh instrument serves the same purpose as foreign travel; it shows things in unusual combinations. The gain is more than a mere addition; it is a transformation » (cité par Ihde, 2009).

[10] « On ne "dessine" plus un graphique une fois pour toutes. On le "construit" et on le reconstruit (on le manipule) jusqu'au moment où toutes les relations qu'il recèle ont été perçues » (Bertin, 1983).

multiplication de représentations fabriquées, nous n'avons pas trouvé de résultats aberrants.

Malgré ces indices de fiabilité, nous sommes conscients de nombreuses limites des méthodes numériques. Haruspex repère et fabrique la totalité de liens binaires formés entre deux nœuds du corpus à partir de la liste d'expressions spécifiques validées. Mais, il est absurde de croire que cet ensemble de relations numériques calculées englobe la totalité des connaissances historiennes qui peuvent être déduites de ce même corpus. Trois arguments au moins le montrent. Le premier est numérique : seul un sous-ensemble de cette totalité numérique est visualisé par une requête. Le deuxième est linguistique : il existe des complexes signifiants qui incorporent plus de deux éléments, et ce sont sans doute les plus nombreux. Le troisième est historique : le corpus s'inscrit dans un contexte qui le contient et dont il ne représente qu'une infime partie par la taille et par la nature (toute réalité n'est pas discursive) alors que le travail historique va et vient entre corpus et contextes.

Administration de la preuve

Haruspex fabrique une représentation numérique du corpus avec un bon degré de précision et de fiabilité sémantique. Il dessine une cartographie de la mémoire collective en traçant des inhomogénéités sociales et épistémiques. Ce faisant, il joue, pour les sciences de l'homme, un rôle instrumental comparable aux mesures expérimentales pour les sciences de la nature.

Pourtant aucun équivalent des méticuleuses calibrations des appareils de mesures des sciences de la nature n'existe pour les sciences de l'homme. Certes, des corpus de test existent en traitement automatique de la langue (TAL)[11] ainsi que des thésaurus et d'autres taxonomies lexicales. Ces corpus sont destinés à l'analyse de textes « tout-venant » dans une perspective *distant reading* d'une masse de données textuelles.

La calibration d'outils d'analyse, en revanche, fait défaut. Ceci rend indispensable la connaissance préalable d'un corpus. L'administration de la preuve en humanités numériques articule interprétations, représentations

[11] Les étalons disponibles, qu'ils soient anglophones (MUC, ACE, DUC) ou francophones (ESTER, ester2) sont des corpus annotés permettant d'évaluer la performance d'une extraction d'entités nommées, initialement prévus pour l'armée, les renseignements.

et quantifications de corpus. Les commentaires critiques de tiers et les développements de méthodes alternatives par d'autres équipes constituent des rouages-clés de cette machinerie argumentative complexe sur laquelle s'appuient les humanités numériques.

Conclusion et perspectives

La conclusion reprend les trois perspectives présentées dans le résumé : pragmatique, heuristique et réflexive.

Tout d'abord, la dimension pragmatique concerne la fiabilité de la méthode numérique d'analyse du corpus. Cette approche n'a pas donné lieu à des résultats aberrants. Au contraire, la représentation globale du corpus de la figure 1, pour laquelle la seule intervention a consisté à fixer des paliers de visualisation, a pu être interprétée de manière conforme à notre connaissance qualitative du corpus et historique des communautés scientifiques correspondantes. Ceci est aussi le cas pour l'analyse spécifique de la sous-structure épistémique de la figure 4. Notre méthode numérique semble donc adaptée à l'analyse de corpus textuels de sciences humaines et sociales. Ceci ne signifie pas pour autant que toute représentation puisse trouver une signification, que toute interprétation puisse être visualisée ou que toute interprétation soit univoque.

Ensuite, notre méthode peut ouvrir des perspectives de recherche sur les archives orales. L'heuristique se joue au moins à trois niveaux. Premièrement, l'interprétation de relations *surprenantes* permet de braquer le regard sur un angle mort ou d'ouvrir une voie non explorée. La possibilité de dé-corréler les registres de langage, organisant les expressions en catégories (disciplinaires, identitaires ou organisationnelles) est une voie intéressante pour faire apparaître des relations insoupçonnées : *clusters*, chaînes, paires de nœuds et nœuds isolés. Deuxièmement, le tracé de nouvelles représentations suite à un questionnement historique peut confirmer quantitativement des résultats qualitatifs. Le clivage de la mémoire collective entre deux « pères fondateurs » (figure 3) a fourni un cas d'étude satisfaisant. Le filtrage des mots-clés par des registres sémantiques est prometteur, notre essai pour le registre épistémique s'étant avéré concluant. Troisièmement, une piste plus porteuse encore nous semble être la comparaison entre la structure globale du corpus (figure 1) et l'une de ses sous-structures, notamment épistémique (figure 4). Ceci

pourrait modifier l'usage et la signification de l'outil numérique Haruspex : plus que dans l'interprétation d'une image statique, l'heuristique pourrait se trouver plus fondamentalement dans l'écart entre deux images numériques, c'est-à-dire dans les relations entre différentes structures mémorielles.

Enfin, notre pratique réflexive nous a appris que l'interaction entre le numérique et les humanités est d'autant plus efficace que numériciens et historiens peuvent dialoguer librement et à égalité. De telles interactions interdisciplinaires, constructives et critiques, façonnent d'ailleurs le meilleur garde-fou contre de possibles débordements d'ordre numérique (car à peu près n'importe quoi peut être visualisé) ou historique (car à peu près n'importe quoi peut être expliqué). Cet article marque une étape dans le programme de recherche consacré aux archives orales concernant la « recherche sur les matériaux » depuis les années 1940. Il était indispensable, pour développer Haruspex et tester sa fiabilité sur les archives orales, de choisir un sous-corpus de faible taille, bien connu et malgré tout hétérogène.

À mesure que l'analyse des archives orales sera élargie à d'autres domaines que la chimie du solide, la connaissance qualitative des sous-corpus s'amoindrira ou sera fragmentée, chaque sous-corpus ayant été constitué par des chercheurs différents. Or, nous avons compris, avec ce premier cas d'étude, à quel point une appréhension intime des textes mémoriels et des méthodes numériques était déterminante. La cartographie de la mémoire collective sur les matériaux ne se fera donc pas sans mal. Elle ne se fera pas, quoiqu'il en soit, sans un travail collectif, empirique, durable et interdisciplinaire. Loin des standards des humanités numériques et des méthodes de cartographie de *big data*, Haruspex esquisse une voie étroite, modeste certes, mais résolument humaine et stimulante.

Remerciements

Les auteurs voudraient remercier les deux rapporteurs anonymes de l'article pour leurs critiques et propositions stimulantes ainsi que la rédactrice en chef et la secrétaire de rédaction des *Cahiers François Viète* pour leur travail éditorial, leurs relectures patientes et leurs suggestions pertinentes. Enfin, les auteurs souhaiteraient remercier les personnes in-

terviewées, qui, par leurs témoignages, ont rendu possible la constitution du corpus d'étude.

Références

BENSAUDE-VINCENT Bernadette & TEISSIER Pierre (2015), « Building, Preserving and Using Oral Archives on Materials Research. An Attempt towards the Biography of Research Communities », Communication présentée au Congrès international *10th International Conference on the History of Chemistry (ICHC) - Chemical Biography in the 21st Century*, Aveiro, Portugal, halshs-00183252.

BERTIN Jacques (1983), *Semiology of Graphics*, vol. 94, Madison, University of Wisconsin Press, DOI : 10.1103/PhysRevLett.94.208902.

ENGUEHARD Chantal (1993), « Acquisition de terminologie à partir de gros corpus », dans *Informatique & langue naturelle*, ILN'93, p. 373–384, http://pagesperso.lina.univ-nantes.fr/info/perso/permanents/enguehard/recherche/ana/iln.htm.

FINLAYSON Mark Alan & KULKARNI Nidhi (2011), « Detecting Multi-Word Expressions Improves Word Sense Disambiguation », *Multiword Expressions: from Parsing and Generation to the Real World*, Stroudsburg, Association for Computational Linguistics, p. 20–24.

HOCKEY Susan (2007), « The History of Humanities Computing », dans Susan SCHREIBMAN, Ray SIEMENS & John UNSWORTH (éds.), *A Companion to Digital Humanities*, Oxford/Malden, Blackwell Publishing Ltd, p. 1–19, DOI : 10.1002/9780470999875.ch1.

IHDE Don (2009), *Postphenomenology and Technoscience: The Peking University Lectures*, Albany, SUNY Press, p. 5–23.

JAIN A. K., MURTY M. N. & FLYNN P. J. (1999), « Data clustering: a review », *ACM Computing Surveys*, vol. 31, n° 3, p. 264–323, DOI : 10.1145/331499.331504.

MASTERMAN Margaret (1962), « The Intellect's New Eye », *The Times Literary Supplement*, vol. 284, p. 38–44.

MILNE David & WITTEN Ian H. (2008), « Learning to Link with Wikipedia », dans *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, New York, Association for Computing Machinery, p. 509–518, DOI : 10.1145/1458082.1458150.

MORETTI Franco (2005), *Graphs, Maps, Trees : Abstract Models for a Literary History*, London/New York, Verso.

QUANTIN Matthieu (2018), *Proposition de chaînage des connaissances historiques et patrimoniales*, Thèse de doctorat, École Centrale de Nantes.

QUANTIN Matthieu, HERVY Benjamin, LAROCHE Florent & BERNARD Alain (2016), « Supervised Process of Un-structured Data Analysis for Knowledge Chaining », dans Lihui WANG (éd.), *CIRP Design*, vol. 00, Stockholm, Elsevier, DOI : 10.1016/j.procir.2016.04.123.

ROWLEY Jennifer (2007), « The Wisdom Hierarchy: Representations of the DIKW Hierarchy », *Journal of Information Science*, vol. 33, nº 2, p. 163–180, DOI : 10.1177/0165551506070706.

SALTON Gerard (1983), *Introduction to Modern Information Retrieval*, New York, McGraw-Hill, Inc.

SCHREIBMAN Susan, SIEMENS Ray & UNSWORTH John (éds.) (2004), *A Companion to Digital Humanities*, Malden, Blackwell Publishing Ltd.

SHANNON Claude Elwood (1948), « A Mathematical Theory of Communication », *The Bell System Technical Journal*, vol. 27, nº 3, p. 379–423, DOI : 10.1002/j.1538-7305.1948.tb01338.x.

TEISSIER Pierre (2007), *L'Émergence de la chimie du solide en France (1950-2000) : de la formation d'une communauté à sa dispersion*, Thèse de doctorat, Université Paris 10.

TEISSIER Pierre (2014), *Une histoire de la chimie du solide. Synthèses, formes, identités*, Paris, Hermann.