Haruspex, Outil de Gestion de Connaissances non Structurées

Haruspex, Knowledge Management Tool for Unstructured Data

Matthieu Quantin[1, 2], Benjamin Hervy[1], Florent Laroche[1], Jean-Louis Kerouanton[2]

[1] IRCCyN, UMR_CNRS_6597, École Centrale de Nantes, matthieu.quantin@ec-nantes.fr, benjamin.hervy@univ-nantes.fr, Florent.Laroche@ec-nantes.fr
[2] CFV, EA_1161, Université de Nantes, Jean-Louis.Kerouanton@univ-nantes.fr

RÉSUMÉ. L'objet de cette communication est de proposer une méthode pour l'analyse et l'exploitation de corpus de documents non-structurés ou faiblement structurés. Le terme non-structuré se réfère au concept informatique de données non-décrites, non-marquées explicitement. Aujourd'hui la création de corpus de données numériques (ouverts ou privés) est un phénomène massif. Toujours plus de données sont scannées, photographiées, retranscrites, etc pour être analysées. Les jeux de données (numériques) constituent la matière exclusive, quotidienne du chercheur. Ces jeux de données sont souvent construits spécialement pour les besoins du projet voire collectés par le chercheur lui-même. Ce phénomène demande à être accompagné par une évolution des outils d'analyse: données physiques et données numériques ont des potentiels d'analyse différents. Or le chercheur en SHS est souvent démuni face aux sources non structurées qu'il collecte: articles, scan d'archives, documents OCR, images et métadonnées. La mise en place d'une base de données se résume souvent (au mieux) à un « tableau excel ». Les domaines du *bigdata* et du *data-mining* sont cantonnés à des projets de très grande envergure, pour des données déjà structurées, avec une équipe de soutien logistique conséquente. Un fossé se creuse entre le chercheur en histoire, en archéologie, en sociologie et les « humanités numériques ».
L'outil proposé, intitulé *Haruspex*, vise à réduire ce gap. Il traite des données texte (et images éventuellement) en français ou en anglais, pour produire une base de données orientée graphe, requêtable, contenant les documents liés entre-eux (proximité sémantique). En entrée, divers formats (pdf, txt, odt, latex…) sont pris en charge, le processus se déroule ensuite en 4 étapes :
1. Gestion de corpus: création ou récupération d'éventuelles métadonnées (dates, lieux, étiquetage) pour les documents; concaténation, découpage, regroupements, exclusion, …
2. Indexation sémantique de ce corpus: extraction de mots clés (génériques mais aussi très spécifiques), puis classification de ces mot-clés en catégories (si possible).
3. Modération des résultats précédents par l'utilisateur.
4. Calcul de la « distance sémantique » entre documents à partir de l'indexation modérée.
Les premiers essais dans divers domaines – patrimoine industriel, histoire de la chimie au XXe siècle, histoire du travail dans les colonies et analyse des publication scientifiques – sont concluants aux yeux des chercheurs du domaine concerné.
ABSTRACT. This study presents a method designed to analyse and tap corpus made of unstructured or weakly structured documents. The term *structured* refers to a computer point of view, and means non-described, non explicitly marked up data. Nowadays, digital (open, or private) corpus creation is a massive trend. More and more data is being scanned, photographed, faithfully transposed, etc. to be analysed (among other uses). Digital data set is the exclusive material, daily handled by the researcher. These sets are often specifically designed for a project, even collected by the researcher himself. This trend needs to be accompanied by analytic tools. Actually physical and digital data have different potentials of analysis. Yet, the researcher in humanities often remains powerless facing the unstructured data he collects: articles, scan of archives, OCR documents, media and their metadata. Deploying a database is often limited to an "excel sheet" or some few SQL tables. Big data and data-mining technologies are restricted to large scale project, for already structured text, with a significant IT support team. This opens the gap between historians, archaeologist, sociologist and the "digital humanities".
This tool, nammed *Haruspex*, aims at closing this gap. It processes textual data, eventually combined with pictures, written in french or english, and outputs a graph oriented database. This database contains interlinked documents (semantic closeness). As inputs, several formats (pdf, txt, odt, latex …) are supported. The process is ran through 4 steps:
1. Corpus management: create or extract eventual metadata (date, place, tags) for each document; manipulate them: concatenate, split, gather, exclude…

2. Semantic indexing of the corpus: keyword extraction (generic but also specific) and classification of these keyword in categories (if possible).
3. Results monitoring by the researcher.
4. Computing the “semantic closeness” between documents from the monitored keywords.
First tests of *haruspex* concern several fields of study: shipyards industrial heritage , history of chemistry in the xxth century, labour history in french colonies and contemporary scientific publications studies. These tests convinced the concerned researchers.
MOTS-CLÉS. Graphe, indexation, proximité sémantique, corpus, texte non-structuré.
KEYWORDS. Graph, index, semantic closeness, corpus, unstructured text.

Constat

Contexte de recherche

Le groupe de recherche EPOTEC[1], à la confluence des sciences pour l’ingénieur et de l’histoire des techniques, développe et utilise des outils numériques pour les sciences de l’homme. Ce groupe est composé de membres de l’IRCCyN[2] et du CFV[3] qui collaborent sur des projets liés au patrimoine industriel et technique, et produisent une réflexion épistémologique sur les humanités numériques. Initialement axé sur la modélisation 3D de patrimoine technique: machine à laver le sel de Batz-sur-Mer, poudrerie royale de S^{t} Chamas, Canot DCNS, … l’interdisciplinarité a poussé le groupe à questionner la relation entre sémantique et 3D. À cet égard, le projet *Nantes1900* est emblématique, il couple une base de gestion documentaire à un dispositif de réalité augmenté pour la maquette du port de Nantes en 1900. Dans cette lignée, les liens entre captation de données (3D ou textuelles), analyse et valorisation sont renforcés par ce travail.

Hypothèse, objectifs et enjeux

Les outils de l’historien ont peu évolué depuis l’informatisation des contenus. Or ces contenus n’ont pas la même nature ni les mêmes potentialités: une page web n’a que peu de points communs avec une page papier.

Haruspex est une chaîne de traitement de corpus textuels (gérant les multimédia comme des annexes), mettant en relation les éléments de ce corpus en fonction de leurs contenus. Des *proximités* entre les documents sont calculées. Ainsi 2 élément du corpus qui traitent de la même chose seront proches, un réseau d’information (graphe) est construit (figure 1).

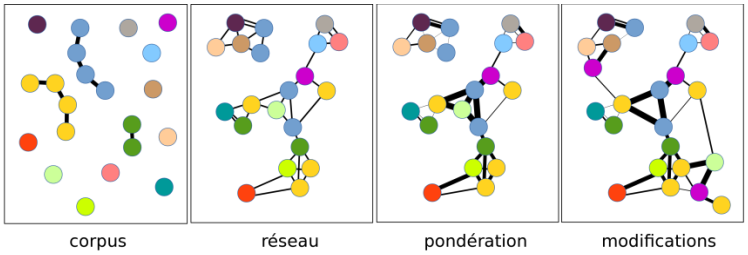

Figure 1. *Illustration des étapes du passage d'un corpus de document vers un réseau d'informations pondérées (crédit : Matthieu Quantin. CC-BY-NC-ND)*

[1] Évolution des Procédé et Objets TEChniques (epotec.fr)

[2] Institut de Recherche en Communications et Cybernétique de Nantes

[3] Centre François Viète

Les contraintes à respecter, issues des sciences humaines, différencient ce travail d'autres approches :

– **Sans modèle préalable** : réfuter toute possibilité de créer *a priori* des catégories de connaissances: ni entités, ni relations, ni vocabulaires externes au corpus. Les standards de description (ontologies (Doerr, 2003) et *linked open vocabularies*[4] du web sémantique) couplés à de puissants thésaurus comme Pactols (Nouvel & Rousset, 2015) offrent l'opportunité (insuffisamment saisie en sciences humaines) de partager des jeux de données, d'exposer des travaux ou de collaborer sur des collections importantes (approche *bigdata*). Cependant, ces technologies ne permettent pas de saisir la complexité du vocabulaire d'un corpus que l'on considère comme unique. De plus, l'absence de modèle préalable écarte le difficile consensus au sein d'une communauté. Notre approche se focalise sur les liens intra-corpus, pour une analyse de celui-ci. La mise en correspondance des concepts et vocabulaires à destination du web de données reste une finalité complémentaire non nécessaire pour les objectifs visés ici.

– **Supervisé** : l'utilisateur garde le contrôle sur les données. *Haruspex* est largement paramétrable. La vitesse et l'exhaustivité de l'algorithme laisse place à l'humain pour modérer les résultats. La qualité des résultats est primordiale.

– **Pour des données non-structurées** : Le processus accepte en entrée des textes faiblement structurés d'un point de vue informatique (.pdf, .html, .doc, .odt, .tex). Dans l'optique de traiter les corpus tels qu'ils existent sans dénaturer l'information contenue (sans pré-traitement).

– **Non entraîné** : l'unicité (supposée) du corpus ne permet pas d'entraîner le logiciel sur d'autre corpus sans risquer une perte de précision dans les données extraites. Chaque corpus est sa seule référence. Il est néanmoins possible d'itérer entre passes d'analyse automatique et supervision.

– **Sans interprétation** : *Haruspex* ne vise pas à interpréter des données, mais à les présenter différemment pour permettre à l'expert d'en produire une (nouvelle) analyse. Principalement basé sur des algorithmes statistiques, l'analyse lexicale complexe prime sur la sémantique, ce travail diffère donc des NER *Named Entity Recognition*.

Un écart se creuse entre d'un coté les grand projets en humanités numériques (avec *BigData*, web sémantique, base de données, etc…) et de l'autre la réalité du terrain dans la plupart des laboratoires et des musées (avec des corpus restreints de textes non-structurés et des besoins de précision pour l'analyse). Un des enjeux d'*Haruspex* est de proposer un outil adapté aux *small data* et méthodes des SHS.

État de l'art

Des outils de textométrie, comme TXM (Heiden, 2010) permettent une analyse statistique des textes, de nombreux outils visent à l'extraction de mot-clé comme (Rocheteau & Daille, 2011), ou à l'extraction de *topic (Guille, Soriano-Morales, & Truica, 2016)*. Ces techniques inspirent la méthodologie présentée ici, et sont parfois combinées avec la proposition pour offrir une analyse complémentaire. Cependant nous avons estimé que l'analyse statistique pure ne rendait pas compte de la dimension « réseau d'information » sous-jacente à un corpus, ni ne reflétait de hiérarchisation de concepts[5]; l'extraction de mots-clés proposée par d'autres outils n'offrait pas le compromis quantité - précision (*recall - precision*) (Buckland & Gey, 1994) nécessaire; enfin l'extraction de thèmes (*topic modeling*) peut être utile en amont du processus, mais le découpage du corpus en sous-corpus ne constitue pas en soi un outil d'analyse fine pour l'expert. La notion de corpus dans un nouveau paradigme (numérique) est un élément fondateur de notre raisonnement, l'espace du document est bouleversé (Ghitalla, 2000), sa représentation sous forme de

[4] http://lov.okfn.org/dataset/lov

[5] Par exemple *archéologie préventive* est un une sous-catégorie de *archéologie*

fiche (Ardans, 2011) permet de naviguer entre documents selon des liens calculés comme par exemple entre des pages web (Jacomy, Girard, Ooghe, & Venturini, 2016). Dans notre cas nous devons tenir des spécificités du récit historique considérant qu'il n'y a pas d'atome historique et donc pas de niveau de détails qui puisse être absolu (Veyne, 1971), seuls comptent le récit et les mots employés. Enfin, suivant le principe d'entropie maximale (Jaynes, 1957) nous ciblons les ensembles discriminants d'expressions complexes extraites par un algorithme inspiré de ANA (Enguehard, 1993) pour établir des liens entre documents. Enfin le calcul de similarité entre document, basé sur la co-occurrence d'expressions-clés extraites, est inspiré du TF-IDF (Salton, 1983).

Proposition de processus de traitement

L'ensemble du processus se divise en 4 étapes principales articulées autour du corpus. Le schéma (figure 2) représente leur enchainement.

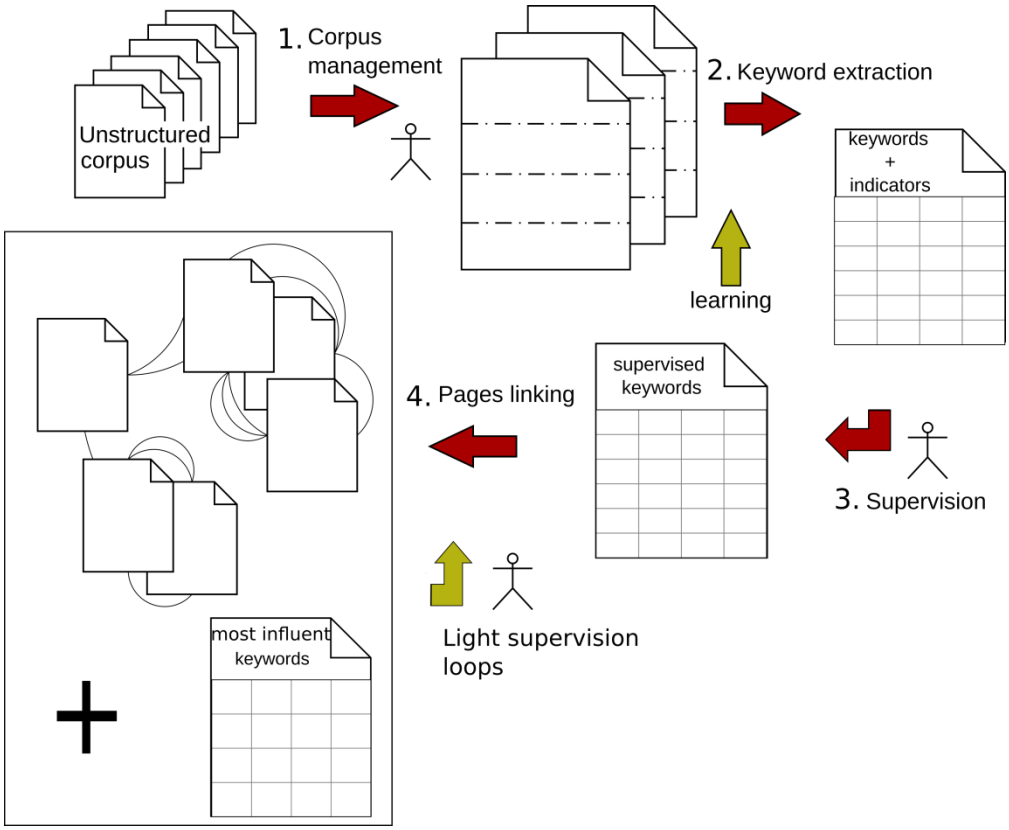

Figure 2. *Schéma descriptif du processus en 4 étapes principales (crédit : Matthieu Quantin. CC-BY-NC-ND)*

Création des nœuds

Il s'agit de transformer le contenu d'entrée en un ensemble homogène du point de vue informatique (étape 1 sur figure 2). Il ne s'agit pas d'une étape d'OCR[6] qui constitue un éventuel préalable au processus. Suivant les paramètres les fichiers d'entrées peuvent être redécoupées en sous-parties (les partie « naturelles » du textes: chapitres, section, sous-section…). Suite à ce découpage nous obtenons l'*unité documentaire* de notre corpus. Un unique document peut donc être considéré comme un corpus de

[6] OCR: *Optical Character Recognition*: reconnaissance automatique de caractères sur une « image de texte » pour en produire du texte.

sous-parties (par exemple un article, une thèse, un mémoire). Par la suite nous appellerons *fiche* cette unité documentaire du corpus.

À chaque fiche sont attribués un identifiant et un nœud correspondant dans une base de données orientée graphe[7]. Les images et les notes de bas de page de chaque fiche sont stockés à part du texte, un nœud spécial (nœud fiche, nœud référence) et un identifiant sont attribués à chacun.

Toutes les fiches sont concaténées en un fichier unique de transfert.

Extraction d'expressions-clés

À partir du fichier créé précédemment, l'objectif est d'extraire des expressions caractéristiques aux textes, discriminante au sein du corpus (étape 2 sur figure 2). Ainsi nous cherchons à être le plus spécifique possible, tout en excluant les expressions isolées dans une seule fiche. Par exemple, dans un corpus de traités de mécanique romaine antique, l'expression *trispaste à double rang de poulies* est préférée au simple *trispaste*. L'algorithme développé pour ce besoin est inspiré d'ANA (Enguehard & Pantera, 1995) dans son fonctionnement par essaimage / expansion.

Cet algorithme est indépendant de vocabulaire extérieur (thésaurus) et n'a pas d'à priori sur les contenus. Il est donc approprié à l'analyse de corpus dans des domaines très spécifiques, en dehors de tout sentier battu: ce qui est souvent le cas des corpus en sciences de l'homme. Des paramètres permettent de le rendre tolérant aux erreurs de frappe ou d'OCR (inversion de caractères). Cette tolérance s'accompagne malheureusement d'une baisse de la précision: confusion entre des mots similaires comme *port* et *porte* par exemple.

À la fin de cette étape, le contenu de chaque fiche est associé à une liste d'expressions-clés.

Post-traitement

Il s'agit de surveiller la production de la machine et d'en améliorer la qualité (étape 3 sur figure 2). Concernant les expressions-clés, nous avons fait le choix de conserver le maximum de propositions possibles pour éviter de « manquer » une expression-clé, ce qui a tendance à générer du bruit : des expressions indésirées se glissent dans les résultats. Afin d'assister l'utilisateur dans la modération des résultats, des indicateurs ont étés créés. Ils permettent de filtrer les expressions-clés selon plusieurs paramètres critiques; par exemple: contient une forme verbale, est composée de mots très communs. Un arbre d'héritage de chaque expression permet d'en tracer la formation et de la ramener à une forme parente si besoin. Ainsi *trispaste à double rang de poulies* est construit à partir de *trispaste* et *poulie*. Certaines descendances ne sont pas utiles, par exemple *trispaste antique* peut-être ramené à *trispaste* (manuellement).

Des requêtes vers des serveurs extérieurs permettent de vérifier si les expressions-clés extraites sont connues par ailleurs. Si une page Wikipédia existe sur le sujet *trispaste à double rang de poulies* alors l'expression sera mise en lien vers cette ressource. Cette fonctionnalité permet de lier les contenus extraits à d'autres silos de données. C'est une forme de standardisation des résultats, mais l'inexistence de ressource extérieure n'implique pas le rejet d'une expression-clé. Nous conservons la spécificité des résultats.

Concernant les nœuds, certaines métadonnées utiles à l'analyse sont extraites automatiquement si disponibles (date, auteur). L'utilisateur peut modérer ces résultats et en ajouter à sa convenance, de

[7] Neo4J (neo4j.com)

manière libre selon les paramètres qu'il juge utile pour l'analyse et ce qu'il est en mesure de renseigner. Il n'y a pas de champ préformaté, puisqu'il ne s'agit pas de partager les données mais d'analyser un corpus.

Création de liens

Chaque co-occurrence d'expression entre deux fiches donnera naissance à un lien entre elles-deux. Ainsi toutes les fiches contenant l'expression *trispaste à double rang de poulies* seront liées entre elles (étape 4 sur figure 2). Ce lien est pondéré en fonction de deux paramètres : la spécificité de l'expression et sa fréquence d'apparition dans les fiches concernées. Il s'agit d'une version améliorée d'un algorithme classique : le tf-idf[8] (Salton, 1983). Le référentiel de calcul de spécificité est le corpus. Ainsi une expression qui apparaît dans toutes les fiches recevra une mauvaise note, les liens qui en sont issus recevront une mauvaise pondération sauf si une forte co-occurrence est repérée entre deux fiches. Par exemple: si toutes les fiches évoquent le terme *géométrie*, elles seront toutes liées entre elles par un lien qui recevra une pondération faible; si parmi elles, deux fiches évoquent beaucoup le terme *géométrie* alors le lien éponyme qui les unit reçoit une meilleure pondération (figure 3).

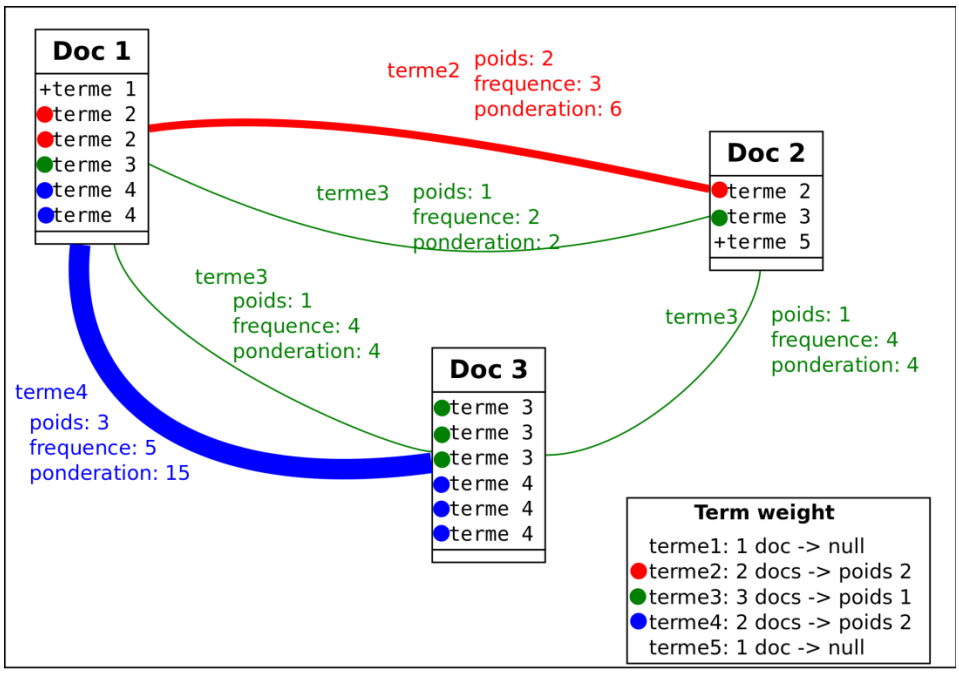

Figure 3. *Exemple de liens entre fiches avec 3 fiches et 3 expressions (crédit : Matthieu Quantin. CC-BY-NC-ND)*

Ainsi nous obtenons un hypergraphe, chaque nœud est relié à d'autres nœuds par de nombreux liens (1 lien par co-occurrence d'expression). Des liens *uniques* entre les fiches sont générés, la pondération de ce lien est la somme de chaque lien *expression*. Ainsi le lien unique entre deux fiches partageant de nombreux sujets sera mieux noté que le lien unique entre des fiches très différentes. Il faudra ensuite en décortiquer les composantes pour comprendre les raisons de ces liaisons ou non-liaisons.

L'inverse de l'apparition d'expression est calculée. Par exemple, si quasi toutes les fiches évoquent *géométrie*, il est alors intéressant de considérer celles qui ne l'évoquent pas. Au même titre que les quelques fiches qui évoquent *trispaste à double rang de poulies* sont fortement liées entre elles car elles ont une connivence sur un sujet spécifique, les fiches qui n'évoquent pas *géométrie* sont fortement liées entre elles.

[8] pour *term frequency inverse document frequency*

Boucles heuristiques

Le résultat, un graphe de fiches liées, est un atlas total, illisible directement. Il convient de faire des choix pour voir ce que l'on veut montrer. Des *vues* de ce graphe total sont alors générées suivant les questions du chercheur. Elles permettent de soulever de nouvelles questions, qui donnent lieu à la création de nouvelles vues. Nous parlons alors de boucle heuristique. Souvent la création de nouvelles vues implique d'injecter de nouvelles données dans le système: par exemple créer de nouvelles catégories d'expression. Afin d'éviter un travail fastidieux sur la liste de toutes les expressions extraites précédemment, il est proposé d'extraire certaines informations du graphe (nœuds ou liens), de les modifier et de les réinjecter. Ainsi on peut choisir de ne modérer que les liens (expressions) les plus influents du graphe, soit une liste de 200 expressions environ (au lieu de la liste globale de 2000 expressions environ), ou bien les liens qui unissent tel et tel documents, ou bien préciser une catégorie en sous-catégories. Ceci évite une lourde modération préalable et se focalise sur les questionnements du chercheur pour créer du détail itérativement.

Cas d'applications

Plusieurs corpus de textes, de thématique et configurations différentes, ont permis de valider la méthode. Cet article ne fait l'objet de l'analyse « sciences humaines » des résultats du traitement. Le tableau 1 compare les différents corpus. Un coefficient de variation élevé indique une forme homogène (faible variation de longueur de texte entre les fiches), une appréciation arbitraire est donnée à l'homogénéité de contenu et à la qualité du texte. La qualité du texte concerne les fautes de frappe, les erreurs de reconnaissance de caractère.

Attribut	CIRP	Chimie s.	HdTCol	Halls A.	Nantes 1900
lang.	en	fr	fr	fr	fr
format	pdf	pdf	xml (ocr)	odt	doc
struct. int.	non	non	non	oui	non
qualité du texte	~	+	-	+	+
nb. fiches	109	38	107	196	440
nb mots	400k	300k	300k	108k	100k
moy nb mots/fiche	3700	7900	2800	551	212
écart type	431	5151	5605	1271	214
coefficient de variation	0.11	0.65	1.77	2.61	1.01
homogénéité contenus	-	+	~	+	-

Tableau 1. *Tableau récapitulatif des statistiques de différents corpus utilisés*

Corpus compliqués

Les corpus les plus compliqués sont *Nantes1900* et l'histoire du travail dans les colonies (HdTCol). Néanmoins certains résultats sont intéressants.

Nantes1900

Ce corpus regroupe les fiches documentant la maquette du port industriel de Nantes en 1900 (Hervy et al., 2014), exposée au musée d'histoire de Nantes. Elles sont courtes, le vocabulaire utilisé (visant le grand public) n'est pas suffisamment développé et précis pour que l'analyse intéresse un spécialiste du domaine. Une mesure semblait prometteuse: comparer le réseau de fiches constitué manuellement ces dernières années et celui produit automatiquement par l'algorithme. Ces réseaux ne sont pas comparables. Une analyse approfondie permettrait étudier ces divergences. Exemple d'expressions extraites: *Conditions de travail des ouvriers*; *Ateliers et Chantiers de la Loire*.

HdTCol

Il s'agit d'un corpus d'archives sur l'histoire du travail dans les colonies. Plusieurs biais se superposent: le choix des archives photographiées, le choix des photographies à l'OCR. Ces choix ont conservé l'amplitude du domaine d'étude, impliquant une forte hétérogénéité de contenu dans le corpus. A cela s'ajoute une forte hétérogénéité de forme (variation de longueur des fiches) créant un biais: plus une fiche est longue plus elle contient d'expression-clé. Enfin l'OCR produit des résultats exploitables à condition d'accepter une augmentation du bruit.

Le corpus ayant bénéficié d'une indexation manuelle, il a été possible de comparer celle-ci aux résultats d'*Haruspex*. Ces résultats sont très positifs: l'ensemble des termes utilisé pour l'indexation sont repris par l'algorithme, le nombre de termes utilisés est 5 fois plus important et beaucoup plus précis: indexation avec des expressions complexes. Quelques fiches (environ 3%) ont été indexées manuellement par du vocabulaire implicite que l'algorithme n'a pas pu repérer (Faux négatif), par exemple certaines fiches sont indexées *syndicat* tandis que ce mot n'apparaît pas dans leur contenu. A l'inverse de très nombreuses fiches évoquant un terme ont été repérées, alors que l'indexation manuelle les avait omises, par exemple seule la moitié des fiches contenant le terme *syndicat* avait été repérée. Aucune analyse du réseau n'a été effectuée par le chercheur. Exemple d'expressions extraites : *Syndicat des ouvriers de l'usine Rodler*; *Populations montagnardes du sud indochinois*.

Corpus fragmenté

Halls A.

Le corpus est constitué de deux mémoires traitant respectivement d'une parcelle du port industriel de Nantes et d'une entreprise ayant occupé cette parcelle. La taille des fiches (sous-parties des mémoires) fluctue mais le contenu est précis. Certaines fiches prévalent sur les autres par leur poids (nombre de mots), ce qui peut mener à des biais d'interprétation. Les expression extraites sont de bonne qualité car l'homogénéité de contenu est forte.

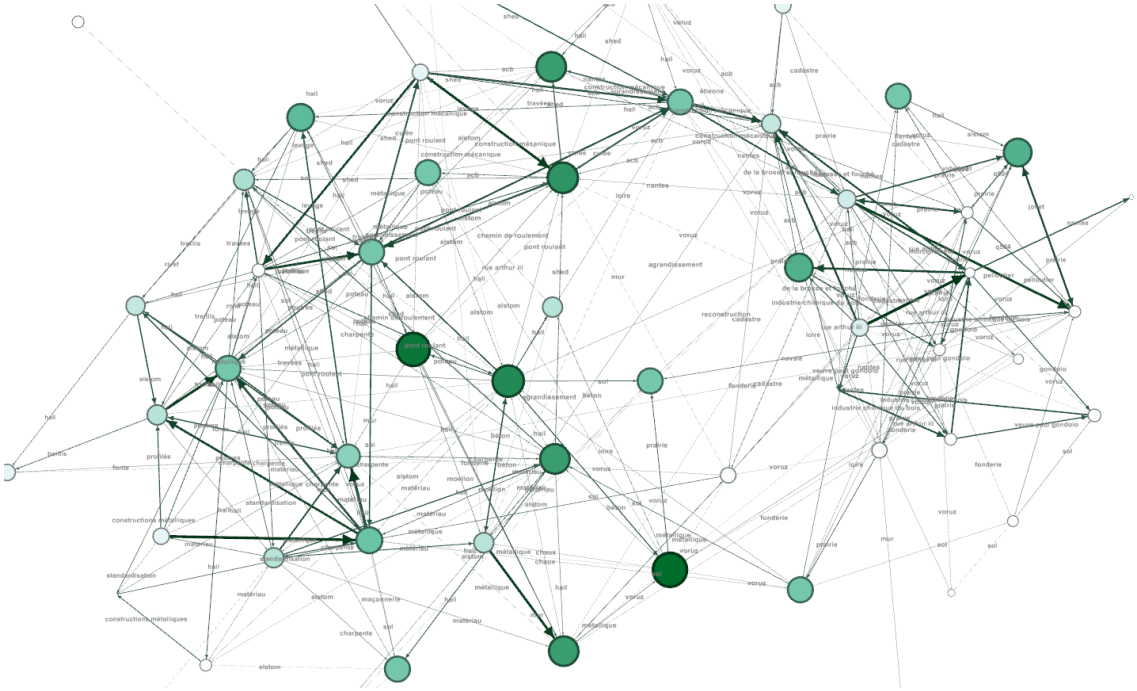

Figure 4. *Principales relations unissant les différentes sous-parties de 2 mémoires traitant du même sujet (crédit : Matthieu Quantin. CC-BY-NC-ND)*

L'objectif de cette expérience est double : (1) rompre la juxtaposition de deux linéarités (mémoires) pour produire un graphe de connaissance; (2) corréler ces contenus aux fiches de *Nantes1900* pour créer du détail dans un réseau existant.

Le premier point est accompli, chaque sous partie des corpus devient une fiche, la linéarité des discours est rompue et laisse place à des fiches interconnectées. On remarque (figure 4) que l'ordre initial (linéaire) n'est pas totalement aboli. Le second objectif (corrélation avec *Nantes1900*) permet au musée d'intégrer des contenus issus de productions universitaires. La définition du réseau de connaissances est localement augmentée: au lieu d'une seule fiche, un autre graphe prend place et se connecte à quelques éléments voisins (figure 5). Malheureusement, les expressions les plus spécifiques ne trouvent pas d'équivalent dans *Nantes1900.* Exemple d'expressions extraites : *torpilleurs de 2 400 tonnes*; *Ateliers et Chantiers de la Loire*. On remarque que la dernière expression est commune avec le corpus de *Nantes1900.*

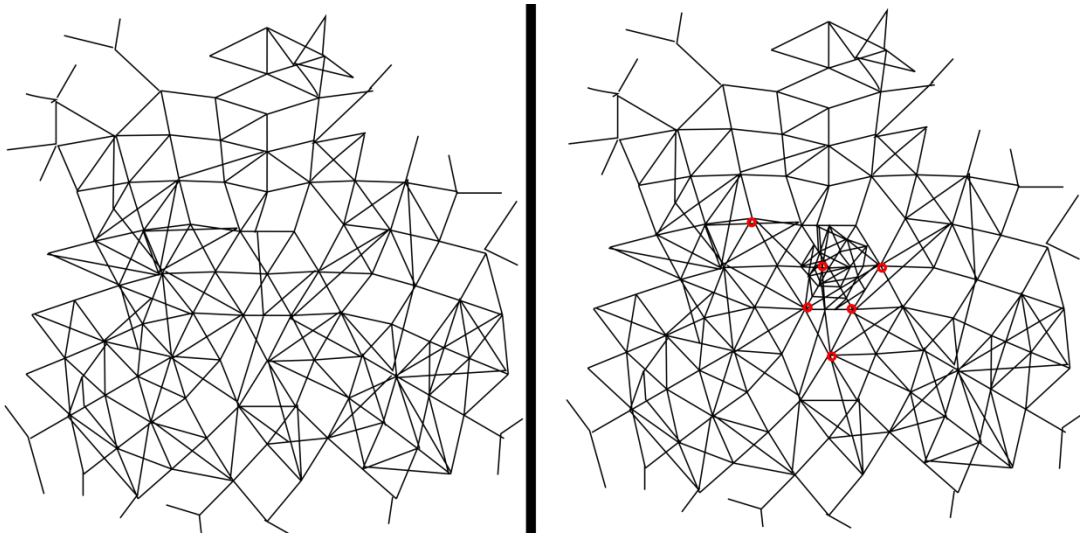

Figure 5. *Schéma illustrant l'augmentation de définition locale sur un réseau de connaissance établi, en rouge les quelques nœuds auxquels se raccroche le graphe de haute définition (crédit : Matthieu Quantin. CC-BY-NC-ND)*

Corpus homogènes

Il s'agit des corpus dont le coefficient de variation est faible, avec des vocabulaires spécifiques et une qualité du texte correcte voire bonne.

CIRP

Les annales de la conférence CIRP entre 1980 et 2016 (*CIRP annals manufacturing technology*) sont composé de 109 articles en anglais. Ce corpus présente une grande hétérogénéité de contenus : de nombreux domaine de l'ingénierie s'y côtoient, ce qui limite la qualité des expressions extraites.

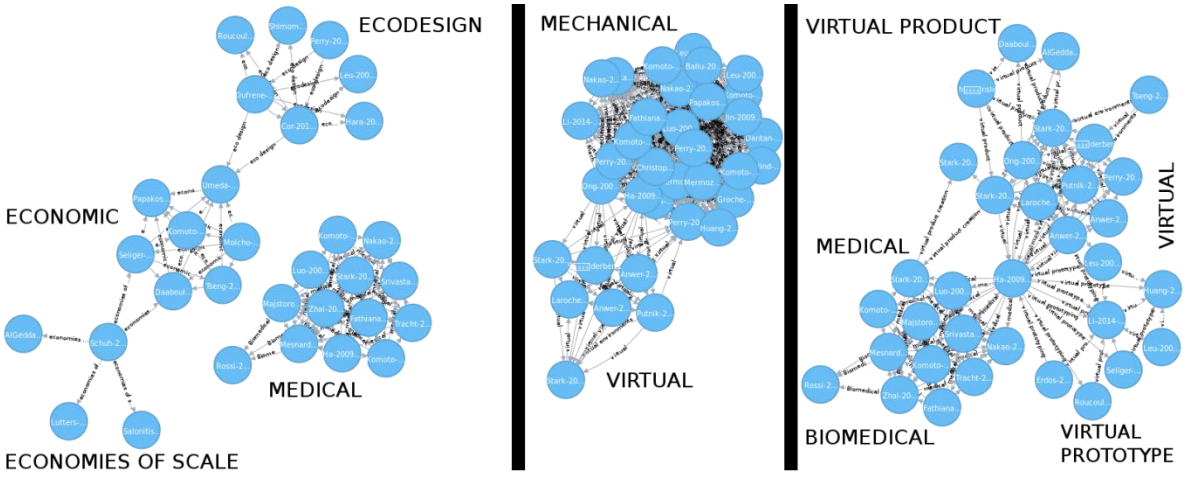

Figure 6. *3 vues du corpus CIRP. La première illustre des domaines disjoints: l'économie et le médical. La seconde le lien ténu entre virtuel et mécanique, la troisième les relations entre médical et virtuel. Certains articles interdisciplinaires sont les pivots de ces relations (crédit : Matthieu Quantin. CC-BY-NC-ND)*

L'intérêt porte sur l'identification de communautés de chercheurs autours de sujets parfois connexes comme le *médical*, le *biomédical* et le *virtuel*; parfois distincts comme le *médical* et *l'économie*. Puisque la publication de cette revue est annuelle, la dimension temporelle peut être une variable: on observe alors la formation de certaines communautés s'émancipant de leur cluster d'origine. C'est le cas du *mécanique*

et du *virtuel* (figure 6). Exemple d'expressions extraites : *synthetic DNA based approach*; *International Journal of Advanced Manufacturing Technology*.

Chimie s.

Il s'agit d'un corpus de retranscriptions d'entretiens en histoire de la chimie du solide. La qualité du texte et l'homogénéité des documents (forme et contenu) sont bonnes. L'objectif de l'historien est d'étudier la formation d'une discipline et d'une communauté scientifique (Teissier, 2007). Les co-occurrences de termes sont un bon indice pour étudier les affinités que les chercheurs entretiennent. La qualité des expressions extraites est bonne.

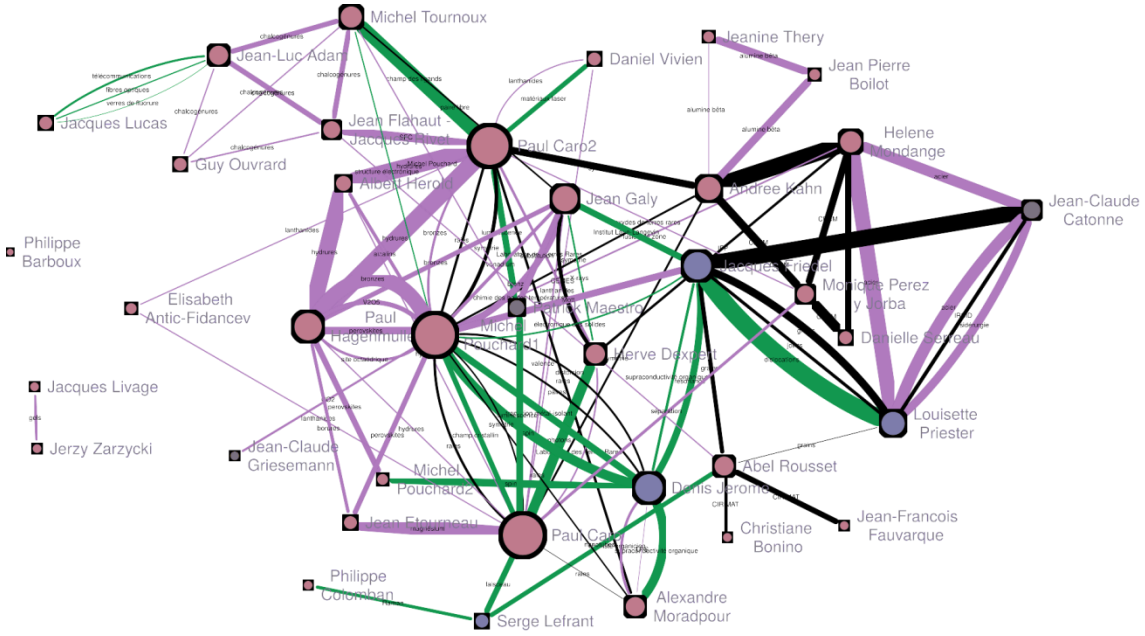

Figure 7. *Liens entre les fiches du corpus. Chaque fiche représente un entretien. Des communautés de recherche émergent, certaines sont influencées par le directeur de thèse, d'autres par le laboratoire actuel. Certains chercheurs sont centraux d'autres marginaux. Les thématiques de recherche sont réparties entre Chimie (rose), Physique (vert) et sciences (noir) (crédit : Matthieu Quantin. CC-BY-NC-ND)*

L'étude de ce corpus a donné lieu à un processus itératif de questions/enrichissements/réponses. Les propriétés heuristiques du graphe ont été mises en avant: les évidences (pour le chercheur) autant que les surprises sont fructueuses. Les unes confortent le chercheur (et l'algorithme) dans ses positions, les autres soulèvent de nouvelles questions. Par exemple, dans un cas les chercheurs d'un même laboratoire sont liés par des objets de recherche communs, dans l'autre cas deux chercheurs au parcours similaire ne sont pas liés (figure 7). Les expressions extraites (qui donneront les liens) sont automatiquement typées dans différentes catégories: sciences (générique), physique, chimie, personnalité, administration, industrie… Ces catégories permettent d'affiner les requêtes et de préciser les vues.

Conclusion

Apports de la proposition

Haruspex est complémentaire aux outils de partage et de mise à disposition des connaissances (web sémantique, *topic modeling*, *named entities recognition*…). La méthode propose un nouveau regard (quantitatif) aux experts d'un corpus.

Par rapport à une indexation manuelle, la rigueur de l'algorithme recensant l'exhaustivité des occurrences apporte de la fiabilité. Ce gain s'accompagne d'une légère perte : les données implicites ne sont pas repérées. Par rapport à d'autres indexations automatiques, cet algorithme présente plusieurs avantages: précision, complexité des marqueurs (expression-clé), marqueurs sans à priori sur la terminologie.

La création d'inférences bas-niveau supervisées caractérise *Haruspex*. Pour optimiser les performances d'analyse, nous allions la rigueur de la machine au travail qualitatif de l'humain. Les inférences de la machine sont réduites à des calculs de proximités entre fiches (paramétrables). Nous évitons ainsi l'écueil des inférences haut niveau, qui sont réservées à l'historien, seul capable d'un véritable travail qualitatif, c'est à dire de réaliser des inférences dites « historiques ». La machine se contente de compter et calculer sans tirer de conclusions.

Menaces et faiblesses

Les outils développés pour *Haruspex* sont encore trop instables et insuffisamment ergonomiques pour permettre une utilisation autonome. Cette faiblesse est aussi une opportunité: elle implique de travailler en binôme ingénieur / historien. L'évolution rapide d'*Haruspex* (développement dynamique) biaise les comparaisons entre différents corpus: les premiers corpus traités n'ont pas été traités avec le même code que les derniers. Des versions stables seraient les bienvenues.

Perspectives

Dans la lignée du projet *Nantes1900*, la documentation de modèles 3D (par fiches liées) pourrait renforcer le lien avec le patrimoine. Cette ambition conforte l'idée d'une passerelle entre recherches universitaires et musées. Une perspective forte réside dans la création de documentation multi-accès. En effet, la structure des données permet de documenter des éléments publiés en proposant des liens vers d'autres items proches (sémantiquement voire géographiquement ou temporellement par exemple). Une forte contrainte de proximité conviendrait plutôt au grand public curieux, une ouverture sur des documents plus éloignés conviendrait à des objectifs de recherche.

L'utilisation de graphe est arbitraire, cette forme permet une heuristique à partir des comptages de l'algorithme, un dialogue homme-machine. D'autres formes de représentation des connaissances restent à explorer de manière complémentaire.

Bibliographie

Ardans. (2011). Ardans Knowledge Maker : Introduction , principes et philosophie implantés dans cet environnement de gestion des connaissances (Vol. 33).

Buckland, M., & Gey, F. (1994). The relationship between Recall and Precision. Journal of the American Society for Information Science, 45(1), 12–19.

Doerr, M. (2003). The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 24(3), 75.

Enguehard, C. (1993). Acquisition de terminologie à partir de gros corpus. Informatique & Langue Naturelle, p.373-384.

Enguehard, C., & Pantera, L. (1995). Automatic natural acquisition of a terminology. Journal of Quantitative Linguistics, 2(1), 27–32.

Ghitalla, F. (2000). L'espace du document num{é}rique. Communication et Langages, 126(1), 74–84.

Guille, A., Soriano-Morales, E.-P., & Truica, C.-O. (2016). Topic modeling and hypergraph mining to analyze the EGC conference history. Conférence Sur l'Extraction et La Gestion Des Connaissances, (i).

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In 24th Pacific Asia Conference on Language, Information and Computation (pp. 389–398). Institute for Digital Enhancement of Cognitive Development, Waseda University.

Hervy, B., Laroche, F., Kerouanton Jlk, J.-L., Bernard, A., Courtin, C., D'Haene, L., … Waels, A. (2014). Augmented historical scale model for museums: from curation to multi-modal promotion. In Laval Virtual VRIC 14 (pp. 10–13). Laval (France).

Jacomy, M., Girard, P., Ooghe, B., & Venturini, T. (2016, May 18). Hyphe, a Curation-Oriented Approach to Web Crawling for the Social Sciences. International AAAI Conference on Web and Social Media. Association for the Advancement of Artificial Intelligence.

Jaynes, E. T. (1957). Information theory and statistical mechanics. Physical Review, 106(4), 620–630.

Nouvel, B., & Rousset, M. (2015). FRANTIQ : faciliter l'interconnexion des données de la recherche en archéologie et sciences de l'Antiquité. In Digital Humanities and Antiquity. Grenoble, FR.

Rocheteau, J., & Daille, B. (2011, November 9). TTC TermSuite: A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora. 5th International Joint Conference on Natural Language Processing (IJCNLP).

Salton, G. (1983). Introduction to Modern Information Retrieval. New York, NY, USA: McGraw-Hill, Inc.

Teissier, P. (2007). L'émergence de la chimie du solide en France (1950-2000) de la formation d'une communauté à sa disparition. Paris 10.

Veyne, P. (1971). Comment on écrit l'histoire (Seuil). Paris.

