Médiévales

Langues, Textes, Histoire

73 | automne 2017

Le texte à l'épreuve du numérique

La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique

Full Text Search in Medieval Manuscripts: Issues and Perspectives of the HIMANIS Project for Electronic Publishing

Dominique Stutzmann, Jean-François Moufflet et Sébastien Hamel

Édition électronique

URL : <https://journals.openedition.org/medievaux/8198>

DOI : 10.4000/medievaux.8198

ISSN : 1777-5892

Éditeur

Presses universitaires de Vincennes

Édition imprimée

Date de publication : 15 décembre 2017

Pagination : 67-96

ISBN : 978-2-84292-801-8

ISSN : 0751-2708

Référence électronique

Dominique Stutzmann, Jean-François Moufflet et Sébastien Hamel, « La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique », *Médiévales* [En ligne], 73 | automne 2017, mis en ligne le 15 décembre 2019, consulté le 11 août 2022. URL : <http://journals.openedition.org/medievaux/8198> ; DOI : <https://doi.org/10.4000/medievaux.8198>

Tous droits réservés

Dominique Stutzmann
Jean-François Moufflet
Sébastien Hamel

La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique

HIMANIS (Historical MANuscript Indexing for user-controlled Search) est un projet de recherche européen associant, sous le pilotage de l'Institut de Recherche et d'Histoire des Textes (IRHT, CNRS, France), la société innovante A2iA (France), la Rijksuniversiteit Groningen (RUG, Pays-Bas) et l'Universitat Politècnica de València (UPV, Espagne). Il est l'un des seize projets retenus parmi 352 lors de l'appel *Heritage Plus Joint Call* de la JPI ou *Joint Programming Initiative* « Cultural Heritage and Global Change : a challenge for Europe »¹. Il vise à l'indexation du texte copié dans les registres de la chancellerie royale française des années 1302-1483, c'est-à-dire les registres conservés aux Archives nationales sous les cotes JJ35 à JJ211, à partir des images produites par leur numérisation. Les enjeux de recherche d'information (données massives et bruitées) permettent de conjointre les enjeux technologiques (reconnaissance de l'écriture manuscrite) et historiques (analyses paléographiques et diplomatiques, recherche sur les institutions, le fonctionnement de la monarchie, la naissance de l'État-nation).

La présente contribution propose un modèle d'accès à l'information dans un corpus de données massives, d'un point de vue tant ergonomique qu'herméneutique. À cette fin, la première partie présente le corpus des registres de la chancellerie royale française retenus par le projet, les outils actuels pour accéder à l'information qu'ils contiennent et leur formalisation en TEI. La deuxième partie problématise l'édition électronique comme « vérité terrain » et « terrain d'apprentissage », en renversant l'approche

1. FP7-ERANET-2013-RTD. Grant Agreement n° 618104. Cf. <http://cordis.europa.eu/project/rcn/191238_fr.html> et <<https://www.era-learn.eu/network-information/networks/jpi-cultural-heritage>>. Description de l'appel : <<https://www.era-learn.eu/network-information/networks/heritage-plus/jpi-cultural-heritage-and-global-change>>.

classique de l'édition critique comme finalité. Enfin, la troisième partie décrit le modèle d'accès proposé, à la fois pour une approche par « indexation » (et non par transcription) et pour une granularité par acte.

Les registres de la chancellerie royale française : un corpus difficile d'accès

Présentation du corpus

Les plus anciens registres de chancellerie remontent à l'époque de Philippe Auguste². Sans entrer dans le débat sur leur origine³, ils attestent assurément d'une mutation profonde des pratiques de conservation et de thésaurisation de la mémoire du gouvernement royal. Révélateurs des méthodes d'administration qui se mettent en place dans un contexte géopolitique profondément bouleversé en ce début de XIII^e siècle, mais aussi de la vision des missions royales, leur visée première est de garder la trace des actes utiles et de connaître de manière circonstanciée les droits comme les engagements du roi. Le registre, compagnon de la chancellerie itinérante, se substitue d'une certaine manière aux chartes dont il est la copie et qui demeurent dans le palais de Paris.

Qualifiés plutôt de « cartulaires-registres » par les diplomatistes, les registres du XIII^e siècle se caractérisent par une certaine hétérogénéité due à la variété des informations qui y sont copiées : il n'y a pas en vérité que des chartes. Les actes sont enregistrés de manière non linéaire, de façon méthodique, répartis entre de grandes rubriques thématiques dont la liste ne cesse de se complexifier au cours du siècle, au point de rendre l'opération de recopie – et donc de reclassement des actes au sein de ces thèmes – intenable à la fin du règne de saint Louis⁴. Cette complexité de l'enregistrement, paradoxalement nourrie par les succès d'une royauté assumant des prérogatives toujours plus larges, explique vraisemblablement les changements du mode de tenue des registres à l'orée du XIV^e siècle, sous le règne de Philippe le Bel. À partir du registre JJ35, qui couvre les années 1302-1305, débute la pratique jugée conventionnelle de l'enregistrement, c'est-à-dire une copie sérielle, continue, d'actes qui se suivent plus ou moins chronologiquement. Les actes d'un registre sont référencés dans une table précédant les cahiers où est copié le texte enregistré (fig. 1). Ces tables

2. *Registrum veterius* (1204-1212), Bibliothèque apostolique vaticane, Ottoboni, lat. 2796.

3. Voir notamment Y. POTIN, *La Mise en archives du trésor des chartes (XIII^e-XIX^e siècle)*, thèse pour le diplôme d'archiviste-paléographe, Paris, 2007, positions : <<http://theses.enc.sorbonne.fr/2007/potin>>.

4. *La nova compilatio*, qui prévoyait la mise au point d'un nouveau cartulaire-registre aux rubriques encore plus nombreuses, n'a semble-t-il jamais dépassé le stade de projet.

[illegible]

Fig. 1. Table du registre JJ37 (1303-1307)
(Paris, Archives nationales de France, JJ37, f. Ar.)

attribuent un numéro à chaque acte et en donnent un court regeste (lui-même copié en marge de l'acte dans les cahiers principaux).

Dans la pratique, il ne faut pas voir une linéarité régulière entre les actes : des désordres chronologiques apparaissent bel et bien, dus à des enregistrements tardifs et décalés dans le temps, à des additions ultérieures, à la reliure parfois hasardeuse des cahiers récupérés auprès des notaires de la chancellerie. Néanmoins, le corpus visé par le projet HIMANIS offre une homogénéité certaine dans la structure et le mode de tenue, écartant les cartulaires-registres pour ne retenir que ceux dont le mode d'enregistrement est sériel.

Le choix des actes enregistrés évolue aussi clairement par rapport aux origines : sont de plus en plus privilégiés les actes à valeur perpétuelle au détriment des actes d'administration plus courante. D'un enregistrement abordant initialement une large gamme de questions, on passe à un enregistrement plus sélectif. Le Parlement et la Chambre des comptes prennent également le relais de la chancellerie. Parmi les actes que cette dernière continue d'enregistrer, une forte proportion concerne désormais des mesures visant des personnes particulières : les fameuses lettres de rémission, accordant la grâce royale à des personnes condamnées, prennent une part importante, mais les mesures d'anoblissement en relèvent aussi. Au-delà de ces mesures individuelles, témoins de l'emprise de la royauté sur les sujets, les registres n'en demeurent pas moins le reflet de la politique des Capétiens et des Valois. Ceux de la fin du XV^e siècle, par exemple, contiennent les actes relatifs aux (re)fondations de foires et marchés, dans le contexte d'un royaume dont l'activité commerciale s'intensifie à la sortie de la guerre de Cent Ans⁵. Enfin, pour achever la comparaison avec les premiers registres, la masse d'information enregistrée devient de plus en plus considérable au fil des décennies. Cette masse explique le net déséquilibre entre les registres du XIV^e et ceux du XV^e siècle, tant en ce qui concerne les outils permettant de les exploiter que le traitement qu'en a fait l'historiographie. Les premiers bénéficient de travaux d'inventaires encore assez nombreux et précis, tandis que les seconds ne peuvent être approchés que par des réalisations plus partielles ou plus générales, à quelques exceptions près. Et si les registres de Charles VII ont fait l'objet d'une étude récente⁶, ceux de Louis XI restent encore largement méconnus et sous-exploités.

5. Les actes relatifs aux foires et marchés font l'objet d'une étude et d'une édition par les étudiants de l'Université Paris 7, sous la direction de Judicaël Petrowiste.

6. H. LARCHER-MAURIN, « *Tam Parisius quam alibi* ». *Unité et pluralité de la chancellerie royale au temps de Charles VII*, thèse pour le diplôme d'archiviste-paléographe, Paris, 2008, positions : <<http://theses.enc.sorbonne.fr/2008/larcher>>.

Outils d'accès

Le projet HIMANIS est donc aussi une opportunité pour approfondir nos connaissances sur ce vaste ensemble de registres et, à terme, pour favoriser de nouvelles études grâce à des outils qui faciliteront l'accès à leur contenu. Car pour l'heure, la recherche d'information dans les registres de la chancellerie est complexe et souvent frustrante : les outils actuels ne dispensent pas de la lecture continue des volumes pour suppléer à la pauvreté des rubriques et tables ainsi qu'à l'incomplétude des entreprises d'édition et d'inventaire (fig. 2).

Fig. 2. Couverture géographique et chronologique des inventaires et éditions

Avant de décrire les opérations menées en vue d'une lecture intégrale ou du moins de l'indexation automatique du plein texte, il s'avère utile de présenter synthétiquement l'état actuel des outils de recherche et leur disponibilité numérique. Il en est de trois sortes : les inventaires analytiques systématiques, rendant compte de l'ensemble des actes des registres décrits (table 1) ; les inventaires analytiques sélectifs portant sur un choix de documents en raison de leur thème ou de leur portée géographique (table 2) ; les éditions partielles (table 3).

Table 1. Inventaires systématiques

IDENTIFICATION	COUVERTURE	INVENTAIRE ANALYTIQUE	FORMAT	INDEX
IR421-IR424 et IR 429(JJ581/1 à JJ581/4)	JJ1-JJ264	Inv. ms. de JJ 1 à JJ 264, dressé au début du XVIII ^e s., révisé et complété par A. LONGNON et A. COULON, 1880-1900, copie en quatre vol.	Vol. 1 à 4 : Papier- Vol. 1 à 4 : EAD (vol. 2 non publié en ligne)	<i>Index ms. des noms de personnes de JJ A-J et JJ 1-263</i> Renvoi au registre JJ. Si le nom ne figure pas dans les analyses, il faut dépouiller entièrement le registre JJ considéré pour retrouver le nom recherché.Papier
IR33-IR34	JJ99, JJ167, JJ245B, JJ249B- JJ266	Acta omissa. Inv. somm. ms. de la série JJ, actes non mentionnés dans l'inv. de A. Longnon et A. Coulon, 1898-1900, 227 p.	Papier	Acta omissa. Index ms. sur fiches de la série JJ, noms géographiques et de personnes de l'inv. de A. Coulon, 1900 : env. 8 500 fiches[papier, fiches ms., consultables en microfiches]
Registres du Trésor des Chartes, dir. R. FAWTIER ; J. GLÉNISSON et J. GUEROUT, J. VIARD, H. JASSEMINE et A. VALLÉE, Paris, 1958-1999				
IR43	JJ37-JJ50	t. I : Philippe le Bel (1300-1314)	EAD avec index matière uniquement	Index des noms : papier Index matière : .doc
IR44	JJ 50-JJ60	t. II : Fils de Philippe le Bel (1314-1328). Partie 1 : Louis X le Hutin et de Philippe V le Long (1314-1321)	Papier	Index en cours pour les noms de lieux
IR45	JJ61-JJ64	t. II : Fils de Philippe le Bel (1314-1328). Partie 2 : Charles IV le Bel (1322-1328)	Papier	Index en cours pour les noms de lieux
IR46	JJ65A-JJ69	t. III.1 : Philippe de Valois (1328-1349)	Papier	Papier
IR47	JJ70-JJ75	t. III.2 : Philippe de Valois (1336-1346)	EAD sans index	Papier
IR48	JJ 76-JJ79B	t. III.3 : Philippe de Valois (1339-1350)	EAD sans index	Papier
Inventaires dactylographiés ou manuscrits : Jean II le Bon et Charles, dauphin et duc de Normandie, lieutenant du roi ou régent, puis roi (1364-1367)				
IR 49	JJ80	Registre de chancellerie de Jean II le Bon (1350-1351). Inv. anal. ms. de JJ/80, par Y. Lanhers, 1945, 365 p.	.doc	.doc
IR50	JJ81	Registre de chancellerie de Jean II Bon (1351-1353). Inv. anal. dact. de JJ/81, par Y. Lanhers, 1974 et 1985, 504 p.	.odt	Papier
IR51	JJ82	Registre de chancellerie de Jean II le Bon (1353-1355). Inv. anal. dact. de JJ/82, par A. Vallée, 1970, 432 p.	Papier	Papier
IR52	JJ83	Registre de chancellerie de Jean II le Bon (mars 1355). Inv. anal. dact. de JJ/83, par A. Vallée, 1985, 22 p.	Papier.	Papier

IR53	JJ84	Registre de chancellerie de Jean II le Bon et de Charles, dauphin et duc de Normandie, lieutenant du roi ou régent (1354-1357). Inv. anal. ms. de JJ/84, par S. Clémencet, 1968, 288 p.	Papier	Papier
IR54	JJ84-JJ85	Registres de chancellerie de Charles, dauphin et duc de Normandie, lieutenant du roi ou régent (1356-1359). Inv. anal. ms. de JJ/85 et 86, par P. Lue et S. Clémencet, 1953, 187 p.	Papier	Papier
IR55	JJ87-JJ88	Registres de chancellerie de Charles, dauphin et duc de Normandie, lieutenant du roi ou régent (1357-1360). Inv. anal. ms. de JJ/87 et 88, par S. Clémencet, 1957, 129 p.	Papier	Papier
IR56	JJ89	Registre de chancellerie de Charles, dauphin et duc de Normandie, lieutenant du roi ou régent, et de Jean II le Bon (1356-1361). Inv. anal. ms. de JJ/89, par S. Clémencet, 1959, 215 p., deux vol. reliés en un	Papier	Papier
IR5127	JJ90 (part.)	Registre de chancellerie de Charles, dauphin et duc de Normandie, lieutenant du roi ou régent (1358-1360). Inv. anal. ms. de JJ/90, par S. Clémencet, 1960, 172 p. en deux vol.	Papier	Papier
IR57	JJ91	Registre de chancellerie de Jean II le Bon (1361-1363). Inv. anal. ms. de JJ/91, par S. Clémencet, 1962, 158 p.	Papier	Papier
IR58	JJ92-JJ93	Registres de chancellerie de Charles, dauphin et duc de Normandie, lieutenant du roi ou régent (1361-1363) et de Jean II le Bon (1362-1363). Inv. anal. ms. de JJ/92 et 93, par S. Clémencet, 1963, 180 p.		
IR59	JJ94-JJ95	Registres de chancellerie de Jean II le Bon et Charles V (1363-1364). Inv. anal. ms. de JJ/94 et 95, par S. Clémencet, 1964, 99 p.		
IR50000	JJ96-JJ98	--	--	Inv. anal. ms. sur fiches.xls
Index				
IR1810	JJ129	Registre de Charles VI (1386). Inv. anal. ms. sur fiches dans l'ordre alphabétique : 1. des noms de personnes, 2. des noms géographiques, 3. des noms de matières, par L. Douët d'Arcq, 1860 : env. 15 000 fiches.	Papier	Papier
--	JJ172-175	Registres d'Henri VI. Projet d'inventaire analytique par C. Gut, 2010		Index des noms de personnes et de lieux, des destinataires, des secrétaires et des personnes assistant à la décision de l'acte.doc

Table 2. Inventaires sélectifs, thématiques et géographiques

IDENTIFICATION	COUVERTURE	INVENTAIRE SÉLECTIF	FORMAT	INDEX
IR35	JJ34, JJ40, JJ4-JJ232	Anoblissements (1308-1499). Inv. anal. ms. des lettres d'anoblissements contenues dans les registres JJ/34, 40, 44 à 232, par Dom Carpentier. 1753 : copie faite en 1882 du ms. lat. 18345 de la Bibl. nat.	Papier<http://gallica.bnf.fr/ark:/12148/btv1b10720871k>	Papier
IR39	JJ185-JJ263B	--	--	Table des noms de lieux dans les actes de concessions ou confirmations de privilèges, foires et marchés : env. 500 fiches
IR36	JJ35-JJ266	<i>La Gascogne dans les registres du Trésor des chartes</i> [XIV ^e -XVI ^e s.], par Ch. Samaran, Paris, 1966, in-8 ^e , XVI-308 p.	Papier<http://gallica.bnf.fr/ark:/12148/bpt6k6243594l>	Papier
IR37	JJ35-JJ235	<i>Le Languedoc et le Rouergue dans le Trésor des chartes</i> [1302-1502], par Y. Dossat, A.-M. Lemasson et Ph. Wolff, Paris, 1983, 637 p.	Papier<http://gallica.bnf.fr/ark:/12148/bpt6k6244424l>	Papier
IR38	JJ80-JJ235	<i>Les Pays de la Loire moyenne dans le Trésor des chartes</i> : Berry, Blésois, Chartrain, Orléanais, Touraine, 1350-1502, par B. Chevalier, Paris, 1993, 644 p.	Papier<http://gallica.bnf.fr/ark:/12148/bpt6k62483417>	Papier

Table 3. Éditions partielles

	COUVERTURE	NOMBRE D'ACTES ÉDITÉS	ÉDITION	FORMAT
Poitou	JJ35-JJ232	1744	P. GUÉRIN, <i>Recueil des documents concernant le Poitou contenus dans les registres de la chancellerie de France</i> , 1883-1938, 14 t.	Édition TEI de l'École nationale des chartes < http://corpus.enc.sorbonne.fr/actesroyauxdupoitou/ >
Paris (1328-1350)	JJ65A-JJ78	444	J. VIARD, <i>Documents parisiens du règne de Philippe VI de Valois (1328-1350) : extraits des registres de la chancellerie de France</i> , Paris, 1899-1900	< http://gallica.bnf.fr/ark:/12148/bpt6k6439256v > ; < http://gallica.bnf.fr/ark:/12148/bpt6k6439677d >
Paris (1420-1436)	JJ171-JJ175	176	A. LONGNON, <i>Paris pendant la domination anglaise (1420-1436) : documents extraits des registres de la Chancellerie de France</i> , Paris, 1878	< http://gallica.bnf.fr/ark:/12148/bpt6k64171303 >
Normandie (1420-1436)	JJ172-JJ175	229	P. LE CACHEUX, <i>Actes de la chancellerie d'Henri VI concernant la Normandie sous la domination anglaise (1422-1435), extraits des registres du Trésor des Chartes aux Archives nationales</i> , Rouen, 1907-1909, 2 t.	< https://archive.org/details/actesdelachance02vigooq >< https://archive.org/details/actesdelachance01vigooq > Océrisation publiée sur le site de l'Université de Caen : < http://www.crisco.unicaen.fr/IMG/pdf/Le_Cacheux_2_Daphne.pdf >< http://www.crisco.unicaen.fr/IMG/pdf/Le_Cacheux_2_Cecile.pdf >.
Amiens	JJ65-JJ201	Environ 90 (380 avec les actes du Parlement)	E. MAUGIS, <i>Documents inédits concernant la ville et le siège du bailliage d'Amiens : extraits des registres du Parlement de Paris et du Trésor des Chartes</i> , Paris 1908-1921, 3 t.	< http://gallica.bnf.fr/ark:/12148/bpt6k57347287 >< http://gallica.bnf.fr/ark:/12148/bpt6k5440952w >< http://gallica.bnf.fr/ark:/12148/bpt6k5440927t >

Tables élaborées par D. Stutzmann et J.-F. Moufflet à partir de la liste des inventaires de la salle des inventaires virtuelle, consultée le 31 mai 2016, <https://www.siv.archives-nationales.culture.gouv.fr/siv/POG/FRAN_POG_02c614ua0mj6-1my9e763zwp1q>.

En ce qui concerne les inventaires systématiques, tous les registres ont fait l'objet d'un dépouillement ancien, mais extrêmement sommaire, qui s'est contenté de compiler les tables des actes qu'ils contiennent. Des quatre volumes concernés, trois sont consultables dans la salle des inventaires virtuelle : le tome 1 couvrant les cotes JJA-JJN et JJ1-JJ79, le tome 3 couvrant les cotes JJ156-JJ211 et enfin le tome 4 pour JJ212 à JJ264⁷. Quant aux autres instruments de recherche, alors que la série des registres proprement dits commence avec la cote JJ35, le chercheur dispose d'inventaires publiés pour les registres JJ37 à JJ79B (mais dépourvus d'index pour les registres JJ50 à JJ64⁸), et d'inventaires dactylographiés ou manuscrits pour les registres JJ80 à JJ96.

Les inventaires géographiques réalisés aux Archives nationales et couvrant la Gascogne, le Languedoc, le Rouergue ainsi que la Loire moyenne sont plus aisément exploitables, car ils ont été numérisés par la Bibliothèque nationale de France et leur contenu est mis à disposition sous forme de fichier texte dans la bibliothèque numérique Gallica grâce à un logiciel OCR. Les Archives nationales ne sont certes pas restées inactives, mais, si certains de ces inventaires ont été numérisés, voire convertis dans le format EAD (*Encoded Archival Description*)⁹, ils ne sont pas encore à la disposition du public ; la salle des inventaires virtuelle ne permet pas pour l'heure d'exploiter des index matières et personnes¹⁰, et les index n'ont pas

7. Tome 1 : <https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?udId=&consIr=&irId=FRAN_IR_000421&frontIr=&auSeinIR=false> ; tome 3 : <https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?udId=&consIr=&irId=FRAN_IR_000423&frontIr=&auSeinIR=false> ; tome 4 : <https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?udId=&consIr=&irId=FRAN_IR_000424&frontIr=&auSeinIR=false>. https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?udId=&consIr=&irId=FRAN_IR_000421&frontIr=&auSeinIR=falsehttps://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?udId=&consIr=&irId=FRAN_IR_000423&frontIr=&auSeinIR=falsehttps://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?udId=&consIr=&irId=FRAN_IR_000424&frontIr=&auSeinIR=false

8. Les index de ces registres, issus des règnes de Louis X, Philippe V et Charles IV, n'ont pas encore été publiés. Jean Guérout travaille néanmoins sur la finalisation des index des noms de lieux.

9. Le format de document électronique EAD (*Encoded Archival Description*) est un format basé sur le langage XML qui permet de structurer des descriptions de manuscrits ou de documents d'archives.

10. Plus précisément, les personnes qui ne sont pas identifiées comme des producteurs et qui ne bénéficient donc pas de notices d'autorité dans le référentiel *ad hoc* ne sont pas indexées par le système – cela représenterait d'ailleurs, compte tenu du nombre faramineux de noms d'individus que l'on trouve dans les archives, une masse de données considérable à gérer par le système d'information. Les noms de lieux sont quant à eux enregistrés dans

toujours été intégrés lors des opérations de rétroconversion massive des inventaires anciens, même si une recherche plein texte peut maigrement compenser l'absence d'index informatisé.

L'état des inventaires et éditions et de leur conversion met en lumière des enjeux heuristiques forts. La non-rétroconversion des index tient sans doute aussi de la difficulté à modéliser et exploiter les renvois dans le système structuré, mais « à plat », de l'EAD, dépourvu de lien systémique à des notices d'autorité ; elle fait cependant perdre les informations d'identification, autant pour les personnes que pour les lieux et les sujets. Une tension sous-jacente et peu problématisée existe ainsi entre des recherches historiques cumulant des informations à l'unité et d'autres modes d'exploitation en *distant reading*, qu'il s'agisse de statistique historique comprenant une exploitation des index ou de linguistique de corpus dans ses versants diplomatiques.

Un format unique pour les gouverner tous

Dans le détail, chacun de ces inventaires, avec ou sans ses index, présente de légères variations de format. Dans ce contexte, un premier pan du travail consiste en la création d'un format homogène et unique pour les inventaires et les éditions afin de mettre à un même niveau, au-delà de la différence essentielle entre un inventaire et une édition de texte, des informations de même nature (fig. 3).

En particulier, nous faisons le choix de considérer l'acte comme niveau de granularité fondamental, que nous disposions ou non d'une édition. Pour cette partie de l'édition électronique des inventaires et éditions de texte, nous faisons le choix de la TEI, seul format qui permette de combiner les métadonnées sur les textes, y compris sur leur tradition (original, registre, éditions), avec les transcriptions et textes des éditions en les annotant et en les enrichissant (variantes, abréviations, etc.). Dans les parties supérieures du document, nous utilisons la racine <TEI> pour les inventaires unifiés (par exemple, inventaires géographiques) et la racine <teiCorpus> pour les inventaires où se retrouvent des subdivisions par volume (inventaires systématiques), avec un élément <TEI> par volume, ainsi que pour l'édition en plusieurs volumes de Paul Guérin. Dans les deux cas, un volume est considéré comme un ensemble textuel <text> regroupant les textes que sont les actes, encodés comme <text> dans un élément <group>.

un référentiel dédié. L'étape suivante consiste à faire le lien entre un composant <c> d'un inventaire en EAD et les entités pertinentes de ce référentiel.

LE TEXTE À L'ÉPREUVE DU NUMÉRIQUE

```
<text xml:id="xxx"> <!--OR <text xml:id="with another ID"> -->
<front>
<head>9999</head> <!-- Identification number of the abstract/text within the inventory/edition -->
<docDate>
<date when="1111-11-11">1111</date>
<origPlace>Place Name</origPlace>
<docDate>
<argument>
<p>Argument</p>
<quote type="extrasigillum">Extra sigillum notes</quote>
</argument>
<div type="tradition">
<listWit sortKey="copie">
<witness>
<msDesc>
<msIdentifier>
<repository>Archives nationales</repository>
<idno>JJ000</idno>
<!-- Register shelfmark with JJ at the beginning, with additional letter A, B, C etc. if needed, without space -->
<msIdentifier>
<msDesc>
<locus scheme="folio">1r</locus>
<locus scheme="folio">1r-1v</locus>
<locus scheme="folio">1r-2r</locus>
<idno>1</idno> <!-- Identification number within the register -->
<lang cert="unknown"> <!-- for abstracts without indication -->
<lang>lat</lang> <!-- lat, fr, occ -->
</witness>
<listWit>
<listWit sortKey="edition">
<witness n="a" ana="edited">P. Guérin, <title>Archives historiques du Poitou</title>, 11, p. 1-2</witness>
</listWit>
</div>
<idp>
<index>
<term>[Index entry]</term>
</index>
</front>
<body>
<p>Text transcript if available</p>
</body>
</text>
```

Fig. 3. Exemple d'encodage des informations concernant un acte

Paris, Archives nationales de France, IR53, fol. 124v-125r.

Fig. 4. Inventaire manuscrit du registre JJ84

Inventaire analytique manuscrit, établi en 1968 par S. Clémencet, du registre de chancellerie de Jean II le Bon et de Charles, dauphin et duc de Normandie, lieutenant du roi ou régent (1354-1357) [Paris, Archives nationales de France, JJ84].

Dans le cadre du projet HIMANIS, et en nous appuyant sur le format proposé par l'École nationale des chartes pour l'édition des actes¹¹, nous assumons volontairement le choix inverse de celui qui a été fait par la même institution de publier un texte ancien en EAD, fût-il un inventaire d'archives comme celui de Saint-Denis¹². En effet, nous considérons ici que l'inventaire est un regroupement de métadonnées sur des textes, même si nous ne disposons pas (encore) de ces derniers.

De ce fait, nous ne décrivons pas les actes originaux. Il faut néanmoins noter ici une différence substantielle : nous utilisons `<msDesc>` à l'intérieur de `<witness>` pour bénéficier de toute la finesse d'encodage native, ici manifestée par l'utilisation de `<msIdentifier>` avec `<repository>` et `<idno>`¹³.

Le travail sur la formalisation des inventaires et leur édition électronique, entamé dès le début du projet, a fourni des résultats spectaculaires. En particulier, le partenaire A2iA a entraîné un « reconnaisseur », c'est-à-dire une suite logicielle adaptée pour transcrire spécifiquement une écriture, pour les inventaires manuscrits des registres JJ84 à JJ95, de sorte que ces inventaires, autrement hors d'accès du public hors de Paris, sont désormais structurés et vont pouvoir être publiés en ligne comme tous les autres inventaires des Archives nationales (fig. 4). Bien qu'utile, ce n'est néanmoins qu'un aspect secondaire et instrumental pour exploiter au mieux le potentiel d'apprentissage et automatiser l'évaluation des résultats à partir de données validées. Il en est de même pour le texte lui-même.

L'édition électronique comme « vérité terrain » du projet européen HIMANIS

Le projet HIMANIS considère l'édition électronique d'abord de façon instrumentale, comme « vérité terrain » et « terrain d'apprentissage », en mettant en crise l'approche classique de l'édition critique comme finalité. Elle s'insère désormais dans un cycle plus riche des sources et des métadonnées. Pour expliquer ce renversement apparent, il faut décrire le consortium HIMANIS, ses objectifs appliqués aux volumétries de

11. O. GUYOTJEANNIN *et al.*, « Un acte diplomatique en TEI », *École des chartes*, <<http://developpements.enc.sorbonne.fr/diple/schema/acte>>; F. GLORIEUX et V. JOLIVET, « Actes royaux du Poitou, schéma de balisage », *École des chartes, Corpus*, 2012, <<http://corpus.enc.sorbonne.fr/actesroyauxdupoitou/schema>>.

12. F. CLAUD, « The Digital Edition of the Medieval Charters of the Abbey of Saint-Denis : first Results and Prospects », *Digital Medievalist*, 8 (2013), <<http://www.digitalmedievalist.org/journal/8/claud/>>.

13. Sur ce point, nous divergeons du modèle proposé par l'École nationale des chartes, qui utilise `<orgName>` avec `@type="repository"` et `<num>` avec `@type="idno"`.

données massives, ainsi que les méthodes d'apprentissage de l'intelligence artificielle.

Le patrimoine culturel au défi des données massives

HIMANIS a proposé la numérisation et l'indexation du texte contenu dans les registres de la chancellerie royale pour permettre exploitations scientifiques et utilisations novatrices, tout en soulignant combien les registres constituent un patrimoine matériel profondément lié à l'idéologie monarchique et centralisatrice française, en étant le reflet le plus concret d'une tradition où l'État central intervient dans la vie des individus sur tout le territoire.

Le projet transdisciplinaire conjoint des enjeux technologiques et en humanités. Pour la vision par ordinateur, les principaux sont les *Key Word Spotting* (recherche par mots-clés), indexation, identification de scripteurs et recherche par négociation de contenu. Pour les sciences humaines, le premier enjeu est d'organiser l'accès à l'intégralité des documents conservés et de faciliter la recherche historique et linguistique dans ce vaste et complexe ensemble. Au-delà de cette première dimension, les véritables enjeux couvrent l'analyse de données massives, l'analyse de réseaux et la linguistique de corpus, tant en général qu'appliquées à l'histoire des institutions et aux sciences fondamentales de l'histoire telles que la diplomatie et la paléographie. Ces deux disciplines doivent en effet permettre une meilleure compréhension du fonctionnement de l'administration royale et de ses stratégies pour étendre le droit du roi.

Un corpus complexe

S'attaquer aux données massives du fonds des registres représente une tâche complexe. La volumétrie impressionne : le corpus utilisé comprend les registres proprement dits, c'est-à-dire les volumes cotés JJ35 à JJ211 des Archives nationales couvrant les années 1302 à 1483. Il pourra même être complété par dix-neuf registres, originaux ou copies, et huit formulaires, numérisés par la Bibliothèque nationale de France. Débutée avant le projet et couvrant les registres jusqu'à JJ91, la numérisation des Archives nationales est maintenant achevée, y compris pour trente-deux registres qui ont dû être traités pour améliorer l'apparence des feuillets, voire déreliés compte tenu de leur épaisseur. L'ensemble comprend déjà 70 000 images. Si l'intérêt d'une recherche en plein texte est d'autant plus évident que le corpus est vaste, les difficultés sont également accrues.

Nous avons déjà dit que le corpus est très faiblement décrit par les inventaires disponibles sous forme informatique, notamment pour la deuxième moitié de l'ensemble. Il faut maintenant en souligner la diversité

structurelle. Du point de vue linguistique, le fonds est bilingue et contient principalement des actes en latin et en français ; mais des textes en occitan, flamand et navarrais se rencontrent également¹⁴. L'exploitation d'un corpus multilingue est toujours délicate. La complexité est ici accrue par la pratique médiévale des vidimus et actes insérés, qui peut transformer un acte royal en une mosaïque multilingue : ainsi, sous Philippe VI, 46 % des actes sont concernés et contiennent entre un et dix actes insérés. En outre, la variété diplomatique des actes est grande, allant des lettres de rémission aux mandements, en passant par les amortissements, anoblissements, et confirmations, tout en intégrant souvent des copies de procurations et pouvoirs.

Les objectifs ainsi que les caractéristiques du corpus, multilingue, non homogène, et assez peu documenté, ont déterminé les partenaires à mettre en place une stratégie originale, où l'édition électronique est utilisée pour nourrir l'intelligence artificielle.

*L'édition électronique comme étape préliminaire :
format, enrichissement, ergonomie*

La possibilité d'entraîner une « machine à lire » a été testée et développée par deux des partenaires d'HIMANIS dans un projet antérieur : ORIFLAMMS (« *Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts* », financement ANR/Cap Digital, 2013-2016, réf. ANR-12-CORP-0010)¹⁵. Celui-ci visait à étudier la variabilité de l'écriture selon la formalité et les supports et, pour ce faire, se fondait sur un « alignement texte-image » automatique et exhaustif, plutôt que sur une annotation sélective, afin d'éviter le « biais de confirmation » et la surreprésentation des phénomènes rares¹⁶.

14. Le latin domine d'abord (80 % d'actes en latin) jusqu'en octobre 1330, date à partir de laquelle le français (80 % d'actes en français) le supplante. La latin redevient largement majoritaire sous Jean II. À partir de Charles V une pratique linguistique se met en place, principalement en fonction de la nature des actes enregistrés. Cf. S. LUSIGNAN, *La Langue des rois au Moyen Âge. Le français en France et en Angleterre*, Paris, 2004.

15. Agence Nationale de la Recherche, « Projet ANR. Corpus, données et outils de la recherche en sciences humaines et sociales (Corpus) 2012 : projet ORIFLAMMS », 2013, <http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-12-CORP-0010> ; Consortium Oriflamms, « Compte rendu final du projet ORIFLAMMS/ORIFLAMMS Final report », Billet, *Écriture médiévale & numérique | Écritures médiévales et lecture numérique. Carnet du projet ORIFLAMMS (Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts)*, 4 janvier 2017, <<http://oriflamms.hypotheses.org/1592>>.

16. D. STUTZMANN, « Ontologie des formes et encodage des textes manuscrits médiévaux. Le projet ORIFLAMMS », *Document numérique*, 16/3 (2013), p. 81-95, doi :10.3166/DN.16.3.69-79 ; D. STUTZMANN *et al.*, « From Text and Image to Historical Resource : Text-Image Alignment for Digital Humanists », *Digital Humanities 2015*, Sydney, 2015, <<http://>

Le principe en est le suivant : on utilise des logiciels de reconnaissance de l'écriture manuscrite (HTR, *Handwritten Text Recognition*), donc des technologies d'intelligence artificielle et des réseaux neuronaux profonds, mais, au lieu d'en attendre une transcription, on utilise une transcription qui existe déjà et on crée des modèles statistiques qui optimisent la production automatique d'un texte pour qu'il soit le plus identique possible à la transcription « humaine » fournie au préalable¹⁷. Celle-ci est appelée « vérité terrain » (*ground truth*) : comme son nom l'indique, cette « vérité terrain » est censée être un échantillon représentatif de textes tels qu'ils sont en réalité sur le terrain (par opposition au « modèle »).

Si cette « vérité terrain » peut être augmentée par des référentiels externes, tels des dictionnaires linguistiques, des référentiels d'entités nommées (noms de lieux et de personnes, concepts) et des outils de lemmatisation, elle constitue la brique essentielle de l'apprentissage et contient de fait un premier modèle linguistique (dictionnaire de mots, fréquences, etc.).

Dans le cadre d'HIMANIS, nous avons instrumentalisé ainsi l'édition électronique des *Actes royaux du Poitou*¹⁸, fondée sur l'édition donnée par Paul Guérin¹⁹, et nous l'avons utilisée comme vérité terrain pour entraîner une machine et obtenir un effet levier qui permette de se rapprocher d'une édition de l'ensemble du corpus, c'est-à-dire, au moins, d'indexer les mots présents. Cette édition, avec ses douze tomes et ses 1 744 actes au total, non contente d'être la seule disponible, a l'avantage de couvrir l'ensemble du corpus jusqu'à la fin du Moyen Âge.

L'opération d'alignement entre une édition électronique et les images constitue une base de savoir où l'on restitue une information qui a été perdue dans le processus historique de l'édition électronique, à savoir la forme graphique qui a été vue par l'éditeur scientifique et transformée en chaîne de caractères. Cette information désormais recrée est propagée pour interpréter les pages et actes inédits.

Pour favoriser l'alignement texte-image, une préparation des données est nécessaire : établir une concordance entre les textes édités et les images numériques sur lesquels ils apparaissent. Ces renvois sont enregistrés par un attribut @facs inséré dans l'élément <pb/>. Lorsque

dh2015.org/abstracts/xml/STUTZMANN_Dominique_From_Text_and_Image_to_Histor/STUTZMANN_Dominique_From_Text_and_Image_to_Historical_R.html>.

17. T. BLUCHE, D. STUTZMANN, C. KERMORVANT, « Automatic Handwritten Character Segmentation for Paleographical Character Shape Analysis », *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016, p. 42-47, doi :10.1109/DAS.2016.74.

18. <<http://corpus.enc.sorbonne.fr/actesroyauxdupoitou/>>.

19. P. GUÉRIN et L. CELIER, *Recueil des documents concernant le Poitou contenus dans les registres de la chancellerie de France*, 14 vol., Poitiers, 1881, <<http://gallica.bnf.fr/ark:/12148/bpt6k209478j>>.

l'acte ne couvre pas toute la page, l'attribut @facs renvoie à une <zone> déclarée dans une section <facsimile> entre le <teiHeader> et le <text>. Ensuite, pour faciliter l'apprentissage de la lecture, la transcription a été enrichie des indications de passage de ligne par l'élément <lb/> dans près de 300 actes, parmi les 474 donnés par Paul Guérin dans ses trois premiers tomes et correspondant à la tranche chronologique des registres JJ35-JJ91, déjà numérisée par les Archives nationales en début de projet et sur laquelle s'est fondé l'apprentissage.

Cet investissement en temps n'est pas négligeable. En outre, nous avons également enrichi vingt-sept transcriptions en marquant les abréviations, avec l'espoir que la lecture automatique en serait facilitée et permettrait en retour d'enrichir les éditions à venir et d'étudier les évolutions du système abrégatif de façon diachronique. En pratique, un premier alignement a été produit avant l'indication des sauts de ligne ; il a été intégralement corrigé pour offrir une meilleure base d'apprentissage et c'est à cette occasion qu'ont été insérés les éléments <lb/> dans l'édition²⁰.

L'effort minutieux d'enrichissement de l'édition électronique a permis d'en voir de nombreuses faiblesses, allant d'erreurs de transcription à des choix éditoriaux très invasifs, tout en validant son intérêt pour entraîner l'intelligence artificielle²¹. À cet égard, il est maintes fois apparu que les sciences humaines entretiennent un rapport ambigu avec la qualité de leur production : puisque celle-ci ne peut pas être strictement évaluée, l'impression d'une perfection supposée est peut-être donnée trop facilement à l'extérieur ; en tout cas, les partenaires ingénieurs attendent une précision de 100 % alors que la réalité médiévale – à commencer par les abréviations – fait dépendre la perfection des attentes et objectifs de l'édition. Ainsi, les partenaires s'attendaient à une édition qui indique de façon exhaustive les abréviations, car – du moins le pensions-nous initialement – seule la production d'une chaîne de caractères reflétant celle présente sur les images pouvait être attendue des machines avec restitution selon les règles et statistiques d'abrègement. Or, pour des raisons ergonomiques et cognitives, il est impossible de fournir un tel corpus ! L'apprentissage des abréviations de façon à enrichir automatiquement les transcriptions à des fins d'analyse paléographique reste un problème non résolu, mais il se trouve que la machine a finalement appris à lire en « lecture globale » en restituant les

20. Il faut souligner ici que les technologies utilisées par A2iA et UPVLC exploitent uniquement l'information ligne par ligne (même quand on dispose d'un alignement au niveau des mots, c'est la ligne qui sert à apprendre).

21. Les autres éditions d'actes n'ont pas été utilisées, parce qu'elles portaient sur des volumes non numérisés en début de projet. Le formulaire d'Odart de Morchesne (Paris, BnF, fr. 5024), dont le microfilm a été numérisé par la BnF et dont le texte a été édité par l'École nationale des chartes (<<http://elec.enc.sorbonne.fr/morchesne/>>), pourra servir à mesurer la qualité d'une transcription générée automatiquement.

abréviations ; le problème de la qualité et de sa définition reste crucial pour tous les projets en sciences humaines.

Il faut aussi prendre en compte les questions d'ergonomie. En effet, cette dernière tient une place capitale car elle conditionne à la fois la masse et la qualité des données, puisqu'un outil ergonomique permet de multiplier les données correctement formatées avec un investissement moindre en temps. Dans le projet HIMANIS, tant pour l'annotation préalable des zones concernées par l'édition de Paul Guérin qu'ensuite, pour corriger l'alignement, nous avons utilisé le logiciel Transkribus²² développé par le projet Transcriptorium et continué par le consortium READ dont fait partie l'Universitat Politècnica de València, partenaire d'HIMANIS (fig. 5). Ce logiciel permet un alignement très précis et très descriptif (indication des abréviations, des mots suscrits ou rayés...) et, fondé sur le format PAGE, offre des possibilités d'exporter le résultat en différents formats (METS, TEI, PDF, RTF, Excel).

Ayant, dans un premier temps, instrumentalisé l'édition électronique et subverti ce sommet de l'activité d'éditeur de texte pour en faire le marchepied d'une simple « indexation », il convient maintenant de clarifier les concepts et le cycle des métadonnées.

« Indexer », « analyser », « éditer », « évaluer » : le cycle des métadonnées et travaux en cours

Le mot « index » est polysémique. S'il désigne toujours un accès tabulaire à une information textuelle linéaire, l'index est à la fois le niveau zéro de l'édition et son degré ultime. C'est l'étape finale qui permet de créer un sens supplémentaire dans une édition et de révéler ce qui n'y est que sous-jacent : quand l'éditeur fournit un index des citations et réminiscences bibliques d'un texte, il ouvre une porte vers une compréhension renouvelée, non seulement de l'œuvre, mais aussi de la pensée de l'auteur. L'index, en ce sens, parachève et illustre la compréhension de l'édition, et est l'aboutissement du travail ecdotique et de la critique textuelle. Les inventaires « analytiques » d'archives portent en leur nom leur finalité d'indexation puisque l'acte ultime de l'analyse est l'identification. Pourtant, l'index est aussi un degré minimal. C'est dans ce sens qu'il est au fondement même des humanités numériques en la forme précise de l'*Index Thomisticus* de Roberto Buso (S. J.). L'index est l'état d'un texte en ruine dont les morceaux éparpillés ont été rassemblés et triés, souvent alphabétiquement : toute compréhension en est retirée, mais il contient un potentiel de lectures nouvelles par les raccourcis, le contexte et les rapprochements (fig. 6).

22. <<https://transkribus.eu/Transkribus/>>.

Image reproduite avec l'autorisation des Archives nationales.

Fig. 5. Interface du logiciel d'édition et d'alignement Transkribus (juin 2016)

Interface et logiciel : © 2013-201xDEA. Transkribus is part of the READ project and has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement no. 674943. Document représenté : Paris, Archives nationales de France, JJ65A, f. 1r.

Fig. 6. Index des mots produit par Monk pour le registre JJ35 au 18 mai 2016

Dans le projet HIMANIS, « l'index » que nous créons est, de même, une porte d'accès au texte, à la fois plus précise et moins intelligente. Il vise en effet à retrouver les occurrences d'un terme choisi dans l'immense continent des registres de la chancellerie royale (fig. 7).

Cet index est en-deçà du sens des actes royaux et de l'identification des acteurs et des lieux, mais ouvre la voie à de nouvelles analyses. Il donne évidemment accès à des parties du fonds non inventoriées, mais surtout il modifie les questions possibles : d'un côté, il crée un accès à la textualité de la source, complètement abolie par l'opération d'inventoriage et d'analyse ; de l'autre, fondé sur l'analyse graphique, il permet des recherches sur la forme même des mots (fig. 8).

Pour établir un index pertinent et efficace pour la recherche, en préservant la valeur sémantique des textes, les partenaires du projet HIMANIS cherchent à exploiter des données d'autorité et des référentiels, ainsi que des outils linguistiques.

Référentiels et outils linguistiques

Les référentiels linguistiques, tout d'abord, ont un potentiel important pour améliorer la reconnaissance et l'indexation du texte en régularisant et normalisant des termes que les outils automatiques auraient mal lus et dont ils auraient mutilé l'un ou l'autre caractère. Les indices de fréquence permettent en outre de calculer la « confiance » d'un mot de façon bien plus pertinente, de façon à offrir un meilleur service de recherche. La lemmatisation, utile aussi bien en latin qu'en ancien et moyen français, affine non seulement le calcul de pertinence et de confiance des termes indexés, mais permet surtout une recherche plus intelligente, par radical, indépendamment des formes déclinées ou conjuguées, tout en ouvrant la voie à des analyses lexicométriques qui tiennent compte de la construction syntaxique. Pour HIMANIS, ces outils doivent couvrir et le français et le latin. Un premier référentiel est constitué par le texte des éditions qui fournissent un premier « dictionnaire », c'est-à-dire, dans le contexte des sciences de l'ingénieur, une liste de mots avec des statistiques de fréquence. Les référentiels externes au corpus comprennent à la fois des dictionnaires définissant les termes, mais sans fréquences (Du Cange, *NGML* pour le latin ; Godefroy, Tobler-Lommatzsch et *DMF 2015* pour le français)²³, et des outils de lemmatisation (TreeTagger et Colatinus pour le latin ; LGeRM

23. Du Cange (<<http://ducange.enc.sorbonne.fr>>) ; Godefroy (<<http://www.eonet.ne.jp/~ogurisu/Fr/Godefroy.html>>) ; Tobler-Lommatzsch (<<http://www.uni-stuttgart.de/lingrom/stein/tl/allgemeinf.htm>>) ; *Novum Glossarium Mediae Latinitatis* (<www.glossaria.eu/ngml/>) ; *Dictionnaire du Moyen Français* (DMF 2015, <<http://atilf.atilf.fr/dmf.htm>>).

Monkweb © 2005-2014 Lambert Schomaker.

fig. 7. Visualisation des occurrences d'un mot dans l'interface du logiciel MONK

Fig. 8. Exemple de recherche par mot « abbatisse »

Différentes couleurs, ici reproduites en niveaux de gris, marquent l'indice de confiance de chaque occurrence encadrée (interface de mai 2017).

pour le français)²⁴. Pour le latin, à titre de comparaison, nous disposons également des fréquences des mots dans la *Patrologie latine* et dans les *CBMA*²⁵.

Ces outils, malgré leur potentiel heuristique fort, n'ont pas encore fait l'objet d'implémentation dans la chaîne de traitement d'HIMANIS et il est impossible de fournir ici une évaluation de leurs apports.

Entités nommées

Un second domaine où l'indexation et les analyses disponibles peuvent améliorer l'indexation par mot d'HIMANIS est celui des entités nommées, d'abord pour les noms de lieux et de personnes, mais aussi pour les concepts et matières.

Plusieurs référentiels sont issus des travaux menés aux Archives nationales sur les registres de la chancellerie royale : il s'agit simplement, une fois ocrisés, des index publiés en appendice des inventaires systématiques et géographiques. Les partenaires d'HIMANIS ont travaillé à les nettoyer, puis à les rétroconvertir et, en reprenant le modèle proposé en EAD, à recréer des liens structurels entre les références des actes et les termes indexés grâce aux éléments <index> et <term> (nous n'avons pas traité de façon structurelle les renvois qui ne sont pas les mêmes d'un inventaire à l'autre). Grâce à l'excellente qualité de la reconnaissance d'écriture manuscrite produite par A2iA, nous avons également pu récupérer les analyses et entrées d'index géographiques des inventaires manuscrits. Ces index ont l'avantage de porter directement sur les volumes étudiés et de contenir les identifications absentes des regestes. Néanmoins, ils mélangent noms de lieux et noms de personne – alors que le traitement adéquat des anthroponymes comprenant un nom de lieu est une difficulté commune dans les études médiévales ; les inventaires géographiques ne couvrent pas tout le territoire ; surtout, tous ces index ne donnent que la forme actuelle et ne peuvent donc enrichir les dictionnaires de formes pour la reconnaissance de texte²⁶. Ils servent ainsi, d'une part, d'étalon pour mesurer le degré de précision de l'indexation automatique en recherche par mot en fournissant

24. TreeTagger (<<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>>) ; Collatinus (<<http://outils.bibliissima.fr/collatinus/>>) ; LGeRM (<<http://www.atilf.fr/LGeRM/>>).

25. CBMA (<<http://www.cbma-project.eu/>>). Nous remercions Nicolas Perreaux de nous avoir communiqué ces instruments.

26. Les entrées d'index étant reliées à un numéro d'acte de l'inventaire et l'analyse indiquant cote, numéro d'acte et folio, il semblait envisageable d'utiliser l'information selon laquelle un mot « ressemblant » à l'entrée d'index se trouve « dans les environs » de telle image (parce qu'un acte peut s'étendre sur plusieurs pages). Un tel processus se révèle à la fois trop flou et trop contraignant pour être implémenté dans l'analyse de l'intelligence artificielle.

non pas une « vérité terrain », mais une liste minimale des occurrences à identifier, sans toutefois en fournir la forme précise ; d'autre part, leur rétroconversion sous forme informatique prépare des visualisations de données par cartes et frises chronologiques, reliant les toponymes entre eux et aux dates où on les trouve mentionnés.

Parmi les référentiels externes, outre des données universelles telles que GeoNames et Wikipedia (ou plutôt DBpedia) qui permettent d'obtenir les coordonnées géographiques, nous désirons utiliser la série des dictionnaires topographiques²⁷. Ceux-ci ont été convertis dans un format XML selon une DTD spécifique par le Comité des travaux historiques et scientifiques. Même s'ils ne couvrent pas non plus toute la zone géographique concernée par les registres médiévaux de la chancellerie royale, ils présentent l'intérêt de fournir les formes anciennes et les formes latines.

Comme les outils linguistiques, les référentiels d'entités nommées ont été préparés dans une première phase du projet, mais ne sont utilisés que dans la seconde phase.

Des données à l'utilisateur : crowdsourcing, granularité, ergonomie et pertinence d'accès

L'énergie consacrée aux référentiels d'entités nommées, anthropo- et toponymiques, est un investissement évidemment nécessaire, car les communautés d'utilisateurs, grand public, généalogistes et historiens médiévistes, vont naturellement chercher les informations de cette nature. Si cet aspect se vérifie dans toutes les communautés en ligne utilisatrices d'archives numérisées²⁸, il ne fait que prolonger le besoin qui présidait à la réalisation des inventaires analytiques.

Pourtant, il est un autre aspect où les attentes de l'utilisateur nécessitent un investissement particulier : celui de la pertinence de la granularité d'accès (fig. 9).

En effet, tout utilisateur d'édition d'actes au format papier sait qu'il lui faut d'abord déterminer si l'index qu'il utilise renvoie à la page ou au numéro d'acte. L'index de l'édition de Paul Guérin renvoie aux pages ; ceux des inventaires analytiques renvoient à l'acte. La même question se pose pour le texte des registres. Pour l'ordinateur, ces mots sont d'abord sur une *image*, donc sur une *page*. Or, il est complexe, même pour les parties

27. <<http://cths.fr/dico-topo/>>.

28. *Nouveaux usages, nouveaux usagers : quels contenus, quels services allons-nous offrir ?*, Actes des rencontres annuelles de la section Archives départementales (RASAD) de l'Association des archivistes français, *Gazette des archives*, 227 (2012) ; H. WREDE et ICARUS (INTERNATIONAL CENTRE FOR ARCHIVAL RESEARCH), « What is the Topotheque ? Topotheque Portal », *Topotheque. Our History, our Archive*, 2015, <<http://www.topothek.at/en/what-is-the-topotheque/>>.

inventoriées, d'établir quelles parties de chaque image appartiennent à chaque acte. Contrairement à ce que l'on pourrait croire, dans un contexte où les espaces entre les actes (quand il y en a) sont partiellement occupés par la copie des mentions hors teneur, voire par les rubriques et analyses postérieures, il est difficile de « découper » une page en actes. En outre, l'affectation de « zones d'images » à des « actes » nécessite de disposer de la liste des actes et des informations complètes sur l'emplacement du début et de la fin de l'acte, ainsi que d'une concordance entre le nom des images et la foliotation effective. L'infrastructure doit donc être à plusieurs niveaux : l'indexation des mots renvoie à des coordonnées sur une image, qui renvoient à une page ; ces coordonnées doivent être confrontées à celles des actes pour savoir auquel de ceux-ci attribuer le mot.

La granularité d'accès et ses rapports avec les besoins de l'utilisateur peuvent également motiver à interroger d'autres niveaux d'information. Si l'acte unitaire est le niveau pertinent, les vidimus et actes insérés pourraient faire l'objet d'un traitement particulier pour accroître la pertinence heuristique : ici, comme pour l'édition de texte, des choix d'accès différents sont théoriquement possibles²⁹. De même, le diplomate désirera un accès à une autre granularité textuelle et graphique, où le texte, central sur la page, cèdera sa place à des informations marginales, mais cruciales pour l'étude de la chancellerie : titres, rubriques, annotations, mentions de collation, traitement des mentions hors teneur.

La granularité d'accès doit avoir un corollaire ergonomique et cognitif : un environnement qui permette la navigation par acte, où chacun serait muni des informations déjà disponibles ou rétroconverties (analyse, édition, indexation scientifique et indexation automatique). C'est l'objectif qu'HIMANIS s'est fixé dès le début du projet, mais, s'il semble le plus évident et le plus naturel qui soit, l'on comprendra aisément qu'il soit extrêmement difficile à atteindre, car il est impossible à automatiser et nécessite un contrôle par un personnel qualifié, capable au minimum de repérer la foliotation médiévale en chiffres romains, de déterminer si un acte se poursuit sur la page suivante, et de contrôler que numérotation et contenu concordent avec la teneur et la numérotation (dont on sait qu'elle est souvent erronée) des éditions et inventaires les plus anciens (fig. 9).

Le choix est fait, pour la fin du projet, de donner accès à l'ensemble des informations, même celles qui n'auront pas été vérifiées. Cela vaut à la fois pour la segmentation en actes et pour l'indexation elle-même. Cette intervention est déjà intégrée par l'infrastructure MONK qui permet aux utilisateurs de vérifier les entrées d'index et de les corriger le cas échéant, tant en segmentation qu'en valeur (fig. 10).

29. École nationale des chartes, *Conseils pour l'édition de textes médiévaux, fascicule II, Actes et documents d'archives*, Paris, 2001, p. 211-219.

Fig. 9. Granularité d'accès et environnement de visualisation :
maquette présentant un acte sur deux pages et les informations concernant l'acte

Monkweb © 2005-2014 Lambert Schomaker.

Fig. 10. Interface de travail de MONK : correction de l'index et visualisation des différentes formes indexées

Retour sur l'indexation et l'édition électronique

Avec ou sans l'intervention des utilisateurs, et avec la possibilité de « propager » les lectures et améliorations³⁰, donc de repérer les autres formes-mots qui auraient été indexées de façon similaire et de leur attribuer le nouveau label, il nous faut maintenant revenir à une différence essentielle entre l'indexation et l'édition. En effet, l'indexation, bien que plus pauvre sémantiquement, se présente de façon plus riche. Il ne s'agit pas d'un texte « à plat », mais d'un « treillis », d'une multiplicité d'interprétations possibles pour chaque mot sur l'image, avec un indice de confiance pour chacune de ces hypothèses.

Il est possible de restituer un texte linéaire en collant bout à bout les hypothèses avec le meilleur score, mais c'est s'exposer à des erreurs nombreuses et sensiblement moins interprétables pour l'utilisateur qu'un modèle de négociation de contenu où l'on choisit directement le degré de bruit informationnel que l'on accepte dans sa requête (voir fig. 10, où la bonne indexation est en seconde position). Les deux accès présentent des fonctionnalités et des intérêts différents : la lecture linéaire, même fautive, privilégie le sens, tandis que l'indexation donne la priorité aux atomes d'information et prépare la re-création d'un sens à partir des co-occurrences et réseaux lexicaux et sémantiques.

Même dans le cas d'une restitution linéaire du texte indexé, et même si l'identification des toponymes peut être préparée automatiquement, l'on reste loin d'une édition critique numérique ou d'une édition savante électronique (*scholarly digital edition*)³¹. Il manque aussi bien le repérage et la collation textuelle des témoins que toute la critique textuelle et historique.

Évaluer l'incertitude : précision, rappel et utilisabilité

Si l'on ne vise pas à une édition critique, il reste néanmoins nécessaire d'évaluer la qualité de l'indexation ou du texte restitué. C'est là

30. R. McNICHOLL et T. MILES-BOARD, « TranScriptorium : Computer-Aided, Crowd-Sourced Transcription of Handwritten Text (for Repositories) », dans *10th International Conference on Open Repositories (OR2015)*, 2015 ; TranScriptorium et J. SÁNCHEZ PEIRÓ, « Interactive-Predictive Handwritten Text Recognition. TranScriptorium », *TranScriptorium*, <<http://transcriptorium.eu/interactive-predictive-handwritten-text-recognition/>>; J. ANDREU SANCHEZ *et al.*, « ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset », dans *Proceedings of International Conference on Frontiers in Handwriting Recognition*, 2016, p. 630-635.

31. P. SAHLE, « Scholarly Digital Editions. Complete List A-Z », 2008, <http://www.uni-koeln.de/~ahz26/vlet/vlet_a-z.html>; P. SAHLE, *Digitale Editionsformen : zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*, 3 vol, Norderstedt, 2013 ; E. PIERAZZO, *Digital Scholarly Editing : Theories, Models and Methods*, Farnham (Surrey)/Burlington, 2015.

que la conjonction des différentes disciplines apporte des nouveautés. En effet, il est de règle dans les sciences de l'ingénieur de mesurer la qualité des résultats, en particulier en termes de « précision » et de « rappel ». La précision, en matière d'indexation, est le pourcentage d'occurrences correctes parmi celles identifiées comme répondant à une requête ; le rappel est le pourcentage d'occurrences correctement identifiées parmi l'ensemble des occurrences dans le corpus. Cette approche mesurée et quantifiée n'étonne guère dans les humanités numériques, mais modifie profondément l'approche classique de l'ecdotique où la perfection supposée ne laisse guère de place à la notion d'évaluation, d'abord, ni à celle d'incertitude, ensuite.

Mesurer la performance de l'indexation impose de constituer une vérité terrain représentative. Or, il n'est pas possible de définir la représentativité *a priori*, et la vérité terrain utilisable pour évaluer la qualité de l'indexation pose des problèmes de méthode. Il s'agit en effet, d'une part, des éditions qui ont précisément servi à l'entraînement et ne sont donc pas une mesure neutre, puisque la machine a pu « apprendre par cœur » les passages concernés, et, d'autre part, des index des inventaires, qui présentent non seulement des formes normalisées et modernisées, mais sont en outre constitués d'entités anthro- et toponymiques moins bien reconnues car elles ne peuvent pas être prédites par des modèles linguistiques.

L'évaluation par la double mesure de précision et de rappel est un indicateur de performance clair, mais difficile à établir. En outre, sa signification exacte pour la recherche historique demeure à déterminer. La précision, par exemple, est d'assez peu de conséquence car les historiens dispensés de la charge de lire 70 000 pages trieront volontiers les occurrences fausses ; en revanche, le rappel a des conséquences importantes si l'on vise à l'exhaustivité ou, par exemple, à l'identification de l'émergence d'un phénomène par ses premières apparitions. Dans de nombreux cas où l'historien cherchera un fait par un seul acte, la mesure moyenne du rappel sera indifférente à la recherche ponctuelle, qui ne connaîtra d'abord que l'alternative entre l'occurrence trouvée et le silence. Pourtant, face au silence, de nouvelles armes sont fournies et des stratégies de recherche pourront être mises en œuvre, que l'on pourrait qualifier d'« agiles », selon le terme des développements informatiques et de gestion de projet. Contrairement à un index où le silence est la fin de la recherche, la requête par mot offre des moyens de contourner un silence ponctuel, dû à un mot mal reconnu, en cherchant des formes proches ou des mots pouvant apparaître dans le même contexte.

Autre indicateur de performance, l'indice de confiance appliqué à chaque indexation permet de proposer des lectures plus riches, non limitées au résultat avec le meilleur score ; mais comment l'utiliser pour évaluer la performance et quelles sont ses implications sur l'utilisabilité ?

Ces questions appliquées à la reconnaissance des scribes sont encore plus délicates. Si Jean Guérout a proposé d'identifier 109 mains différentes et actives dans les registres JJ37 à JJ50, le tableau en est précédé par la remarque suivante : « On sait toute la difficulté que présente l'étude des écritures d'une chancellerie et l'on peut être amené à se demander si les résultats des distinctions qu'elle amène à faire entre les mains sont d'une solidité à toute épreuve³². » Quel usage faire de cette vérité terrain ? La reconnaissance de scribes dans les sources médiévales fait désormais l'objet d'une bibliographie établie³³. Il n'en demeure pas moins qu'en l'absence de preuve sur caractères internes, le paléographe reste dans l'attribution ; il relève de sa responsabilité et de son jugement critique d'établir les frontières entre le même et l'autre. Sur la distinction des mains, il n'y a donc pas, à proprement parler, d'évaluation quantifiée possible. En revanche, les technologies de *writer identification* sont aussi les mieux à même de mesurer objectivement le « changement graphique » (*cambio grafico*) dans un milieu social homogène tel que celui des clercs de chancellerie.

C'est une étape nouvelle dans le cycle des métadonnées qui émerge, entre la bibliothèque, les archives numériques et l'édition électronique. La fusion, l'intégration et la mise à disposition d'instruments de recherche améliorés, en particulier la numérisation des inventaires manuscrits et dactylographiés, doivent accompagner la numérisation des sources comme l'exploitation des éditions et offrir une navigation entre sources numérisées, instruments de recherche et éditions. Le tout se voit enrichi de l'accès au « (plein) texte » comme à un nouveau niveau de granularité (celui du « texte » de l'acte). Dans cet univers où le texte indexé d'une source médiévale est une métadonnée parmi d'autres, le projet HIMANIS permet ainsi de voir une étape inattendue de l'édition électronique qui, *horribile dictu*, remet à

32. J. GLÉNISSON et J. GUÉROUT, *Registres du trésor des chartes : inventaire analytique, I : Règne de Philippe le Bel*, éd. R. FAWTIER et A. CHAMSON, Paris, 1958, p. XLVIII-LV.

33. J. SMIT, « Paleography and the Digital Middle Ages : Experiences with the Groningen Intelligent Writer Identification System (GIWIS) », New York, 2010 ; Id., « Meten is weten ? De toepassing van het Groningen Intelligent Writer Identification System (GIWIS) op Hollandse kanselarij-oorkonden, 1299-1345 », *Bulletin de la Commission royale d'Histoire*, 176/2 (*Chancelleries princières et Scriptoria dans les anciens Pays-Bas, X^e-XV^e siècles/ Vorstelijke kanselarijen en Scriptoria in de Lage Landen, 10de-15de eeuw*, 2010), p. 343-60 ; V. CHRISTLEIN et al., « Automatic Writer Identification in Historical Documents : A Case Study », *Zeitschrift für digitale Geisteswissenschaften*, 2 (2016) ; A. A. BRINK et al., « Writer Identification Using Directional Ink-Trace with Measurements », *Pattern Recognition*, 45/1 (2012), p. 162-171, doi :10.1016/j.patcog.2011.07.005 ; L. SCHOMAKER et M. BULACU, « Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26/6 (2004), p. 787-98, doi :10.1109/TPAMI.2004.18.

l'honneur la livraison brute et potentiellement erronée d'un texte, telle celle des érudits d'avant le tournant philologique et la construction de l'ecdotique comme science. Cela devient aujourd'hui possible et se justifie car le texte n'est pas seulement livré de façon linéaire et figée, mais sous forme d'index malléable, parce qu'ainsi il ne se donne pas pour autre chose qu'il n'est ; il tient compte du travail critique passé et pourra accueillir la critique future.

Cet ensemble se place dans l'univers des données ouvertes et liées, qui offre la possibilité de visualisations cartographiques des actes par toponymes indexés, et ouvre la voie à des réexploitations à l'extérieur des interfaces du projet. La publication des métadonnées créées sous licence libre, tant l'indexation des mots que les identifications, n'épuise pas la question des droits de propriété intellectuelle qui portent ici aussi sur des technologies produites par des équipes dont certaines sont privées et en concurrence dans un marché commercial. Les défis technologiques encore ouverts portent grandement sur l'intégration de modèles linguistiques et la gestion adéquate d'un corpus bilingue, avec une reconnaissance de la langue des documents ajoutée à la reconnaissance du texte afin de proposer des résultats plus pertinents.

Si l'indexation est une étape transitoire vers une édition électronique qui ne viendra peut-être jamais, elle pose la question dès maintenant de transformer les *big data* en informations pertinentes pour l'historien médiéviste. L'exploitation des référentiels de noms de lieux et de personnes n'évite pas les « faux positifs » et rend complexe une analyse de réseaux, où les nombreuses homonymies médiévales imposent une prudence extrême, car la double menace sur les humanités numériques est toujours la même : répéter des évidences ou tirer des conclusions faussées. Nous devons apprendre à interpréter l'incertitude ou à l'utiliser d'une façon consciente et mesurée pour expliciter ce que « vraisemblablement » veut dire et rendre le raisonnement historique davantage « traçable ». Cela impose à l'historien, plus encore que dans l'analyse qualitative, de définir la pertinence de ses sources pour la question traitée, la qualité des hypothèses conditionnelles et l'applicabilité des méthodes heuristiques.

Dominique Stutzmann – IRHT-CNRS, UPR 841

Jean-François Moufflet – Archives nationales de France

Sébastien Hamel – IRHT-CNRS, UPR 841

La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique

HIMANIS (Historical MANuscript Indexing for user-controlled Search) est un projet de recherche européen, associant, sous le pilotage de l'IRHT (CNRS, France), la société innovante A2iA (France), la Rijksuniversiteit Groningen (Pays-Bas) et l'Universitat Politècnica de València (Espagne). Il vise à l'indexation du texte des registres de la chancellerie royale française des années

1302-1483, conservés aux Archives nationales sous les cotes JJ35 à JJ211, à partir des images produites par leur numérisation. Les enjeux de recherche d'information (données massives et bruitées) permettent de conjointre les enjeux technologiques (reconnaissance de l'écriture manuscrite) et historiques (analyses paléographiques et diplomatiques, recherche sur les institutions, le fonctionnement de la monarchie, la naissance de l'État-nation). La présente contribution propose un modèle d'accès à l'information dans un corpus de données massives, d'un point de vue tant ergonomique qu'herméneutique. À cette fin, après une présentation du corpus, des outils actuels pour accéder à l'information qu'ils contiennent et de leur formalisation en TEI, elle problématise l'édition électronique comme « vérité terrain » et « terrain d'apprentissage », en renversant l'approche classique de l'édition critique comme finalité. Enfin, elle décrit le modèle d'accès proposé, à la fois pour une approche par « indexation » (et non par transcription) et pour une granularité par acte.

Archives nationales – chancellerie royale – édition électronique – HIMANIS
– humanités numériques – recherche en plein texte.

Full Text Search in Medieval Manuscripts: Issues and Perspectives of the HIMANIS Project for Electronic Publishing

HIMANIS (Historical MANuscript Indexing for usercontrolled Search) is a European research project involving the innovative A2iA company (France), the Rijksuniversiteit Groningen (the Netherlands) and the Universitat Politècnica de València (Spain). It aims at indexing the full manuscript text of the Royal French Chancery registers from their digital images from 1302 to 1483, kept in the National Archives under the call numbers JJ35 to JJ211. The challenges for information retrieval in these registers (massive and noisy data) combine both technological (recognition of handwriting) and historical issues (paleographic and diplomatic analyzes, research on institutions, the functioning of the monarchy, national state genesis) issues. This paper proposes a model for accessing information in a massive body of data, both from an ergonomic point of view and from an hermeneutic one. To this end, after presenting the corpus, the current tools for accessing information contained in the corpus and their formalization in TEI format, it challenges electronic publishing as a "ground truth" and as "training data", by aiming to reverse the classical approach of critical editions. Finally, it describes the proposed access model, both for an "indexing" approach (which is not a transcription) and for a granularity per act.

Digital Edition – Digital Humanities – Full Text Search – HIMANIS – National Archives (France) – Royal French Chancery.