

ÉCOLE NATIONALE DES CHARTES

---

**Gwenaëlle Patat**

*Licenciée ès lettres modernes et histoire*

*Diplômée du master « Mondes médiévaux »*

# L'étude des livres d'heures à la lumière du numérique

Le cycle de vie des données et des  
métadonnées – Analyser, Modéliser,  
Structurer, Visualiser.

Mémoire pour le diplôme de master  
« Technologies numériques appliquées à l'histoire »

2020



# Résumé

Ce mémoire est le fruit d'une réflexion autour de problématiques abordées lors de mon stage à l'IRHT dans le cadre du projet HORAE, projet qui propose d'étudier les pratiques religieuses de la fin du Moyen Âge à travers les livres d'heures. Le stage s'est concentré sur la production de données de qualité pour la recherche historique, ce qui nous a amené à réfléchir aux stratégies de développement pour la structure, la modélisation, la conversion et la présentation ergonomique des données et des métadonnées, avec le souci constant de garantir leur interopérabilité et leur pérennité. La participation au projet HORAE, qui s'inscrit pleinement dans les Humanités numériques, nous a permis de cerner les apports, les limites et les promesses d'évolution du numérique dans la recherche en sciences humaines et sociales.

**Mots-clés :** Formats et structuration automatique des données et métadonnées ; livres d'heures ; usages liturgiques ; dévotion au Moyen-Âge ; gestion et management de projet en Humanités numériques ; Reconnaissance automatique de texte ; *Handwritten Text Recognition* ; *Optical Character Recognition* ; Interopérabilité ; XML-TEI ; XSLT ; ODD ; Schematron ; Python ; *Machine learning* ; Ergonomie ; Modélisation ; Base de données relationnelles.

**Informations bibliographiques :** Gwenaëlle Patat, *L'étude des livres d'heures à la lumière du numérique. Le cycle de vie des données et des métadonnées – Analyser, Modéliser, Structurer, Visualiser*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Jean-Baptiste Camps et Dominique Stutzmann, École nationale des Chartes, 2020.



# Remerciements

Mes remerciements vont tout d'abord à l'équipe de l'IRHT qui m'a accueillie et m'a intégrée à ses réunions quotidiennes : bien évidemment mon tuteur Dominique Stutzmann, qui a eu la patience de stimuler ma réflexion et de me faire part de sa rigueur et de son exigence. Je remercie également les deux autres membres de l'équipe HORAE, Laura Lebarbey et Louis Chevalier, pour avoir répondu avec enthousiasme à mes doutes et à mes questions, ainsi que Sergio Torres et Ian Johnson pour leur aide et leurs conseils techniques. Je souhaite également remercier l'équipe de Teklia pour m'avoir apporté des précisions sur le projet et les problématiques relatives au *machine learning*.

Je tiens à remercier l'ensemble des professeurs et intervenants qui nous ont formés lors de cette année de master, en particulier mon tuteur Jean-Baptiste Camps, ainsi que Thibault Clérice, Ariane Pinche, Gautier Poupeau et Rémy Delmotte pour avoir répondu à mes questions et apporté leur soutien durant le stage. Merci également à Laura Albiero, avec qui j'ai eu la chance d'échanger lors d'une formation continue en septembre dernier.

Enfin, je remercie chaleureusement Catherine Patat, Alain Patat et Marie Millarec pour leur relecture, ainsi que l'ensemble de la promotion 2020 dans laquelle la solidarité et la stimulation intellectuelle n'ont pas manqué, en particulier Jean-Luc Mirepoix et Chloë Fize pour leur aide et leur soutien sans faille.



# Bibliographie

## Manuscrits, livres d'heures, liturgie et dévotion au Moyen Âge

ANR (AAPG), IRHT, TEKLIA et LS2N, *HORAE Heures : Reconnaissance de l'écriture manuscrite, catégorisation automatique, éditions Hours - Recognition, Analysis, Editions*, 2017.

BAROFFIO (Giacomo), « Testo e musica nei libri d'ore », *Rivista Italiana di musicologia*, XXXIV (2011), p. 19-87, URL : [https://www.academia.edu/39811280/Testo\\_e\\_musica\\_nei\\_libri\\_dore](https://www.academia.edu/39811280/Testo_e_musica_nei_libri_dore) (visité le 24/07/2020).

BURROWS (Toby), HYVÖNEN (Eero), RANSOM (Lynn) et WIJSMAN (Hanno), « MANUSCRIPT STUDIES A Journal of the Schoenberg Institute for Manuscript Studies », *Manuscript Studies*, 3 (2019), URL : [https://repository.upenn.edu/mss\\_sims/vol3/iss1/13](https://repository.upenn.edu/mss_sims/vol3/iss1/13) (visité le 06/05/2020).

CAVET (Laurent), *Les Heures de la Vierge : identification liturgique et origine du manuscrit*, 2015, URL : <https://irht.hypotheses.org/643> (visité le 26/04/2020).

COQUET (Michèle), « Alfred Gell, Art and Agency. An Anthropological Theory », *L'Homme*, 157 (2007), p. 261-263, URL : <http://journals.openedition.org/lhomme/5658> (visité le 20/06/2020).

DEROLEZ (Albert), *The Palaeography of gothic manuscript books, from the Twelfth to the Early Sixteenth Century*, Cambridge, 2003.

DRIGSDAHL (Erik), *Tutorial - Hours of the Virgin Hore Beate Marie Virginis - Index to a Selection of Uses*, en, 2002, URL : <http://manuscripts.org.uk/chd.dk/use/index.html> (visité le 14/05/2020).

— *Introduction and Tutorial Books of Hours*, en, 2005, URL : <http://manuscripts.org.uk/chd.dk/tutor/index.html> (visité le 14/05/2020).

HEIKKILA (Tuomas) et ROOS (Teemu), « Quantitative methods for the analysis of medieval calendars », *Digital Scholarship in the Humanities*, 33–4 (2018), p. 766-787, DOI : 10.1093/lhc/fqy007.

HERMAN (Nicholas), *Le livre enluminé, entre représentation et illusion*, Paris, 2018.

- 
- LEBIGUE (Jean-Baptiste), *Initiation aux manuscrits liturgiques*, 2007, URL : <https://cel.archives-ouvertes.fr/cel-00194063> (visité le 27/01/2020).
- *Les usages liturgiques*, 2016, URL : <https://irht.hypotheses.org/2484> (visité le 30/03/2020).
- *Les règles de préséances entre les offices*, 2017, URL : <https://irht.hypotheses.org/2473> (visité le 30/03/2020).
- LEGENDRE (Olivier), SAUTEL (Jacques-Hubert), HEID (Caroline), BOURLET (Caroline) et FLAMAND (Jean-Marie), *Livret du stage d'initiation au manuscrit médiéval (domaine latin et roman)*, 2006, URL : <https://cel.archives-ouvertes.fr/cel-00139917v4> (visité le 23/04/2020).
- LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927.
- ORNATO (Ezio) et ALII, *La face cachée du livre médiéval, L'histoire du livre vue par Ezio Ornato, ses amis et ses collègues*, Rome, 1997.
- PLUMMER (J.) et CLARK (G.), *Beyond Use*, URL : <http://www6.sewanee.edu/beyonduse/> (visité le 20/07/2020).
- PORCHER (Jean), « Le chanoine Victor Leroquais (1875-1946) », *Scriptorium*, 1–1 (1946), p. 170-172, URL : [https://www.persee.fr/doc/scrip\\_0036-9772\\_1946\\_num\\_1\\_1\\_2063](https://www.persee.fr/doc/scrip_0036-9772_1946_num_1_1_2063) (visité le 15/06/2020).
- RUDY (Kathryn M.), *Piety in Pieces, How Medieval Readers Customized their Manuscripts*, Cambridge, 2016.
- SCOTT-STOKES (Charity), *Women's Books of Hours in Medieval England*, Suffolk, 2006.
- STUTZMANN (Dominique), « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? », dans *Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2*, 2011, p. 247-277, URL : <https://halshs.archives-ouvertes.fr/halshs-00596970> (visité le 26/01/2020).
- « Ontologie des formes et encodage des textes manuscrits médiévaux. Le projet ORIFLAMMS », *Document Numérique*, 16–3 (2013), p. 81-95, DOI : 10.3166/DN.16.3.81-95.
- *Les écritures gothiques livresques : classification de Lieftinck-Gumbert-Derolez*, fr, URL : <https://oriflamms.hypotheses.org/quest-ce-que-la-paleographie/les-ecritures-gothiques-livresques-classification-de-lieftinck-gumbert-derolez> (visité le 29/02/2020).

## Réflexions autour des Humanités numériques

BARTSCHERER (Thomas) et COOVER (Roderick), *Switching Codes, thinking through digital technology in the humanities and the arts*, 2011.

BERRA (Aurélien), « Faire des humanités numériques », dans *Read/Write Book 2 : Une introduction aux humanités numériques*, Marseille, 2012, p. 25-43, DOI : <https://doi.org/10.4000/books.oep.238>.

BOILLET (Mélodie), BONHOMME (Marie-Laurence), STUTZMANN (Dominique) et KERMORVANT (Christopher), « HORAE : an annotated dataset of books of hours », dans *The 5th International Workshop on Historical Document Imaging and Processing*, Sydney, 2019 (2019 International Conference on Document Analysis and Recognition (ICDAR)), p. 7-12, DOI : [10.1145/3352631.3352633](https://doi.org/10.1145/3352631.3352633).

DACOS (Marin) et MOUNIER (Pierre), *Humanités numériques État des lieux et positionnement de la recherche française dans le contexte international*, Paris, 2014.

DUMOUCHEL (Suzanne), *Les Humanités Numériques : une nouvelle discipline universitaire ?*, 2020, URL : <https://dhiha.hypotheses.org/1539> (visité le 23/04/2020).

GUILLOT (Céline), HEIDEN (Serge), LAVRENTIEV (Alexei) et MARCHELLO-NIZIA (Christiane), « Constitution et exploitation des corpus d'ancien et de moyen français », *Corpus*, 7 (2009), URL : <http://journals.openedition.org/corpus/1495> (visité le 14/04/2020).

SAOU-DUFRENE (Bernadette), *Heritage and Digital Humanities : How Should Training Practices Evolve ?*, 2014.

SCHREIBMAN (Susan), SIEMENS (Ray) et UNSWORTH (John), *A Companion to Digital Humanities*, 2004, URL : <http://www.digitalhumanities.org/companion/> (visité le 17/07/2020).

SINATRA (Michaël E.) et VITALI-ROSATI (Marcello), dans *Pratiques de l'édition numérique*, Montréal, 2014, chap. 3. Histoire des humanités numériques, p. 49-60, DOI : <https://doi.org/10.4000/books.pum.317>.

WELGER-BARBOZA (Corinne), « Les digital humanities aujourd’hui : centres, réseaux, pratiques et enjeux », dans *Read/Write Book 2 : Une introduction aux humanités numériques*, Marseille, 2012, DOI : <https://doi.org/10.4000/books.oep.244>.

---

## Normes, méthodes et pratiques numériques

URL : <https://www.irif.fr/~carton/Enseignement/XML/Cours/Schematron/index.html> (visité le 15/07/2020).

ANDRÉ (Jacques), « Numérisation et codage des caractères des livres anciens », *Document numérique*, 7-3 (2003), p. 127-142, DOI : 10.3166/dn.7.3-4.127-142.

*dhSegment : A generic deep-learning approach for document segmentation*, 2019, DOI : 10.1109/ICFHR-2018.2018.00011.

BENSAMOUN (Alexandra) et FARCHY (Joëlle), *MISSION INTELLIGENCE ARTIFICIELLE ET CULTURE, Rapport final*, rapp. tech., Conseil supérieur de la propriété littéraire et artistique (CSPLA), 2020.

BLUCHE (Théodore), *Deep Neural Networks for Large Vocabulary Handwritten Text Recognition*, thèse de doct., Université Paris Sud - Paris XI, 2015, URL : <https://tel.archives-ouvertes.fr/tel-01249405> (visité le 11/07/2020).

BOROS (Emanuela), TOUMI (Alexis), ROUCHET (Erwan), ABADIE (Bastien), STUTZMANN (Dominique) et KERMORVANT (Christopher), « Automatic page classification in a large collection of manuscripts based on the International Image Interoperability Framework », dans *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, 2019 (2019 International Conference on Document Analysis and Recognition (ICDAR)), p. 756-762, DOI : 10.1109/ICDAR.2019.00126.

BRUHA (Ivan) et FAMILI (A. Fazel), *Postprocessing in Machine Learning and Data Mining*, New York, déc. 2000, DOI : 10.1145/380995.381059.

CAMPS (Jean-Baptiste), Lecture, 2017, URL : <https://halshs.archives-ouvertes.fr/cel-01706530> (visité le 20/04/2020).

DAILLE (Béatrice), HAZEM (Amir), KERMORVANT (Christopher), MAARAND (Martin), BONHOMME (Marie-Laurence), STUTZMANN (Dominique), CURRIE (Jacob) et JACQUIN (Christine), « Transcription automatique et segmentation thématique de livres d'heures manuscrits », *TAL*, 60-3 (2019), p. 13-36.

FIZE (Chloë), *Archives et numériques, un traitement intelligent des données historiques : application à la recherche et au patrimoine privé*, mém. de mast., École nationale des Chartes, 2020.

GABAY (Simon), KHEMAKHEM (Mohamed) et ROMARY (Laurent), *Les catalogues et GROBID. Doctorat. Du catalogue aux humanités numériques : quelles méthodes pour quels résultats ?*, 2018, URL : <https://hal.archives-ouvertes.fr/cel-01951107> (visité le 19/06/2020).

GATOS (B.), STAMATOPOULOS (N.) et LOULOUDIS (G.), « Handwriting Segmentation Contest », dans *10th International Conference on Document Analysis and Recognition*, 2009, p. 1393-1397, DOI : 10.1109/ICDAR.2009.245.

HAROLD (Elliote Rusty), MEANS (W. Scott), ENSARGUET (Philippe) et AL., *XML en concentré*, Paris, 2005.

*HEURIST : UNE BASE DE DONNÉES GÉNÉRIQUE POUR LES SCIENCES HUMAINES ET SOCIALES*, URL : <https://paris-timemachine.huma-num.fr/heurist-une-base-de-donnees-generique-pour-les-sciences-humaines-et-sociales/> (visité le 10/05/2020).

ISO et IEC, *Information technology — Document Schema Definition Languages (DSDL) — Part 3 : Rule-based validation — Schematron*, 2016, URL : [https://standards.iso.org/ittf/PubliclyAvailableStandards/c055982\\_ISO\\_IEC\\_19757-3\\_2016.zip](https://standards.iso.org/ittf/PubliclyAvailableStandards/c055982_ISO_IEC_19757-3_2016.zip) (visité le 17/05/2020).

LEBARBEY (Laura), *Reconstituer et étudier une collection ancienne au-delà de la pluralité des bases et des formats*, mém. de mast., École nationale des Chartes, 2015.

LOPEZ (Patrice) et ROMARY (Laurent), *Automatic Key Term Extraction from Scientific Articles in GROBID*, SemEval 2010 Workshop, Uppsala, Sweden, 2010, eprint : [inria00493437](https://hal.inria.fr/inria-00493437), URL : <https://hal.inria.fr/inria-00493437/document/> (visité le 22/06/2020).

MAZZIOTTA (Nicolas), « Traiter les abréviations du français médiéval. Théorie de l’écriture et pratiques d’encodage », *Corpus*, 7 (2009), URL : <http://journals.openedition.org/corpus/1517> (visité le 28/01/2020).

MEIMOUN (Thomas), URL : <https://www.quantmetry.com/intelligence-artificielle-data-science-automatisation-machine-learning-suffisait/> (visité le 14/07/2020).

MIOULET (Luc), *Reconnaissance de l’écriture manuscrite avec des réseaux récurrents*, thèse de doct., Université de Rouen, 2015, URL : <https://hal.archives-ouvertes.fr/tel-01301728> (visité le 28/07/2020).

OYALLON (Edouard), *Analyzing and Introducing Structures in Deep Convolutional Neural Networks*, Theses, Paris Sciences et Lettres, 2017, URL : <https://hal.archives-ouvertes.fr/tel-02353134> (visité le 20/08/2020).

ROBERTSSON (Eddie), *An Introduction to Schematron*, 2003, URL : <https://www.xml.com/pub/a/2003/11/12/schematron.html> (visité le 12/06/2020).

ROMARY (Laurent) et LOPEZ (Patrice), « GROBID - Information Extraction from Scientific Publications. » *ERCIM News*, 100 (2020), URL : <https://hal.inria.fr/hal-01673305> (visité le 23/06/2020).

RONDEAU DU NOYER (Lucie), *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la Revue des autographes de la librairie Charavay*, mém. de mast., École nationale des Chartes, 2019.



# Introduction

« Il y a là, semble-t-il, une injustice à réparer, une lacune à combler », écrit Victor Leroquais à propos de la rareté de travaux érudits sur le sujet des livres d'heures<sup>1</sup>. Objets de dévotion privée, ils permettent pourtant de saisir une part des sentiments, aspirations, craintes et espérances de ceux qui les possèdent<sup>2</sup>. La variété de leur composition les a peut-être rendus plus difficiles à saisir que les autres livres liturgiques, comme le missel pour la messe ou le bréviaire pour l'office. L'émergence de la notion d'« Humanités numériques » et sa diffusion depuis le début du XXI<sup>e</sup> siècle ouvrent de nouvelles perspectives dans l'étude de ce type d'objet. C'est dans ce contexte que s'inscrit le projet HORAE, pour *HOurs - Recognition, Analysis, Editions*, acronyme qui renvoie à la désignation des livres d'heures par leurs contemporains<sup>3</sup>.

Le projet est porté par trois partenaires qui rassemblent leur compétences autour de l'étude des livres d'heures numérisés afin de répondre à des enjeux technologiques et anthropologiques. Les trois membres du projet sont l'Institut de recherche et d'histoire des textes (IRHT), la société Teklia et le Laboratoire des Sciences du Numérique de Nantes (LS2N). L'IRHT, grâce à qui le stage dont découle le présent mémoire a été possible malgré les conditions sanitaires exceptionnelles, est une unité de service et de recherche (UPR 841) propre du CNRS et fondée en 1937. Rattachée à l'Institut des Sciences humaines et sociales, elle se consacre à la recherche fondamentale sur les manuscrits médiévaux et les imprimés anciens. Centre de compétences de renommée internationale, l'IRHT réunit des spécialistes de disciplines telles que la philologie, la lexicographie, la paléographie ou la codicologie, afin de dater et localiser les manuscrits, identifier et établir leurs textes, l'histoire de leur production, de leur circulation, de leur audience et de leurs usages de l'Antiquité au début de la Renaissance.

L'IRHT<sup>4</sup> est par ailleurs un pionnier dans l'utilisation du numérique au service de la recherche. Plus précisément, la section de paléographie latine<sup>5</sup> impulse depuis 2012 des

1. Victor LEROQUAIS, *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. II.

2. *Ibid.*

3. Cf. <https://www.irht.cnrs.fr/?q=fr/recherche/les-programmes-de-recherche/horae>

4. <https://www.irht.cnrs.fr/>

5. L'IRHT est en effet composé de 13 sections de recherche : Arabe ; Codicologie, Histoire des bibliothèques et heraldique ; Diplomatique ; Grecque et Orient chrétien ; Hébraïque ; Humanisme ; Latine ;

---

projets de pointe en matière d'analyse des écritures, de lecture et d'analyse des sources par ordinateur, sous l'appellation de « Paléographie numérique ». Le projet ORIFLAMMS<sup>6</sup>, qui s'est étendu de 2013 à 2016, visait ainsi à étudier les écritures du Moyen Âge central et tardif (XII<sup>e</sup> - XV<sup>e</sup> siècles), et le multilinguisme médiéval. L'enjeu était d'analyser l'évolution des systèmes et formes graphiques des écritures d'un temps long (le Moyen Âge) selon leur contexte de production (écritures usuelles, diplomatiques ou livresques) et leur langue (latin ou vernaculaire)<sup>7</sup>. Dans la même lignée de la reconnaissance d'écriture manuscrite, le projet HIMANIS(2015-2017)<sup>8</sup> a permis la recherche plein texte dans les registres de la chancellerie royale des XIV<sup>e</sup> et XV<sup>e</sup> siècles conservés aux Archives nationales. Ces différents projets ont montré les possibilités offertes par les nouvelles technologies dans la reconnaissance et l'analyse des écritures anciennes, et fournissent donc des acquis pour le projet HORAE.

Teklia<sup>9</sup>, entreprise active depuis 2014, est spécialisée dans la reconnaissance automatique des documents manuscrits et imprimés. Elle développe des solutions inspirées des techniques du *machine learning* et du *deep learning*<sup>10</sup>. Elle propose des outils d'analyse de documents via la classification des pages, la détection des lignes et la reconnaissance d'écritures manuscrites, comme nous l'avons testé avec la plateforme Arkindex<sup>11</sup>.

Le LS2N<sup>12</sup>, unité mixte de recherche (UMR 6004) créée en 2017, est le résultat de la fusion de deux autres UMR : l'institut de Recherche en Communications et Cybertétique de Nantes et le Laboratoire d'Informatique de Nantes Atlantique. Il a pour tutelles et partenaires l'université de Nantes, le CNRS, l'École Centrale de Nantes, l'IMT Atlantique et l'Inria. Son objectif est de faire progresser significativement la visibilité de la recherche en Cybertétique et Informatique à Nantes. Il se structure autour de 5 pôles de recherche : Conception et conduite de systèmes Robotique, procédés, calcul ; Science des données et de la décision ; Signaux, images, ergonomie et langues ; Science du logiciel et des systèmes distribués. Dans le cadre du projet HORAE, le LS2N apporte son expertise en traitement automatique du langage naturel, et permet donc de livrer des analyses lexicales au sein des livres d'heures.

Que sait-on des livre d'heures, dont Victor Leroquais déplore le manque d'études approfondies ? Selon ce dernier, qui dressa un nombre de catalogues conséquents sur les

---

Lexicographie latine ; Manuscrits enluminés, Paléographie latine ; Papyrologie, Romane et Sciences du Quadrivium.

6. *Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts*.

7. Dominique 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.16.3.81-95, p. 81.

8. *Historical MANuscript Indexing for user-controlled Search*.

9. <https://teklia.com/>

10. Cette question est davantage abordée dans le chapitre 3

11. Cf. sous-partie 3.1.2

12. <https://www.ls2n.fr/>

livres liturgiques conservés dans les bibliothèques publiques françaises<sup>13</sup>, ils empruntent dans leur composition aux bréviaires<sup>14</sup>. Les points communs entre le bréviaire et le livre d'heures sont notamment la présence d'un calendrier, du petit office de la Vierge, des psaumes de la Pénitence, des litanies, des suffrages et de l'office des morts. Toutefois, la principale différence du livre d'heures avec les autres livres liturgiques est qu'il est indépendant du cycle liturgique, des fêtes de l'année chrétienne, des martyrs et des saints, des anniversaires des dédicaces des églises ou des transactions de reliques ; sa récitation ne dépend que de la dévotion privée. On y trouve ainsi souvent un mélange de latin et de langue vernaculaire. L'éditeur ou le copiste peut alors disposer les éléments comme il le souhaite, ajouter des prières ou même des textes profanes<sup>15</sup>. C'est tout ce qui fait la complexité et la richesse anthropologique des livres d'heures. Ne répondant pas à des normes précises, ils sont propices à la découverte de sensibilités particulières chez leurs contemporains.

Le prêtre érudit nous délivre un historique de l'apparition du livre d'heures. Son potentiel ancêtre pourrait être les *libelli precum*, livres de dévotion à l'usage des laïcs au Haut Moyen Âge<sup>16</sup>. Si les fidèles empruntent aux moines et aux prêtres leurs pratiques de piété, la récitation de l'office divin n'est pas compatible avec les tâches quotidiennes. Ils se concentrent donc sur les petits offices et les prières supplémentaires, textes suffisamment courts et fixes, à l'exception de l'office de la Vierge pour le temps de l'Avent et de Noël<sup>17</sup>. Il faut rappeler que jusqu'au XIII<sup>e</sup> siècle, le livre de prières des fidèles est le psautier<sup>18</sup>. Aux psaumes s'ajoutent au fil du temps des litanies, des prières, l'office des morts, le petit office de la Vierge, des suffrages. Les premiers livres d'heures apparaissent donc comme des extensions du psautier<sup>19</sup>. C'est à partir du XIV<sup>e</sup> siècle que les offices et prières se détachent du psautier pour former à proprement parler ce que l'on nomme livre d'heures<sup>20</sup>. Se développant aux XIV<sup>e</sup> et XV<sup>e</sup> siècles, les livres d'heures sont souvent de petites dimensions et richement décorés. Ce type de livres se distingue visuellement par ses marges larges et ses bords enluminés.

---

13. Cf. sous-partie 1.1.1 et <https://data.bnf.fr/fr/documents-by-rdt/12444975/te/page1>.

14. Le bréviaire est l'un des livres de l'office le plus complet, dont les pièces sont articulées autour du temporal, du sanctoral et du commun des saints. Le temporal se réfère au calendrier de l'année liturgique, du premier dimanche de l'Avent au premier dimanche après la Pentecôte. Le sanctoral désigne les formulaires propres à la célébration des saints les plus importants. Quant au commun des saints, il renvoie à des textes génériques pour célébrer des saints sans formulaires propres. Ils y sont classés par catégories : les apôtres, les évangélistes, le commun des martyrs, des confesseurs, puis des saintes vierges.

15. V. LEROQUAIS, *Les livres d'heures manuscrits de la Bibliothèque nationale....*, p. VI.

16. *Ibid.*, p. IX.

17. *Ibid.*, p. X.

18. Un psautier est un livre de l'office recueillant les 150 psaumes répartis sur toute la semaine. On peut y trouver d'autres textes comme des cantiques ou des hymnes. Ils sont parfois suivis de litanies des saints.

19. *Ibid.*

20. *Ibid.*, p. XII.

---

Si Victor Leroquais est une référence par son approche minutieuse et ses observations régulières des livres liturgiques, ayant détecté plus de 200 usages différents pour l'office de la Vierge, le livre d'heures demeure un objet mystérieux qui fascine par la richesse de ses enluminures. L'histoire de l'art<sup>21</sup> et les études de genre autour de la dévotion féminine<sup>22</sup> ont fait du livre d'heures une source historique à part entière. L'historien de l'art danois Erik Drigsdahl est ainsi à l'origine d'une mise en ligne du contenu varié de livres d'heures répertoriés<sup>23</sup>. Il y distingue 11 grandes parties dans l'ordre suivant : le calendrier, les Évangiles, les prières *Obsecro te* et *O Intemerata*, les Heures de la Vierge, l'Office de la Croix, l'Office de l'Esprit Saint, les Sept Psaumes de la Pénitence, l'Office des morts, les suffrages, quelques prières<sup>24</sup> et des textes accessoires, du psautier de saint Jérôme aux Heures de la Conception en passant par les Heures de la Passion du Christ. Son site fournit les textes élémentaires et les variantes textuelles rencontrées, ainsi qu'une liste des usages liturgiques repérés<sup>25</sup>.

Une autre base de données en ligne met à disposition plus de 1000 manuscrits. *Beyond Use* se concentre davantage sur des manuscrits français ou néerlandais. On peut consulter sur le site une bibliographie, dont beaucoup de catalogues d'exposition<sup>26</sup>. La base permet ainsi de dater et de localiser des livres d'heures de la fin du Moyen Âge à partir des variantes textuelles. Elle a été initiée dans les années 1970 par John Plummer, professeur à Princeton University et conservateur à la Pierpont Morgan Library, puis elle est enrichie à partir de 1981, par Gregory Clark, professeur d'histoire de l'art. C'est dans ce contexte que s'inscrit le projet HORAE, *HOurs - Recognition, Analysis, Editions*, où le numérique sert une approche renouvelée de l'étude du livre d'heures, comme l'indiquent les trois volets de son titre.

Le terme *Recognition* fait ainsi référence aux technologies développées dans le cadre de la reconnaissance automatique d'écritures manuscrites afin d'identifier les textes des livres d'heures en partant du plus grand nombre possible de sources. En effet, avec plus de 10 000 témoins conservés, les livres d'heures offrent une masse de données importantes et nécessaires au développement efficace de méthodes propres au *machine learning* et au *deep learning*<sup>27</sup>. Pour des raisons d'interopérabilité et d'accessibilité, le projet se concentre

---

21. Nicholas HERMAN, *Le livre enluminé, entre représentation et illusion*, Paris, 2018.

22. Charity SCOTT-STOKES, *Women's Books of Hours in Medieval England*, Suffolk, 2006.

23. Erik DRIGSDAHL, *Introduction and Tutorial Books of Hours*, en, 2005, URL : <http://manuscripts.org.uk/chd.dk/tutor/index.html> (visité le 14/05/2020).

24. Comme les Quinze joies de notre Dame ou les Sept requêtes à notre Seigneur dans les manuscrits français.

25. Id., *Tutorial - Hours of the Virgin Hore Beate Marie Virginis - Index to a Selection of Uses*, en, 2002, URL : <http://manuscripts.org.uk/chd.dk/use/index.html> (visité le 14/05/2020).

26. <http://www6.sewanee.edu/beyonduse/index.php>.

27. Cette partie du projet est davantage développée dans le chapitre 3.

autour des images de livres d’heures numérisées disponibles avec l’API<sup>28</sup> IIIF<sup>29</sup>. Dans le cadre du stage, ce pan du projet a été concrétisé par le travail d’annotation et d’alignement de textes afin d’établir une vérité terrain qui sert de modèle à l’apprentissage machine.

Le volet *Analysis* est inhérent au projet, à la fois pour les aspects technologiques et historiques. Il s’agit d’analyser les résultats des technologies utilisés, les adapter au besoin, mais aussi d’analyser la structuration des livres d’heures et leur contenu. Lors du stage, l’analyse s’est davantage concentrée sur les données et métadonnées utilisées. Les données manipulées se sont concrétisées autour des livre d’heures numérisés à annoter, mais aussi autour de celles à structurer et à importer dans la base de données relationnelle Heurist, soit les usages liturgiques, les contenus textuels ou les manuscrits témoins<sup>30</sup>. Quant aux métadonnées, qui désignent les données sur les données, elles se sont cristallisées lors du stage autour de la structuration et de l’encodage semi-automatisé d’un catalogue de notices de livres d’heures conservés à la Bibliothèque nationale de France, afin de produire une nouvelle vérité terrain et d’ajouter des témoins dans la base de données, bien qu’ils ne soient pas tous numérisés<sup>31</sup>. L’analyse des données découle aussi de leurs possibles visualisations, via la production de cartes et de graphes, que ce soit sur le contenu des livres d’heures et les liens entre les différentes pièces liturgiques, que sur la circulation des témoins et des usages liturgiques.

L’*Edition* recouvre la rétroconversion en XML-TEI de catalogues de notices, selon un format interopérable et normalisé, comme celui de Victor Leroquais<sup>32</sup> lors du stage, mais aussi une anthologie des textes inédits édités durant le projet, afin d’apporter une pierre à l’édifice de l’histoire des lectures dévotes du XIII<sup>e</sup> au XVI<sup>e</sup> siècle.

Les trois grands volets du projet mettent en avant les problématiques de la qualité des données pour produire des connaissances historiques véridiques ; de leur normalisation selon des normes et standards définis dans le domaine des humanités numériques afin de garantir leur intéropérabilité pour favoriser les échanges et les réutilisations ; de l’ergonomie de leur présentation pour l’utilisateur final. Le fil rouge du stage et de ce présent mémoire est ainsi la transformation des données et des métadonnées, de leur structure à leur présentation ergonomique, avec toutes les questions de conversion, de fluidité et d’équilibre que cela engendre. L’objectif principal est ainsi de produire des données de recherche de qualité. Comment y parvenir ?

---

28. *Application Programming Interface*. Le terme désigne une interface par laquelle un logiciel offre des services à un autre logiciel. Un exemple célèbre est celui de l’API Google Maps utilisée par les sites d’institutions, entre autres, pour pouvoir les localiser plus facilement.

29. *International Image Interoperability Framework*. Il s’agit d’afficher les objets numériques de manière standardisée sur le Web afin de les rendre consultables, manipulables et annotables par n’importe quelle application compatible. Cf. <https://doc.biblissima.fr/introduction-iiif>.

30. Les questions relatives à la modélisation et à l’import des données sont davantage exposées au chapitre 2.

31. Cf. chapitre 1.

32. V. LEROQUAIS, *Les livres d’heures manuscrits de la Bibliothèque nationale....*

Le premier chapitre se concentre alors autour de la question de la structuration des métadonnées, du format à utiliser et des possibles automatisations de leur encodage. Ensuite, dans un deuxième chapitre est exposé toute une réflexion à propos de la modélisation des données, de leurs possibles imports dans une base de données relationnelle, et des visualisations qui en résultent pour générer des analyses historiques fructueuses. Enfin, le dernier chapitre aborde davantage les techniques et enjeux de l'apprentissage machine, pan du projet mené en symbiose avec les autres membres de l'équipe, ce qui nous a permis d'observer la gestion et le management d'un projet en humanités numériques.

# Chapitre 1

## Structuration semi-automatique des métadonnées : allier quantité et qualité

Si le projet HORAE a pour objectif de s'appuyer sur un maximum de livres d'heures numérisés selon le protocole IIIF, les catalogues de manuscrits permettent de saisir la distorsion approximative du corpus par rapport aux sources connues et disponibles. Il est en effet important en histoire, et particulièrement en histoire médiévale, d'avoir conscience des sources qui ont été transmises et de celles qui ont été perdues. Parmi les catalogues de notices de livres d'heures qui ont été soigneusement dressés, celui de Victor Leroquais<sup>33</sup> concerne les manuscrits conservés à la Bibliothèque nationale de France<sup>34</sup>.

Ces notices ont été océrisées dans le cadre du projet afin de pouvoir récupérer plus facilement les informations et les structurer selon un protocole défini. L'enjeu est en effet d'encoder ces métadonnées selon des standards internationaux afin de favoriser l'échange et l'interopérabilité des données, tout en respectant leur richesse descriptive et leurs particularités.

### 1.1 Analyse du document source : des données semi-structurées

Avant d'opérer un quelconque encodage, il est important d'observer soigneusement le document source océrisé à structurer. Cette étape est cruciale, car elle permet de définir quelle structure choisir pour l'encodage des métadonnées, mais aussi quelles difficultés

---

33. *Ibid.*

34. Les deux volumes sont disponibles en ligne. Un supplément est publié en 1943 pour les acquisitions récentes qui n'avaient pas été prises en compte dans les deux premiers volumes. Le lien vers la notice bibliographique suivante : <https://catalogue.bnf.fr/ark:/12148/cb30797672n> donne accès aux trois volumes numérisés sur Gallica, disponibles depuis 2016 car entrés dans le domaine public.

seront à prendre en compte lors de l'automatisation de la tâche.

### 1.1.1 Le travail minutieux d'un chanoine passionné

Victor Leroquais (1875-1946), ordonné prêtre en 1900, était si passionné par les textes liturgiques qu'il se fit mettre en congé pour les étudier. Il entraît pour la première fois à la Bibliothèque nationale en 1912. Cela signe le début de 30 années d'études prolifiques qui débouchèrent sur 20 volumes recouvrant non seulement des catalogues de livres d'heures, mais aussi de sacramentaires<sup>35</sup> et de bréviaires<sup>36</sup>.

Le chanoine autodidacte a pour habitude d'ajouter à ses notices descriptives une introduction résumant les résultats de ses enquêtes<sup>37</sup>. Son introduction au catalogue des livres d'heures conservés à la BnF montre leurs spécificités, leurs variétés et leur difficile définition qui fait justement leur intérêt en tant que source historique. Ils reflètent ainsi la diversité des usages liturgiques, des dévotions collectives et particulières<sup>38</sup>. Son étude permet donc de dresser un premier tableau sur la structure des livres d'heures.

L'abbé Leroquais distingue donc trois grands types de texte au sein de leur contenu<sup>39</sup> :

- les textes essentiels, c'est-à-dire ceux empruntés au bréviaire, tels que le calendrier, le petit office de la Vierge, les psaumes de la pénitence<sup>40</sup>, les litanies<sup>41</sup>, les suffrages et l'office des morts ;
- les textes secondaires, qui se retrouvent dans la plupart des livres d'heures, comme les fragments des Évangiles, la passion selon saint Jean, les prières à la Vierge *Obsecro te* et *O Intemerata*, les Heures et l'Office de la Croix, les Heures et l'Office du Saint-Esprit, les Quinze joies de la Vierge et les Sept Requêtes à Notre-Seigneur ;
- les textes accessoires, soit les 15 psaumes graduels, les Heures en l'honneur des différents saints, les oraisons diverses, les prières pour la journée chrétienne, les prières de la messe, le psautier de saint Jérôme, les 10 commandements et quelques autres pièces variables.

Cependant, le texte qui constitue l'élément principal du livre d'heures lui semble être les Heures de la Vierge, qui contiennent au moins six offices en son honneur : la Nativité le 8 septembre, la Présentation le 21 novembre, l'Annonciation le 25 mars, la Visitation

---

35. Si la définition du livre d'heures est complexe, le sacramentaire contient les textes et prières des sacrements et cérémonies où la présence d'un ministre ordonné prêtre ou évêque est requise.

36. Recueil à l'usage du clergé régulier et séculier contenant l'ensemble des textes nécessaires à l'office divin.

37. Jean PORCHER, « Le chanoine Victor Leroquais (1875-1946) », *Scriptorium*, 1-1 (1946), p. 170-172, URL : [https://www.persee.fr/doc/scrip\\_0036-9772\\_1946\\_num\\_1\\_1\\_2063](https://www.persee.fr/doc/scrip_0036-9772_1946_num_1_1_2063) (visité le 15/06/2020).

38. V. LEROQUAIS, *Les livres d'heures manuscrits de la Bibliothèque nationale....*, p. I-LXXXV.

39. *Ibid.*, p. XIV.

40. Aussi appelés « Psaumes pénitentiaux », ils sont au nombre de sept et expriment la douleur de l'homme face à ses péchés. Ils sont généralement placés après l'office de la Vierge. Cf. *Ibid.*, p. XX-XXI.

41. Prières composées d'une antienne, d'un verset et d'une oraison que l'on récite après les Vêpres ou les Laudes, en l'honneur de Dieu ou des saints. Cf. *Ibid.*, p. XXI.

## 1.1. ANALYSE DU DOCUMENT SOURCE : DES DONNÉES SEMI-STRUCTURÉES

---

le 2 juillet, la Purification le 2 février et l'Assomption le 15 août.

On trouve à leur suite un office abrégé propre aux livres d'heures, l'*Officium parvum beatae Mariae virginis*, aussi appelé, quand son titre apparaît, *Cursus, Hore beatissime virginis Marie, Hore beate Marie secundum usum romanum*<sup>42</sup>. Il se divise en 7 parties qui correspondent à la liturgie des heures, et comprend un nocturne<sup>43</sup> et trois leçons<sup>44</sup> :

- Matine et Laudes
- Prime
- Tierce
- Sexte
- None
- Vêpres
- Complies

Les psaumes et l'unique nocturne varient selon les jours de la semaine. Il est commun au dimanche, lundi et jeudi<sup>45</sup>; au mardi et vendredi<sup>46</sup>; et au mercredi et samedi<sup>47</sup>.

Entre le calendrier et l'office de la Vierge s'intercalent quasi systématiquement les quatre textes suivants :

- les fragments des évangiles de saint Jean, saint Luc, saint Matthieu et saint Marc
- la Passion selon saint Jean
- les prières *Obsecro te* et *O Intemerata*

Ces deux dernières prières s'adressent à la Vierge. Dupliquée en de multiples versions, l'oraison *Obsecro te* est en général récitée pour appeler la Vierge sur son lit de mort, d'où la miniature fréquente représentant la Vierge aux côtés d'un mourant<sup>48</sup>. Une des plus anciennes versions de *O Intemerata* est pour sa part destinée à la Vierge et à saint Jean<sup>49</sup>.

Parmi les pièces apparaissant rarement, on retrouve :

- les psaumes graduels récités pour les fidèles, les trépassés ou les défunt récemment décédés ;
- les offices en l'honneur des différents saints ;
- les heures abrégées pour chacun des jours de la semaine<sup>50</sup>.

Victor Leroquais souligne que, dans les inventaires, les livres d'heures peuvent

---

42. *Ibid.*, p. XVII.

43. Élément des matines contenant des psaumes et des leçons.

44. Destinées à l'office des matines, elles se réfèrent à des passages d'origine scripturaire ou patristique.

45. *Domine, Dominus noster... Caeli enarrant... Domini est terra*, cf. *Ibid.*

46. *Eructavit... Deus noster refugium... Fundamenta...*, cf. *Ibid.*

47. *Cantate Domino... Dominus regnavit... Cantate Domino*, cf. *Ibid.*

48. J. PORCHER, « Le chanoine Victor Leroquais (1875-1946) »..., p. XXXIII-XXIV.

49. V. LEROQUAIS, *Les livres d'heures manuscrits de la Bibliothèque nationale*..., p. XXV.

50. Leur répartition la plus commune est la suivante : la Trinité le dimanche, les défunt le lundi, l'ensemble des saints et le Saint-Esprit le mardi ou le mercredi, le Saint Sacrement le jeudi, la Passion le vendredi et la Vierge le samedi. Cf. *Ibid.*, p. XXVIII.

être qualifiés de différentes manières, notamment par des qualificatifs comme « grandes heures » ou « petites heures ». D'après les descriptions et les manuscrits conservés, il apparaît que ces adjectifs s'appliquent à la taille du manuscrit. Il existe donc une différence de langage entre les liturgistes et ceux qui dressent les inventaires. Les uns se réfèrent au contenu abrégé, les autres au format du contenant<sup>51</sup>.

Dans son introduction, le chanoine détaille les illustrations les plus courantes selon les parties des livres d'heures étudiées. Si les calendriers contiennent pour la plupart des représentations des travaux des champs, des signes du zodiaque ou des saints fêtés, les sujets d'enluminures les plus nombreux sont ceux relatifs à la vie de la Vierge et du Christ<sup>52</sup>. Si les choix des sujets dépeints n'ont pas toujours de liens directs avec les parties de l'office, Leroquais rappelle que le livre d'heures s'inscrit à l'origine dans la continuité du psautier, recueil illustrant les principales scènes de la vie du Christ<sup>53</sup>. Les suffrages constituaient la partie la plus riche et la plus variée d'un point de vue iconographique, avec des galeries de tableaux des martyrs et des saints<sup>54</sup>.

Les 313 notices qui suivent l'introduction particulièrement éclairante de Leroquais témoignent ainsi de la richesse et de la variété des livres d'heures, dont la majorité des témoins appartient au XV<sup>e</sup> siècle, et une petite partie aux XII<sup>e</sup>, XIII<sup>e</sup> et XIV<sup>e</sup> siècles<sup>55</sup>.

Après avoir résumé les analyses textuelles découlant des notices, il est temps d'analyser leur structure formelle pour réfléchir à un encodage des métadonnées.

### 1.1.2 Des notices papiers aux notices océrisées

Il s'agit ici de comprendre la structure sous-jacente des notices de Victor Leroquais qui répondent à une logique bien définie. Pour mieux évaluer la cohérence de cette structure, trois notices parmi les 313 ont été plus minutieusement observées : la première (notice 1), la dernière (notice 313) et une vers le milieu du recueil (notice 112)<sup>56</sup>. Cette première phase d'observation est importante pour, d'une part, comprendre les métadonnées et choisir l'encodage sémantique le plus approprié, d'autre part, repérer quels éléments sont structurants pour favoriser le processus d'automatisation.

Les notices débutent par un numéro d'ordre composé d'un chiffre puis d'un point<sup>57</sup>,

---

51. *Ibid.*, p. XX.

52. Pour une description plus détaillée des sujets de décoration selon les parties du texte concernées, cf.*Ibid.*, p. XLIII-LXXXIV.

53. *Ibid.*, p. XLIV.

54. *Ibid.*, p. XLVIII.

55. *Ibid.*, p. LXXXIV.

56. Une reproduction des trois notices papiers ici plus amplement étudiées est disponible en annexes, section A.1.1.

57. La seule exception est la notice « 311-312. CREDO DU SIRE DE JOINVILLE ET LIVRE D'HEURES. XIII<sup>e</sup> SIÈCLE FIN, ET XIV<sup>e</sup> SIÈCLE FIN », qui ne comprend pas un mais deux numéros séparés par un tiret. Étant donné qu'il s'agit de la seule exception, il est plus prudent de l'ajouter manuellement dans le document transformé, afin de ne pas créer de potentielles erreurs dans la récupération

## **1.1. ANALYSE DU DOCUMENT SOURCE : DES DONNÉES SEMI-STRUCTURÉES**

---

d'un résumé du contenu en lettres capitales suivi d'un point et d'une date. Sur une deuxième ligne, dans une taille de police plus petite, est indiqué le lieu de conservation, dans le cas présent la Bibliothèque nationale, puis la cote. Les paragraphes qui suivent détaillent le contenu du manuscrit. Le premier paragraphe peut tantôt résumer le contenu du manuscrit s'il présente une particularité (cf. notice 1), tantôt détailler une partie précise du manuscrit en partant du début de l'ouvrage (cf. notices 112 et 313).

Les paragraphes qui suivent listent les parties structurantes du manuscrit en précisant la foliation, les titres, quelques extraits de textes<sup>58</sup> et les rubriques. Ces paragraphes ont pour particularité d'être divisés par des tirets séparant soit des parties différentes, soit des citations. Les tirets sont ainsi souvent placés avant un numéro de page ou de folio, comme pour indiquer un changement de page, mais cela n'est pas systématique. Ils indiquent en fait un déplacement dans le texte, également à l'intérieur d'une même page. Par ailleurs les folios ne sont pas toujours indiqués par des numéros, on trouve parfois des lettres capitales (cf. notice 1 p. 1).

Ces premiers paragraphes sont parfois suivis d'un ou de plusieurs paragraphes relatant l'historique du manuscrit, l'usage liturgique dans lequel il s'inscrit, des hypothèses sur son contexte de création, ses probables possesseurs, sa datation. Cette première partie de la notice se conclut par un saut de ligne, et la taille de police plus petite des paragraphes suivants indique le début d'une nouvelle partie qui concerne davantage la description matérielle des manuscrits.

Cette description matérielle débute toujours par des indications concernant le support et les dimensions de l'objet livre. Cette partie s'ouvre systématiquement sur le matériau employé : « Parch. », « Papier » ou « Vélin ». S'ensuit une description de la décoration, qui peut s'allonger sur plusieurs paragraphes (cf. notice 1), ou bien être circonscrite entre deux tirets, notamment lorsque la décoration est absente (cf. notice 313). Cette partie se conclut par une mention du type de reliure introduite par « Rel. » ou « Demi-reliure ». La description de la reliure peut être suivie, dans un même paragraphe, de quelques références bibliographiques, mais cela est loin d'être systématique.

Cependant, quelques irrégularités sont à prendre en compte. Certains numéros de notices ne sont pas suivis de point. C'est notamment le cas des notices 29 et 46. La notice 265 ne contient pas de mention de date à la fin de son titre. Si ces informations sont trop exceptionnelles pour être prises en compte dans un code de transformation, il est utile de les connaître pour rétablir le bon encodage de ces informations dans le document transformé. C'est notamment le cas pour les numéros de folios, de pages ou de notices, où les erreurs les plus récurrentes sont la lettre « O » à la place du chiffre zéro, ou les lettres « l », « I » ou « i » à la place du chiffre un. S'il est trop dangereux de corriger ces erreurs

---

des titres par ailleurs.

58. Lorsque Victor Leroquais donne des citations, il s'agit souvent de textes qui n'étaient pas encore édités, soit des variantes qui peuvent donner des renseignements sur le type d'usage liturgique.

directement sur le document source, il est possible de le faire dans un deuxième temps dans le document encodé, car les informations sont ainsi plus faciles à localiser. Il faut toutefois prendre en compte ces exceptions récurrentes dans le code de transformation, afin de ne pas perdre la récupération de certaines informations.

Dans le cadre du projet, les notices ont été océrisées. L'océrisation, terme dérivé du sigle anglophone OCR<sup>59</sup>, consiste à traiter une image via un logiciel de reconnaissance de caractères afin d'obtenir un fichier texte. Il est ainsi possible de faire de la recherche plein-texte et de récupérer plus facilement les informations. Les notices ont alors été récupérées avec le logiciel de traitement de texte Microsoft Word.

Le processus d'océrisation peut toutefois engendrer un certain nombre d'erreurs, liées à la qualité du document initial, aux polices employées, aux notes ou à la forme du texte, notamment. Il est donc essentiel d'avoir conscience de ce taux d'erreurs avant tout encodage automatisé, afin de les corriger si elles sont répétitives, puis être attentif aux exceptions qui pourraient échapper à tout processus d'automatisation. Parmi les erreurs les plus récurrentes et les plus faciles à corriger, on retrouve :

- « I)'une autre main » au lieu de « D'une autre main »
- « Arnen » ou « Anien » au lieu de « Amen »
- « Pareil. » ou « Parch. » au lieu de « Parch. »
- « Eel. » au lieu de « Rel. »
- « Inill. » ou « min. » au lieu de « mill. »

Les erreurs plus ponctuelles qui on été corrigées, notamment grâce à une lecture attentive des trois notices choisies pour l'observation, signalent qu'il est possible que quelques coquilles typographiques subsistent dans l'ensemble des notices<sup>60</sup>. Il est également important de supprimer les réclames qui se sont insérées dans le texte du document océrisé, afin d'éviter l'encodage d'informations parasites lors de la transformation.

Toutefois, la répétition de certaines variations, difficiles à harmoniser directement sur le document source, au risque de créer des erreurs ailleurs dans le document, sont à prendre en compte dans le code de transformation. Par exemple, si la plupart des tirets séparant des citations ou des parties structurantes sont longs, certains sont plus courts et correspondent aux tirets des mots composés<sup>61</sup>.

Dans les notices papiers de Leroquais, les paragraphes concernant le contenu précis

---

59. *Optical Character Recognition*.

60. On peut citer le remplacement de « ccriture » par « écriture » ou de « Jean-Gaiéas Visconti » par « Jean Galéas Visconti » à la page 4 de la première notice, les mentions « 318v<>, 321, 322v<> » où les chevrons correspondent au symbole du dégré, ou encore « la première moitié du xvi® siècle » où l'exposant a été mal transcrit.

61. Cf. notice 1 p. 1 : « 39. « Die Iovis. Ad matutinas de sacramento. » - 43 à 45. Heures de la Croix. » et p. 3 : « 435- « Ordo ad catecumenum faciendum. » ». Sur les tirets, une autre erreur plus ponctuelle repérée est la transformation d'un tiret long en deux petits tirets juxtaposés, comme on peut le voir avant les références bibliographiques de la notice 104.

## **1.2. DÉFINIR LE DOCUMENT CIBLE : TROUVER L'ÉQUILIBRE ENTRE CONTRAINTE ET ADAPTATION**

---

du manuscrit commencent par « Fol. », « Feuillet » ou un chiffre. Les paragraphes concernés ne débutant pas de cette manière dans le document océrisé sont à fusionner avec le paragraphe précédent. L'observation des quelques différences entre les notices papiers et les notices océrisées permettent d'établir ce qu'il faut penser à corriger dans la structuration des notices, quelque soit la technique d'automatisation choisie.

De cette première phase d'observation résulte deux arbres de décision. Un arbre de décision est un outil méthodologique qui permet de définir les décisions à prendre selon la situation qui se présente à soi. Il résume le résultat obtenu selon la méthode employée. Il permet de décrire sa réflexion sur trois conditions :

- la présentation du document source ;
- ce que l'on cherche à créer ;
- la méthode employée pour la transformation.

Le premier prend la forme de l'arborescence d'une structure TEI<sup>62</sup> en partant de la typographie présente dans le document océrisé. Toutefois, cet arbre a pour défaut de refléter davantage la structure du document cible que celle du document source. La deuxième version de l'arbre se veut donc plus proche de la structure des données à encoder, afin de mieux percevoir ce qui peut être automatisé et ce qui ne peut pas l'être<sup>63</sup>.

La mise en forme d'arbres de décision nécessite toutefois d'avoir défini ce qui est attendu en format cible. ce format doit rendre compte de la richesse des données afin de mieux les visualiser et de faciliter les comparaisons.

## **1.2 Définir le document cible : trouver l'équilibre entre contrainte et adaptation**

Définir la structuration électronique des notices de livres d'heures établies par Lerouais implique le choix d'un modèle et d'un format suffisamment contraignants pour garantir l'interopérabilité<sup>64</sup> et la pérennité des fichiers. Ce choix doit également permettre assez de souplesse pour s'adapter aux variantes possibles dans des notices papiers établies hors des standards informatiques actuels. Il est ainsi crucial de réfléchir dans un premier temps à un format qui réponde à ces exigences, puis de personnaliser un schéma d'encodage afin de favoriser l'homogénéité des fichiers.

---

62. Cf. annexes section A.1.3. Nous y reviendrons plus explicitement dans la section 1.2.

63. Cf. annexes section A.1.3.

64. L'interopérabilité est un principe d'adaptation des systèmes informatiques entre eux afin de favoriser l'échange des données, d'où la constitution de standards et de normes.

### 1.2.1 Normaliser la description des manuscrits

Il s'agit de trouver le format standardisé le plus structurant d'un point de vue sémantique et le plus interopérable d'un point de vue informatique. Le choix s'est donc naturellement orienté vers la TEI<sup>65</sup>, communauté scientifique définissant des recommandations pour l'encodage de document textuel en langage de balisage XML<sup>66</sup>. La *Text Encoding Initiative*, fondée en 1987<sup>67</sup>, vise à définir un langage commun pour partager et mutualiser les textes encodés numériquement. Ses recommandations sont regroupées dans des *Guidelines*, ensemble de balises rassemblées dans des modules thématiques, d'attributs et de valeurs d'attributs qui servent à décrire un document source, tant dans sa forme que dans son contenu<sup>68</sup>. L'usage de ces balises, appelées « éléments » dans la terminologie de la TEI, est réglementé dans les « Recommandations pour l'encodage et l'échange de textes électroniques » disponibles en ligne<sup>69</sup>. Chacun des éléments possède un sens précis et son contenu sémantique peut être encore affiné par l'ajout d'attributs dont l'usage est tout aussi réglementé<sup>70</sup>. Outre de représenter l'intérêt d'associer un texte à ses métadonnées, un texte codé en TEI peut être affiché avec une typographie relevant de l'édition professionnelle tout en permettant de faire des recherches ciblées via le balisage sémantique<sup>71</sup>.

Dans le cas des notices de livres d'heures, le chapitre le plus approprié est celui concernant la description des manuscrits<sup>72</sup>. Résultant de la cinquième proposition de la TEI en 2007 (TEI P5), le guide de l'encodage des sources manuscrites est le fruit d'un processus de réflexion né en 1999. Le projet européen MASTER<sup>73</sup>, rassemblant des chercheurs spécialisés, des bibliothécaires et des catalogueurs, s'est concentré sur l'apport de standards pour les informations relevant de la codicologie<sup>74</sup> et de la philologie<sup>75</sup>. Ce projet débouche en 2001 sur la DTD Master, une personnalisation de la TEI complétée par l'ajout d'un élément msDescription, ainsi que d'autres éléments spécialisés qui représentent un enrichissement des possibilités pour la description des manuscrits.

L'encodage des notices selon les standards des TEI Guidelines facilite la collecte

---

65. *Text Encoding Initiative*.

66. *Extensible Markup Language*.

67. Laura LEBARBEY, *Reconstituer et étudier une collection ancienne au-delà de la pluralité des bases et des formats*, mém. de mast., École nationale des Chartes, 2015, p. 31-34.

68. « TEI : History. », Consulté le 5 août 2020. <https://tei-c.org/about/history/>.

69. Intitulée P5, la version courante des recommandations de la TEI est disponible à l'adresse suivante : <https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/index.html>.

70. Lucie RONDEAU DU NOYER, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la Revue des autographes de la librairie Charavay*, mém. de mast., École nationale des Chartes, 2019, p. 25.

71. Jacques ANDRÉ, « Numérisation et codage des caractères des livres anciens », *Document numérique*, 7-3 (2003), p. 127-142, DOI : 10.3166/dn.7.3-4.127-142, p. 128.

72. « 10 Manuscript Description - The TEI Guidelines. » Consulté le 20 avril 2020, <https://tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>.

73. *Manuscript Access through Standards for Electronic Records*.

74. Étude de l'aspect matériel du manuscrit.

75. Étude du contenu et de l'expression linguistique d'un document textuel.

## 1.2. DÉFINIR LE DOCUMENT CIBLE : TROUVER L'ÉQUILIBRE ENTRE CONTRAINTE ET ADAPTATION

---

et la comparaison d'informations au sujet des livres d'heures qui servent de sources au projet HORAE. Les notices, vouées à être intégrées dans une base de données descriptive, doivent donc être finement encodées tout en respectant des standards reconnus dans la base de données. Quel modèle d'encodage a donc été retenu ?

### 1.2.2 Des données structurées pour mieux exploiter les informations

Le modèle d'encodage définitif se doit de respecter un schéma XML bien précis défini via un fichier ODD<sup>76</sup>. Ce type de fichier permet de personnaliser ses choix d'encodage en formalisant des contraintes, telles que l'enchaînement spécifique d'éléments imbriquées ou bien la détermination de valeurs d'attributs, par exemple. Toutefois, toute modification apportée à un schéma doit permettre la validité d'un document lorsqu'il est conforme aux spécifications originelles de la TEI<sup>77</sup>. S'il on veut restreindre des valeurs d'attributs, il est important de conserver celles autorisées par la TEI en plus de celles personnalisées. La restriction est pertinente uniquement si elle permet une homogénéisation. S'il peut être utile que le schéma ODD soit contraignant, il faut toujours se demander, avant de supprimer un élément, s'il n'a pas lieu d'être maintenu. Par exemple, la balise <note> qui peut toujours être nécessaire pour une explication apportée. Toutefois, son usage peut être restreint grâce à par exemple des attributs, une classification, une sous-structuration ou une interdiction des balises <p>. Il faut alors se poser la question de l'équivalence des éléments : à quoi servent-ils ? Dans quelle situation en fait-on usage ?

Le document ODD a été tout d'abord généré via le processus oddbyexample<sup>78</sup> à partir de l'encodage de notices de manuscrits dans le cadre du projet ECMEN<sup>79</sup>. Le corpus à partir duquel a été généré le fichier ODD a donc pour ambition d'être le plus représentatif possible de la production écrite médiévale en s'appuyant sur le catalogue des manuscrits datés du fonds français. À la différence des notices de livres d'heures de Leroquais pour le projet HORAE, le catalogue d'ECMEN ne contient pas uniquement des notices mais aussi des extraits du contenu des manuscrits qui sont mis en regard avec leurs pages numérisées sur Gallica. Cela implique une définition du format cible quelque peu différente, mais l'ODD est ici avant tout utile pour la définition du <msDesc> structurant la description

---

76. *One Document Does it all.*

77. Jean-Baptiste CAMPS, Lecture, 2017, URL : <https://halshs.archives-ouvertes.fr/cel-01706530> (visité le 20/04/2020).

78. Le principe est de générer automatiquement un fichier ODD à partir d'un corpus encodé en TEI. Le processeur analyse l'ensemble des éléments et attributs du fichier de départ et les compare avec ceux propres à la TEI P5. Il élimine alors les balises et leurs attributs non utilisés et conserve ceux présents dans le document.

79. Écritures médiévales et outils numériques. Projet mené par Katja Monier, sous la direction de Dominique Stutzmann, il vise à analyser les évolution de l'écriture du XII<sup>E</sup> AU XV<sup>E</sup> SIÈCLES POUR MIEUX ÉCLAIRER LES MODES DE PRODUCTION ET DE COMMUNICATION DU MOYEN ÂGE.

du manuscrit tant du point de vue du contenu que du contenant.

Dans la définition du schéma, il faut avoir à l'esprit que les données héritées ne sont jamais parfaites, et rarement complètes. Il s'agit donc de favoriser l'amélioration de la saisie des données tout en ne bloquant pas l'avancement. L'idée est ainsi d'intégrer des avertissements via le langage de validation schematron afin de favoriser l'homogénéisation de la structuration des données. Toutefois, dans le cadre d'une généralisation de ce modèle à d'autres encodages de catalogues de notices, lorsqu'il s'agit de caractéristiques spécifiques à un projet, il est important de se contenter d'avertissements et de ne pas créer des contraintes qui puissent provoquer des invalidités abusives, et donc de rendre le document TEI invalide.

Ces avertissements peuvent être significatifs dans le cas d'une absence d'éléments au sein d'une séquence, mais uniquement sur les éléments considérés comme essentiels pour l'analyse propice au projet. Ainsi, quand les données ne sont pas connues, on peut imaginer la suggestion d'un attribut *type* avec comme valeur *absent*. Leur présence est également intéressante pour encourager la normalisation de certains attributs, soit pour la cohérence de l'encodage, par exemple si l'attribut *ref* ne renvoie pas à un *xml :id* répertorié, soit pour systématiser des informations considérées comme incontournables dans le cadre du projet. On peut par exemple penser à l'attribut *xml :lang* dans l'élément *<textLang>*, avec comme valeur le code langue correspondant à la norme ISO 639<sup>80</sup>, soit la valeur « *fro* » pour indiquer le moyen français ou la valeur « *frm* » pour signaler l'ancien français. Toutefois, on peut rétorquer que la normalisation de cette information n'est pas pertinente dans le cadre du projet ECMEN qui répertorie des manuscrits du fonds français, donc la majorité des titres sont en français. La manipulation représente alors une perte de temps pour les catalogueurs. Ces conditions restent intéressantes davantage dans les notices qu'au sein même du catalogue, notamment dans le cadre du projet HORAE qui s'attache à montrer les différents usages au sein de l'Occident médiéval. On voit ainsi que la définition de contraintes d'encodage est indissociable du projet dans lequel elles s'inscrivent.

Dans le cadre du projet ECMEN, une règle schematron a ainsi été implémentée pour l'attribut *class* associé à la plupart des balises *<msItem>*. Les valeurs de l'attribut *class* renvoient à des catégories définies dans le *<teiHeader>* général du document. On peut donc utiliser une règle schematron et un chemin Xpath qui permettent, dans le cas d'un attribut *class* associé à un *<msItem>*, de sélectionner une valeur d'attribut obligatoirement définie dans l'attribut *xml :id* de l'élément *<category>* du *<teiHeader>* :

```
<elementSpec ident="msItem" mode="change">
  <constraintSpec ident="class" scheme="isoschematron">
```

---

80. ISO 639 est une norme internationale de l'Organisation internationale de normalisation qui définit des codes pour la représentation des noms de langues.

## 1.2. DÉFINIR LE DOCUMENT CIBLE : TROUVER L'ÉQUILIBRE ENTRE CONTRAINTE ET ADAPTATION

---

```
<!-- Restriction des valeurs de l'attribut "class"
aux "xml:id" déclarés dans les éléments <category>.
La valeur de l'attribut "class" renvoie en effet
aux catégories de livres définies dans l'élément
"category". Message d'alerte si l'attribut "class"
n'apparaît pas. -->
<constraint>
<!--<rule
xmlns="http://purl.oclc.org/dsdl/schematron" context="//tei:msItem"
role="warning">-->
<assert
xmlns="http://purl.oclc.org/dsdl/schematron"
test="substring-after(@class, '#') = ancestor::tei:
teiCorpus/tei:teiHeader/
tei:encodingDesc/tei:classDecl/tei:taxonomy
/tei:category/@xml:id">
Please select a valid class</assert>
<!--</rule>-->
</constraint>
</constraintSpec>
[...]
</elementSpec>
```

Si la règle schematron suivante ne permet pas l'auto-complétion des valeurs d'attributs restreintes dans l'attribut *class*<sup>81</sup>, elle enclenche l'affichage d'un message d'erreur en cas de référence erronée.

Dans l'objectif d'établir des règles d'encodage qui s'appliquent à plusieurs projets, et notamment à HORAE, on peut jouer sur des contraintes relatives au typage des données dans les valeurs d'attributs. Par exemple, on autorise uniquement des chiffres pour les attributs *quantity*<sup>82</sup> et *columns*<sup>83</sup> grâce à la contrainte suivante :

```
<datatype>
<dataRef key="teidata.numeric"/>
</datatype>
```

---

81. L'auto-complétion est implémentable via la modification des règles CSS, des frameworks et des actions au sein du logiciel Oxygen XML editor, afin de créer une interface ergonomique pour les chercheurs qui travaillent sur des catalogues et des données bibliographiques. Le résultat de ces modifications est disponible en mode auteur.

82. L'attribut *quantity* est utile pour indiquer le nombre de folios et les dimensions du manuscrit.

83. L'attribut *columns* sert à indiquer le nombre de colonnes présentes sur une page.

Parmi les autres contraintes utiles au projet HORAE se trouve la suppression de l'attribut *status*<sup>84</sup> dans l'élément <msDesc> ou la définition d'une séquence d'éléments au sein d'un élément parent, notamment au sein des éléments <msContents> et <msItem><sup>85</sup>. En effet, un problème d'utilisation sémantique avait été noté dans le catalogue relatif au projet ECMEN à propos des balises <msItem>. Certains contiennent parfois plusieurs œuvres :

```
<msContents>
    <summary>Léonard L'Arétin, La première guerre punique,
    traduction de Jean le Begue; Tite-Live,
    Histoire de Rome, traduction de Pierre
    Bersuire</summary>
    textLang>Français</textLang>
    <msItem class="#cHistoriographie">
        <author corresp="Jean Lebègue (1368-1457)"
            ref="http://catalogue.bnf.fr/ark:/12148
            /cb13541385j"/>
        <author>Léonard Arétin</author>
        <author corresp="Tite-Live (0059?
            av. J.-C.-0017)"
            ref="http://catalogue.bnf.fr/ark:/12148
            /cb11886799m"/>
        <author corresp="Bersuire, Pierre (129.?-1362)"
            ref="http://catalogue.bnf.fr/ark:/12148
            /cb13322092g"/>
        <title>La première guerre punique</title>
        <title>Troisième et quatrième
            décades de Tite Live</title>
    </msItem>
</msContents>
```

Il en a été conclu que pour l'encodage des notices de Leroquais, il est plus pertinent de mettre plusieurs <msItem> par <msContents> si cela est nécessaire, car l'élément est créé pour décrire un œuvre dans son individualité au sein du contenu intellectuel d'un manuscrit<sup>86</sup>. La personnalisation de l'ODD a alors été affinée. Il est donc possible de mettre autant de <msItem> que souhaités au sein de l'élément <msContents> :

84. Comme son nom l'indique, il sert à décrire le statut du document décrit, s'il est publié, obsolète, à l'état de brouillon, ... cf. <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-att.docStatus.html>.

85. Les contraintes définies dans le document ODD sont visibles en annexes dans la section A.1.4.

86. Cf. <https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-msItem.html>.

## 1.2. DÉFINIR LE DOCUMENT CIBLE : TROUVER L'ÉQUILIBRE ENTRE CONTRAINTE ET ADAPTATION

---

```
<elementSpec ident="msContents" mode="change">
<!-- Définition d'une séquence d'éléments à l'intérieur de l'élément
<msContents>, qui est parfois vide. Il est préférable d'avoir
un élément "msItem" par œuvre au sein du manuscrit, et de ne
pas les multiplier au sein de l'élément qui prend son sens
dans l'individualité de ce qu'il décrit. -->
<content>
    <sequence preserveOrder="true">
        <elementRef key="summary"
            minOccurs="0" maxOccurs="1" />
        <elementRef key="textLang"
            minOccurs="0" maxOccurs="1" />
        <elementRef key="msItem"
            minOccurs="0" maxOccurs="unbounded" />
    </sequence>
</content>
```

Une séquence d'éléments a ensuite été réglementée à l'intérieur du <msItem>, bien qu'il faille laisser la possibilité d'ajouter plusieurs <title> et plusieurs <author> dans le cas de traductions et de versions linguistiques différentes. On peut ainsi ajouter un attribut *xml :lang* à ces deux éléments pour plus de précision :

```
<elementSpec ident="msItem" mode="change">
    [...]
<!-- Définition d'une séquence d'éléments dans l'élément <msItem>.
Aucun des éléments ci-dessous n'apparaissent systématiquement.
Par ailleurs, on peut retrouver plusieurs fois la séquence
dans un même <msItem>. Le titre et l'auteur peuvent
apparaître plusieurs fois pour signaler d'éventuelles
traductions dans plusieurs langues. -->
<content>
    <sequence preserveOrder="true">
        <elementRef key="locus" minOccurs="1" maxOccurs="1"/>
        <elementRef key="author" minOccurs="0" maxOccurs="unbounded"/>
        <elementRef key="title" minOccurs="0" maxOccurs="unbounded"/>
        <elementRef key="rubric" minOccurs="0" maxOccurs="1"/>
        <elementRef key="incipit" minOccurs="0" maxOccurs="1"/>
        <elementRef key="quote" minOccurs="0" maxOccurs="unbounded"/>
        <elementRef key="explicit" minOccurs="0" maxOccurs="1"/>
        <elementRef key="colophon" minOccurs="0" maxOccurs="1"/>
```

```

<elementRef key="finalRubric" minOccurs="0" maxOccurs="1"/>
<elementRef key="note" minOccurs="0" maxOccurs="unbounded"/>
</sequence>
</content>
```

Une autre règle schematron a également été implémentée car utile à plusieurs projets, notamment à HORAE. Il s'agit de faire en sorte que l'élément provenance contienne au moins l'élément <orgName>, <placeName> ou <persName> et l'élément <p> :

```

<elementSpec ident="provenance" mode="change">
  <constraintSpec ident="provStruct" scheme="isoschematron">
    <constraint>
      <!--<rule xmlns="http://purl.oclc.org/dsdl/schematron"
context="//tei:provenance" role="warning">-->
        <assert xmlns="http://purl.oclc.org/dsdl/schematron"
test ="(child::tei:orgName
          or child::tei:placeName
          or child::tei:persName)
          and child::tei:p">
          Provenance must contain (orgName or placeName
          or persName) and (p)</assert>
      <!--</rule>-->
    </constraint>
  </constraintSpec>
```

L'intérêt de cette méthode est la génération d'un message personnalisé en cas de non-conformité à la règle, ce qui n'est pas le cas d'une simple définition de séquence dans l'élément parent <elementSpec><sup>87</sup>. La balise <assert> affiche le message si la valeur de son attribut *test* est « false ». On peut également utiliser la balise <report>, à la place de <assert>, mais elle affiche le message si la valeur de son attribut *test* est égale à « true ». La balise <rule> n'est pas obligatoire<sup>88</sup>.

Pour que la règle schematron soit associée à un fichier ODD, la documentation du logiciel Oxygen précise qu'en plus de l'association du fichier au format rng<sup>89</sup>, il faut associer le schéma des règles schematron afin d'obtenir les associations suivantes :

```
<?xml-model href="percent.rng" type="application/xml"
schematypens="http://relaxng.org/ns/structure/1.0"?>
```

87. Les règles schematron offrent en effet plus de fonctionnalités dans la définition des contraintes. Cf. <https://en.wikipedia.org/wiki/Schematron>. Les possibilités sont d'autant plus fines grâce à l'utilisation de chemins Xpath, langage de requête pour localiser une portion d'un document XML.

88. Eddie ROBERTSSON, *An Introduction to Schematron*, 2003, URL : <https://www.xml.com/pub/a/2003/11/12/schematron.html> (visité le 12/06/2020).

89. Relax NG (Regular Language for XML Next Generation). C'est un langage de description de document XML qui permet de définir des contraintes quant à la structure d'un document XML.

## 1.2. DÉFINIR LE DOCUMENT CIBLE : TROUVER L'ÉQUILIBRE ENTRE CONTRAINTE ET ADAPTATION

---

```
<?xml-model href="percent.rng" type="application/xml"
schematypens="http://purl.oclc.org/dsdl/schematron"?>
```

Toutefois, la règle schematron affiche une erreur en rouge et non pas un simple avertissement en jaune, comme souhaité au départ.

Une fois le schéma d'encodage défini selon des règles bien déterminées, dans un fichier qu'il est possible d'associer au document final souhaité pour vérifier sa conformité avec le schéma, il reste à établir un modèle d'encodage en TEI pour les notices de livres d'heures. Dans un projet en humanités numériques, il est indispensable d'avoir conscience des informations à disposition et de comment les exploiter. Le projet HORAE est ainsi construit sur une méthode originale qui consiste à obtenir une bonne information à partir d'une information imparfaite. L'objectif est d'être le plus proche possible du document source en conservant toutes les données, la perte de données serait une faute grave. L'étape de la définition du document cible est d'autant plus cruciale qu'elle permet de dresser la liste des points à vérifier lors de la transformation. Trois notices ont donc été sélectionnées pour établir le modèle : la notice 1<sup>90</sup>, la notice 112<sup>91</sup> et la notice 313<sup>92</sup>.

Dans un premier temps, il faut vérifier les métadonnées du catalogue de notices pour bien remplir le <teiHeader> de l'ensemble du document. Cette partie du document est délicate car il existe des débats sur la place à accorder à l'édition électronique. Dans un projet, il est important de savoir reconnaître toutes les contributions à leur juste valeur. Dans le cadre du projet HORAE, l'édition électronique du catalogue de Victor Leroquais s'inscrit sous sa paternité, sa contribution scientifique est essentielle. L'édition électronique répond également aux critères du Consortium TEI qui édicte tout un panel de balises avec une sémantique bien précise, et ce pour répondre à une exigence d'interopérabilité. Il faut donc prendre garde à ce qui est contenu dans les balises <edition><sup>93</sup> et <responStmt><sup>94</sup>. La personne qui encode se doit donc de se demander où elle se situe dans la chaîne d'interactions que recouvre la responsabilité intellectuelle. Le principal rôle de l'encodeur est d'expliciter l'implicite du texte sur lequel il travaille, de révéler sa structure. Le <teiHeader> doit donc refléter correctement les métadonnées disponibles tout en signalant les différents acteurs de la publication à leur juste place<sup>95</sup>.

En ce qui concerne la structure générale des notices, il est mieux d'encadrer chaque notice par des balises <TEI>, comme pour le projet Ecmen, afin que le fichier ODD

---

90. V. LEROQUAIS, *Les livres d'heures manuscrits de la Bibliothèque nationale....*, p. 1-7.

91. *Ibid.*, p. 213-232.

92. *Ibid.*, p. 303-304.

93. La balise se réfère aux particularités de l'édition d'un texte. Cf. <https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-edition.html>.

94. Cette balise est relative à la notion de responsabilité quant au contenu intellectuel d'un texte, d'un enregistrement ou d'une publication en série. Cf. <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-respStmt.html>.

95. Le modèle établi pour l'encodage des notices de livres d'heures est disponible en annexes à la section A.1.4.

définie puisse un maximum s'appliquer. Pour rendre le document valide avec TEI P5, cela implique de conclure chaque notice de la manière suivante, même si nous ne disposons pas pour l'instant du corps du texte :

```
<text>
  <body>
    <p/>
  </body>
</text>
```

Ensuite, il est important de respecter la structure mise en place par Leroquais. Par exemple, l'ancienne cote, qui pourrait logiquement prendre sa place dans la balise <altIdentifier>, est ici citée avec la description du feuillet de garde, ce qui signifie que l'ancienne cote doit être encodée avec la description des premières pages du manuscrit, soit une partie de celui-ci, ce qui rapproche davantage l'information d'un <msItem>. Par ailleurs, le projet ne requiert pas la connaissance des anciennes cotes, donc il n'est pas essentiel de les rendre explicites. En ce qui concerne les en-têtes en majuscule des notices, l'information est éclatée entre un numéro de notices, un résumé du contenu et une date. Ce qui donne sur le papier « 1. LIVRE D'HEURES ET MISSEL FRANCISCAINS. 1380 », devient en XML<sup>96</sup> :

```
—      <TEI n="1">
—      <summary>LIVRE D'HEURES ET MISSEL FRANCISCAINS.</summary>
—      <origDate>1380</origDate>
```

Le numéro de la notice a finalement été placé dans un attribut « n » au sein de la balise <TEI> car cela permet de pointer immédiatement le numéro de la notice sur laquelle on se trouve.

L'enjeu le plus problématique de la structuration des notices de Leroquais est la mise en valeur de la succession et de l'imbrication des <msItem><sup>97</sup>. La composition des livres d'heures est en effet réputée complexe, avec des pièces qui se succèdent et qui s'imbriquent selon les niveaux. Ces <msItem> pourraient être à terme dotés de l'attribut *class* qui renvoie à des classes répertoriées dans le cadre du projet HORAE, sur le modèle du projet ECMEN. De manière générale, un paragraphe correspond à un <msItem>. Pour les questions d'imbrication et de subdivisions de <msItem>, il est envisageable de jouer sur les tirets longs et les numéros de folios.

En ce qui concerne les folios, il est important de les référencer dans des attributs sur lesquels il est possible de jouer dans le cadre d'une automatisation. Les folios sont

---

96. Cf. Notice 1.

97. La balise <msItem>, dans ce cas précis, fait référence à une partie d'une partie dans une partie de la notice.

## 1.2. DÉFINIR LE DOCUMENT CIBLE : TROUVER L'ÉQUILIBRE ENTRE CONTRAINTE ET ADAPTATION

---

ainsi encodés non seulement au sein des <msItem>, mais aussi des différents types de citations, sur le modèle suivant<sup>98</sup> :

```
<msItem>
  <locus from="54r" to="108r">54 à 108</locus>
  <title>Heures de la Vierge</title>
  <quote><locus n="101r">101.</locus>Chi apres sensicut l'offisse (sic)
  de Vadvent...</quote>
  <note>le début manque.</note>
</msItem>
```

Les <locus> dans les <msItem> sont intéressants en ce qu'ils répondent à des enjeux codicologiques. On peut, par exemple, comparer le calibrage des heures de la Vierge par rapport à la prière *Obsecro Te* pour savoir le taux d'abrévement relatif, et donc la capacité d'alphanumerisation du lecteur. Le <locus> est ainsi rendu obligatoire pour les feuillets extrêmes d'un <msItem>. Il se trouve alors être le premier enfant du <msItem>.

Les autres enfants sont les citations, qui peuvent être de cinq types différents. Voici leur ordre d'apparition au sein d'un <msItem> :

1. les <rubric><sup>99</sup>, indiquées en italique dans le texte de Leroquais, constituent logiquement la première citation ouvrant le <msItem>.
2. les <incipit>, signalés par un guillemet ouvrant et se terminant la plupart du temps par des points de suspensions.
3. les <quote>, citations dans le corps du texte débutant et se finissant par trois points de suspension.
4. les <explicit>, citation signalant la fin du texte dont il est question, et se finissant logiquement par une balise fermante.
5. les <finalRubric>, la rubrique de fin en italique dans le document source.

Enfin, les <msItem> peuvent se conclure par une balise <note>, utile pour tous types de commentaires sur la partie étudiée. Elle est même la seule balise du <msItem>, lorsque la notice signale des feuillets blancs, manquants, ou des lacunes, comme suivant<sup>100</sup> :

---

98. Cf. Notice 112.

99. En XML-TEI, la balise <rubric> n'est pas entendue dans son sens codicologique, elle signale le début d'un texte. Cf. <https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-rubric.html>. En codicologie, le terme désigne l'intitulé d'un texte ou d'une de ses parties mis en valeur par l'emploi d'encre de couleur, ou de lettres d'un type ou d'un module spécial, ou par tout autre procédé. Dans les manuscrits liturgiques latins, l'encre rouge est quelquefois utilisée pour les titres des pièces, les règles de célébration, les paroles de la consécration ou l'indication détaillée des gestes à accomplir. Cf. [http://codicologia.irht.cnrs.fr/theme/liste\\_theme/333#tr-65](http://codicologia.irht.cnrs.fr/theme/liste_theme/333#tr-65).

100. Cf. Notice 112.

```
<msItem>
  <note>Lacune entre 115 et 116.</note>
</msItem>
```

La question s'est posée d'encoder cette information dans un `<msItem>`. Du point de vue structurel, il s'agit bien d'un `<msItem>`, car nous sommes dans une partie autonome du manuscrit détachée des autres. De plus, cette partie est syntaxiquement indiquée par Leroquais comme les autres `<msItem>`, c'est-à-dire entre deux tirets longs. Si on choisit les tirets comme moyens de découper l'information lors de l'automatisation, la mention des lacunes se retrouveront bien dans un `<msItem>`.

En ce qui concerne la partie contextuelle et historique de la notice, il faut distinguer la balise `<origin>`<sup>101</sup>, qui renvoie à la création du manuscrit, de la balise `<provenance>`<sup>102</sup>, qui fait référence à l'appartenance du manuscrit après sa création et avant son acquisition, ainsi que de la balise `<acquisition>`<sup>103</sup>, qui contient des informations sur les circonstances de l'entrée du manuscrit dans l'institution de conservation qui le détient. Dans le cadre de l'automatisation de l'encodage, on peut imaginer pouvoir jouer sur des mots-clés comme « usage » pour la partie `<origin>` ou « possesseurs » pour la partie `<provenance>`.

Le format cible étant défini, il est temps de réfléchir aux moyens d'y arriver. Si différentes solutions sont possibles, le choix se porte sur la solution la plus rapide et la plus efficace selon les moyens humains et techniques dont on dispose.

### **1.3 Encoder les métadonnées : une opération entièrement automatisable ?**

Si certaines informations, notamment celles relatives aux références bibliographiques, à la description matérielle ou à l'historique du manuscrit, sont aisées à encoder automatiquement, les informations à propos du contenu du manuscrit présentent davantage d'irrégularités tout en devant être encodées dans une structure qui reflète la complexité des livres d'heures. Il est donc important de déterminer ce qui peut être encodé automatiquement de ce qui peut être affiné à la main. Pour réaliser cette opération semi-automatisée, plusieurs méthodes sont possibles. Il faut alors établir les avantages et inconvénients de chacune d'entre elles. Dans un projet en humanités numériques, le choix final de la méthode employée dépend de plusieurs facteurs, parmi lesquels le format des données initiales, la finesse exigée, l'environnement logiciel, les moyens humains et financiers

---

101. Cf. <https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-origin.html>.

102. Cf. <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-provenance.html>.

103. Cf. <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-acquisition.html>.

## **1.3. ENCODER LES MÉTADONNÉES : UNE OPÉRATION ENTIÈREMENT AUTOMATISABLE ?**

---

à disposition.

### **1.3.1 Structurer les notices avec XSLT**

Dans le cadre du stage, le choix de la méthode à employer s'est porté sur XSLT<sup>104</sup>. Langage de programmation fonctionnel, il permet de transformer, via une feuille de style XSL, un document au format XML en un autre document XML, mais il peut également produire d'autres formats en sortie, tels HTML ou PDF<sup>105</sup>. Le processeur XSLT lit le document d'entrée en suivant son arborescence et l'ordre d'apparition des éléments, ce qui signifie que les règles définies lors de la transformation sont activées dans l'ordre de rencontre des éléments, et que les éléments parents sont transformés avant les éléments enfants<sup>106</sup>. Le choix d'utiliser XSLT se justifie donc par sa vocation à produire des documents XML conformes à la TEI tout en constituant un langage aisément utilisable à prendre en main via le logiciel Oxygen XML Editor.

Étant donné que le langage de programmation XSLT accepte en document source essentiellement des documents au format XML, il a fallu convertir les notices océrisées dans ce format. Pour faciliter la récupération d'information selon la structure cible définie, les fonctionnalités du logiciel de traitement de texte Word ont tout d'abord été mobilisées. En effet, le découpage des notices en différents styles, chacun associé à une des grandes parties constitutives de la notice, a permis de jouer sur les critères structurants de mise en forme. Par exemple, tous les paragraphes contenant le mot « usage » de police de caractère de taille 16 appartiennent au style propre à la partie contextuelle et historique<sup>107</sup>.

La structure mise en valeur par les styles et leur ordre d'apparition sont les suivants :

1. le titre de la notice en rouge correspondant à l'élément <summary> ; dans ce même titre, on trouve dans un encadré le numéro de la notice, correspondant à l'attribut *n* de l'élément <TEI>, et la date attribuée au manuscrit, correspondant à l'élément <origDate> ;
2. le lieu de conservation et la cote du manuscrit en violet, correspondant respectivement aux éléments <repository> et <idno> ;
3. les paragraphes relatifs au contenu du manuscrit laissés tels quels et correspondant au <msContents> ; on a toutefois fait ressortir en jaune les informations relatives aux calendriers<sup>108</sup> ;

---

104. *Extensible Stylesheet Language Transformations*.

105. Elliotte Rusty HAROLD, W. Scott MEANS, Philippe ENSARGUET et al., *XML en concentré*, Paris, 2005, p. 519.

106. *Ibid.*, p. 164.

107. Une reproduction de la structuration des trois notices modèles avec les styles du logiciel Word est disponible en annexes, section A.2.1.

108. Les calendriers constituent en effet un <msItem> à part : les diviser sur les tirets longs créeraient des <msItem> fautifs. L'intérêt de les détacher des autres éléments du contenu est donc de pouvoir créer

4. les paragraphes concernant l'histoire du manuscrit en bleu et correspondant à l'élément <history>;
5. le paragraphe révélant des informations propres à la codicologie et correspondant à l'élément <objectDesc> en vert foncé;
6. les paragraphes concernant la décoration du manuscrit et correspondant à l'élément <decoDesc> en gris;
7. les débuts de paragraphe correspondant à la description de la reliure et concernant l'élément <bindingDesc> en saumon;
8. les fins de paragraphes indiquant des références bibliographiques et correspondant à l'élément <listBibl> en vert clair.

Le document word ainsi structuré, il suffit de le convertir au format XML via le service web Oxgarage Conversion proposé par le site de la TEI<sup>109</sup>.

Si l'on veut encore améliorer la bonne récupération des informations, il est possible de corriger dans le document d'entrée au format XML certaines erreurs liées à l'océrisation. Par exemple, un travail de restructuration des paragraphes, éclatés par des sauts de page créant des sauts de ligne, a été accompli. Ainsi, dans les notices papiers de Leroquais, les paragraphes concernant le contenu du manuscrit commencent par « Fol. », « Feuillet » ou un chiffre. Ceux ne débutant pas de cette manière dans le document océrisé sont à fusionner avec le paragraphe précédent. Toutefois, le cas des paragraphes débutant par un chiffre est problématique, car certains sont mal découpés sur les chiffres, comme le montre l'exemple ci-dessous :

```
<p rend="Histoire">  
L'office de la Vierge et celui des morts sont ceux de Rome ; le  
calendrier est franciscain ainsi que les litanies et le  
<hi rend="italic">Confiteor.</hi> La présence de  
saint Prosdocimc et de sainte Justine dans  
le calendrier semble désigner Padoue comme  
lieu d'origine du manuscrit ;  
toutefois, il convient de noter que les deux  
saints ne figurent ni dans les  
litanies ni dans le sanctoral. Je ne saurais  
dire s'il y a lieu d'attacher une  
signification spéciale à l'invocation : « Pro  
ministro » dans les litanies (fol.
```

---

une règle xsl qui leur est propre dans le code de transformation XSLT. Toutefois, cela peut aussi engendrer un déplacement de leur ordre d'apparition au sein de la description du contenu du manuscrit par rapport à leur place dans les notices originales.

109. Cf. <https://oxgarage.tei-c.org/>. On a ici sélectionné une conversion du format Microsoft Word (.docx) vers le format TEI P5 XML Document.

### 1.3. ENCODER LES MÉTADONNÉES : UNE OPÉRATION ENTIÈREMENT AUTOMATISABLE ?

152). Les différentes formules de prières ont été rédigées au masculin ; celle du fol. 428 v° semblerait indiquer que le volume a été transcrit pour un personnage dont le nom commençait par **<hi rend="italic">** Ma.**</hi>** La table pascale du fol. D v° donne la date du manuscrit :**</p>** **<p rend="Histoire">**1380. L'écriture et la décoration sont italiennes; les fautes d'orthographe sont assez fréquentes ainsi que les erreurs de transcription.**</p>**

Il faut donc vérifier et corriger manuellement ce dernier cas.

Au total, 1656 paragraphes concernant les contenus des manuscrits ne commencent ni par 'F', ni par un chiffre. Les paragraphes océrisés à corriger sont donc les suivants :

- Ceux débutant par un tiret ;
- Ceux débutant par une minuscule ;
- Ceux débutant par un guillemet ouvrant ;
- Ceux débutant par des points de suspension ;
- Ceux commençant par une majuscule mais qui concernent une citation centrée qui a été mise en exergue<sup>110</sup>.

Une partie a pu être corrigée automatiquement via une feuille xsl, notamment ceux commençant par un tiret, un guillemet ou des points de suspensions, mais une autre partie, trop dangereuse à automatiser, a été corrigée à la main, plus particulièrement les 172 paragraphes fautifs débutant par une minuscule<sup>111</sup>. De plus, il est possible de supprimer certains paragraphes parasites, indiquant des sauts de page par exemple, comme celui-ci :

```
<p>
<hi style="color:red; font-size: 14pt; font-weight:bold;" rend="ERROR"> </hi><note place="margin" type="conversion" resp="#teitodocx"><hi rend="docxError">unable to handle picture here, no embed or link</hi></note>
</p>
```

On retrouve 605 paragraphes de ce type. De même, 494 paragraphes vides ont été supprimés. Leur suppression rend les codes de transformation d'autant plus efficaces.

110. C'est notamment le cas dans la notice 112. Cf. Annexes Figure A.8.

111. Les codes xslt de repérage des paragraphes problématiques et de leur fusion avec le paragraphe précédent sont disponibles en annexes section A.2.1.

Le code XLST a tout intérêt à être le plus générique et le plus simple possible, afin de faciliter sa manipulation et sa réutilisation. En effet, complexifier une regex<sup>112</sup> pour quelques exceptions risque de générer des erreurs dans la récupération de l'information. Il est donc plus pertinent de diviser les problèmes, de les résoudre séparément, de faire des extractions étape par étape pour contrôler les éventuelles sources d'erreurs, et même parfois d'agir en plusieurs passes. De plus, XSLT étant un langage fonctionnel, il est plus approprié de faire une règle par élément pour une sortie attendue. Il s'agit donc d'avancer du plus général au plus spécifique en ayant conscience qu'il reste toujours des exceptions.

Dans le déroulé du code disponible en annexes<sup>113</sup>, une des premières étapes est donc de construire l'ossature des notices. Dans cette structuration, où l'on retrouve chaque balise nécessaire, on fait appel à des apply-templates<sup>114</sup> définis plus bas dans le code, afin que chaque élément de la notice suive les mêmes règles, sachant que chaque élément doit être traité indépendamment ensuite dans le code. Par ailleurs, la génération des éléments est conditionnée par la présence de l'information dans les notices d'origine grâce à une condition indiquée par une balise <xsl :if>, afin de ne pas générer de balises vides<sup>115</sup>. Il faut également faire attention aux balises de récupération des informations du document source, qui peuvent être encodées soit dans une balise <p>, soit dans une balise <hi>, notamment quand elles ne concernent qu'une partie du paragraphe et non pas un paragraphe entier, comme c'est le cas pour les informations relatives à la décoration ou aux reliures.

Les régularités dans la structure des notices ont permis de construire des règles simples. C'est notamment le cas pour le numéro des notices, le contenu général des manuscrits, leur cote, les informations relatives à la codicologie, à la décoration, à l'histoire et aux références bibliographiques. Pour la structuration des informations codicologiques, il a par exemple été possible de jouer sur la récurrence des termes « Parchemin », « Papier », « Vélin », « colonnes » ou « longues lignes ».

Les principales difficultés se sont concentrées dans la structuration des différentes parties thématiques composantes du manuscrit. En effet, ces parties structurantes, les <msItem> en TEI, peuvent tout autant concerner un paragraphe entier, mais pas systématiquement, que ce qui est entre des tirets longs, mais ce n'est pas une règle absolue.

---

112. Contraction de l'expression *regular expression*, une regex est un ensemble de motifs décrivant des chaînes de caractères bien précises afin de les sélectionner plus facilement dans un large corpus. Par exemple, dans le cadre de notre encodage semi-automatisé, les regexs doivent prendre en compte les éventuels sauts de ligne coupant, entre autres, une citation, pour bien récupérer l'ensemble de la citation concernée.

113. Cf. Les deux codes de transformation sont disponibles en annexes section A.2.2.

114. Cette instruction permet de sélectionner un ensemble de noeuds dans le document d'entrée pour leur appliquer les modèles appropriés.

115. C'est notamment le cas pour les informations relatives à la décoration du manuscrit, à ses dimensions et au nombre de feuillets.

### **1.3. ENCODER LES MÉTADONNÉES : UNE OPÉRATION ENTIÈREMENT AUTOMATISABLE ?**

---

Jouer sur l'un ou l'autre cirrière implique donc soit de sous-diviser les parties du manuscrit, soit de les sur-diviser, et dans les deux cas, de créer des *<msItem>* fautifs. Le plus judicieux reste toutefois de les diviser sur les tirets, car cela permet de structurer plus facilement les autres informations que contiennent les *<msItem>*, soit la foliation, les titres et les différents types de citation.

Il paraît alors possible de les regrouper lors d'une deuxième passe, d'une part en créant un attribut *corresp* qui associe un identifiant à chaque paragraphe, ce qui fait que les *<msItem>* appartenant au même paragraphe ont la même valeur dans l'attribut *corresp*, et d'autre part en jouant sur les numéros de folios. Il s'agit alors de faire une condition mathématique selon laquelle si le numéro de folio est inférieur ou strictement égal au numéro de folio précédent, alors l'information est à imbriquer dans le *<msItem>* précédent. Au contraire, s'il est supérieur, on crée un nouvel *<msItem>*<sup>116</sup>. Utiliser de telles conditions implique toutefois que les numéros de folios soient correctement encodés, détectés comme des chiffres et non pas des chaînes de caractères<sup>117</sup>. De plus, elles ne prendraient pas en compte les folios indiqués par des lettres capitales.

Les informations relatives au contenu du manuscrit ont donc été divisées sur les tirets. Pour chaque *<msItem>* hypothétique généré par le code, on crée une variable « *locus\_full* » qui récupère le numéro de folio selon les différentes situations observées dans le document source : les chiffres ou lettres majuscules qui suivent « *Fol* » ou « *fol* » ; les numéros qui coupent une citation et qui se trouvent donc entre des points de suspension ; les numéros précédés de « *P* »<sup>118</sup> ; et enfin, si les cas précédents ne sont pas rencontrés, tout ce qui précède un point, car il s'agit de la forme la plus récurrente au sein des paragraphes divisés sur les tirets longs. Les numéros de folios sont ensuite associés, dans les attributs *n*, *from* et *to*, à un « *v* » ou à un « *r* » selon qu'il s'agit d'un verso ou d'un recto.

À partir de la première variable « *locus\_full* », on crée une variable « *after\_locus* » qui, en excluant le numéro de folio de la sélection et en prenant tout ce qui suit, facilite le découpage des informations restantes selon la structuration souhaitée. Ainsi, tout ce qui ne relève pas d'une citation, donc ce qui n'est pas entre guillemet ou suivi de ponts de suspension, relève de la balise *<title>*. Si l'information est dans la majorité des cas correctement encodée, cette condition a toutefois pour défaut de récupérer parfois trop d'informations, dont certaines relèvent davantage de la balise *<note>*. Ensuite, dans la continuité des observations réalisées auparavant quant à la structure des différents types de citation (*<incipit>*, *<quote>*, *<explicit>*), nous avons établi des règles pour celles dont

---

116. Cf. <https://haypo.developpez.com/tutoriel/xml/xslt/programmation/#LV-A..>

117. Ce qui peut être problématique dans notre encodage, car les attributs *n*, *from* et *to* associent le numéro de folio à un « *r* » ou un « *v* » selon qu'il s'agit d'un recto ou d'un verso. De plus, il faudrait avant cette étape corriger les dernières éventuelles erreurs d'océrisation, principalement les « *O* » à la place des zéros, ou les « *l* », « *I* » ou « *i* » à la place du chiffre « *1* ».

118. Les manuscrits indiqués avec des pages plutôt que des folios restent relativement rares dans notre document source. Ce cas est présent dans les 7 notices suivantes : 83, 149, 185, 186, 290, 300, 310.

la structure était la plus sûre<sup>119</sup>. Les citations structurées différemment, qui demandent de la compréhension et qui ne contiennent pas de signes typographiques distinctifs, sont à encoder lors d'une vérification du document de sortie et de l'intégrité des données. Dans ces conditions, il est apparu plus prudent de récupérer l'ensemble des informations contenues entre les tirets, soit le <msItem> hypothétique, au sein de la balise <note>, afin que la vérification des données soit plus aisée et la compréhension de leur structuration plus évidente.

Cette division des <msItem> sur les tirets pose la question de la récupération des informations en italique. En effet, comme dit plus haut, l'italique indique dans les notices ce qui relève des balises <rubric> ou <finalRubric>. Or, jouer sur l'italique n'est pas possible directement dans la règle qui définit les <msItem>, car la fonction *tokenize()*, qui permet de diviser le texte initial sur les tirets s'applique sur des chaînes de caractères, et ne conserve que le texte du document source et non pas son encodage. L'information de l'italique ne peut donc se récupérer qu'à l'échelle du paragraphe. Pour pouvoir plus facilement associer l'information relevant de la balise <rubric> au bon <msItem>, nous avons eu l'idée de générer deux feuilles de sortie, l'une récupérant l'ensemble des informations sauf l'italique, l'autre récupérant uniquement l'italique au sein de <msItem> découpés sur les paragraphes. On peut ensuite associer les informations en italique au msItem> correspondant en les fusionnant sur les identifiants communs au paragraphe via la valeur de l'attribut *corresp*. Cette opération de fusion est possible avec les langages python<sup>120</sup> ou XSLT<sup>121</sup>, mais il est aussi très simple de passer par XQuery<sup>122</sup> via la requête suivante :

```
let $tout := doc("reprise/tout.xml")
let $rubricDoc := doc("reprise/rubric.xml")
return
  for $msItem in $tout//msItem[@corresp]
    let $corresp := $msItem/@corresp/string()
    return
      for $rubric in $rubricDoc//msItem[@corresp/string()=$corresp]//rubric
        return
          insert node $rubric
          as last into $msItem
```

On crée d'abord deux variables. La première désigne le document de sortie récupérant toutes les informations sauf l'italique, la deuxième désigne le document de sortie

119. Cf. section 1.2.2.

120. Cf. <https://stackoverflow.com/questions/15921642/merging-xml-files-using-pythons-elementtree>.

121. Cf. [http://www.journaldunet.com/developpeur/ressource/xml/xml\\_merge.shtml](http://www.journaldunet.com/developpeur/ressource/xml/xml_merge.shtml) et <http://pageperso.lif.univ-mrs.fr/~pierre-alain.reynier/XML/files/cours3.pdf>.

122. XQuery est un langage de requête permettant d'extraire des informations d'un ou de plusieurs documents au format XML, mais aussi d'effectuer des opérations complexes pour forger de nouveau documents ou fragments de documents XML.

### **1.3. ENCODER LES MÉTADONNÉES : UNE OPÉRATION ENTIÈREMENT AUTOMATISABLE ?**

---

récupérant uniquement l’italique au sein des `<msItem>`. Ensuite, on stocke la valeur des attributs `corresp` du premier document au sein d’une variable du même nom. Enfin, on copie dans le premier document chaque balise `<rubric>` dans le deuxième document qui est dans un `<msItem>` contenant la même valeur d’attribut `corresp`. Si la balise `<rubric>` est alors placée en dernière position au sein du `<msItem>`, ce qui n’est pas sa place appropriée<sup>123</sup>, cela indique qu’une vérification est nécessaire. En effet, l’italique peut désigner aussi bien une `<rubric>` qu’une `<finalRubric>`. De plus, l’italique désigne parfois des titres d’ouvrages cités en commentaire ou bien des termes en langue étrangère ou ancienne, en l’occurrence en latin, ce qui implique une relecture avisée pour les replacer ou les supprimer si elles sont fautives. Étant divisées sur les paragraphes et non sur les tirets, les balises `<rubric>` se répètent dans tous les `<msItem>` divisés sur les tirets qui ont le même paragraphe, ce qui facilite la vérification par rapport à ce qui est contenu dans la balise `<note>`<sup>124</sup>.

Pour les calendriers, traités comme des `<msItem>` à part, les règles du code sont semblables, avec les citations en mois. Les titres sont dans leur majorité bien encodés. Les seules exceptions concernent les titres qui ne suivent pas immédiatement l’information de foliotation, mais il est trop imprudent de changer la condition pour quelques exceptions à la règle. Toutefois, deux tâches sont à avoir à l’esprit lors de la phase de vérification. La première consiste à vérifier une éventuelle répétition d’informations<sup>125</sup>. La deuxième repose dans la vérification de l’enchaînement des folios au sein de la notice pour que le calendrier soit à sa juste place au sein du manuscrit.

Une fois la transformation opérée, il est important de tout d’abord vérifier l’intégrité des données du document de sortie, et qu’aucune information n’ait été perdue. Deux méthodes de vérification se complètent. La première est une lecture continue via le mode auteur du logiciel d’Oxygen, afin de rendre la lecture plus lisible et d’identifier la cohérence du fichier produit. La deuxième est une lecture ciblée en sélectionnant l’affichage des informations quelque peu déplacées par rapport au document d’origine<sup>126</sup>, puis en comparant le document source au document produit grâce aux fonctionnalités de Word<sup>127</sup>. Toutefois, le document de sortie présente davantage d’informations en double que de pertes.

La relecture a aussi permis de noter des exceptions qui n’avaient pas été repérées

---

123. Selon le schéma ODD défini, la balise `<rubric>` se situe après le titre et avant l’incipit s’il y en a un.

124. Comme dit plus haut, la balise `<note>` contient l’ensemble des informations sélectionnées entre les tirets, soit le `<msItem>` hypothétique.

125. Lorsque les informations relatives aux calendriers se trouvent dans le document source entre des balises `<hi rend=“Calendrier_Car”>` qui sont elles-mêmes contenues dans une balise `<p rend=“Texte du Corps (2)”>`, la même portion de texte est répétée, car les règles structurant les autres `<msItem>` récupèrent les informations au sein de la balise `<p rend=“Texte du Corps (2)”>`.

126. C’est notamment le cas pour la date donnée au manuscrit, qui se trouve sur la papier dans le titre et en XML TEI vers la fin de la notice.

127. Dans l’onglet « Révision », il est possible de comparer deux documents. Toutefois, étant donné la taille des fichiers manipulés, il est mieux de les diviser pour éviter un temps de chargement trop long.

lors de la première phase d’observation, mais qu’il est nécessaire de corriger à la main. Elles sont en effet trop exceptionnelles pour être incluses dans le code de départ. Certains numéros de folios signalant des citations ou un nouveau contenu au sein du manuscrit ne sont ainsi pas systématiquement séparés par un tiret et suivi d’un point, ce qui rend leur récupération plus complexe. De plus, certains folios sont indiqués comme « bis ». Nous avons donc choisi de les normaliser dans les attributs correspondant sous la forme suivante : « xxx\_bis\_r ». Par ailleurs, certains manuscrits présentent des variations de mise en page, tantôt avec deux colonnes, tantôt à longues lignes<sup>128</sup>. Selon les recommandations des TEI Guidelines<sup>129</sup>, ces variations sont inscrites de la manière suivante :

```
<layoutDesc>
    <layout columns="1 2">1 col. ou 2 col.</layout>
</layoutDesc>
```

Si XSLT est peut-être le langage le plus approprié pour encoder un document xml vers un autre document xml, nous voyons que le travail de reprise manuelle est indispensable à la bonne structuration des notices selon le document cible défini<sup>130</sup>. Il est donc intéressant d’observer les autres possibilités d’encodage semi-automatisé.

### 1.3.2 Structurer les notices avec Python

Utiliser Python<sup>131</sup> présente l’intérêt de partir directement du document Word océrisé et observé dans les arbres de décisions sans avoir à le transformer vers un autre format. La première étape consiste alors à délimiter les notices en elles-mêmes dans leur intégralité, puis leurs structures les plus englobantes (titre, contenu, description matérielle et références bibliographiques), pour structurer progressivement le général vers le particulier.

Ce découpage des notices en grandes parties structurelles est possible grâce à l’utilisation de regexs, comme en XSLT<sup>132</sup>. Toutefois, le document océrisé étant une chaîne unifiée de caractères, certains opérateurs, comme l’accent circonflexe (« î ») pour signaler un début de ligne et le dollar (« \$ ») pour signaler une fin de ligne, ne peuvent fonctionner. De plus, le texte présentant de riches combinaisons alpha-numériques, procéder avec

---

128. Cf. notice 12.

129. Cf. <https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-layout.html>.

130. Pour une vision des résultats de l’encodage avec la transformation via XSLT, cf. Annexes section A.2.3.

131. Python est un langage de programmation interprété, multi-paradigmes et multiplateformes. Contrairement à XSLT, ce n’est pas un langage de programmation dit « fonctionnel », qui se déroule comme une suite d’évaluations de fonctions, mais il est qualifié comme « orienté objet », ce qui signifie que le programme peut être considéré comme une collection d’objets en interaction.

132. Pour utiliser des expressions régulières en Python, cf. documentation du package re à l’adresse suivante : <https://docs.python.org/fr/2.7/library/re.html#re.match>.

### 1.3. ENCODER LES MÉTADONNÉES : UNE OPÉRATION ENTIÈREMENT AUTOMATISABLE ?

---

des regexs trop génériques est dangereux. Voici donc le code pour séparer les différentes notices et associer leur titre à leur contenu<sup>133</sup> :

```
import re
import docx2txt

notices = docx2txt.process('/Users/gwenaellepatat/Desktop/Stage_TNAH/
MémoireHORAE/Catalogue_VL/noticesTestsoocr.docx', 'r')

#Capture des titres de notices dans l'ordre stockées dans un dictionnaire
livres_heures={}
for char in range(1, 320):
    try:
        titre=re.search(r"\n"+'((1|I){1,2})?' + str(char) +
        "\." + '.*?[A-ZÉÈÙÀ]{2,}.*?\n', notices).group(0)
        livres_heures[char]=titre
        #print(titre)
    except: #Exceptions utiles pour détecter les possibles
        #dénominations irrégulières
        #print(char)
        continue

#Utilisation des valeurs du dictionnaire, soit les titres de notices,
pour diviser le texte et associer le contenu à chaque titre
regexPattern = '|'.join(map(re.escape, livres_heures.values()))
contenu=re.split(regexPattern, notices)

#Zip de la liste pour créer une liste de tuples contenant les titres
et leur contenu
structure_notices = list(zip(livres_heures.values(), contenu[1:]))

for char in structure_notices:
    fichier = open("/Users/gwenaellepatat/Desktop/Stage_TNAH/
MémoireHORAE/Catalogue_VL/notice_Leroquais.txt", "a")
    fichier.write('<TEI> Titre : %s \n Content : %s </TEI>
\n\n' % (char[0], char[1]))
    fichier.close()
    #print('<TEI> Titre : %s \n Content : %s </TEI> \n\n'
% (char[0], char[1]))
```

---

133. Une copie du code dans son intégralité est disponible en annexes, section A.2.2.

Le premier package importé permet d'utiliser des opérations propres aux expressions régulières, le deuxième de lire un document word. Ensuite, on ouvre le document source pour pouvoir le lire et le parcourir. On peut alors stocker les titres des notices associées à un numéro d'ordre (celui de l'ordre de lecture qui correspond bien aux numéros attribués par Leroquais) dans un dictionnaire<sup>134</sup>, où le titre est la valeur et le numéro de notice la clé. On fait donc une boucle qui parcourt les 313 notices et qui permet de sélectionner les titres via une regex. Ces titres sont ensuite stockés comme une entrée du dictionnaire. Les clauses « *try* », « *except* » et « *continue* » permettent de ne pas arrêter l'exécution du code même si des exceptions sont rencontrées. L'utilisation de la fonction *print()*, ici mise en commentaire, est utile pour repérer les éventuelles irrégularités.

La regex pour récupérer les titres fonctionne ici pour les erreurs d'océrisation, notamment les « I » et « l » à la place de 1 dans certains numéros de notices, mais pas pour la seule exception qui prend la forme suivante : « 311-312. CREDO DU SIRE DE JOINVILLE ET LIVRE D'HEURES. XIII<sup>e</sup> SIÈCLE FIN, ET XIV<sup>e</sup> SIÈCLE FIN ». En effet, la notice ne comprend pas un, mais deux numéros séparés par un tiret. Comme il s'agit de la seule exception, il est plus prudent de la corriger à la main, afin de ne pas créer d'erreurs dans la récupération des titres par ailleurs.

Les titres capturés sont ensuite utilisés pour diviser le texte en notices et associer à chaque titre son contenu. Dans la variable « *regexPattern* », on fait des titres stockés dans le dictionnaire une suite d'itérables (fonction *map()*) joints par un *pipe* (fonction *join*) dans laquelle les métacaractères, à l'exception des lettres ASCII<sup>135</sup>, des nombres et de tirets du bas, ont été échappés (fonction *re.escape*). La variable *contenu* désigne alors ce qui dans le document Word se situe entre les titres. On crée ensuite une liste de tuples<sup>136</sup>, où un tuple contient le titre de la notice et en regard son contenu.

On itère alors sur la liste de tuples en générant un document au format .txt où chaque notice est encadrée par une balise TEI. Le texte peut ainsi être parcouru item par item, ce qui permet d'utiliser des regexs séquentielles moins complexes et de baisser le taux d'erreurs<sup>137</sup>.

Pour diviser les informations au sein du contenu ainsi récupéré, il faut faire une liste de listes en passant par un dictionnaire. La liste principale sert donc à la fois d'entrée et de sortie dans le code. En effet, on ajoute systématiquement un nouvel élément à partir d'une itération sur le contenu, de la manière suivante :

#### *#Capture des descriptions matérielles des manuscrits*

---

134. En python, un dictionnaire est un objet conteneur qui associe des clés et des valeurs.

135. *American Standard Code for Information Interchange*. L'acronyme désigne une des normes informatiques de codage de caractères les plus influentes.

136. Un tuple permet de créer une collection ordonnée de plusieurs éléments. Contrairement aux listes, il n'est pas mutable et n'est plus modifiable après sa création, ce qui est un gage de sécurité dans le cas de notre travail de structuration.

137. Une présentation des résultats obtenus avec Python est disponible en annexes, section A.2.3.

### 1.3. ENCODER LES MÉTADONNÉES : UNE OPÉRATION ENTIÈREMENT AUTOMATISABLE ?

---

```
for index, char in enumerate(structure_notices):
    try:
        physDesc=re.search(r"(Pareil\.|Parch\.){1}((.*?)\n)*(Rel\.
|Rcl\.|Demi\|-reliure){1}(.*?)( -)?", char[1]).group(0)
        # char[1] est le contenu dans notre tuple
        structure_notices[index].append(physDesc)
        #Modification de la liste principale
        #print(physDesc)

    except :
        print(index)
        Continue

#Capture de la bibliographie
for index, char in enumerate(structure_notices):
    try:
        additional=re.search('(Rel\.|Rcl\.|Demi\|-reliure){1}(.??
\.\\s+\\-\\s+([A-ZÉÀÙ]{1},.*))\\n', char[1]).group(3)
        # char[1] est le contenu dans notre tuple
        structure_notices[index].append(additional)
        #Modification de la liste principale
        #print(additional)

    except :
        print(index)
        Continue
```

Chaque tuple dans la liste étant numéroté grâce à la fonction `enumerate()`, on peut boucler sur le numéro d'indexation et les caractères associés. On fait ensuite une regex capturant la portion de texte souhaitée au sein du contenu de la notice, puis on ajoute la nouvelle entrée à la liste principale de départ. Il faut toutefois faire attention dans la formulation des regexs, car certains noms dans les références bibliographiques sont en petites capitales dans le document Word, et cette typographie a été transformée en lettres minuscules dans le fichier text en structuration.

Cependant, l'idée de passer par python pour automatiser la structuration des notices présente certains désavantages. Outre que certains passages de code peuvent être plus longs à penser, et donc moins rentables, qu'une structuration à la main via l'utilisation de styles sous Word, il faut éviter le format .txt en sortie. Il est plus judicieux, pour conserver la structure de la liste finale, de créer un fichier de sortie au format json à partir duquel on pourrait générer un document au format final souhaité, en XML TEI<sup>138</sup>.

---

138. Il est toutefois possible de générer un document au format xml depuis Python. Cf. <https://www.fil.univ-lille1.fr/~marvie/python/chapitre4.html>.

De plus, Python est souvent utilisé pour faire des transformations vers texte brut, ce qui implique la perte des métadonnées de style. La technique utilisée ici ne peut en effet plus s'appuyer sur des critères comme la taille de police. Si l'on souhaite explorer les possibilités offertes par Python, on peut toutefois s'intéresser au package `python_docx` qui permet de séparer le texte en italique et en gras du texte normal, ou obtenir la taille de la police, mais cela peut tout de même donner des résultats déstructurés depuis le document océrisé. Python reste tout de même intéressant pour jouer davantage sur les numéros de folios dans la structuration des `<msItem>`. Avec une condition « if » et une fonction `startswith()`, on peut séparer le contenu à partir des séries de folios en début de paragraphe. Une fois les informations séparées sur les numéros de folios, informations que l'on peut assimiler à des `<msItem>`, il devient pertinent de jouer sur les métadonnées de style, car elles sont davantage isolées dans une portion de texte restreinte. On peut ensuite utiliser les fonctions `range()`<sup>139</sup> et `.format()`<sup>140</sup> pour faire s'imbriquer les `<msItem>` selon les numéros de folios.

Si nous n'avons pas pu aller au bout de l'expérimentation en utilisant Python, il est probable que la structuration automatisée des informations implique une relecture attentive et une correction à la main, notamment pour les citations à l'intérieur des `<msItem>`. L'utilisation de langage de programmation ne semble pas empêcher ici une structuration davantage semi-automatisée qu'automatisée. Il est alors intéressant de s'interroger sur le potentiel des techniques offertes par l'apprentissage machine et dans quelles mesures elles peuvent s'adapter au cas des notices de livres d'heures établies par Leroquais.

### 1.3.3 Les possibilités offertes par l'apprentissage machine

L'apprentissage machine, plus couramment appelé « *machine learning* », est une technologie d'intelligence artificielle qui consiste à entraîner un modèle grâce à un grand nombre de données à analyser, ce que l'on nomme le *Big Data*. Il ne s'agit donc pas de langages de programmation comme Python ou XSLT, mais d'algorithmes, afin de découvrir des répétitions dans plusieurs flux de données et d'affiner les capacités de reconnaissance de la machine. Le *machine learning* permet ainsi de fournir des prédictions en se fondant sur des statistiques. Plus le nombre de données est important, plus l'apprentissage machine est efficace<sup>141</sup>. La notion de *Big Data* se fait alors toute relative dans le cadre de l'encodage des 313 notices de livres d'heures établies par le chanoine Victor Leroquais.

L'emploi du *machine learning* pour l'automatisation d'encodages de catalogues a déjà été expérimenté. L'un de ces travaux est notamment celui de l'encodage automatisé

---

139. Cf. <https://docs.python.org/fr/3/tutorial/controlflow.html>.

140. Cf. <https://www.geeksforgeeks.org/python-format-function/>.

141. Cf. <https://ia-data-analytics.fr/machine-learning/>.

### 1.3. ENCODER LES MÉTADONNÉES : UNE OPÉRATION ENTIÈREMENT AUTOMATISABLE ?

de catalogues de ventes d'autographes numérisées<sup>142</sup>. Une des premières étapes préalables à l'utilisation de l'apprentissage machine a été la transformation des catalogues numérisés en fichier PDF recouverts d'une couche de texte<sup>143</sup>. Le *machine learning* s'est ensuite effectué grâce à la mise au point d'un logiciel spécifique, GROBID-dictionaries. Il constitue un sous-projet de GROBID<sup>144</sup>, solution pour extraire les informations bibliographiques dans les articles scientifiques<sup>145</sup>. Le schéma suivant résume le fonctionnement du modèle d'apprentissage<sup>146</sup> :

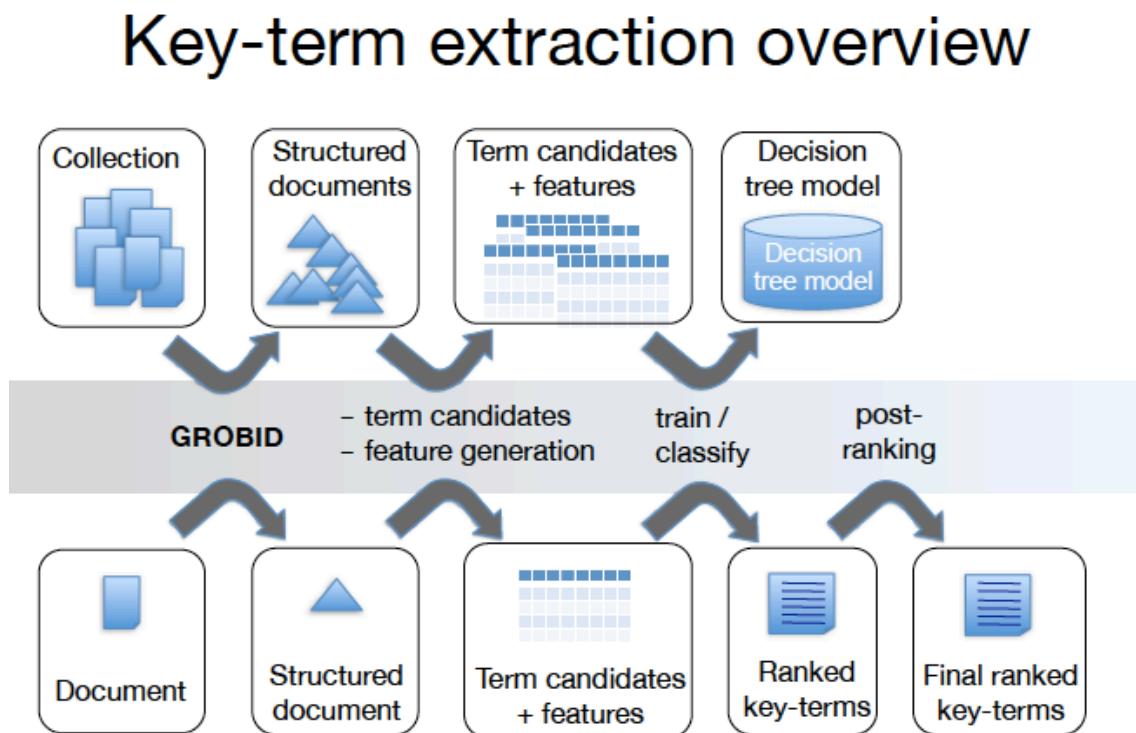


FIGURE 1.1 – Extraction des termes clés avec GROBID

À partir d'un document donné, le logiciel repère sa structure, composé dans le cadre des articles scientifiques d'un *header*, soit le titre et un résumé, du corps du texte, soit une introduction, des sections et une conclusion, puis d'une liste de références, avec des articles, des revues et des livres, entre autres. GROBID se concentre ensuite sur les termes équivalents et les *features*, soit la présence de références bibliographiques dans le *header*

142. L. RONDEAU DU NOYER, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la Revue des autographes de la librairie Charavay...*

143. *Ibid.*, p. 31.

144. La documentation de l'API Web est disponible à l'adresse suivante : <https://grobid.readthedocs.io/en/latest/>.

145. Laurent ROMARY et Patrice LOPEZ, « GROBID - Information Extraction from Scientific Publications. » *ERCIM News*, 100 (2020), URL : <https://hal.inria.fr/hal-01673305> (visité le 23/06/2020).

146. P. LOPEZ et L. ROMARY, *Automatic Key Term Extraction from Scientific Articles in GROBID*, SemEval 2010 Workshop, Uppsala, Sweden, 2010, eprint : [inria-00493437](https://hal.inria.fr/inria-00493437), URL : <https://hal.inria.fr/inria-00493437/document/> (visité le 22/06/2020), p. 27.

et le corps du texte, ainsi que la position de la première occurrence d'un mot dans le document<sup>147</sup>. Il s'agit enfin, en s'inspirant d'un arbre de décision, comme nous l'avons établi dans le cadre de l'encodage des notices de livres d'heures, d'entraîner la classification et la typologie des termes clés.

GROBID-dictionaries consiste donc à transposer le principe de GROBID aux informations lexicales via une librairie java d'apprentissage supervisé<sup>148</sup>. Le logiciel GROBID-dictionaries permet alors de traiter, d'extraire des informations textuelles de dictionnaires numérisés enregistrés au format PDF, et de structurer automatiquement les informations en XML-TEI<sup>149</sup>. L'intérêt de ce logiciel est de s'adapter aux entrées lexicales par l'apprentissage supervisé, et donc de pouvoir traiter toute ressource numérisée prenant une forme encyclopédique. Pour obtenir l'encodage de sortie souhaité, il faut donc entraîner le modèle selon une structure bien définie des entrées de catalogue. Un fichier XML respectant les standards de la TEI sert donc d'entraînement pour améliorer la performance du logiciel d'apprentissage machine, ce qui signifie que le fichier d'entraînement doit être encodé avec soin manuellement<sup>150</sup>.

Si l'on fait une analogie avec le catalogue de notices de Leroquais, les documents nécessaires à l'apprentissage machine seraient donc d'une part l'établissement des trois notices modèles en XML TEI, d'autre part un fichier .rawtxt qui contiendrait le texte extrait du document source à encoder, soit les notices océrisées. Pour adapter le modèle au cas des notices de livres d'heures, plusieurs niveaux d'entraînement seraient envisageables. Le premier consisterait à établir une segmentation bien précise de l'ensemble du document à encoder, de la structure de la page à l'entrée lexicale. Cela impliquerait de prendre en compte la variété des pages, car une notice peut s'étaler sur plusieurs pages comme recouvrir moins d'une page entière. L'objectif de cette première étape est de séparer le corps du texte des éléments annexes, comme les réclames ou les numéros de page<sup>151</sup>. Par lignes de commandes, on crée ensuite les données d'entraînement à partir de données annotées<sup>152</sup>.

Après avoir segmenté la structure générale de la page, la deuxième étape se concentre sur la segmentation du corps du texte, soit dans notre cas les différentes parties constitutives de la notice, celles qui apparaissent systématiquement (le numéro et le titre de la

---

147. L. ROMARY et P. LOPEZ, « GROBID - Information Extraction from Scientific Publications. »..., p. 15.

148. L'installation de GROBID-dictionaries se fait à partir du terminal. Cf. (Simon GABAY, Mohamed KHEMAKHEM et L. ROMARY, *Les catalogues et GROBID. Doctorat. Du catalogue aux humanités numériques : quelles méthodes pour quels résultats ?*, 2018, URL : <https://hal.archives-ouvertes.fr/cel-01951107> [visité le 19/06/2020], p. 14).

149. L. RONDEAU DU NOYER, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la Revue des autographes de la librairie Charavay...*, p. 34.

150. *Ibid.*, p. 39.

151. S. GABAY, M. KHEMAKHEM et L. ROMARY, *Les catalogues et GROBID. Doctorat. Du catalogue aux humanités numériques : quelles méthodes pour quels résultats ?...*, p. 21-22.

152. *Ibid.*, p. 23-25.

### **1.3. ENCODER LES MÉTADONNÉES : UNE OPÉRATION ENTIÈREMENT AUTOMATISABLE ?**

---

notice, le résumé et le détail du contenu, les éléments concernant la reliure, etc.), comme celles qui n'apparaissent pas à chaque fois (l'historique ou les références bibliographiques par exemple). De même, les données d'entraînement sont créées et annotées.

Enfin, pour l'analyse des dictionnaires, les troisième et quatrième niveaux d'entraînement s'appuient sur les entrées lexicales, leurs formes et leurs sens. Si l'on adapte le modèle aux notices de livres d'heures, il faudrait créer et annoter autant de données d'entraînement que de balises TEI définies dans le format cible, afin que le modèle reconnaisse ce qui concerne le <msContents>, le <physDesc>, l'<history> ou l'<additional> .

Toutefois, l'expérience de l'encodage automatisé des catalogues de ventes montre que l'apprentissage machine n'est pas une science exacte. Malgré la quantité et la qualité des données, les résultats d'encodage n'ont pas toujours été pertinents. Pour améliorer les résultats, il a été fait appel au *feature engineering*. Le *feature* renvoie à une représentation numérique d'un aspect de la donnée brute, préparée pour l'algorithme, afin de servir d'intermédiaire entre celles-ci et les modèles utilisés. Le *feature engineering* consiste donc à déterminer quels sont le ou les *features* les plus pertinents à extraire des données brutes disponibles pour arriver aux meilleurs résultats. Cette technique doit être contextuelle pour être efficace, dépendant à la fois des données à traiter, du modèle d'apprentissage utilisé et de l'objectif final. Dans cette exemple, le *feature engineering* s'est traduit par le remplacement d'un modèle unigramme par un modèle bigramme, appliqué aux mêmes données d'entraînement et d'évaluation. Un modèle unigramme associe une étiquette à une unité lexicale fournie en fonction des seules caractéristiques de cet élément. Un modèle bigramme prend lui en compte l'étiquette appliquée à l'unité lexicale précédente pour améliorer l'exactitude de l'étiquetage. Cette étape a considérablement amélioré le modèle dans le cas des catalogues de ventes d'autographes<sup>153</sup>.

*In fine*, la technologie développée par le logiciel GROBID s'appuie sur les éléments de mise en page, le texte et la ponctuation, ce qui signifie que la qualité de l'océrisation est cruciale<sup>154</sup>, mais cette remarque vaut pour tout processus d'encodage que l'on souhaite automatiser. Par ailleurs, plus la structure du document cible est complexe (enchaînement et imbrication des balises TEI), plus l'entraînement et l'apprentissage du modèle seront longs et complexes. Mettre au point une automatisation de l'encodage des notices de livres d'heures établies par Victor Leroquais demanderait donc du temps, ainsi que les moyens financiers et humains adéquats pour trouver des solutions le plus rapidement et efficacement possible. Si l'aventure peut s'avérer complexe et ardue, la création d'un logiciel d'encodage automatique de catalogues de notices de manuscrits, s'appuyant sur des modèles d'encodage en XML TEI défini en amont, pourrait par la suite représenter

---

153. L. RONDEAU DU NOYER, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la Revue des autographes de la librairie Charavay...*, p. 42-44.

154. S. GABAY, M. KHEMAKHEM et L. ROMARY, *Les catalogues et GROBID. Doctorat. Du catalogue aux humanités numériques : quelles méthodes pour quels résultats ?...*, p. 59.

un gain de temps considérable, bien que la vérification de l'intégrité des données et les éventuelles corrections manuelles restent essentielles pour s'assurer de la qualité des données.

Si la structuration de catalogues de notices peut se faire au format XML TEI, structurer des données en masse complexes et liées entre elles implique d'utiliser une base de données relationnelles.

# Chapitre 2

## De l'importance de la modélisation des données

La modélisation, opération qui consiste à schématiser un modèle complexe afin de mieux « mesurer les effets sur ce système des variations de tel ou tel de ses éléments composants »<sup>155</sup>, est une étape essentielle à tout import de données dans une base appropriée. En effet, elle permet de structurer les données, d'établir leurs relations et la cardinalité de leurs relations, c'est-à-dire le nombre d'occurrences minimal et maximal d'une association par rapport à chaque occurrence d'une entité. Pour le projet HORAE, il s'agit de modéliser les usages liturgiques, qui désignent les éléments liturgiques propres à un lieu. Ils peuvent donc être aussi variés que le kaléidoscope de livres d'heures conservés.

### 2.1 Modéliser les usages : restituer une réalité complexe

Modéliser, c'est d'abord comprendre ce que l'on doit représenter, les données qui sont à manipuler. On peut alors dresser un modèle conceptuel qui définit les données, leur mode d'évolution dans le temps et leurs relations. La deuxième étape consiste à le transformer en un modèle logique relationnel pour que les données deviennent des objets manipulables dans la base de données relationnelle. Enfin, des tests pour l'implémentation dans la base ont été effectués, afin de définir quelle est la solution la plus intéressante intellectuellement, applicativement et ergonomiquement pour l'import des données.

#### 2.1.1 Qu'est-ce que l'usage liturgique ?

Si le cursus correspond aux chants, lectures et oraisons choisis par une communauté religieuse pour célébrer les offices et la messe des fêtes du sanctoral et du temporal, ces

---

155. Cf. <https://www.cnrtl.fr/definition/modélisation>.

différentes manières de célébrer le culte dépendent de ce que l'on appelle l'usage liturgique. Ce terme désigne l'ensemble des pratiques particulières au culte d'une communauté, relative à un chapitre, une abbaye, un diocèse ou un ordre religieux. De fait, la mention du cursus d'un lieu sert à évoquer les choix de pièces qui caractérisent les heures de la Vierge à l'usage de ce lieu, par exemple. Si l'usage ne modifie pas la structure même de l'office<sup>156</sup> et de la messe<sup>157</sup>, il se distingue par<sup>158</sup> :

- le culte rendu à des saints particuliers ;
- la préséance entre les offices<sup>159</sup> ;
- le choix ou la création de pièces liturgiques dans les propres du temps et des saints ;
- des cérémonies particulières, comme des processions.

Dans le cadre du projet HORAE, il s'agit de déceler les dévotions les plus intimes à partir des usages présents dans les livres d'heures. On peut caractériser ces usages grâce à plusieurs critères, en premier lieu la rubrique. Si ces mentions deviennent plus fréquentes à la fin du Moyen Âge, il convient de les regarder avec un œil critique. L'usage inscrit n'est en effet pas toujours celui de la destination liturgique. Ainsi, dans les livres d'heures, les usages de l'office des morts ou du petit office de la Vierge les plus répandus sont ceux de Rome, Paris, Sarum<sup>160</sup> et Utrecht. Ils n'indiquent pas pour autant de manière systématique la destination du manuscrit<sup>161</sup>.

Un autre critère est celui des calendriers. On peut en effet y trouver les fêtes des saints locaux, mais aussi la fête d'anniversaire de la dédicace de l'église locale et la fête de ses reliques ; la mention d'obits<sup>162</sup> et de fondations, bien qu'il faille distinguer celles dues à la copie primitive et celles ajoutées ultérieurement ; le rite des offices, ou encore les jeux d'encre mettant en valeur telle ou telle fête. Toutefois, l'examen du calendrier doit être mis en regard avec le contenu du manuscrit qui confirme ou infirme l'usage qui ressort du calendrier<sup>163</sup>.

Un autre pan du livre d'heures intéressant à observer pour caractériser les usages sont les litanies. Il s'agit d'une liste de saints invoqués. On prête alors particulièrement

---

156. L'office désigne un ensemble de lectures et de prières prévues pour un moment précis.

157. Célébration collective de l'office du jour.

158. Jean-Baptiste LEBIGUE, *Les usages liturgiques*, 2016, URL : <https://irht.hypotheses.org/2484> (visité le 30/03/2020).

159. L'une des principales difficultés dans la récitation de l'office des heures et la célébration de la messe provient de la superposition du temporal et du sanctoral, c'est-à-dire lorsqu'une ou plusieurs célébrations issues du temps de l'année liturgique ou de celui des saints se superposent le même jour. Ces cas sont résolus grâce à l'établissement de préséances entre les divers offices, susceptibles toutefois de présenter de nombreuses variantes selon les usages. Cf. Id., *Les règles de préséances entre les offices*, 2017, URL : <https://irht.hypotheses.org/2473> (visité le 30/03/2020).

160. Aussi appelé usage de Salisbury, c'est une synthèse des usages en vogue en Grande-Bretagne et du rite pratiqué à Rouen.

161. Id., *Les usages liturgiques...*

162. Un obit est une messe célébrée par fondation pour un défunt à la date anniversaire de son décès.

163. *Ibid.*

## 2.1. MODÉLISER LES USAGES : RESTITUER UNE RÉALITÉ COMPLEXE

---

attention aux saints locaux, mais aussi à la catégorie dans laquelle ils se trouvent (martyrs, apôtres ou confesseurs)<sup>164</sup>.

Les autres éléments révélateurs sont les propres du sanctoral<sup>165</sup>, les messes et offices votifs, parfois liés à des reliques locales ou les précisions toponymiques pour les processions<sup>166</sup>.

Toutefois, il faut distinguer l'usage et la destination, d'autant plus dans les livres dédiés à la dévotion privée. En effet, le destinataire peut vouloir s'inspirer d'un usage qui n'est pas propre à son lieu de vie. Ainsi, le choix et l'ordre des chants dans le petit office de la Vierge et l'office des morts est propre à un diocèse, à un ordre monastique ou religieux, voire à une abbaye ou à une collégiale, mais l'usage employé ne révèle pas une indication d'origine. Les usages de Rome et de Paris sont par exemple trop répandus pour indiquer une provenance stricte de Paris ou de Rome. C'est pourquoi, dans l'étude des livres d'heures, il est préférable d'employer l'expression « selon l'usage de », à moins que le manuscrit ne fournisse la preuve explicite de sa destination à tel lieu, ordre ou institution<sup>167</sup>.

Une fois que l'on a saisi ce qu'il faut modéliser, il s'agit de s'interroger sur comment décrire un usage dans une base de données. Lorsqu'il s'agit du cursus, il faut penser simultanément la modélisation de la hiérarchie des objets et de leur ordre de succession, qui constituent deux informations spécifiques à l'usage liturgique. La base de données doit donc répondre à la question suivante : Qu'est-ce qu'une pièce textuelle du livre d'heures, dans sa forme, son contenu et sa place au sein des autres textes, nous dit des usages liturgiques propres à une institution ?

Voici notre premier essai de modèle conceptuel<sup>168</sup> :

---

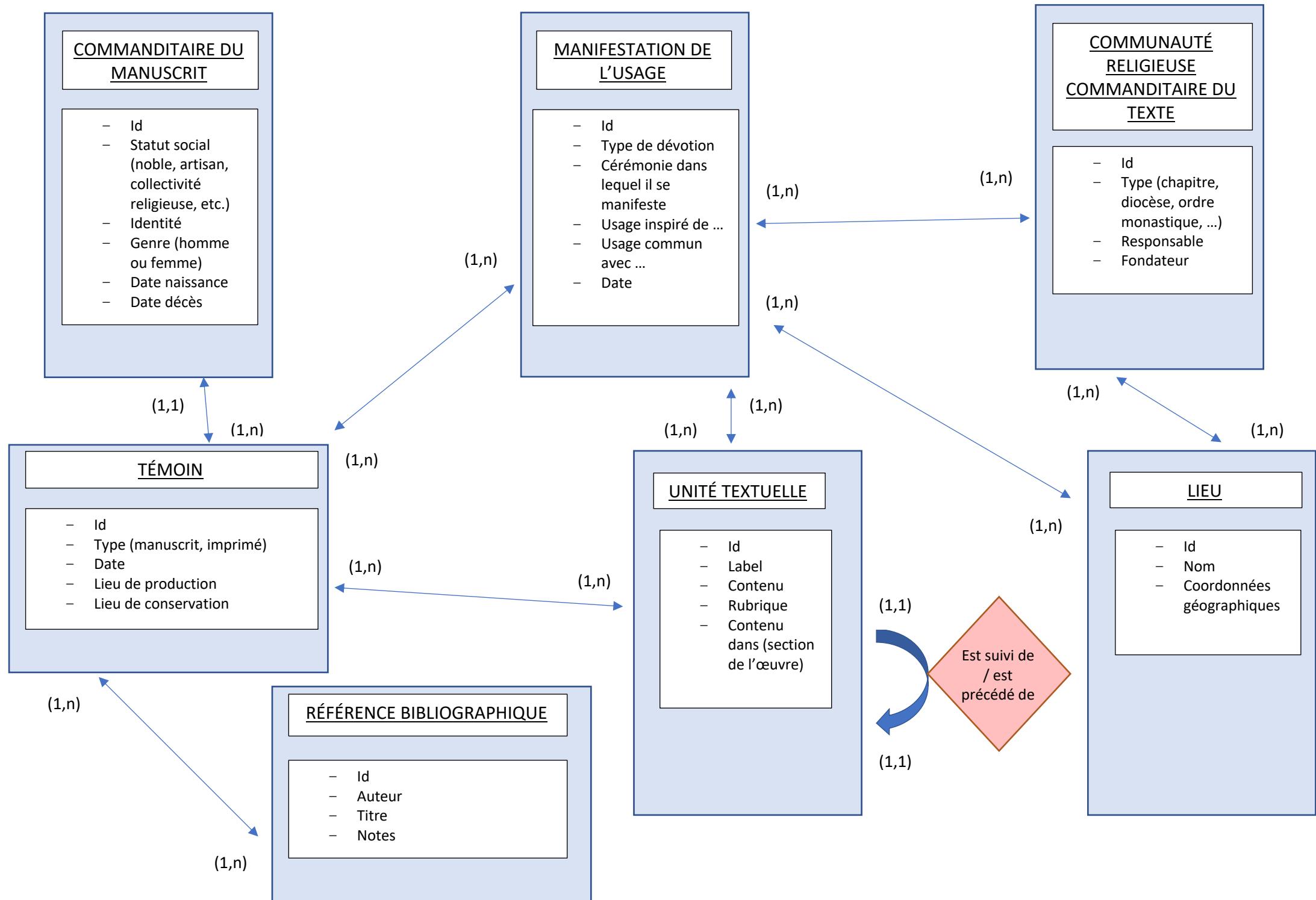
164. *Ibid.*

165. Il s'agit de l'ensemble des fêtes des saints du calendrier liturgique célébrées.

166. *Ibid.*

167. Id., *Initiation aux manuscrits liturgiques*, 2007, URL : <https://cel.archives-ouvertes.fr/cel-00194063> (visité le 27/01/2020), p. 5.

168. Une reproduction est disponible en annexes section B.1.



## 2.1. MODÉLISER LES USAGES : RESTITUER UNE RÉALITÉ COMPLEXE

---

Chaque rectangle correspond à une entité, qui peut contenir plusieurs attributs, ici signalés par des tirets. Il est évident que les attributs sont amenés à être enrichis, et qu'ils ne peuvent être systématiquement renseignés ; tout dépend des données disponibles. Chaque entité possède un identifiant, afin d'être plus facilement manipulable.

Une bonne modélisation doit respecter un certain nombre de règles de normalisation. Ces règles permettent d'éviter d'éventuels problèmes de transaction, la redondance ou l'incohérence des données, ainsi que des contre-performances. Il faut par exemple éviter d'avoir des attributs calculables à partir d'autres, afin de faciliter les mises à jour des données. Les chiffres entre parenthèses de part et d'autre des flèches, qui par ailleurs indiquent une relation entre deux entités, correspondent aux cardinalités. Elles se lisent de la manière suivante : un usage liturgique se manifeste dans une à plusieurs unités textuelles, et une unité textuelle appartient à un ou plusieurs usages. Pour répondre à l'exigence de succession et de hiérarchisation des unités textuelles, nous avons eu l'idée de créer un attribut « contenu dans » et une relation unaire, c'est-à-dire qu'une unité textuelle peut être précédée et suivie d'autre unités textuelles. Les cardinalités sont cruciales pour établir le modèle logique, qui prend la forme suivante<sup>169</sup> :

---

169. Une reproduction de ce modèle est disponible en annexes section B.1.

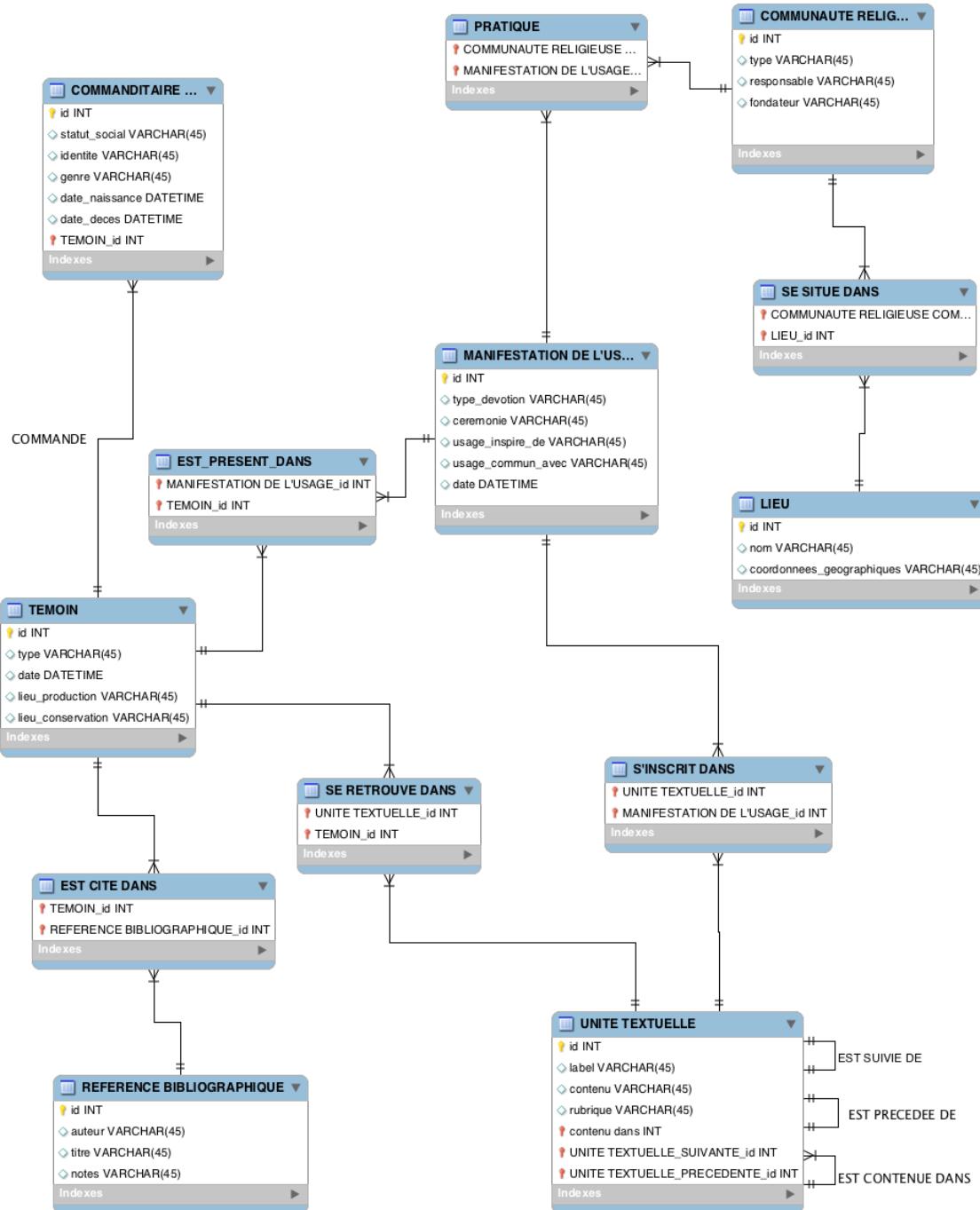


FIGURE 2.1 – Modèle relationnel logique des usages dans les livres d'heures

## 2.1. MODÉLISER LES USAGES : RESTITUER UNE RÉALITÉ COMPLEXE

---

En fait, il s'agit de prendre les deux occurrences les plus grandes des couples de cardinalités. Si les deux correspondent à « n » occurrences, alors une table de relation est créée, prenant comme attribut les identifiants des entités qu'elle relie, identifiants qui deviennent alors des clés étrangères<sup>170</sup>. On peut donc lire le modèle de la manière suivante : un usage est présent dans un ou plusieurs témoins. Une communauté religieuse pratique un ou plusieurs usages. Une unité textuelle s'inscrit dans un ou plusieurs usages. Une communauté religieuse se situe dans un ou plusieurs lieux. Une personne commande un à plusieurs manuscrit(s). Un témoin est cité dans une ou plusieurs référence(s) bibliographique(s). Une unité textuelle se retrouve dans un à plusieurs témoins. Une unité textuelle est suivie, précédée et contenue dans une autre unité textuelle, sachant que ces trois relations de succession et d'imbrication sont manifestées par des clés étrangères. Toutefois, ce modèle présente quelques défauts. Par exemple, les attributs de lieux de production et de conservation au sein de l'entité « Témoin » peuvent être modifiés en des associations vers l'entité « Lieu ».

Quatre tests ont ensuite été effectués quant à l'implémentation des unités textuelles dans la base Heurist. Heurist, base de données opensource MySQL, a justement été pensée pour les données de recherche en sciences humaines et sociales, qui se caractérisent par leur richesse, leur hétérogénéité, leur imperfection et leur forte interconnexion. Ainsi, tout système d'information pour un projet en humanités numériques doit permettre des requêtes complexes, un rendu de résultats interactifs et éditables, un export de données en divers formats, une certaine durabilité, et éventuellement un modèle pour le travail en groupe et parfois la publication sélective de données sur le web<sup>171</sup>. La base est déployée sur les serveurs de la TGIR Humanum, infrastructure de recherche pour les humanités numériques à l'échelle nationale et internationale<sup>172</sup>.

Les quatre tests présentés ci-dessous visent donc à visualiser la solution la plus pertinente pour rendre compte de ce qui fait toute la complexité et la richesse des livres d'heures : le double processus de succession pour les sections et les pièces, ainsi que d'imbrication des pièces liturgiques dans les sections, l'ensemble correspondant à un usage bien précis.

Dans le premier test, le champ « UseItem » revient à une liste sur le séquençage des pièces. Plus on clique sur le lien « UseItem », plus l'arborescence du livre se déploie, jusqu'à arriver au niveau de granularité le plus fin. On peut donc rajouter autant de

---

170. Dans une base de données, une clé étrangère permet de référencer le champ d'une autre table en le désignant par son identifiant.

171. HEURIST : UNE BASE DE DONNÉES GÉNÉRIQUE POUR LES SCIENCES HUMAINES ET SOCIALES, URL : <https://paris-timemachine.huma-num.fr/heurist-une-base-de-donnees-generique-pour-les-sciences-humaines-et-sociales/> (visité le 10/05/2020).

172. Cf. <https://www.huma-num.fr/>.

« UseItem » que l'on souhaite. S'il est spécifique à la relation créée, les œuvres restent génériques.

FIGURE 2.2 – Test d'implémentation n°1

Toutefois, cela n'est pas optimal du point de vue de l'ergonomie. La solution est en effet pratique pour enregistrer la séquence, mais si l'on supprime un « UseItem », les autres pièces qu'il contient sont supprimées.

La solution du deuxième test permet de visualiser, par l'intermédiaire de relations, la succession des sections au sein de l'ouvrage, jusqu'à avoir accès au texte pour la plus fine granularité, soit la dernière relation présentée ci-dessous.

FIGURE 2.3 – Test d'implémentation n°2

Cette idée est visuellement intéressante mais pas suffisamment pertinente du point de vue des imbrications dans les noms donnés aux relations, car plus la granularité est

## 2.1. MODÉLISER LES USAGES : RESTITUER UNE RÉALITÉ COMPLEXE

fine, plus les noms des relations s'allongent. Pour des questions d'intéropérabilité, il est en effet souhaitable d'utiliser le vocabulaire implémenté dans Heurist pour typer les relations. Dans le cas des pièces liturgiques, il est préférable d'utiliser celui de la « séquence » dans la section « temporal », soit les termes suivants :

- EndsBefore ;
- ImmediatelyFollows ;
- ImmediatelyPrecedes ;
- StartsAfter.

En outre, la linéarité ne modélise pas la séquence. Ceci pose le problème de la modélisation appliquée à un logiciel. Si une information n'est pas modifiable dans le logiciel, c'est que le problème vient du modèle sous-jacent qui doit être compatible avec le logiciel utilisé.

La troisième implémentation rend davantage compte de la réalité du texte dans l'usage.

Record View Map-Timeline Map new Custom Reports Export Network Diagram Crosstabs

**Agde** ID:402868

Type 91: UseTest3

Organisation Agde

Office Parvum

RELATED

asSectionHoursOfTheVirgin-Matins-Hymn Quem terra pontus sidera colunt adorant praedicant trinam regentem machinam claustrum Mariae bajulat ; Cui luna sol et omnia deserunt per tempora perfusa caeli gratia gestant pueræ viscera ; Beata mater munere cuius supernus artifex mundum pugillo continens ventris sub arca clausus est ; Beata caeli nuntio fecunda sancto spiritu desideratus gentibus cuius per alvum fusus est ; Jesu tibi sit gloria qui natus es de virgine cum patre et almo spiritu in sempiterna saecula ; Amen - Hymn

asSectionHoursOfTheVirgin-Matins-Psalms Benedicta tu in mulieribus et benedictus fructus ventris tui - Antiphon

FIGURE 2.4 – Test d'implémentation n°3

Si la relation vers la section apparaît, ce n'est pas le cas de la hiérarchie des textes.

La solution la plus satisfaisante est la quatrième. Visuellement le texte apparaît comme dans la source. Si l'on clique dessus, on voit le texte comme une entité avec ses relations aux autres sections, à l'image du deuxième test.

Ce modèle est d'un point de vue intellectuel et applicatif supérieur, car l'on se rapproche du modèle RDF où chaque relation n'est déclarée qu'une seule fois, et donc la

FIGURE 2.5 – Test d’implémentation n°4

hiérarchie une seule fois également<sup>173</sup>. Cependant, il peut être plus pertinent de typer la relation par « ImmediatelyFollows », en partant de l’entité vers l’entité précédente, car dans notre lecture du livre de gauche à droite, on est sûr de ce qui précède mais pas de ce qui suit.

Toutefois, il faudrait ajouter le séquençage des sections en plus de l’ordre affiché au sein des œuvres. Le problème est en effet de représenter à la fois la succession au niveau des pièces et la succession au niveau des sections. Cette dernière idée s’implémente ainsi dans la description de l’usage :

```
Agde(use) :contains :Matins, Agde(use) :contains :Lauds, :contains(1) :pre-
cedes :contains(2)
```

Se pose également la question des usages pour les pièces non-renseignées, ou bien si la source utilisée ne contient pas l’information mais que cette dernière est peut-être présente dans d’autres témoins. Les pièces absentes ou inconnues révèlent en effet la nécessité de choisir un format qui est facile à générer si on met à jour le cursus et qu’on trouve une pièce non-renseignée.

Ainsi, chercher l’inspiration du côté de standards utilisés dans les notices bibliographiques ou pour le web sémantique aide à penser la modélisation et les possibles implémentations du modèle dans la base.

173. L’inspiration du modèle RDF pour cette implémentation est explicitée à la sous-partie 2.1.2.

## 2.1.2 S'inspirer de modèles standardisés

Différents modèles peuvent nous aider à modéliser le cursus. Une des premières sources d'inspiration à laquelle on peut se référer est la modélisation de la séquence par Georg Vogeler. Historien au centre de modélisation de l'information en Humanités à l'Université de Graz, Georg Vogeler mène des recherches en technologies numériques pour les sciences sociales, les arts, les humanités, la structuration des données et l'histoire culturelle. Sa proposition prend la forme suivante :

```
:S1 a :Sequence ; :hasStart :Hour1 ; :hasEnd :Hour2 ; :isDocumentedIn :BookOfHoursX
. :text a rdf :List ; :isDocumentedIn :Book1 ; rdf :first :fragment1 . :fragment1 :references
:Hour1 ; rdf :rest .tf1.t :ft1rdf :first :fragment2; :refferences :Hour2, rdf :restrdf :nil. :Book1 :contains :fragment1, :fragment2. :fragment2 :follows
:fragment1. :fragment1 :refferences :Hour1. :fragment2 :refferences :Hour2.
```

Ce modèle s'appuie sur un standard propre au web sémantique : le triplet RDF. Le web sémantique est en fait une extension du web courant qui vise à donner du sens au contenu des pages web, en majorité créées pour être lues, et permettre leur interprétation par des machines. Un standard comme RDF<sup>174</sup>, doté d'un vocabulaire spécifié par le W3C<sup>175</sup>, sert à attribuer un sens défini à chaque donnée, à créer un réseau d'informations structurées facilement réutilisables. Il s'agit alors d'éviter les redondances, les conversions lourdes, et de permettre la traçabilité des données<sup>176</sup>.

La modélisation de la séquence de Georg Vogeler s'inspire du modèle RDF car on retrouve bien la structure en triplet séparée par des points-virgules. L'information est ainsi formée :

- d'un sujet
- d'un prédicat
- d'un objet

Les sujets sont ici en début de ligne : « S1 » fait référence à un sujet lambda que l'on pourrait retrouver dans le cadre des livres d'heures, tout comme « text », « fragment1 » et « Book1 ». Les prédicats se trouvent en deuxième position : « a » correspond à « rdf:type » et révèle donc la nature du sujet ; les autres prédicats, aisément compréhensibles, indiquent des relations de succession, de hiérarchie, de contenu et de contenant (« hasStart » ; « hasEnd » ; « isDocumentedIn » ; « references » ; « contains » ; « follows »). L'appellation « rdf:first » indique le premier type d'occurrence dans un sujet de type liste, tandis que le prédicat « rdf:rest » signale le reste des autres éléments d'une liste<sup>177</sup>. Quant aux objets, ils sont ici propres à la modélisation de la séquence du cursus, formés d'une suite

---

174. *Ressource Description Framework*.

175. *World Wide Web Consortium*. Organisme de standardisation fondé en 1994, il a pour but de favoriser la compatibilité des différentes technologies du Web.

176. Cf. <https://www.enssib.fr/le-dictionnaire/web-semantique>.

177. Cf. [https://www.w3.org/TR/rdf-schema/#ch\\_first](https://www.w3.org/TR/rdf-schema/#ch_first).

de sections et de pièces liturgiques à l'intérieur d'un livre d'heures qui s'interpénètrent entre elles.

Si l'on adapte cette proposition à une base de données relationnelle, car les données du projet sont ici destinées à être structurées dans la base Heurist et non sur le web sémantique, les sujets et objets des triplets peuvent faire référence à des clés étrangères. On peut également typer les relations avec un vocabulaire spécifique pour préciser où l'on se trouve dans le cursus. C'est justement l'association spécifique entre un usage et une œuvre que l'on nomme cursus.

Il est intéressant pour notre réflexion de croiser ce type de modèle avec le standard FRBR<sup>178</sup>. Il s'agit d'une modélisation conceptuelle de l'information contenue dans les notices bibliographiques. Élaborée en 1997, cette modélisation est née de l'informatisation des données bibliographiques depuis les années 1970. Des programmes de catalogues partagés ont alors vu le jour. Outil de description, un catalogue sert à localiser physiquement des supports de contenus. L'unité d'information primordiale est donc le support. Ainsi, lorsque l'on cherche une œuvre, celle-ci n'apparaît pas comme une unité intellectuelle, mais sous la forme d'une liste des différentes versions de l'œuvre, chaque version ayant sa propre notice distincte dans le catalogue. La particularité du modèle FRBR est d'inverser cette approche : l'œuvre devient le concept central. Les œuvres sont ensuite déclinées en manifestations, qui peuvent recouvrir divers supports comme divers langages. Le but est de créer une notice unique par œuvre qui permet d'accéder à toutes les versions, linguistiques comme d'édition, et à tous les supports de l'œuvre et de ses adaptations, du livre au livre audio en passant par le DVD, par exemple<sup>179</sup>. En voici une schématisation<sup>180</sup> :

---

178. *Functional Requirements for Bibliographic Records*.

179. Cf. <https://www.enssib.fr/bibliotheque-numerique/documents/65520-comprendre-le-modele-frbr-et-ses-extensions.pdf>.

180. Le schéma est extrait du site suivante : <https://fr.slideshare.net/nonue12/crfcb-amu-evolutionscatalogage091213-frbr>.

## 2.1. MODÉLISER LES USAGES : RESTITUER UNE RÉALITÉ COMPLEXE

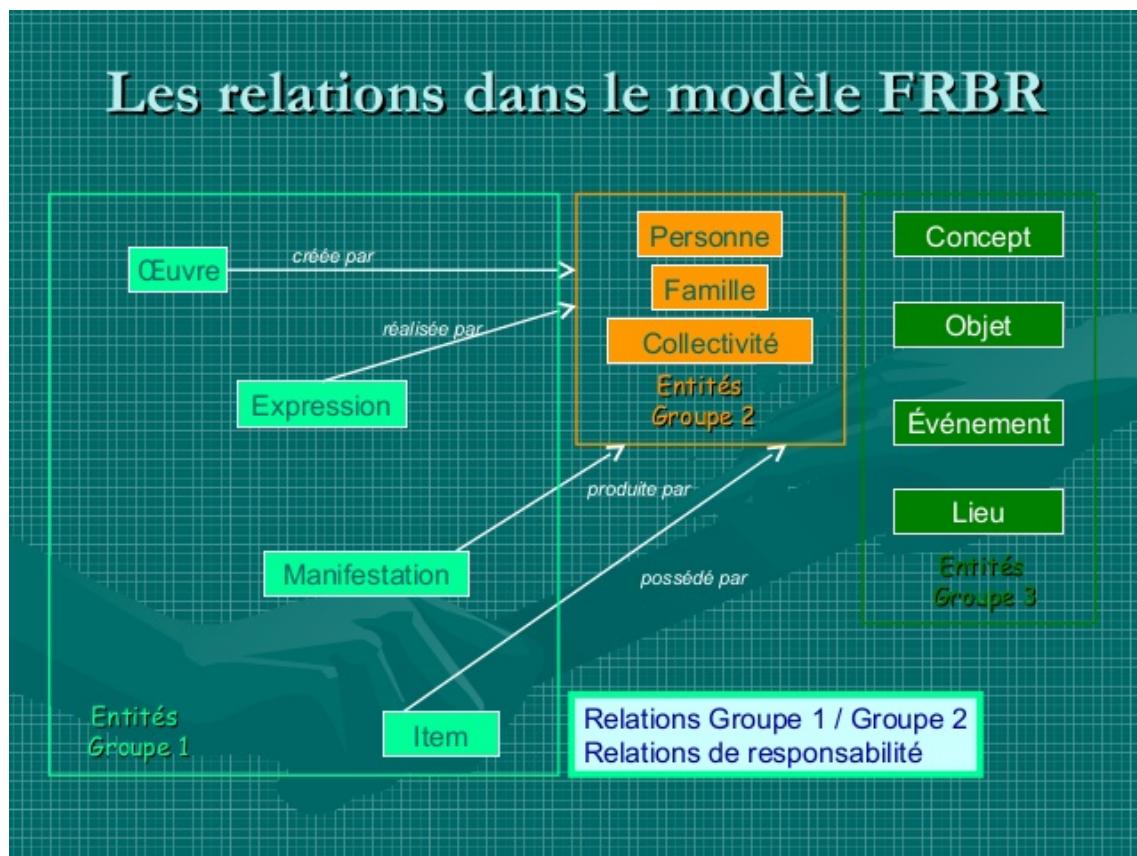


FIGURE 2.6 – Modélisation conceptuelle des usages dans les livres d’heures inspirée du modèle FRBR

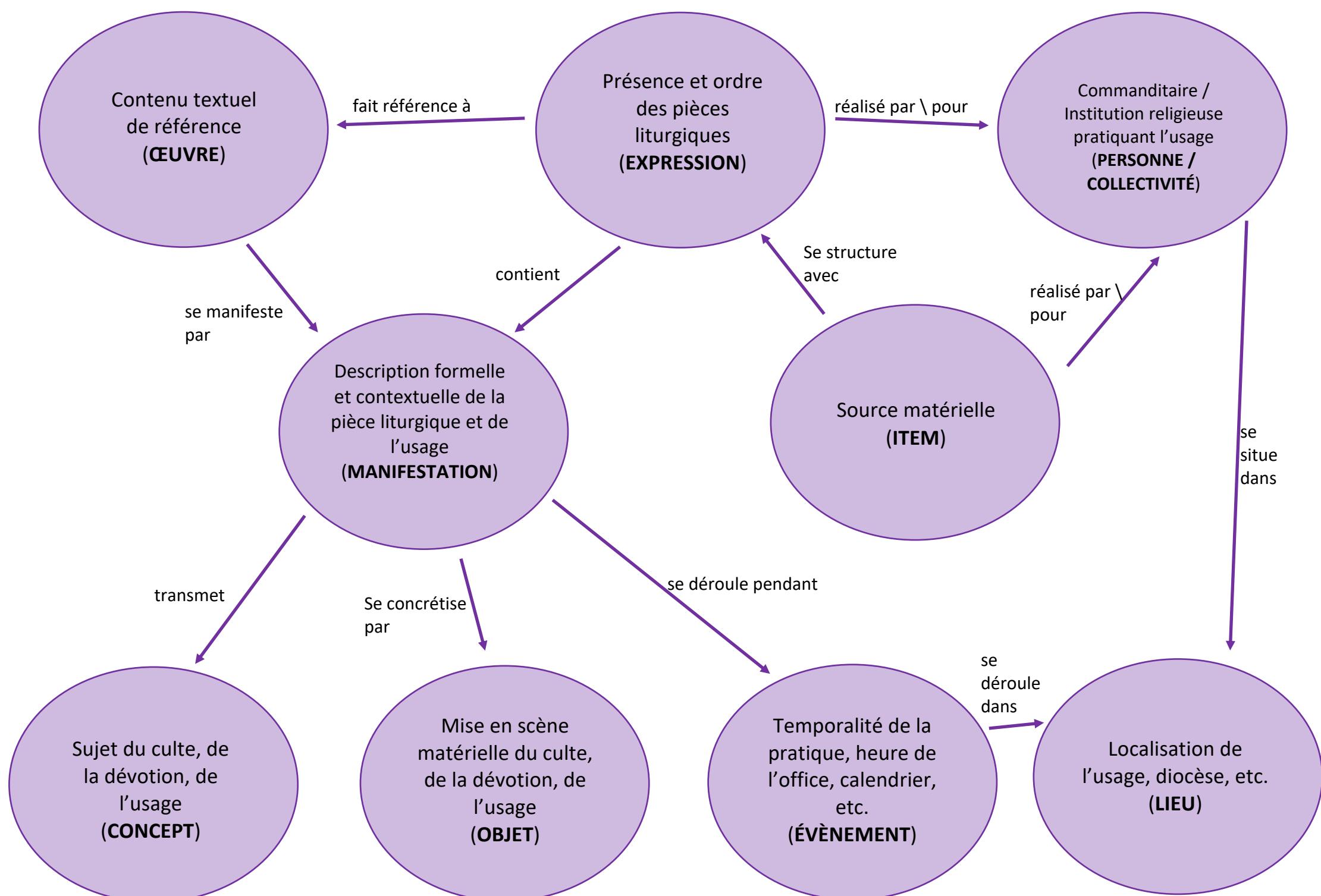
L’entité « Œuvre » se réfère aux caractéristiques de la création abstraite à laquelle se rattache son contenu. L’« Expression » indique les caractéristiques de son contenu intellectuel ou artistique, la « Manifestation » les caractéristiques de la publication à laquelle elle appartient, et enfin l’« Item » ses caractéristiques individuelles en tant qu’exemplaire. On va donc dans la caractérisation de l’œuvre du plus abstrait au plus concret. Ces niveaux d’analyse du document décrit peuvent être mis en relation avec des entités qui sont intervenues dans le processus de création ou de production, mais aussi avec des notions constituant les sujets inhérents à l’œuvre<sup>181</sup>.

Voici une nouvelle modélisation, inspirée à la fois des triplets RDF et du modèle FRBR, adaptée au cas du cursus et des usages résultant de l’analyse des livres d’heures<sup>182</sup> :

---

181. Cf. <https://www.bnf.fr/fr/modeles-frbr-frad-et-frsad>.

182. Une reproduction de ce modèle est disponible en annexes section B.1.



Le modèle tend à rendre compte que, par exemple, les matines de l'usage d'Agde sont une expression de l'œuvre que constitue les matines, et donc que l'expression d'une même œuvre diffère selon les usages. S'inspirer de standards de modélisation propres au monde du web sémantique ou de la bibliothéconomie aide à enrichir la réflexion, mais il est évident que le modèle exposé ci-dessus ne peut être implanté tel quel dans la base de données relationnelles. On peut ainsi reprocher à ce modèle d'être restrictif dans les types de relations, qui ne sont pas nécessairement à sens unique entre les entités, mais soulève également quelques problèmes : l'« Œuvre » est-elle réellement détachable du « Concept » ? L'« Item » et l'« Objet » peuvent-ils être réunis sous une même entité ?

Dans l'optique de l'import des données, c'est la solution suivante qui a été conservée :

FIGURE 2.7 – Modèle d'implémentation de l'entité Use

## 2.1. MODÉLISER LES USAGES : RESTITUER UNE RÉALITÉ COMPLEXE

The screenshot shows the HEURIST v5.22 application interface. The top navigation bar includes Database, Verify, Import, Website, Management, Admin, Help, and a 'hosted by: Huma-Num' logo. The main area displays a detailed view of an 'useitem' entity. The entity type is 'Type 93: useitem'. The 'Work' field contains a Latin text entry: 'Quem terra contus sidera colunt adorant praedicant trinam regentem machinam claustrum Mariae bajulat . Cui luna sol et omnia deserunt per tempora perfusa caeli gratia gestant puerum sub arcu . Beata mater pueri cuius superna arca pugnare ventris sub arcu clausus est . Beata matrem nunc fecunda sancto spiritu desideratus genitibus ciliis per alvum fatus est . Jesu tibi sit gloria qui natus es de virginie cum caele et almo spiritu in semperita saecula ; Amen - Hymn'. Below this, the 'RELATED' section lists 'IsContainedWithin' relationships: 'Hours of the Virgin', 'Matins', and 'Hymn'. It also shows an 'ImmediatelyPrecedes' relationship: 'Benedicta tu in mulieribus et benedictus fructus ventris tui Antiphon AGDE'. The 'LINKED FROM' section includes a 'Referenced by' link to 'Show list below as search results'. At the bottom, there are download links for XML and HTML, and a note about right-clicking to copy URLs. The status bar at the bottom right indicates the record was updated on 2020-04-22 at 14:17:26 (16:17:26 local) and is owned by 'Everyone'.

FIGURE 2.8 – Modèle d’implémentation de l’entité UseItem

FIGURE 2.9 – Modèle d’implémentation de l’entité Work

L’usage (l’entité « Use ») a pour attributs :

- l’organisation à laquelle il est relié<sup>183</sup>
- les références bibliographiques dans lesquelles il est mentionné ;
- son type, « parvum » ou « magnum »<sup>184</sup> ;
- le cursus, qui est constitué d’une succession de pièces liturgiques, soit l’entité « useItem » dans la base.

L’entité « useItem » est quant à elle dotée de l’attribut « work », qui renvoie au contenu textuel de la pièce, et d’un ensemble de relation de succession (« Immediately-Follows » ou « ImmediatelyPrecedes ») et de hiérarchie (« Contains » ou « IsContained-Within »). Si l’on clique sur l’attribut « work » d’un « useItem », s’ouvre la page relative à l’entité « Work » dans laquelle on trouve le niveau de granularité du texte, de « Level 1 » pour les sections les plus englobantes au « Level 4 » pour le niveau le plus fin. On peut lire ensuite le texte lui-même, ou les sections de texte contenues dans une section englobante, ainsi qu’une éventuelle « Note », la référence bibliographique dans laquelle le texte transmis est transcrit (« bibl »), et les témoins dans lesquels on retrouve le texte (« Witness »).

Comment implémenter les données selon ce modèle dans la base heurist ?

183. À noter qu’un usage peut être relié à une ou plusieurs organisations.

184. On peut aussi créer un « subtype » pour déterminer ce qui est de l’ordre du « Parvum » ou du « Magnum » dans l’usage concerné. En effet, le type d’office ne concerne pas tout le cursus mais une ou des office(s) en particulier.

## 2.2 L’impossible import des données structurées en xml : bilan d’une tentative avortée et solutions alternatives

Les données sont stockées dans des tableurs, aux formats excel ou csv<sup>185</sup>. Les colonnes divisent les informations relatives au cursus, avec le nom de l’usage liturgique, les différents niveaux hiérarchiques et œuvres associées, ainsi que le texte qui le compose. La base Heurist offre plusieurs possibilités d’import, parmi lesquels l’usage de CSV, d’une bibliographie Zotero, d’un document au format KML pour les données géospatiales<sup>186</sup>, ou encore aux formats XML ou JSON. Pour l’import du cursus, l’enjeu était de savoir quel import était le plus rapide et le plus pertinent : passer par des données stockées dans un document CSV ou XML ?

### 2.2.1 Définir la structure du format cible en XML

Avant toute structuration des données pour l’import, il faut savoir quelle forme doit prendre le document d’import, mais aussi s’assurer que les données sont homogènes. Ainsi, quand les données sont divisées en plusieurs csv, il faut pouvoir faire des jointures à partir de données cohérentes. Par exemple, les informations relatives au champs « Work »<sup>187</sup>, qui correspond au texte d’une pièce liturgique, étaient dissociées des informations relatives au cursus<sup>188</sup>. Via le logiciel Dataiku, on peut faire cette jointure grâce à une condition d’égalité entre les colonnes « function » dans le document Cursus\_HORAE\_Listes\_import\_heurist.xlsx et « liturgical function » dans le document Export\_stutzmann\_horae\_t65\_Work.csv, ainsi qu’une condition de jointure si les colonnes « Text » ont les mêmes premiers mots. On obtient alors le tableau avec les colonnes suivantes pour tester la solution d’import des données la plus efficiente<sup>189</sup> :

- la colonne « Numérotation », soit le numéro de ligne associée à chaque useItem afin de se retrouver dans les données<sup>190</sup> ;
- « Use », qui correspond à l’usage en vigueur et qui prend le nom de l’organisation à laquelle il est associé ;
- « Office », qui prend la valeur de « Parvum » ou « Magnum » ;

---

185. Signifiant *Comma-Separated Values*, le CSV est un format texte servant à représenter des informations sous forme tabulaire : les sauts de lignes correspondent à des changements de lignes et les virgules (ou points-virgules selon les cas), indiquent la séparation entre les colonnes.

186. *Keyhole Markup Language*. C’est un langage formé de balises destiné à la gestion de l’affichage de données géospatiales dans les logiciels de Système d’Information Géographique (SIG).

187. Document Export\_stutzmann\_horae\_t65\_Work.csv.

188. Document Cursus\_HORAE\_Listes\_import\_heurist.xlsx.

189. Il convient de préciser que chaque ligne du tableau correspond à une pièce liturgique précise au sein d’un usage, soit l’entité appelée « useItem » au sein de la base.

190. Le tableau servant aux tests d’import, UseItem\_Test\_LL.xlsx, contient en effet 4620 entrées.

- Level 1 à 4 qui indiquent les relations de séquence et d'imbrication entre les pièces (« IsContainedWithin ») ;
- title level 1 à 4, indiquant les titres des pièces dont il est question, des Heures de la Vierge aux psaumes ou hymnes en passant par les Matines ou Complies ;
- « function », correspondant à la fonction liturgique, de l'invitatoire au versicule en passant par l'hymne ou le répons ;
- « Text » pour le texte associé à la pièce ;
- H-ID Work 1 à 4, soit les identifiants générés par Heurist pour les entités existantes dans la base ;
- ID HORAE 1 à 4, les identifiants choisis par l'administrateur de la base, afin que les relations entre les entités soient faites au moment de l'import même entre les entités qui n'ont pas encore de *record identifier* générés par la base Heurist ;
- Work title 1 à Work title 4, soit le contenu textuel des œuvres présentes dans l'usage.

À terme, les données doivent également être liées à des ressources extérieures, comme les cotes des sources ou les références bibliographiques dans lesquelles elles sont citées et décrites, notamment dans les catalogues de notices<sup>191</sup>.

Pour tester un import en XML, il a fallu comprendre comment les données à disposition, dans un format csv, allaient être encodées automatiquement dans un format XML, et quelle forme le document d'import allait prendre. Dans un premier temps, nous avons élaboré le document d'import à partir d'un export de données déjà présentes dans la base et concernant le cursus. Le document d'export en XML prend la forme suivante :

```
<?xml version='1.0' encoding='UTF-8'?>
<html xmlns="http://heuristnetwork.org" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://heuristnetwork.org/reference/scheme_hml.xsd">
<database id="0">stutzmann_horae</database>
<query q="t:92" db="stutzmann_horae" depth="all"/>
<dateStamp>2020-04-24T15:24:11+02:00</dateStamp>
<resultCount>1</resultCount>
<records>
<record visibility="viewable" visnote="logged in users"
selected="no" depth="0">
<id>402874</id>
<type id="92" conceptID="0000-92">Use</type>
```

---

191. Ces dernières informations apparaissent dans les colonnes « Sources manuscrites et imprimées », « Bibliographie », « Remarques » et « Reproduction » du tableau de données « Cursus\_HORAE\_Listes.xlsx ».

## 2.2. L'IMPOSSIBLE IMPORT DES DONNÉES STRUCTURÉES EN XML : BILAN D'UNE TENTATIVE AVORTÉE ET SOLUTIONS ALTERNATIVES

---

```
<citeAs>https://heurist.huma-num.fr:443/heurist/
?recID=402874&amp;db=stutzmann_horae</citeAs>
<title>Agde (Use) [2]</title>
<added>2020-04-21 16:34:47</added>
<modified>2020-04-24 13:05:03</modified>
<workgroup id="0">public</workgroup>
<raw>2</raw>
<year>2</year>
<detail conceptID="2-21" name="Organisation"
id="21" basename="Organisation" isRecordPointer=
"true">402689</detail>
<detail conceptID="0000-1011" name="bibl"
id="1011" basename="bibl" isRecordPointer=
"true">402893</detail>
<detail conceptID="0000-1187" name="Office
" id="1187" basename="Office">Parvum</detail>
<detail conceptID="0000-1191" name="Cursus"
id="1191" basename="Cursus" isRecordPointer=
"true">402909</detail>
<detail conceptID="0000-1191" name="Cursus"
id="1191" basename="Cursus" isRecordPointer=
"true">402910</detail>
<detail conceptID="0000-1191" name="Cursus"
id="1191" basename="Cursus" isRecordPointer=
"true">402911</detail>
<detail conceptID="0000-1191" name="Cursus"
id="1191" basename="Cursus" isRecordPointer=
"true">402875</detail>
<detail conceptID="0000-1191" name="Cursus"
id="1191" basename="Cursus" isRecordPointer=
"true">402914</detail>
<detail conceptID="0000-1191" name="Cursus"
id="1191" basename="Cursus" isRecordPointer=
"true">402881</detail>
<detail conceptID="0000-1191" name="Cursus"
id="1191" basename="Cursus" isRecordPointer=
"true">402887</detail>
</record>
<record visibility="viewable" visnote="logged
```

```

in users" selected="no" depth="1">
<id>402689</id>
<type id="4" conceptID="2-4">Organisation</type>
<citeAs>https://heurist.huma-num.fr:443/heurist/
?recID=402689&db=stutzmann_horae</citeAs>
<title>Agde (Organisation)</title>
<added>2020-04-20 18:15:20</added>
<modified>2020-04-20 18:17:59</modified>
<workgroup id="0">public</workgroup>
<detail conceptID="2-2" name="Short name /
acronym" id="2" basename="Short name">Agde</detail>
<detail conceptID="2-134" name="Location
(places)" id="134" basename="Location (place)">
isRecordPointer="true">386187</detail>
<reversePointer id="21" conceptID="2-21"
basename="Organisation" name="Organisation">
402874</reversePointer>
</record>
[...]
<record visibility="viewable" visnote="logged
in users" selected="no" depth="1">
<id>402909</id>
<type id="93" conceptID="0000-93">useItem</type>
<citeAs>https://heurist.huma-num.fr:443/heurist/
?recID=402909&db=stutzmann_horae</citeAs>
<title>Hours of the Virgin</title>
<added>2020-04-24 12:38:53</added>
<modified>2020-04-24 12:39:29</modified>
<workgroup id="0">public</workgroup>
<detail conceptID="0000-1134" name="Work" id="1134"
basename="title (work)" isRecordPointer=
"true">75966</detail>
<reversePointer id="1191" conceptID="0000-1191"
basename="Cursus" name="Cursus">402874</reversePointer>
</record>
```

À partir de ce document d'export, où l'on peut voir un enregistrement relatif à l'usage d'Agde, un deuxième à propos de l'organisation d'Agde et un troisième concernant la section « Hours of the Virgin », il est nécessaire de distinguer ce qui est généré par la base de ce qui est à importer et à structurer dans le document d'import en XML. Il est

## 2.2. L'IMPOSSIBLE IMPORT DES DONNÉES STRUCTURÉES EN XML : BILAN D'UNE TENTATIVE AVORTÉE ET SOLUTIONS ALTERNATIVES

---

donc important de comprendre à quoi correspond chaque balise et chaque attribut. Si les trois premiers attributs de la balise <record>, qui correspond à un enregistrement, sont propres à l'interface d'Heurist (visibility, visnote, selected), l'attribut *depth* indique le niveau depuis lequel est chargée la donnée, donc il augmente de un dans la relation depuis le document sélectionné, selon le schéma suivant :

Agde [usage] (niveau 0, id 402874), Agde [organisation] (niveau 1, id 402689),  
Agde [place] (niveau 2, id 386187) Agde [usage] (niveau 0, id 402874), Be-  
nicta tu [useItem] (niveau 1, id 402881), Benedicta tu [work] (niveau 2, id  
395935), HORAE [bibl] (niveau 3, id 393553), O quam pulchra (niveau 4, id  
399263 = une antienne qui a aussi pour [bibl] = HORAE)

La balise <id> correspond à l'identifiant que l'on veut ici générer automatiquement. La balise <type> renvoie ensuite au type d'entité, dont les attributs *id* et *conceptId* sont générés par la base. La balise <citeAs> indique l'URL<sup>192</sup> vers la données concernée. La balise <title> renvoie au titre donné à l'instance de l'entité, tandis que les balises <added> et <modified> sont des informations d'ajout et de modification des données clairement générées par Heurist et que nous n'avons pas à générer nous-même ; <work-group> semble également être généré par Heurist et renvoyer à des droits relatifs à la base. Quant aux balises <detail> et <reversePointer>, elles sont intéressantes pour l'import car permettent de fixer les relations entre les entités. Ces deux dernières balises pointent en effet vers des entités référencées ailleurs dans la base. Ainsi, la balise <detail> permet de faire des liens entre deux entités, tandis que la relation de séquence et de hiérarchie se fait grâce aux balises <detail> des entités de type « Record relationship ». Il faut donc générer les clés étrangères au moment des relations. Pour les relations de type « many to many » (celles dont le couple de cardinalités est constitué de n,n dans le modèle relationnel logique), elles apparaissent dans les balises <detail> avec l'attribut IsRecordPointer. Pour une relation « one to many » de lien direct (couple de cardinalités 1,n dans le modèle relationnel logique), il faut encoder la balise <reversePointer>. Il faut donc être attentif aux identifiants pour créer des relations vers les entités souhaitées, notamment les UseItem au sein du cursus.

Dans le document d'export en XML, il nous faut distinguer trois sources différentes dans la structuration de l'information :

1. Les données générées par la base Heurist elle-même. S'il est possible de récupérer ces données à partir du document d'export XML, avec des librairies python comme *untangle* pour parser le document, ce qui s'avère inutile car ce sont des informations non nécessaires lors de l'import.

---

192. *Uniform Resource Locator*. Une URL est une adresse vers la localisation d'une ressource du web. Elle est notamment constituée du protocole, du nom de domaine et du chemin.

2. Les données générées automatiquement au moment de d'import, soit les identifiants donnés aux entités, ce qui correspond dans le document d'export XML au texte entre les balises <id> et <detail>.
3. Les données générées depuis le(s) fichier(s) csv qui contiennent les données renseignées et homogènes relatives au projet de recherche.

Il est donc nécessaire de comprendre quels sont les identifiants à générer automatiquement et quels sont ceux inhérents à la base de données Heurist. Les relations vers des entités préexistantes dans Heurist, celles qui ont déjà été importées, notamment de type « Work » et « Organisation », font donc référence aux identifiants déjà implémentés dans la base et n'ont pas à être générés automatiquement au moment de l'import. De plus, chaque « useItem », que ce soit dans une relation de hiérarchie ou de succession, est spécifique à un usage, donc il est important de tous les différencier selon l'usage auquel ils sont liés.

Ce que l'on souhaite importer pour tester l'import, c'est une ou plusieurs instances de type « Use », soit l'usage. L'enregistrement est donc, en général, lié à une entité existante, l'Organisation dans notre cas. L'entité de type « Use » est elle-même reliée à une liste d'enregistrements de type « useItem ». Les relations entre les « useItem » sont implémentées via une table de relation « Record relationship ». Le premier document cible pour l'import ressemble donc à ceci :

```
<?xml version="1.0" encoding="UTF-8"?>
<hml xmlns="http://heuristnetwork.org" xmlns:xsi="http://www.w3.org/2001
/XMLSchema-instance" xsi:schemaLocation="http://heuristnetwork.org/
reference/scheme_hml.xsd">
  <database>stutzmann_horae</database>
  <records>
    <!-- 3 essais de type "Use" : Agen, Aix, Amiens-->
    <record>
      <id>600000</id>
      <type id="92">Use</type>
      <title>Agen</title>
      <detail conceptID="2-21" name="Organisation" id="21"
        basename="Organisation" isRecordPointer="true">402892</detail>
    </record>
    <record>
      <id>600001</id>
      <type id="92">Use</type>
      <title>Aix-en-Provence</title>
      <detail conceptID="2-21" name="Organisation" id="21"
        basename="Organisation" isRecordPointer="true">402893</detail>
    </record>
    <record>
      <id>600002</id>
      <type id="92">Use</type>
      <title>Amiens</title>
      <detail conceptID="2-21" name="Organisation" id="21"
        basename="Organisation" isRecordPointer="true">402894</detail>
    </record>
  </records>
</hml>
```

## 2.2. L'IMPOSSIBLE IMPORT DES DONNÉES STRUCTURÉES EN XML : BILAN D'UNE TENTATIVE AVORTÉE ET SOLUTIONS ALTERNATIVES

---

```
        basename="Organisation" isRecordPointer="true">402960</detail>
    </record>
    <record>
        <id>600002</id>
        <type id="92">Use</type>
        <title>Amiens</title>
        <detail conceptID="2-21" name="Organisation" id="21"
        basename="Organisation" isRecordPointer="true">10004</detail>
    </record>
    <!-- 6 essais de type useItem pour pouvoir faire des relations de
succession et de hiérarchie -->
    <record>
        <id>700000</id>
        <type id="93">useItem</type>
        <title>Quem terra pontus sidera colunt adorant praedicant
trinam regentem machinam claustrum Mariae bajulat ;
Cui luna sol et omnia deserviunt per tempora perfusa caeli
gratia gestant puellae viscera ; Beata mater munere cuius
supernus artifex mundum pugillo continens ventris sub arca
clausus est ; Beata caeli nuntio fecunda sancto spiritu desideratus
gentibus cuius per alvum fusus est ; Jesu tibi sit gloria qui
natus es de virgine cum patre et almo spiritu in sempiterna saecula ;
Amen (Hymn)</title>
        <detail conceptID="0000-9999" name="Use" id="9999" basename="Use"
isRecordPointer="true">600000</detail>
        <detail conceptID="0000-1134" name="Work" id="1134"
basename="title (work)"
isRecordPointer="true">395039</detail>
    </record>
    <record>
        <id>700001</id>
        <type id="93">useItem</type>
        <title>Domine, Dominus noster, quam admirabile est nomen tuum in
universa terra ! quoniam elevata est magnificientia tua super
caelos. Ex ore infantium et lactentium perfecisti laudem propter
inimicos tuos, ut destruas inimicum et ultorem. Quoniam video caelos
tuos, opera digitorum tuorum, lunam et stellas quae tu fundasti. Quid
est homo, quod memor es ejus ? aut filius hominis, quoniam visitas eum ?
Minuisti eum paulominus ab angelis ; gloria et honore coronasti eum ; et
```

```

    constituisti eum super opera manuum tuarum. Omnia subjecisti sub pedibus ejus, oves et boves universas, insuper et pecora campi, volucres caeli, et pisces maris qui perambulant semitas maris. Domine, Dominus noster, quam admirabile est nomen tuum in universa terra ! (Psalm)</title>
<detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
    <record>
        <id>600001</id>
        <type id="93">useItem</type>
        <title>O admirabile commercium creator generis humani animatum corpus sumens de virgine nasci dignatus est et procedens homo sine semine largitus est nobis suam deitatem (Antiphon)</title>
        <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
            <record>
                <id>398123</id>
                <type id="93">useItem</type>
                <title>0 admirabile commercium creator generis humani animatum corpus sumens de virgine nasci dignatus est et procedens homo sine semine largitus est nobis suam deitatem (Antiphon)</title>
                <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                    <record>
                        <id>600002</id>
                        <type id="93">useItem</type>
                        <title>O admirabile commercium creator generis humani animatum corpus sumens de virgine nasci dignatus est et procedens homo sine semine largitus est nobis suam deitatem (Antiphon)</title>
                        <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                            <record>
                                <id>399221</id>
                                <type id="93">useItem</type>
                                <title>O admirabile commercium creator generis humani animatum corpus sumens de virgine nasci dignatus est et procedens homo sine semine largitus est nobis suam deitatem (Antiphon)</title>
                                <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                                    <record>
                                        <id>600003</id>
                                        <type id="93">useItem</type>
                                        <title>Matins</title>
                                        <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                                            <record>
                                                <id>394820</id>
                                                <type id="93">useItem</type>
                                                <title>Matins</title>
                                                <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                                                    <record>
                                                        <id>600000</id>
                                                        <type id="93">useItem</type>
                                                        <title>Matins</title>
                                                        <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                                                            <record>
                                                                <id>399221</id>
                                                                <type id="93">useItem</type>
                                                                <title>O admirabile commercium creator generis humani animatum corpus sumens de virgine nasci dignatus est et procedens homo sine semine largitus est nobis suam deitatem (Antiphon)</title>
                                                                <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                                                                    <record>
                                                                        <id>600001</id>
                                                                        <type id="93">useItem</type>
                                                                        <title>O admirabile commercium creator generis humani animatum corpus sumens de virgine nasci dignatus est et procedens homo sine semine largitus est nobis suam deitatem (Antiphon)</title>
                                                                        <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                                                                            <record>
                                                                                <id>398123</id>
                                                                                <type id="93">useItem</type>
                                                                                <title>O admirabile commercium creator generis humani animatum corpus sumens de virgine nasci dignatus est et procedens homo sine semine largitus est nobis suam deitatem (Antiphon)</title>
                                                                                <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                                                                                    <record>
                                                                                        <id>600002</id>
                                                                                        <type id="93">useItem</type>
                                                                                        <title>O admirabile commercium creator generis humani animatum corpus sumens de virgine nasci dignatus est et procedens homo sine semine largitus est nobis suam deitatem (Antiphon)</title>
                                                                                        <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                                                                                            <record>
                                                                                                <id>399221</id>
                                                                                                <type id="93">useItem</type>
                                                                                                <title>O admirabile commercium creator generis humani animatum corpus sumens de virgine nasci dignatus est et procedens homo sine semine largitus est nobis suam deitatem (Antiphon)</title>
                                                                                                <detail conceptID="0000-9999" name="Use" id="9999" basename="Use">
                                                                

```

## 2.2. L'IMPOSSIBLE IMPORT DES DONNÉES STRUCTURÉES EN XML : BILAN D'UNE TENTATIVE AVORTÉE ET SOLUTIONS ALTERNATIVES

---

```
<detail conceptID="0000-1134" name="Work" id="1134"
basename="title (work)"
isRecordPointer="true">395935</detail>
</record>
<record>
    <id>700005</id>
    <type id="93">useItem</type>
    <title>Hours of the Virgin</title>
    <detail conceptID="0000-9999" name="Use" id="9999" basename="Use"
isRecordPointer="true">600002</detail>
    <detail conceptID="0000-1134" name="Work" id="1134"
basename="title (work)"
isRecordPointer="true">394805</detail>
</record>
<!-- 3 essais de type relations/clés étrangères --&gt;
&lt;record&gt;
    &lt;!-- Matins : Contains : Quem terra pontus... --&gt;
    &lt;id&gt;800001&lt;/id&gt;
    &lt;type id="1"&gt;Record relationship&lt;/type&gt;
    &lt;detail conceptID="2-7" name="Source record" id="7"
basename="Source record pointer"
isRecordPointer="true"&gt;700003&lt;/detail&gt;
    &lt;detail conceptID="2-6" name="Relationship type" id="6"
basename="Relationship type"
termID="3266" termConceptID="2-3266" ParentTerm="Sequence"&gt;
Contains&lt;/detail&gt;
    &lt;detail conceptID="2-5" name="Target record" id="5"
basename="Target record pointer"
isRecordPointer="true"&gt;700000&lt;/detail&gt;
&lt;/record&gt;
&lt;record&gt;
    &lt;!-- Domine, Dominus noster, ... : ImmediatelyFollows :
Benedicta tu... --&gt;
    &lt;id&gt;800002&lt;/id&gt;
    &lt;type id="1"&gt;Record relationship&lt;/type&gt;
    &lt;detail conceptID="2-7" name="Source record" id="7"
basename="Source record pointer"
isRecordPointer="true"&gt;700001&lt;/detail&gt;
    &lt;detail conceptID="2-6" name="Relationship type" id="6"</pre>
```

```

basename="Relationship type"
termID="3266" termConceptID="2-3266" ParentTerm="Sequence">
ImmediatelyFollows</detail>
<detail conceptID="2-5" name="Target record" id="5"
basename="Target record pointer"
isRecordPointer="true">700004</detail>
</record>
<record>
    <!-- Hours of the Virgin : Contains : Ø admirabile commercium... -->
    <id>800003</id>
    <type id="1">Record relationship</type>
    <detail conceptID="2-7" name="Source record" id="7"
basename="Source record pointer"
isRecordPointer="true">700005</detail>
    <detail conceptID="2-6" name="Relationship type" id="6"
basename="Relationship type"
termID="3266" termConceptID="2-3266" ParentTerm="Sequence">
Contains</detail>
    <detail conceptID="2-5" name="Target record" id="5"
basename="Target record pointer"
isRecordPointer="true">700002</detail>
</record>
</records>
</hml>
```

Les usages ici pris en exemples ont été choisis depuis le document UseItem\_Test\_LL.xlsx aux lignes suivantes :

- n°0167 pour Agen ;
- n°0334 pour Aix-en-Provence ;
- n°0553 pour Amiens.

Toutefois, ce premier test a engendré des erreurs au moment de l'import, car tous les champs et toutes les entités n'étaient pas reconnus par Heurist, comme en témoignent les exemples ci-dessous :

## 2.2. L'IMPOSSIBLE IMPORT DES DONNÉES STRUCTURÉES EN XML : BILAN D'UNE TENTATIVE AVORTÉE ET SOLUTIONS ALTERNATIVES

The screenshot shows the HEURIST web application interface. The top navigation bar includes links for Database, Verify, Import, Website, Management, Admin, FAIMS, and Help, along with a 'hosted by Huma-Num' logo. The main content area is titled 'Filtered Result' and displays a list of four items, all of which have 'No data in title fields for this record'. The 'Selected' column shows a checkmark for the first item. Below the list, there is a detailed view for the first item, ID 403024, which is a 'Type 92: Use' record. The view includes fields for 'Original ID' (0-600000), 'Cite as' (XML or HTML), 'Added' (2020-04-29 15:13:53), 'Updated' (2020-04-29 15:13:53), 'Ownership' (Database Managers), and 'Rating' (none). On the left sidebar, under 'Saved Filters', there is a 'My Filters (private)' section with a single entry: 'Recent changes' (All (date order)). Other sections include 'Database Managers (workgroup)', 'HORAE (workgroup)', and 'CORPUS' (Work Level 1). A 'Design schema' button is also visible.

FIGURE 2.10 – Test d’import en XML de l’entité Use

This screenshot shows the same HEURIST interface as Figure 2.10, but for the 'UseItem' entity. The 'Filtered Result' list shows 13 items, all of which have 'No data has been entered in the fields used to construct the title'. The 'Selected' column shows a checkmark for the first item. The detailed view for the first item, ID 403032, is identical to Figure 2.10, showing it is a 'Type 93: useitem' record with the same metadata. The left sidebar shows the same saved filters and entity structures as Figure 2.10.

FIGURE 2.11 – Test d’import en XML de l’entité UseItem

FIGURE 2.12 – Test d'import en XML de l'entité Work

La balise <title> pour nommer les instances ne semble pas avoir de conséquence dans l'implémentation.

Nous avons en fait appris par la suite l'existence d'un template hml servant à l'import des données en xml : Template\_stutzmann\_horae\_20200429152715.hml. Si le document d'export en xml a permis de cerner le fonctionnement de la base Heurist, ce qui est à créer au moment de l'import et ce qui est généré par la base, il est nécessaire de respecter strictement le template indiqué pour réussir l'import. Le titre donné aux entités doit être défini dans la balise suivante :

```
<detail conceptID="2-1" name="Title">TEXT</detail>
```

Le document cible a donc été re-pensé à partir du premier essai, dont voici un extrait<sup>193</sup> :

[...]

```

<record>
    <id>H-ID-600000</id>
    <type conceptId="0000-92">Use</type>
    <detail conceptID="2-1" name="Title">Agen
        (Use) [2]</detail>
    <detail conceptID="2-21" name="Organisation"
        isRecordPointer="true">402892</detail>
</record>
```

193. Vous trouverez le document cible dans son intégralité en annexes, section B.2.1.

## 2.2. L'IMPOSSIBLE IMPORT DES DONNÉES STRUCTURÉES EN XML : BILAN D'UNE TENTATIVE AVORTÉE ET SOLUTIONS ALTERNATIVES

---

[...]

```
<record>
    <id>H-ID-700000</id>
    <type conceptID="0000-93">useItem</type>
    <detail conceptID="0000-1134" name="Work"
        isRecordPointer="true">H-ID-395039</detail>
    <detail conceptID="0000-1188" name="Work (Cursus)">
        Quem terra pontus sidera colunt
            adorant praedicant trinam regentem machinam
            claustrum Mariae bajulat ; Cui luna sol
            et omnia deserviunt per tempora perfusa caeli
            gratia gestant puellae viscera ; Beata
            mater munere cuius supernus artifex mundum pugillo
            continens ventris sub arca
            clausus est ; Beata caeli nuntio fecunda sancto spiritu
            desideratus gentibus cuius
            per alvum fusus est ; Jesu tibi sit gloria qui natus
            es de virgine cum patre et almo
            spiritu in sempiterna saecula ; Amen [Hymn] (useItem)</detail>
    </record>
```

[...]

```
<record>
    <!-- Matins : Contains : Quem terra pontus... -->
    <id>H-ID-800001</id>
    <type conceptID="2-1">Record relationship</type>
    <detail conceptID="2-1" name="Title for relationship">
        Contains | Matins (useItem) &lt;-&gt; Quem
        terra pontus sidera colunt
            adorant praedicant trinam regentem machinam
            claustrum Mariae bajulat ; Cui luna sol
            et omnia deserviunt per tempora perfusa caeli
            gratia gestant puellae viscera ; Beata
            mater munere cuius supernus artifex mundum
            pugillo continens ventris sub arca
            clausus est ; Beata caeli nuntio fecunda
            sancto spiritu desideratus gentibus cuius
            per alvum fusus est ; Jesu tibi sit gloria
            qui natus es de virgine cum patre et almo
            spiritu in sempiterna saecula ; Amen [Hymn]
```

```

        (useItem)</detail>
        <detail conceptID="2-7" name="Source record"
        isRecordPointer="true">H-ID-700003</detail>
        <detail conceptID="2-6" name="Relationship type"
        termID="3262" termConceptID="2-3262"
        ParentTerm="Overlap">Contains</detail>
        <detail conceptID="2-5" name="Target record"
        isRecordPointer="true">H-ID-700000</detail>
    </record>
[...]

```

Une fois le document cible défini, il reste à savoir comment le créer à partir de données classées dans un csv et d'identifiants à générer automatiquement.

## 2.2.2 Structurer les données : du CSV à l'XML en passant par Python

Pour générer un document d'import au format XML, il est nécessaire d'installer la librairie python lxml<sup>194</sup>. Elle permet de créer les balises dont nous avons besoin pour l'import (<hml>, <database>, <records>, <record>, <id>, <type> et <detail>), ainsi que les attributs associés (*conceptID*, *name*, *isRecordPointer*, *termID*, *termConceptID*, *ParentTerm*)<sup>195</sup>. Chaque élément <record> doit correspondre à une ligne d'enregistrement dans le csv, de la manière suivante :

```

with open ("/Users/gwenaellepatat/Desktop/Stage_TNAH/MémoireHORAE
/BaseHeurist/Données/UseItem_Test_LL.csv") as csvfile:
    donnees_Heurist = csv.reader(csvfile, delimiter=';'
        , quotechar='''')
    #id_record = 599999
    for id_record, row in enumerate(donnees_Heurist,
        600000) :
        #id_record += 1
        record = etree.SubElement(records, "record")
        #Sous-éléments de la balise record
        id_node = etree.SubElement(record, "id")
        id_node.text = "H-ID-" + str(id_record)

```

L'identifiant a été choisi à partir du chiffre 600 000 pour ne pas créer d'interférences avec le

---

194. Cf <https://docs.python.org/3/library/xml.etree.elementtree.html>.

195. Les différentes versions du code, qui a évolué avec la compréhension des problèmes et de ce qui est attendu en format cible, sont disponibles en annexes section A.2.2.

## 2.2. L'IMPOSSIBLE IMPORT DES DONNÉES STRUCTURÉES EN XML : BILAN D'UNE TENTATIVE AVORTÉE ET SOLUTIONS ALTERNATIVES

---

test d'import des données en csv. La balise concernant l'enregistrement et son identifiant est donc générée par le code python et dépend du nombre d'entrées dans le csv.

Toutefois l'élément <type> est à créer à la fois depuis le csv et le document d'export XML de la base Heurist, car l'attribut conceptID se réfère à des identifiants propres à la base Heurist :

```
type_node = etree.SubElement(record, "type")
    #Valeurs d'attributs de la balise
    #type dépendant de la base Heurist
    type_node.set("conceptID", exportXML.
        hml.records.record[0].type["conceptID"])
    #Récupération du nom de l'entité dans
    #l'élément type depuis le CSV
    for line in csvfile :
        type_node.text = str(row[1])
```

Le plus délicat est la création des éléments <detail>. Il est certes aisé de récupérer les données adéquates à partir des csv, comme le montre l'exemple ci-dessous :

```
for id_record, row in enumerate(donnees_Heurist, 600000) :

    detail1 = etree.SubElement(record, "detail")
    if type_node.text == "Use" :
        detail1.set("conceptID", "2-1")
        detail1.set("name", "Title")
        detail1.text = str(row[1]) + " (Use) [2]"
```

Il est cependant plus contraignant de créer les relations via l'utilisation de clés étrangères, car il n'est pas possible de toutes les obtenir depuis le document d'export chargé. Cela dépend des données préalablement implémentées dans la base et de leur attribution d'un identifiant par Heurist pour pouvoir faire correctement les liens entre les entités, notamment les entités relevant de l'« Organisation » pour les usages et de « Work » pour les pièces liturgiques. Pour pallier ce problème, il faudrait d'abord générer des entités vides sans attributs et sans relations, pour ensuite récupérer l'identifiant d'Heurist à partir du document d'export XML et les appeler au bon endroit dans le code pour le nouveau document d'import, qui ferait cette fois-ci les jointures. Ces réflexions nous amènent à exposer plus en détail les raisons du choix d'importer les données en CSV.

### 2.2.3 Raisons du choix d'importer les données depuis le format CSV

Pour l'import des données en XML, le template défini pour la base Heurist indique un certain nombre de balises <detail> par entité. Elles peuvent être d'ordre textuel pour indiquer le contenu des données, ou bien elles servent à créer un lien vers une référence numérique pointant vers une autre entité. Toutes les balises <detail> indiquées dans le template ne sont pas obligatoirement remplies au moment de l'import, car nous disposons souvent de données imparfaites. Il est toutefois possible de vérifier leur intégrité sur Heurist.

Le template d'import en XML doit donc être pensé en amont de la modélisation des données, afin que tous les champs qui caractérisent la donnée apparaissent bien via une balise <detail>. En effet, une fois le template implémenté, il n'est plus possible de le modifier et il faut s'y conformer pour la bonne intégration des données dans la base. Ainsi, le template pour l'entité Use permet de faire un lien vers un UseItem :

```
<detail conceptID="2-9" name="Date">DATE</detail>
<detail conceptID="2-21" name="Organisation" isRecordPointer="true">
RECORD_REFERENCE</detail>
<detail conceptID="0000-1011" name="bibl" isRecordPointer="true">
RECORD_REFERENCE</detail>
<detail conceptID="0000-1012" name="Note">MEMO_TEXT</detail>
<detail conceptID="0000-1181" name="UseItem" isRecordPointer="true">
RECORD_REFERENCE</detail>
<detail conceptID="0000-1187" name="Office">TEXT</detail>
```

Cependant, deux types de useItem apparaissent dans le modèle. L'entité useItem avec les champs suivants :

```
<detail conceptID="0000-1012" name="Note">MEMO_TEXT</detail>
<detail conceptID="0000-1134" name="Work" isRecordPointer="true">
RECORD_REF</detail>
<detail conceptID="0000-1188" name="Work (Cursus)">TEXT</detail>
<detail conceptID="0000-1192" name="Sequence">TEXT</detail>
```

Puis on est confronté à l'entité useItem2 qui présente un champ différent, celui de la relation de succession des pièces :

```
<detail conceptID="0000-1012" name="Note">MEMO_TEXT</detail>
<detail conceptID="0000-1134" name="Work" isRecordPointer="true">
RECORD_REFERENCE</detail>
<detail conceptID="0000-1188" name="Work (Cursus)">TEXT</detail>
<detail conceptID="0000-1192" name="Follow">TEXT</detail>
```

## 2.2. L'IMPOSSIBLE IMPORT DES DONNÉES STRUCTURÉES EN XML : BILAN D'UNE TENTATIVE AVORTÉE ET SOLUTIONS ALTERNATIVES

---

Toutefois, l'entité useItem2 n'est référencée ni par le premier useItem, ni par l'entité Use, ce qui pourrait poser des problèmes d'intégrité dans l'import des données si nous choisissons ce format.

Par ailleurs, pour pouvoir faire les renvois vers les bons identifiants, il faut que les *records-id* soient générés préalablement dans la base. La meilleure stratégie à adopter pour effectuer l'import des données en XML est donc d'agir en deux temps :

- Faire un premier import pour générer les identifiants à partir des données brutes ;
- Faire un deuxième import pour faire les liens entre les données avec la récupération des *records-id* générés par Heurist.

L'import en csv, usage par usage<sup>196</sup>, a donc été préféré pour garantir la cohérence des données. En effet, de tout importer dans un même csv pourrait engendrer une insertion des pièces liturgiques dans le désordre. Contrairement à d'autres projets, l'ordre des données n'est pas indifférent. Le projet HORAE présente ainsi une demande rare, celui d'un ordre dans les *recordPointer*, car le modèle est structuré en hiérarchie comme dans un arbre XML. L'ordre des pièces doit refléter l'ordre de présence dans les manuscrits. En effet, chaque œuvre a un attribut de séquence pour ce qui précède et ce qui suit avec des renvois vers les œuvres concernées. L'ordre n'est donc ici pas déterminé selon l'identifiant ou la date mais selon la position de la pièce.

De plus, il s'agit de faire les liens entre les usages et les pièces. On crée donc dans un premier temps les entités useItem dotées des bonnes associations avec les entités Work et les incipits, puis on importe dans un deuxième temps le cursus, soit les relations de succession et de hiérarchie entre les pièces liturgiques. Toutefois, la création de relations entre les useItem a tout de même désordonné la succession des textes au sein des Use. La solution ponctuelle a donc été de donner un attribut de séquence numérique au titre pour replacer les pièces dans le bon ordre de succession.

*In fine*, le principal obstacle rencontré lors de l'import des données est une question de méréologie<sup>197</sup>, car chaque useItem a deux types de relation :

- La section dans laquelle elle se trouve ;
- L'ordonnancement (ce qui précède ou ce qui suit).

Il est alors nécessaire de simplifier le modèle pour l'import, tout en lui faisant refléter la complexité de la réalité décrite. Il faut ainsi d'abord vérifier que l'import fonctionne avant de créer la structure. Cette opération fait se croiser trois thématiques importantes dans le traitement des données :

- La récupération des données héritées ;
- La sécurisation des données ;
- Les contraintes à définir pour garantir l'intégrité des données.

---

196. Chaque usage a été séparé dans des csv différents, pour les raisons expliquées ci-dessus.

197. Logique traitant les relations des parties à un ensemble.

Ainsi la contrainte sur la structuration des données doit-elle permettre leur interopérabilité. D'un point de vue méthodologique, il est important d'identifier ce qui est générique et ce qui est spécifique. La priorité doit être accordée au générique, dans la mesure du possible. Il faut donc trouver les solutions les plus génériques possibles pour tenir compte des nouveautés, c'est-à-dire identifier ce qui est nouveau et le placer sous la bonne étiquette générique.

La vérification de l'intégrité des données peut se faire rapidement par un export de la base en xml, puis, grâce à une feuille xsl, on applique au document la fonction *count()* pour voir les identifiants des œuvres qui pointerait vers plus de 2 éléments, ce qui signifierait une erreur, car chaque identifiant doit être unique.

En ce qui concerne le travail de désambiguïsation de l'Office de la Vierge, il a consisté à compléter le travail de Victor Leroquais et Éric Drigsdahl. En effet, s'ils ont su relever les particularités liturgiques de chaque usage - nous en disposons d'environ 215 - nous n'avons accès qu'aux incipits des différentes pièces, ce qui n'est pas suffisant pour les identifier. Il s'agit donc de reprendre les manuscrits étudiés pour compléter la base de textes disponibles. Cela a donné lieu à deux manières de signaler les particularités dans la base de données :

- « *Not identified* » pour les pièces restées mystérieuses<sup>198</sup> ;
- « *Expected but not attested* » pour les pièces absentes du manuscrit mais qui sont constitutives de l'usage dans lequel elles s'inscrivent<sup>199</sup>.

Toutefois, un problème subsiste. Si l'on découvre une pièce nouvelle, cela provoque un décalage dans la séquence, ce qui a des conséquences sur les autres pièces par ricochet. Chaque ajout d'une pièce implique la modification de deux autres. Le problème est le même pour les pièces à supprimer. Le projet HORAE montre donc le potentiel mais aussi les besoins d'adaptation du logiciel proposé, ce qui implique un dialogue constant avec les techniciens responsables du fonctionnement de la base de données.

L'import des données offre ainsi de nouvelles possibilités en visualisation des usages liturgiques, que ce soit du point de vue de leur géolocalisation que des pratiques de dévotion.

---

198. Dans la plupart des cas, il s'agit de bréviaires imprimés uniques, avec un seul exemplaire attribué, ce qui rend leur vérification impossible.

199. L'absence des pièces est mesurée à partir du travail sur les témoins d'Éric Drigsdahl. Cela concerne les pièces présentes dans l'usage mais pas dans le manuscrit, pour une raison quelconque. Cela est plus particulièrement le cas pour les antennes de certaines heures, les hymnes ou les psaumes.

## 2.3. POSSIBILITÉS EN VISUALISATION DES DONNÉES

---

### 2.3 Possibilités en visualisation des données

En sciences humaines et sociales, l'utilisation de bases de données relationnelles est le tremplin à la création de visualisations de systèmes complexes, d'archives textuelles ou d'œuvres multimédias. Cela peut se concrétiser par la publication d'une carte ou d'un site web. On peut penser au travail d'Éric Drigsdahl qui a mis en ligne différents usages localisés des Heures de la Vierge, d'Amiens à York en passant par Rennes<sup>200</sup>. La définition de ces usages accessibles en ligne fournissent ainsi une référence dans l'étude des données du projet, mais elle offre aussi des possibilités d'enrichissement avec les nouvelles analyses résultant du projet. Ainsi, une base de données relationnelle bien conçue ne contient pas une liste d'informations statiques, mais tout un ensemble de relations ontologiques générant de nouvelles affirmations dans un domaine<sup>201</sup>.

Le projet HORAE s'appuie sur un foisonnement de livres d'heures, d'une part décrits dans les notices, d'autre part numérisés selon le protocole IIIF et disponibles en lignes. L'approche quantitative sert ici une analyse qualitative afin de mieux visualiser le regroupement de manuscrits présentant les mêmes caractéristiques textuelles, avec l'ordre des différentes parties (petit office de la Vierge, offices votifs, suffrages, prières), mais aussi l'ordre des unités textuelles qui permet de repérer les usages liturgiques. Les visualisations possibles des données permettent donc de cerner la diffusion et la circulation des textes dévotionnels et liturgiques transmis par les livres d'heures du Moyen Âge pour mieux comprendre la culture et la foi des XIII<sup>e</sup>-XVI<sup>e</sup> siècles<sup>202</sup>. Le projet montre donc que ce que l'on appelle « humanités numériques » est avant tout une « pratique de recherche », selon les mots d'Aurélien Berra<sup>203</sup>, une méthodologie.

Une utilisation raisonnée du numérique pourrait ainsi répondre au souhait d'Ezio Ornato quant à la discipline de la codicologie, qu'il décrit comme une champ de réflexion prenant en considération tous les aspects du livre, de « la nature des matériaux utilisés jusqu'aux modalités de conservation, de lecture et de diffusion, en passant par toutes les phases du processus de fabrication »<sup>204</sup>. Il parle donc de codicologie quantitative, nécessaire face à des données numériques, afin d'accroître la performance des méthodes de datation, de localisation et d'attribution.<sup>205</sup> Ces méthodes se fondent sur l'observation

---

200. E. DRIGSDAHL, *Tutorial - Hours of the Virgin Hore Beate Marie Virginis - Index to a Selection of Uses...*

201. Susan SCHREIBMAN, Ray SIEMENS et John UNSWORTH, *A Companion to Digital Humanities*, 2004, URL : <http://www.digitalhumanities.org/companion/> (visité le 17/07/2020), Partie II, Chapitre 15.

202. Cf. présentation du projet sur le site de l'IRHT : <https://www.irht.cnrs.fr/?q=fr/recherche/les-programmes-de-recherche/horae>.

203. Aurélien BERRA, « Faire des humanités numériques », dans *Read/Write Book 2 : Une introduction aux humanités numériques*, Marseille, 2012, p. 25-43, DOI : <https://doi.org/10.4000/books.oep.238>.

204. Ezio ORNATO et alii, *La face cachée du livre médiéval, L'histoire du livre vue par Ezio Ornato, ses amis et ses collègues*, Rome, 1997, p. vii.

205. *Ibid.*, p. 41-47.

des techniques de fabrication du livre, croisant certaines variables comme la dimension des feuillets ou la disposition du texte<sup>206</sup>. Ainsi, les manuscrits à longues lignes sont globalement plus petits que ceux à deux colonnes<sup>207</sup>. Cette affirmation se vérifie dans les notices de Leroquais, où la majorité des manuscrits à longues lignes n'excèdent pas les 299 mm de dimensions<sup>208</sup>, tandis que ceux à deux colonnes peuvent dépasser les 300 mm de dimensions<sup>209</sup>; Toutefois, cela n'exclut pas les manuscrits à deux colonnes de format plutôt petit<sup>210</sup>.

### **2.3.1 Géolocaliser les manuscrits et les usages**

Dans une étude sur ce que recouvre le terme « Humanités numériques » en France<sup>211</sup>, et plus particulièrement sur la question des mutations du support de l'écrit, Aurélien Berra rappelle que ce sont les usages qui font vivre les textes, et donc que les pratiques culturelles sont indissociables des techniques de production des manuscrits et imprimés, de leur profusion et de leur diffusion<sup>212</sup>.

En terme de géolocalisation de manuscrits du Moyen Âge et de la Renaissance, un projet réunissant l'IRHT, l'Université d'Oxford, de Pennsylvanie, et l'Université d'Aalto d'Helsinki, *Mapping Manuscript Migrations*, rassemble depuis 2017 des données issues de sources variées afin d'analyser sur la plus large échelle possible l'histoire et la provenance des manuscrits<sup>213</sup>. Le projet est né de la nécessité de rassembler la prolifération de données numériques dispersées dans des catalogues et bases de données diverses<sup>214</sup>. Un projet d'une telle ampleur soulève des questions méthodologiques sur la visualisation des données : comment représenter la complexité des métadonnées relatives aux manuscrits ? Le *Linked data*, traduit par l'expression Web des données<sup>215</sup>, permet alors de produire des cartes et des graphes de réseaux afin de favoriser de nouvelles façons de penser l'histoire et la provenance des manuscrits<sup>216</sup>.

---

206. On trouvera une méthodologie d'observation des manuscrits médiévaux et de rédaction de notices appropriées plus précisément dans J.B. LEBIGUE, *Initiation aux manuscrits liturgiques...*

207. E. ORNATO et alii, *La face cachée du livre médiéval, L'histoire du livre vue par Ezio Ornato, ses amis et ses collègues...*, p. 51.

208. V. LEROQUAIS, *Les livres d'heures manuscrits de la Bibliothèque nationale....*, Notice 145, p. 305.

209. *Ibid.*, Notice 5, p. 9.

210. *Ibid.*, Notice 62, p. 150.

211. Le terme est dérivé de la notion de *digital humanities*, davantage formalisée dans le monde anglo-saxon.

212. A. BERRA, « Faire des humanités numériques »...

213. Cf. <https://mappingmanuscriptmigrations.org/en/>.

214. Toby BURROWS, Eero HYVÖNEN, Lynn RANSOM et Hanno WIJSMAN, « MANUSCRIPT STUDIES A Journal of the Schoenberg Institute for Manuscript Studies », *Manuscript Studies*, 3 (2019), URL : [https://repository.upenn.edu/mss\\_sims/vol3/iss1/13](https://repository.upenn.edu/mss_sims/vol3/iss1/13) (visité le 06/05/2020), p. 249-250.

215. Le terme est inventé et défini par Tim Berners-Lee, alors directeur du W3C. Il désigne la publication de données structurées sur le Web, non pas sous la forme de silos de données isolés les uns des autres, mais reliés entre eux pour constituer un réseau global d'informations qui rend les données plus visibles.

216. *Ibid.*, p. 251.

## 2.3. POSSIBILITÉS EN VISUALISATION DES DONNÉES

---

De plus, la localisation des calendriers des livres d'heures reste un défi majeur. Établir des caractéristiques communes entre certains calendriers permet de définir des groupes issus d'aires géographiques et de périodes particulières<sup>217</sup>. Si la majorité des calendriers contiennent des célébrations communes quant à la Vierge Marie, moins une fête est commune en-dehors d'un groupe, plus elle est caractéristique du groupe qui la célèbre<sup>218</sup>.

Si le calendrier constitue donc un indice majeur pour la datation et la destination d'un manuscrit, il n'est pas le seul indicateur. Ainsi, dans les livres d'heures, les lignes vides sont rares dans un calendrier, ce qui aboutit parfois à la notation de fêtes de saints fautives. C'est plutôt la convergence de plusieurs signes distinctifs qui permet d'avancer des hypothèses sur un usage. Les litanies des saints sont ainsi un autre indice pour la localisation d'un manuscrit. Il reste toutefois difficile de localiser les livres d'heures car chaque section peut mêler des usages différents. Par exemple, l'office de la Vierge peut être à l'usage de Paris mais le calendrier à l'usage d'Arras.

Cette complexité à inscrire le livre d'heures dans un usage précis s'explique par l'origine même de son apparition, et par sa place au sein des autres livres liturgiques. La liturgie votive devient de plus en plus prégnante à partir du XIII<sup>e</sup> siècle. Elle se structure sur un rythme hebdomadaire, mais son usage est local. Inspirée de cette pratique, la liturgie dévotionnelle se développe, sous une forme plus intime et moins communautaire, donnant une place particulière aux oraisons de saints.

Toutefois, les offices ne reflètent pas forcément les lieux de production et de destination, ce qui n'est pas le cas des autres livres liturgiques dont l'usage reflète la destination. Il faut alors distinguer l'usage de la destination, d'autant plus pour ce type de livre voué à la dévotion privée. Si l'on retrouve parfois des notes de possession ou des armoiries, le destinataire du manuscrit peut s'inspirer d'un usage qui n'est pas nécessairement lié à la localisation de son lieu de vie<sup>219</sup>.

Les possibles visualisations des données du projet HORAE s'inscrivent dans ce cadre. En ce qui concerne la géolocalisation, la base Heurist, grâce à des *records pointers*, soit les clés étrangères, et des *relationships makers*, les tables de relation, permet de choisir entre la cartographie des lieux de conservation des manuscrits et celle des usages. Il faut alors faire un *ruleset*, une série de règles, pour afficher les lieux en question. Une version 6 est ainsi en cours de développement pour pouvoir visualiser les données sous forme de

---

217. Tuomas HEIKKILA et Teemu ROOS, « Quantitative methods for the analysis of medieval calendars », *Digital Scholarship in the Humanities*, 33-4 (2018), p. 766-787, DOI : 10.1093/11c/fqy007, p. 781.

218. *Ibid.*

219. Ces propos s'inspirent des séances de la formation continue délivrée par la spécialiste des manuscrits liturgiques Laura Albiero, « Les livres liturgiques manuscrits et imprimés : principes de catalogage ».

tables. Pour lors, il est possible de voir les lieux de conservation des manuscrits étudiés sur une carte :

## 2.3. POSSIBILITÉS EN VISUALISATION DES DONNÉES

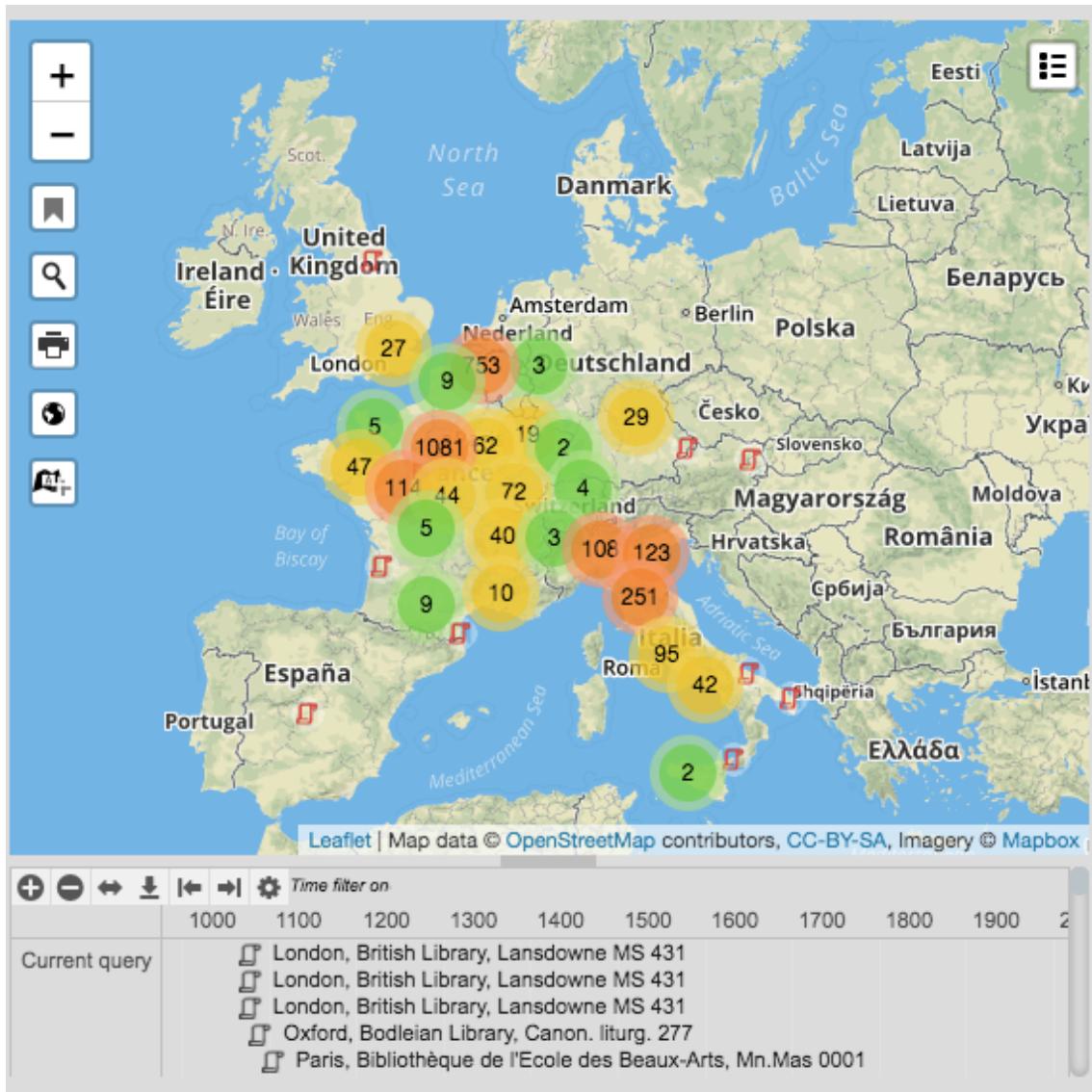


FIGURE 2.13 – Cartographie des lieux de conservation des manuscrits étudiés

Ainsi que les lieux des différentes organisations évoquées :

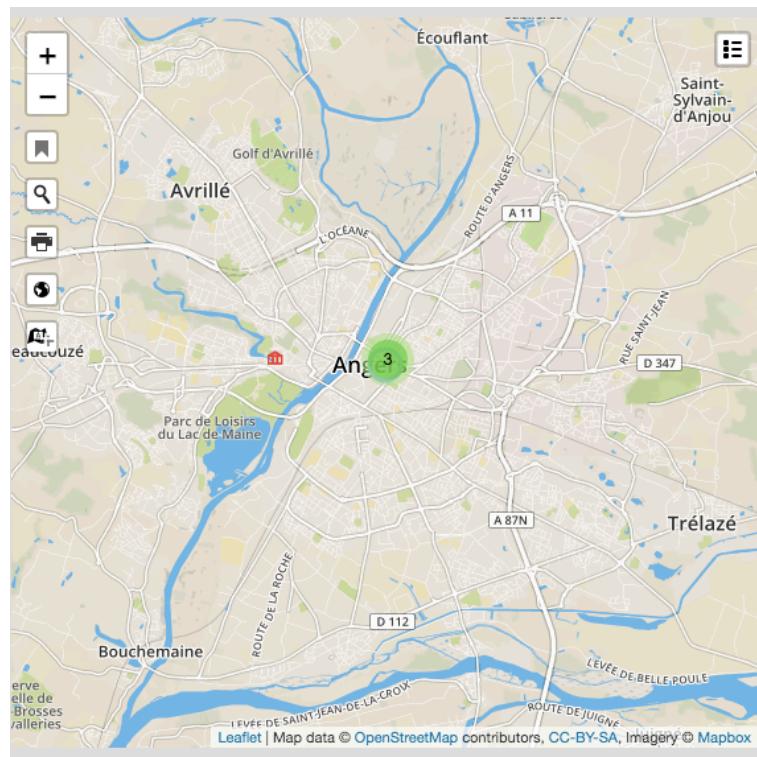


FIGURE 2.14 – Cartographie d'une organisation identifiée dans les usages des témoins : l'exemple d'Angers

## 2.3. POSSIBILITÉS EN VISUALISATION DES DONNÉES

---

### 2.3.2 Ce que nous disent les textes sur les dévotions...

L'historienne Kathryn M. Rudy s'est intéressée à la manière dont les propriétaires de livres médiévaux, et plus particulièrement de livres d'heures du XV<sup>e</sup> siècle aux Pays-Bas, modifient le contenu de leurs livres pour refléter des situations changeantes, pas tant politiques que religieuses, économiques ou sociales<sup>220</sup>. Les ajouts apportés à un livre peuvent ainsi témoigner d'une certaine force émotionnelle et sociale<sup>221</sup>. Elle étudie donc les phénomènes de modifications d'un manuscrit, qui peuvent prendre la forme d'ajouts ou de soustractions<sup>222</sup> de contenu, tout comme de changements dans l'organisation<sup>223</sup>. Ces observations donnent des informations sur la manière dont les gens utilisaient et considéraient les livres de dévotion. L'historienne propose alors une nouvelle approche codicologique qui vise à comprendre quand un propriétaire de manuscrit complet et fini choisit de faire des ajouts pour signifier son appropriation : qu'est-ce qu'il gagne à le faire et qu'est-ce qui a rendu cela possible dans la structure des manuscrits<sup>224</sup> ?

Kathryn M. Rudy divise les modifications apportées aux livres d'heures en deux catégories : celles qui impliquent une nouvelle reliure et celles qui n'en impliquent pas. Parmi la première catégorie, l'on retrouve des changements comme la correction du texte, l'ajout de texte entre les interstices ou sur les folios restés blancs<sup>225</sup>, l'agrandissement de décorations existantes, l'ajout de nouvelles illustrations, l'ajout de morceaux de parchemin sur les pages restées blanches, voire d'autres objets. Pour la deuxième catégorie, il s'agit de feuilles avec du texte ou des images. Ces images peuvent relever des offices les plus communs, ou bien d'indulgences, de portraits et de détails personnels, parfois provenant d'un autre manuscrit. Ce sont parfois des cahiers entiers qui sont ajoutés. Cela est rendu possible grâce à l'assemblage modulaire des manuscrits qui laissent plusieurs zones de parchemins blanches. En effet, le modèle pouvait être partagé entre plusieurs scribes auxquels était attribuée une portion de texte à copier<sup>226</sup>. Cette méthode de fabrication montre qu'un livre est pensé comme une construction progressive, un ensemble de fragments successivement ajoutés.<sup>227</sup> L'historienne analyse l'ajout de textes ou d'images à des manuscrits considérés comme complets comme des désirs relevant de différentes natures<sup>228</sup> :

---

220. Kathryn M. RUDY, *Piety in Pieces, How Medieval Readers Customized their Manuscripts*, Cambridge, 2016, p. 1-2.

221. *Ibid.*, p. 2.

222. Parmi les phénomènes observés de contenu enlevé, on retrouve des traces de folios arrachés, de cahiers retirés, de textes grattés ou d'images dégradées.

223. *Ibid.*, p. 3.

224. *Ibid.*, p. 5.

225. L'historienne est notamment confrontée à l'ajout d'informations familiales, de documents légaux, de gloses, d'informations sur les calendriers, de références à des circonstances particulières, d'explications pour faire du livre un outil didactique, ou de prières.

226. E. ORNATO et alii, *La face cachée du livre médiéval, L'histoire du livre vue par Ezio Ornato, ses amis et ses collègues...*, p. 89.

227. K. M. RUDY, *Piety in Pieces, How Medieval Readers Customized their Manuscripts...*, p. 9.

228. *Ibid.*

- un désir de personnaliser le livre ;
- un désir de prendre possession des nouveaux sujets textuels ou visuels disponibles ;
- un désir de vouer sa dévotion à de nouvelles fêtes et cultes ;
- un désir d'ostentation de son pouvoir financier ;
- un désir de rehausser la décoration du manuscrit ;
- un désir de systématiser la décoration ;
- un désir de fixer des images volantes, parfois données en cadeaux.

Ces désirs peuvent évidemment être mêlés.

Il est également intéressant d'analyser les dévotions qui transparaissent dans les livres d'heures à partir du genre de leur commanditaire. Charity Scott-Stokes s'est par exemple intéressée aux livres d'heures produits en Angleterre et commandés par des femmes. Rappelant que le livre d'heures constitue un des livres les plus populaires auprès des laïcs de la fin du Moyen Âge, ils sont fondamentaux pour comprendre la piété médiévale, car son utilisation est principalement vouée au foyer. Ils sont donc révélateurs d'une dévotion ressentie dans l'intimité<sup>229</sup>. Elle observe entre autres une tendance un peu plus importante à célébrer les saintes féminines, en commençant par la Vierge et sainte Anne ; puis viennent les pénitentes du Nouveau Testament, comme Marie-Madeleine, et les vierges martyrs, notamment Catherine et Marguerite<sup>230</sup>. Ces saintes sont d'autant plus invoquées qu'elles sont associées au soulagement de douleurs et de maladies spécifiques. Ainsi, on fait appel à sainte Marguerite lors des accouchements, à sainte Apolline en cas de rage de dents, ou à sainte Suzanne pour protéger les femmes de la calomnie, des commérages et de la médisance<sup>231</sup>.

Dans le cadre du projet HORAE, il est encore trop tôt pour exposer d'éventuelles découvertes quant aux usages liturgiques et aux pratiques de dévotion. Il est toutefois possible de voir les relations entre les différentes pièces composant un livre d'heures sous forme de graphes :

---

229. C. SCOTT-STOKES, *Women's Books of Hours in Medieval England...*, p. 1.

230. *Ibid.*, p. 14.

231. *Ibid.*, p. 15.

## 2.3. POSSIBILITÉS EN VISUALISATION DES DONNÉES

---

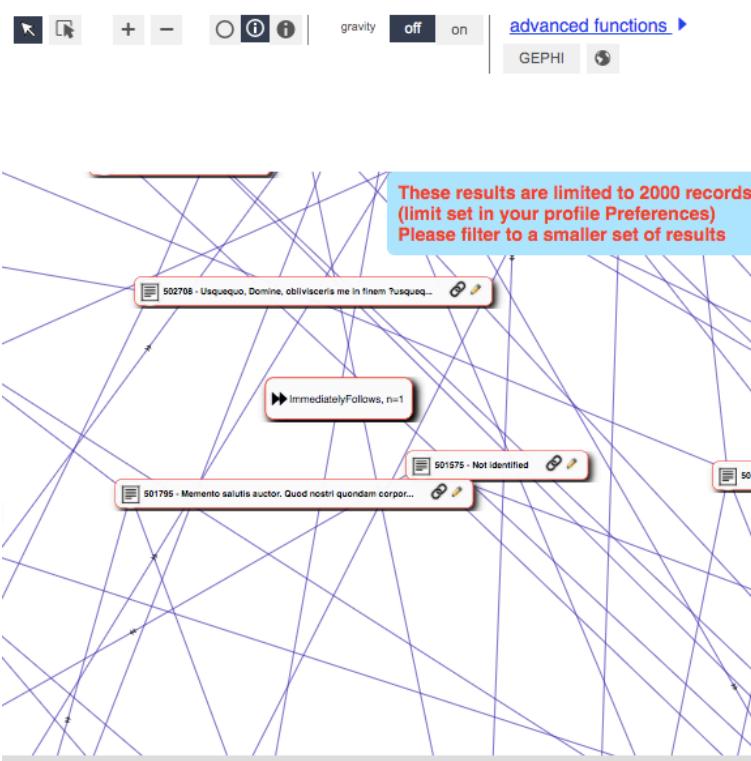


FIGURE 2.15 – Extrait de graphe sur les relations entre les différentes sections d'un livre d'heures

Ce qui ressort à présent des hypothèses formulées est la confirmation des analyses de Victor Leroquais quant aux variantes d'un même usage au sein de plusieurs témoins créés pour des commanditaires différents. Pour étudier plus amplement les dévotions, il est nécessaire de faire des liens entre les usages et les prières, ce qui est possible avec une plus grande prise de recul par rapport aux données amassées. En effet, l'idée est d'opérer un travail de comparaison des textes pour mieux déterminer les influences et les étendues des différents types de dévotion.

Bénéficier de données de qualité et pouvoir les visualiser est une première étape pour avancer des conclusions historiques, mais l'aventure ne s'arrête pas là. La reconnaissance automatique de textes et d'images et l'utilisation de l'intelligence artificielle permet de comparer des données en masse, et à terme, de définir les structures types de livres d'heures selon les usages liturgiques.

# **Chapitre 3**

## **L’annotation pour l’apprentissage machine : ce que le numérique apporte à l’analyse des sources**

L’utilisation de l’intelligence artificielle dans le cadre du projet HORAE, rendue possible grâce à la collaboration de différents acteurs dans les domaines de l’histoire et du numérique, montre la complémentarité des savoirs en humanités et en technologies. En effet, il s’agit de rendre le savoir implicite en humanités explicite, afin de faire de ce savoir la clé de voûte de l’apprentissage machine, favorisant l’analyse de données en masse et leur formalisation. À partir d’un certain nombre de manuscrits annotés manuellement, processus selon lequel sont repérés les débuts de section de différents niveaux d’un panel de livres d’heures, une vérité terrain est établie. Le travail d’annotation, réalisé par les membres d’une institution spécialisée dans la recherche fondamentale sur les manuscrits médiévaux et les imprimés anciens, l’IRHT en l’occurrence, s’est opéré sur une interface développée par la *start-up* Teklia. Les exigences relatives au projet ont ainsi permis au prestataire d’améliorer la qualité de ses services, tout en soulevant de nouveaux défis technologiques en terme de reconnaissance de texte, s’appuyant sur la vérité terrain définie préalablement. Les résultats technologiques obtenus interrogent alors sur les possibilités offertes par l’intelligence artificielle en terme d’analyse et de compréhensions de la structure des livres d’heures.

### **3.1 Gestion et management d’un projet en humanités numériques**

De la manière de gérer et manager un projet, et pas seulement en humanités numériques, découle sa réussite. Durant tout le projet, la coopération est continue, afin que tous les acteurs partagent la même vision managériale, soit les mêmes envies, et la même vision

procédurale. Cette collaboration constante se concrétise par la planification de réunions régulières et l'affirmation claire et précise des échéances et des objectifs à atteindre, et ce dès le début du projet. Ces éléments se retrouvent dans des documents tels que la note de cadrage ou le cahier des charges. Les divers échanges et réunions entre l'IRHT, Teklia et le LS2N ont ainsi permis de clarifier les tâches et attentes de chacun, ce qui a notamment abouti à la mise en service d'une interface d'annotations conçue en partie dans le cadre du projet HORAE : Arkindex<sup>232</sup>.

### **3.1.1 La collaboration IRHT, Teklia et LS2N**

Le cahier des charges du projet HORAE témoigne de l'importance d'impliquer chaque acteur du projet tout au long de ce dernier tout en définissant des tâches propres à chacun<sup>233</sup>. Il rappelle dans un premier temps les objectifs du projet, son originalité et sa pertinence par rapport à l'état de l'art, ainsi que la méthodologie adoptée et la gestion des risques. Cette dernière partie se doit de témoigner de la conscience des potentielles déceptions du projet, et comment il est possible de les contourner, afin de rassurer les organismes financeurs. Pour HORAE, quatre grandes thématiques se détachent en terme d'apports et de risques :

1. « Conséquences méthodologiques de la transdisciplinarité ». Il s'agit de mesurer l'efficacité d'un partenariat transdisciplinaire face à la gestion de données incertaines. En effet, les niveaux de qualité des données produites et analysées ne sont pas abordés de la même façon par les professionnels du numérique, qui cherchent à mesurer la précision d'une méthode d'analyse de document ou d'HTR<sup>234</sup>, et par les chercheurs en humanités et l'éditeur de textes médiévaux qui ont besoin de données parfaites et qui ne peuvent quantifier l'incertitude. Le choix des partenaires d'HORAE est donc de s'appuyer sur des données « bruitées » issues de l'HTR pour les associer à des unités textuelles de référence. Le cahier des charges prend l'exemple suivant :

Par exemple, le verset « *In principio erat verbum et verbum erat apud Deum* » sera reconnu comme étant le premier verset de l'Évangile de Jean, même si on n'a que la séquence « *principio ... verbum ... Deum ...* » en tenant compte de la faible distance séparant les mots. Ainsi, à partir de données fausses ou incomplètes, on peut arriver à une conclusion

---

232. L'interface de traitement automatique de document a vocation à se développer également au service d'autres projets. Cf. Chloë FIZE, *Archives et numériques, un traitement intelligent des données historiques : application à la recherche et au patrimoine privé*, mém. de mast., École nationale des Chartes, 2020

233. AAPG ANR, IRHT, Teklia et LS2N, *HORAE Heures : Reconnaissance de l'écriture manuscrite, catégorisation automatique, éditions Hours - Recognition, Analysis, Editions*, 2017.

234. *Handwritten Text Recognition*.

### 3.1. GESTION ET MANAGEMENT D'UN PROJET EN HUMANITÉS NUMÉRIQUES

---

entièremment correcte.<sup>235</sup>

De plus, pour vérifier les résultats des outils, des visualisations des données sont produites à chaque étape pour vérifier leur qualité et éventuellement les corriger, d'où l'importance de l'ergonomie pour la présentation des données.

2. « Analyse de document et reconnaissance d'écriture ». Il est ici question de mettre à profit le large corpus de manuscrits similaires par leurs structures et leur contenu pour développer les modèles de *deep learning*. Leur développement est favorisé par la ré-utilisation de corpus mobilisés dans de précédents projets<sup>236</sup>. L'autre enjeu est d'adapter les librairies *open source*<sup>237</sup> existantes aux manuscrits médiévaux. Le principal risque serait la production de résultats de qualité insuffisante pour identifier les textes, risque faible au vu des résultats acquis lors des précédents projets.
3. « Segmentation et identification des textes ». Les membres du projet y expliquent une méthode descendante de segmentation, qui consiste à à corriger la segmentation au fur et à mesure de la reconnaissance de textes identifiés.
4. « Formats, visualisation, exploitation, pérennisation ». Cette partie rappelle l'utilisation de formats standards et de logiciels *open source*.

Une deuxième grande partie revient sur l'organisation du projet les moyens mis en œuvre. Cette partie présente tout d'abord les principaux acteurs du projet, puis les cinq grandes tâches du projet, qui ne s'opèrent pas de manière successives mais croisées, selon les principes de la méthode dite « Agile ». Elles recouvrent la constitution des données, les phases de recherche et développement, et la coordination. À chaque tâche est rappelé un calendrier fixant des livrables et des échéances à atteindre, ainsi qu'une évaluation des risques. Les cinq étapes sont les suivantes :

1. « Corpus, vérité terrain et modèle de données ».

Principalement menées par l'IRHT, ces actions consistent à recenser les sources de livres d'heures disponibles en ligne et accessibles au format IIIF. Le projet prévoit d'impulser la numérisation de 99 de manuscrits microfilmés à la BnF

---

235. *Ibid.*, p. 9.

236. Il s'agit des projets Oriflamms et Himanis. Le premier projet, acronyme de *Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts*, vise à analyser l'évolution des systèmes et formes graphiques des écritures au Moyen Âge selon leur contexte de production et leur langue. Une exposition des enjeux et des résultats du projet est disponible à l'adresses suivante : <https://oriflamms.hypotheses.org/1592>. Le deuxième projet, signifiant *Historical MANuscript Indexing for user-controlled Search*, en partenariat avec les Archives nationales, a abouti à la création d'un moteur de recherche en plein texte dans les registres de la chancellerie royale. Une description plus ample du projet est disponible sur le site des Archives nationales, à l'adresse suivante : <http://www.archives-nationales.culture.gouv.fr/himanis>.

237. Le terme *open source* se réfère à l'accès à des codes sources que l'on peut librement réutilisés et redistribués. Ils sont souvent le fruit d'une collaboration entre programmeurs.

et plus de 500 à l'IRHT, qui dispose déjà de trois bases de données recensant des livres d'heures<sup>238</sup>. L'objectif est de créer un lot de plus de 120 000 images extraites de plus de 400 manuscrits. Les livrables concernent la mise à disposition des images numérisées.

La deuxième sous-étape est la constitution du corpus textuel. Il s'agit de donner les moyens d'identifier les textes, ce que l'on appelle la « vérité terrain ». Cette tâche bénéficie d'une double approche. D'une part, avec l'établissement d'unité textuelles de référence, formées d'un identifiant, d'un incipit et d'une version normalisée. D'autre part, grâce à la création de tables décrivant les principales structures des sous-sections, sections et parties d'un livre d'heures. Ce corpus est amené à s'accroître avec la découverte de textes non prédictibles. Les annexes de Victor Leroquais sont ainsi océrisées pour détecter les poèmes et prières inédits. Les livrables concernent donc la fixation de textes de références et de tables de structures.

La troisième sous-étape concerne des problématiques plus amplement abordées lors du stage : le modèle de données et l'unicité du format afin de garantir l'interopérabilité. Si l'affichage des manuscrits se fait selon le protocole IIIF, leur texte est encodé selon un modèle bien précis respectant la TEI, soit un schéma défini dans un document ODD comme nous l'avons présenté pour l'encodage des métadonnées, en l'occurrence des notices de livres d'heures conservées à la BnF. La dernière sous-étape est la publication d'une interface graphique afin de saisir, visualiser et corriger les données si nécessaire, affiner la table de structure et les textes de référence<sup>239</sup>. On peut par exemple penser au mode auteur du logiciel Oxygen XML Editor qui permet de voir et modifier les données affichées de manière plus claire et limpide, sans balises, avec une présentation personnalisable grâce à la modification de la CSS et des frameworks du logiciel<sup>240</sup>.

## 2. « Analyse et reconnaissance de documents ».

Les tâches que recouvrent cette section concernent plus exclusivement la société

---

238. Parmi les bases de données constituées à l'IRHT et contenant des livres d'heures, on peut consulter la base Medium (<http://medium.irht.cnrs.fr/>), qui enregistre tous les manuscrits des bibliothèques françaises et étrangères dont l'IRHT détient des reproductions ; la base Iter Liturgicum Italicum (<https://liturgicum.irht.cnrs.fr/>), qui donne accès aux données relatives aux livres liturgiques d'origine italienne conservés en Italie et ailleurs, ou bien signalés dans les catalogues des maisons de vente ; la base INITIALE (<http://initiale.irht.cnrs.fr/>), qui recense les manuscrit enluminés conservés principalement dans les bibliothèques publiques de France hors la Bibliothèque nationale ; et enfin la base Bibale (<http://bibale.irht.cnrs.fr/>), créant des liens entre les notices pour rassembler les données de provenance de bibliothèques françaises.

239. *Ibid.*, p. 11-13.

240. Des tests de personnalisation de la CSS et des frameworks d'Oxygen sont disponibles dans les livrables techniques.

### 3.1. GESTION ET MANAGEMENT D'UN PROJET EN HUMANITÉS NUMÉRIQUES

---

Teklia. Une des premières étapes est d'opérer la classification des différents types de pages (couverture, page blanche, calendrier texte, texte illustré, etc.). L'objectif est de fournir une interface de visualisation des classes de page avec des fonctionnalités permettant de les regrouper selon leur type.

L'étape suivante est de détecter les lignes de texte. Si les livres d'heures présentent un cas plutôt simple au vu de leur homogénéité, soit des lignes horizontales de même hauteur alignées verticalement, certains éléments graphiques peuvent perturber la localisation des lignes, comme les lettrines et les illustrations. L'objectif principal est de fournir une interface de visualisation des zones de texte avec leur détection automatique.

L'étape logique qui en découle est la reconnaissance automatique de l'écriture manuscrite. Cette reconnaissance s'appuie sur des réseaux de neurones récurrents. Des librairies performantes sont disponibles en *open source* comme la librairie *TensorFlow* de Google<sup>241</sup>. L'objectif est donc de présenter une interface de visualisation des textes avec une fonctionnalité de reconnaissance automatique des textes.

Enfin, la dernière étape consiste à développer une interface graphique pour traiter le corpus. Il s'agit de pouvoir appliquer la chaîne de traitement à tous les manuscrits et de présenter les résultats dans une interface web<sup>242</sup>. Cette étape se concrétise dans la mise à disposition de l'interface Arkindex utilisée pendant le stage.

#### 3. « Segmentation et détection extrinsèque ».

Cette partie est davantage menée par le Laboratoire des Sciences du Numérique de Nantes. La première étape consiste à segmenter les textes du livre d'heures. Cette segmentation se fait à trois niveaux, selon les sections, sous-sections et textes élémentaires. Il s'agit, à l'aide d'algorithmes, de mesurer une cohésion lexicale locale et globale adaptée aux livres d'heures, de calculer leur distribution au sein des sections et sous-sections, et de repérer les langues. Les livrables se concrétisent sous la forme de rapports sur la segmentation des livres d'heures et sur les algorithmes de segmentation.

La deuxième étape s'apparente à l'identification des unités textuelles de référence. Cette identification est possible grâce à des expérimentations sur des empreintes, qui correspondent à des suites de n-grammes<sup>243</sup>. Les empreintes sont

---

241. Outil d'apprentissage automatique fondé sur l'apprentissage profond, *TensorFlow* est autant utilisé dans le milieu commercial que dans le monde de la recherche. Son code source est ouvert depuis 2015. Il est par exemple de plus en plus performant dans la reconnaissance de la parole.

242. *Ibid.*, p. 13-14.

243. Un n-gramme est une sous-séquence d'éléments construite à partir d'une séquence donnée. Il

ensuite comparées selon des tests de mesures de similarité. Les livrables prennent la forme de rapports et de spécifications sur la détection extrinsèque des textes élémentaires.

Il s'agit ensuite de consolider la « table des matières » des manuscrits étudiés, en mettant à jour les textes non-repérés auparavant.

Puis, vient le processus d'alignement des livres d'heures à partir d'un livre d'heures de référence. D'une part, le macro-alignement consiste à mettre en correspondance la structuration des textes avec les textes élémentaires de référence, afin d'identifier les textes non-référencés, les parties communes et les différences. D'autre part, le micro-alignement s'attache à mettre en regard les textes élémentaires entre livres d'heures afin de mieux visualiser les différences, les éventuelles fautes du programme de reconnaissance du document, mais aussi les possibles variations linguistiques. Les livrables correspondent à la visualisation des alignements.

Enfin, comme pour les grandes étapes précédentes, cette tâche se conclut par la création d'une interface de visualisation qui permet de valider les alignements de textes élémentaires détectés<sup>244</sup>.

#### 4. « Crédit de savoirs à l'ère numérique : visualisations et interprétations ».

Cette séquence de tâches concerne l'IRHT. Il s'agit de favoriser l'émergence de nouveaux savoirs grâce à l'intégration des résultats en vue d'une expérience utilisateur de qualité. Pour y parvenir, il faut coordonner et intégrer les différents développements en visualisation des données. Cette étape est exigeante en terme de fonctionnalités pour l'utilisateur final, notamment dans l'objectif d'enrichir les métadonnées et d'interagir avec les données. Un travail de veille sur les outils disponibles est donc indispensable.

La deuxième étape consiste à publier les données de la manière la plus pérenne possible.

La publication des données permet ensuite de mesurer les similarités et de visualiser les réseaux textuels. Il s'agit de repérer les textes inédits à partir des textes non alignés, de comparer les différentes structures des livres d'heures, d'identifier des usages liturgiques et de repérer des réseaux textuels via la visualisation de textes transmis ensemble et dans le même ordre. Cette étape nous rappelle l'importance de produire des données de qualité et de détenir des informations valides pour préserver la véracité des conclusions historiques.

---

s'agit en fait d'établir des probabilités d'apparition d'une lettre ou d'un mot selon les textes élémentaires donnés.

244. *Ibid.*, p. 14-15.

### 3.1. GESTION ET MANAGEMENT D'UN PROJET EN HUMANITÉS NUMÉRIQUES

---

La production de ces savoirs permet de diffuser une histoire des lectures dévotes du XIII<sup>e</sup> au XVI<sup>e</sup> siècles via l'écriture de deux monographies, ainsi que de la publication d'une anthologie de textes édités durant le projet<sup>245</sup>.

5. « Administration et communication ». Cette section rappelle les outils de communication pour le suivi du projet par les différents acteurs et la stratégie de diffusion<sup>246</sup>. Les métadonnées générées par le projet sont ainsi publiées sous licence libre sur Github<sup>247</sup>.

La dernière grande partie du cahier des charges concerne les retombées du projet. Le caractère innovant du projet est en effet de proposer l'automatisation de la transcription et de l'identification de textes transmis par les manuscrits médiévaux. Par ailleurs, il participe à un changement majeur dans la transmission et la valorisation du patrimoine culturel. Ainsi, les conséquences dans le milieu scientifique sont triples<sup>248</sup> :

- L'adaptation de librairies logicielles libres et *open access* à l'analyse d'images et de textes de manuscrits médiévaux ;
- L'analyse de dérivations textuelles en linguistique computationnelle ;
- Le changement de l'approche en catalogage des manuscrits médiévaux et l'augmentation des connaissances du patrimoine textuel, de la liturgie, des échanges culturels de la fin du Moyen Âge en humanités.

Du point de vue culturel, HORAE s'inscrit dans la volonté des pouvoirs publics de rendre plus immédiatement accessible le patrimoine via la numérisation. Toutefois, la présence en ligne ne suffit pas pour être visible, et les objets culturels numérisés sont parfois peu exploités et expliqués. L'expertise scientifique et technologique développée lors du projet peut alors offrir de nouveaux moyens pour structurer les collections et les mettre à disposition du public de manière éclairée, compréhensible et interactive<sup>249</sup>.

Le projet HORAE reflète la « révolution numérique » qui traverse nos sociétés. Le numérique est ici l'outil au centre de la création et du partage des savoirs, tout en favorisant auprès du grand public la découverte par sérendipité<sup>250</sup>.

Lors du stage, le management du projet s'est cristallisé lors des diverses réunions entre les principaux acteurs. Ces réunions régulières sont essentielles pour s'informer et informer à propos des livrables acquis et de ceux qu'il reste à fournir. Dans le projet HORAE, le client, l'IRHT, institution spécialisée dans les manuscrits anciens, s'engage

---

245. *Ibid.*, p. 15-16.

246. *Ibid.*, p. 16-17.

247. Cf. <https://github.com/oriflamms/HORAE>.

248. *Ibid.*, p. 17-18.

249. *Ibid.*, p. 18.

250. *Ibid.*, p. 18-19.

à fournir des données historiques de qualité à exploiter, tandis que la *start-up* Teklia, son prestataire, garantit la mise à disposition d'outils numériques nécessaires pour l'analyse des sources, selon les objectifs définis dans le cahier des charges. Lors du stage, la collaboration entre le client et son prestataire s'est essentiellement focalisée sur le perfectionnement de l'interface Arkindex, afin qu'elle soit la plus efficace et la plus ergonomique possible.

### **3.1.2 Arkindex : une interface d'annotation en construction**

Arkindex est une plateforme de traitement automatique des documents, de la classification des pages à la reconnaissance d'écritures manuscrites en passant par la détection des lignes. Nos premiers pas sur Arkindex se sont effectués autour du travail d'annotation, c'est-à-dire le repérage des débuts de section d'un corpus de livre d'heures, selon un protocole sur lequel nous reviendrons. L'équipe de l'IRHT s'est tout d'abord heurtée à des problèmes liés à l'ergonomie de l'interface, notamment à propos de l'impossibilité de zoomer sur les pages numérisées et du déplacement du rectangle de sélection selon la résolution de l'écran.

Pour régler le problème de la sélection de la zone d'annotation qui dépend de la résolution de l'écran, des tests ont été effectués avec des résolutions différentes afin de stocker les données en fonction. Le résultat final doit être que l'alignement de la ligne du texte corresponde à la sélection, sans que la résolution soit en jeu.

Par ailleurs, les problèmes de performance dus au trop grand nombre de requêtes sur la base de données ont été résolus grâce à l'installation d'un système en clusters. Cela signifie le partitionnement des données sur plusieurs serveurs afin de gagner en rapidité dans le temps de réponse.

Un corpus spécifique pour l'inter-annotation<sup>251</sup> a été préparé avec l'ensemble des classes à gérer<sup>252</sup>. On en compte environ 4500. L'objectif est de pouvoir faire des recherches avancées dans les classes et la base de données, ce qui implique un travail de développement important du côté de Teklia pour accroître la performance d'Arkindex. En effet, à la différence du logiciel Mirador<sup>253</sup>, on doit pouvoir faire une recherche dans une liste.

Il convient de revenir sur le terme de « classe ». La classe est un terme générique

---

251. Il s'agit d'un ensemble de 10 manuscrits communs à plusieurs annotateurs de l'IRHT pour établir la vérité terrain sur les sections à l'intérieur des livres d'heures.

252. Les classes correspondent à tous les contenus d'annotation, qu'il s'agisse des textes ou des niveaux hiérarchiques. Le contenu des classes sera plus explicitement abordé lors de l'explication du protocole d'annotation dans la section 3.2.

253. Cf. <https://projectmirador.org/>. Mirador est un visualiseur web qui offre des fonctionnalités avancées de zoom, de comparaison et d'annotation d'images en haute résolution, indépendamment du type de document ou de la bibliothèque numérique qui les héberge. Il permet d'afficher dans une interface commune des documents numériques provenant d'entreposés d'images différents et compatibles avec les protocoles IIIF.

### 3.1. GESTION ET MANAGEMENT D'UN PROJET EN HUMANITÉS NUMÉRIQUES

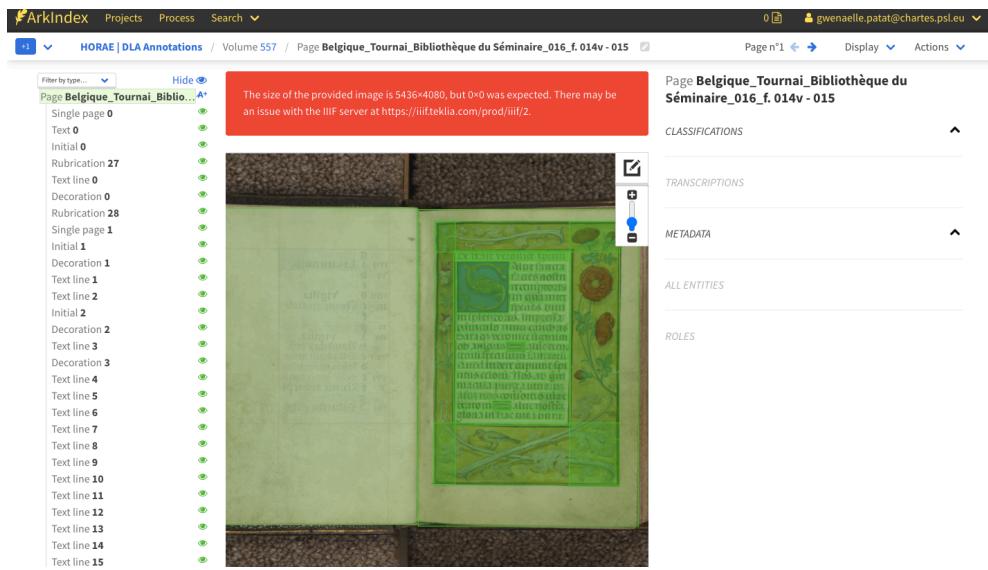


FIGURE 3.1 – Exemple des différentes classes de segmentation d'une page de livre d'heures sur l'interface Arkindex

propre aux tâches du *data mining*<sup>254</sup>, avec d'une part, la classification, et d'autre part, le *clustering*. La classification consiste à déterminer le nombre de catégories à traiter *a priori*. Chaque élément est donc associé à une classe. Quant au *clustering*, le partitionnement de données, il fait référence à un ensemble d'éléments dont on doit optimiser le regroupement. Il s'agit en fait d'analyser les données en les regroupant dans différents paquets homogènes, c'est-à-dire partageant des caractéristiques communes. Les données sont alors regroupées dans des ensembles et des sous-ensembles reliés entre eux.

Pour la détection de ligne, le modèle est entraîné à partir du corpus de livres d'heures, selon les principes du *deep learning*. Grâce à l'insertion en image d'entrée des caractéristiques d'une ligne, la machine génère une image de prédiction au niveau pixel. Le *deep learning* est en effet un type d'intelligence artificielle dérivé du *machine learning*, où la machine est capable d'apprendre par elle-même, sans règles pré-déterminées par un programme. Il s'appuie sur un réseau de neurones artificiels qui s'entraînent à identifier des objets, lettres ou visages, selon les données de départ, qui sont donc essentielles à l'entraînement du modèle. Plus les données de départ sont variées, plus le système sera performant. Dans le cas d'HORAE, la reconnaissance des lignes est une étape primordiale, car elle influe sur la qualité de la reconnaissance de texte. Il faut alors tester le modèle selon les types de lignes, qui peuvent être plus ou moins droites sur l'image selon la qualité de la numérisation. En effet, les qualités de numérisation sont disparates, concernant

254. Le *data mining* désigne le procédé pour trouver des corrélations entre de nombreuses bases de données relationnelles. Ces corrélations peuvent prendre la forme d'associations, des modèles au sein desquels un événement est lié à un autre événement ; d'analyse de séquence, soit des situations au sein desquelles un événement mène à un autre événement plus tardif ; de classification et de *clustering* comme dans le cadre du projet HORAE ; ou de prédiction.

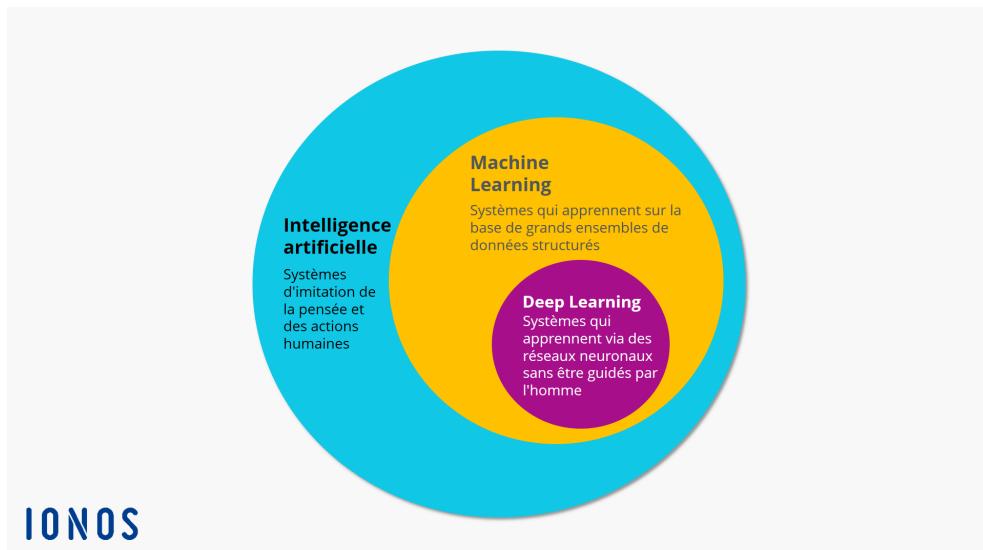


FIGURE 3.2 – L’imbrication du *deep learning*, du *machine learning* et de l’intelligence artificielle. Cf. MARKETING (Search Engine), *Quelles sont les différences entre le Deep learning et le Machine learning ?*, 2020, URL : <https://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/deep-learning-vs-machine-learning/> (visité le 28/08/2020)

parfois le document lui-même, parfois simplement le microfilm.

Le développement de l’interface inter-annotateur Akindex montre l’importance de faire une interface test dans un premier temps pour ensuite comprendre les points à améliorer. En effet, meilleure est l’ergonomie, meilleure est la qualité des données. La mise en place de l’interface Arkindex a permis d’appliquer le protocole d’annotation afin d’établir une véritable terrain pour entraîner la reconnaissance automatique de textes et de sections au sein des livres d’heures.

## 3.2 Le protocole d’annotation pour guider l’apprentissage machine

Pour entraîner le modèle qui doit restituer des données de qualité, une véritable terrain a été établie par les membres de l’IRHT. Elle a été établie grâce à un protocole expliqué dans un manuel d’annotation qui rappelle les niveaux hiérarchiques et les pièces que l’on peut retrouver dans un livre d’heures. Quatre niveaux hiérarchiques ont donc été établis selon les parties structurelles des livres d’heures. Le premier niveau est composé des principales parties du livre d’heures, et contient 74 classes principales. Voici les plus fréquentes :

- *Calendar*
- *Gospel Lections*

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

---

- *Hours of the Cross*
- *Hours of the Holy Spirit*
- *Hours of the Virgin*
- *Hours of the Virgin (Advent)*
- *Hours of the Virgin (Temporale)*
- *Litany of Mary*
- *Mass*
- *O intemerata*
- *Obsecro Te*
- *Office of the Dead*
- *Office of the Dead (Temporale)*
- *Penitential Psalms and Litany*
- *Prayers*
- *Psalter*
- *Suffrages*

Les classes de niveau 2 représentent logiquement une subdivision des contenus des sections de niveau 1. Si les classes de niveau 2 désignant des prières correspondent aux prières les plus couramment présentées dans les livres d'heures<sup>255</sup>, les autres prières sont à annoter au niveau 1, avec la classe « Prayers ». Ce tableau résume l'imbrication des niveaux 1 et 2 :

---

255. Il s'agit des prières suivantes : Les *Quinze Joies de la Vierge*, Les cinq psaumes pour Marie (« *Magnificat anima mea Dominum* » ; *Ad Dominum, cum tribularer, clamavi* » ; « *Retribue servo tuo, vivitica me et custodiam sermones tuos* » ; « *In convertendo Dominus captivitatem Sion* » ; « *Ad te levavi oculos meos qui habitas in coelis* »), les *Sept requêtes à notre Seigneur*, les *Sept vers de saint Bernard*.

Niveau 1	Niveau 2
Calendar	January February March April May June July August September October November December
Gospel Lections	Gospel of John Gospel of Luke Gospel of Matthew Gospel of Mark Passion (St John)
Hours of the Cross Hours of the Holy Spirit Hours of the Virgin Hours of the Virgin (Advent) Hours of the Virgin (Temporale)	Matins Lauds Prime Terce Sext None Vespers Compline Multiple hours
Mass	Preparation Readings Oblation Consecration Communion Conclusion
Office of the Dead Office of the Dead (Temporale)	Vespers Matins Lauds
Prayers	Fifteen Joys of the Virgin Five Psalms for Mary Seven Things which Please God Verses of St Bernard

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

---

Penitential Psalms and Litany	Penitential Psalms Litany
Credo Litany of Mary O intemerata Obsecro Te Office in the chapter Paschal table	
Suffrages	Angeli Anna Antonius Apollonia Barbara Blasius Catharina Christophorus Elisabeth Eutropius Georgius Geraldus Gregorius Jacobus Johannes Baptista Johannes evangelista Laurentius Margareta Maria Magdalena Martinus Martyres Michael Multiple hours Nicolaus Omnes apostoli Omnes sancti Paulus Petrus Petrus et Paulus Roch Sebastianus Stephanus Trinitas Veronica

TABLE 3.3 – Imbrication des niveaux 1 et 2 selon la structure des livres d'heures

Il convient de faire la distinction entre les Heures de la Vierge, les Heures de la

Vierge pour le temps de l'Avent<sup>256</sup> et les Heures de la Vierge dans le Temporal<sup>257</sup>. Elles dépendent de la partie de l'année liturgique. On peut s'appuyer sur le texte et les rubriques pour les distinguer, mais les plus courantes sont les Heures de la Vierge.

Le troisième niveau, ignoré des copistes, signale l'ordonnancement du cursus dans chaque office, et constitue donc une sous-partie des niveaux 2. On y retrouve les pièces liturgiques suivantes :

- *Canticle*<sup>258</sup> ;
- *HSL*<sup>259</sup> ;
- *Hymn*<sup>260</sup> ;
- *Invitatory*<sup>261</sup> ;
- *Invocation*<sup>262</sup> ;
- *Lessons*<sup>263</sup> ;
- *Orationes*<sup>264</sup> ;
- *Preces*<sup>265</sup> ;
- *Psalm*<sup>266</sup> ;
- *Short lesson*<sup>267</sup>.

Toutefois, les heures abrégées de la Croix et du Saint-Esprit ne sont pas annotées avec le niveau 3. Il faut donc annoter le versicule d'invocation dans la section de niveau 2 correspondante, sachant que ces heures ne contiennent pas de laudes. Les annotations de niveau 3 ne concernent en fait que les Heures de la Vierge et l'Office des morts, les deux sections caractéristiques des livres d'heures.

Quant aux pièces de niveau 4, le plus fin niveau de granularité défini dans le cadre du projet, elles désignent principalement les textes élémentaires, dont l'accès est encore à implémenter dans Arkindex. En effet, les pièces de ce niveau sont les plus fécondes en

---

256. Le temps de l'Avent désigne la période qui s'étend du quatrième dimanche précédent Noël à la veille de la fête. Il correspond au début de l'année liturgique.

257. Le temporal désigne l'organisation de l'année liturgique, du premier dimanche de l'Avent au premier dimanche après la Pentecôte. Il s'oppose au sanctoral, qui désigne la célébration des saints.

258. Chant extrait de l'Ancien ou du Nouveau Testament selon où on se trouve dans le cursus.

259. Abréviation pour « *Hymn - Short lesson* », ces deux sections constituent une unité au sein des laudes, des vêpres et des complies.

260. Chant poétique pouvant être rythmé, métrique ou rythmique.

261. L'invitatoire peut désigner l'antienne en introduction des matines ainsi que le psaume qu'il l'accompagne, mais aussi l'antienne en elle-même se concluant par les mots « *Venite adoremus* ».

262. Il s'agit d'une formule introductory de l'office, composée d'un versicule et sa réponse, empruntés aux psaumes 50 et 69 lors des matines et 69 pour les autres heures. Seul l'office des morts n'en possèdent pas.

263. Lecture d'origine scripturaire ou patristique.

264. Conclusion de l'office comprenant une suite d'oraisons et la salutation.

265. Les *preces* désignent la partie de l'office et de la litanie contenant des formules de supplication.

266. Pièce poétique extraite du livre biblique des psaumes. On en trouve 150 dans la tradition latine.

267. Appelée capitule en français, il s'agit d'une courte lecture extraite de la Bible.

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

---

surprises et en découvertes, et sont donc amenées à être ajournées. Elles recouvrent les différents chants, oraisons, prières ou lectures. Par exemple, l'annotation avec la catégorie « *test-class* » a permis de mettre en avant les saints qui n'étaient pas encore répertoriés dans les suffrages et de les ajouter au sein des classes<sup>268</sup>.

L'annotation de ces différentes sections, et plus particulièrement de leur incipit, s'explique par leur caractère révélateur de l'usage liturgique employé. En effet, l'identification de l'usage ne nécessite pas un relevé complet des textes, mais les indices issus du petit office de la Vierge doivent être corroborés par ceux issus de l'office des morts et de la présence de saints locaux dans le calendrier et les litanies. Il suffit alors de relever les incipits des pièces de niveaux 2 et 3 dans le petite office de la Vierge et l'office des morts, puis de les comparer avec ceux établis par le chanoine Leroquais dans Paris, Bibl. nat. de France, n. acq. lat. 3162 (pochette microfilm IRHT n° 43251)<sup>269</sup>.

L'objectif étant d'entraîner le modèle pour la reconnaissance automatique de texte, et *in fine* des sections associées, des stratégies d'évolution ont été confirmées lors de la réunion du 17 juin dernier entre les trois partenaires du projet. Le travail consiste alors à ajouter des transcriptions, mêmes imparfaites<sup>270</sup>, aux psaumes pénitentiels, tout en alignant le texte dans le logiciel Transkribus pour un corpus d'environ 200 images. En effet, les serveurs IIIF ne renvoient pas systématiquement la taille d'image souhaitée, ce qui peut créer des décalages avec les détections de ligne.

En ce qui concerne la création du modèle linguistique, prise en charge par le LS2N, il s'agit de faire apprendre plusieurs fois la Bible comme des textes différents à chaque fois, car les livres d'heures, comme les livres liturgiques de manière générale, contiennent beaucoup de références bibliques. L'idée d'intégrer un dictionnaire de latin dans l'apprentissage a été discutée. En effet, le grand nombre d'abréviations et de variantes orthographiques dans les livres d'heures risque de gêner la détection des caractères si la langue latine est trop normalisée. Toutefois, fournir un texte sans langue détectée est absurde. La solution serait d'utiliser les méthodes du *postprocessing*, afin que la machine soit capable de détecter les anomalies linguistiques, c'est-à-dire ce qui diffère de la langue normalisée. Le *postprocessing*, traduit comme le « post-traitement », est une des composantes de la

---

268. Si les prières à la Vierge *Obsecro te* et *O Intemerata* auraient dû logiquement faire partie du niveau 4, elles ont été placées en niveau 1 du fait de leur présence quasi systématique dans les livres d'heures.

269. J.B. LEBIGUE, *Initiation aux manuscrits liturgiques...*, p. 105-106.

270. La méthodologie, dont on a rappelé les principes dans la section 3.1, consiste à utiliser des données imparfaites, c'est-à-dire qu'un texte de référence est plaqué sur le texte correspondant, même s'il contient d'éventuelles variantes textuelles, afin de parvenir à des conclusions historiques correctes, soit la bonne identification des sections des livres d'heures malgré les variantes textuelles. Les textes de référence établis pour les psaumes pénitentiels sont présentés en annexes, section C.2.

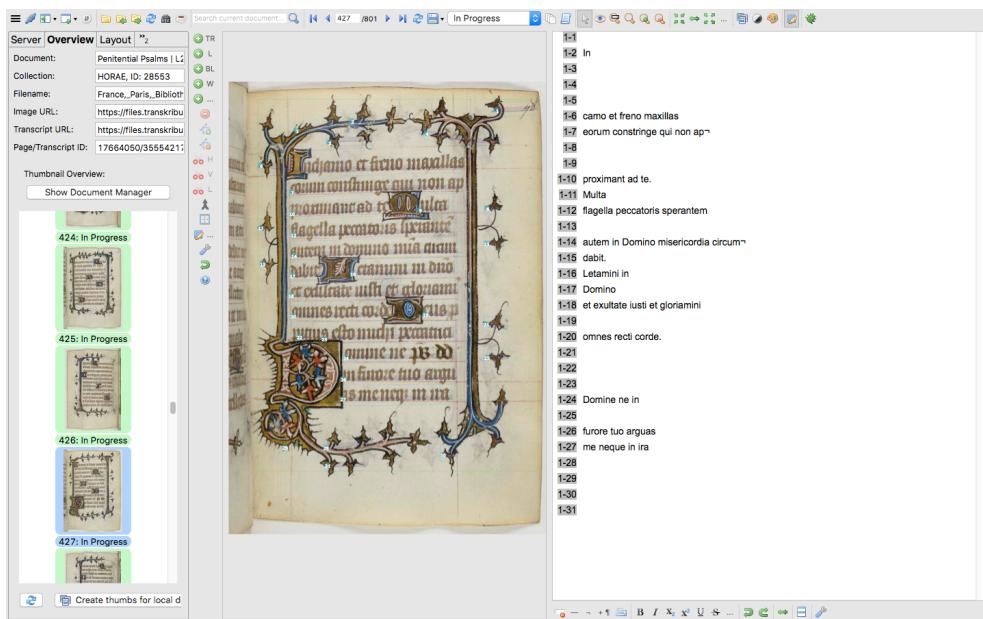


FIGURE 3.3 – Exemple d'un alignement d'un extrait de psaumes pénitentiels sur Transkribus, extrait du manuscrit ms. lat. 01403, f.78r, conservé à la Bibliothèque nationale de France.

*Knowledge Discovery in Databases*, soit la découverte de savoirs dans de larges bases de données<sup>271</sup>. Cette découverte de savoirs est possible grâce à cinq étapes<sup>272</sup> :

1. Déterminer les contours de la problématique, spécifier le but de la découverte afin de choisir le domaine d'application le plus pertinent.
2. Formaliser la représentation de l'objet et collecter les données selon cette représentation.
3. Les données étant souvent incomplètes ou « bruitées »<sup>273</sup>, en somme imparfaites, il est important de les structurer et de les ordonner pour qu'elles deviennent exploitables. Ce travail est appelé « *preprocessing* ».
4. Vient ensuite l'étape du *data mining* pour donner du sens aux données, à l'aide de méthodes statistiques, de réseaux de neurones et d'algorithmes.
5. le *postprocessing* constitue la dernière étape, qui consiste à vérifier la cohérence des traitements effectués, et au besoin, adapter le modèle existant pour diminuer le taux d'erreurs. Le savoir peut alors être interprété et intégré dans celui existant.

Il faut ici rappeler que l'enjeu de ce processus n'est pas d'améliorer la transcription du texte, mais sa détection. On peut ensuite identifier les termes de référence pour le plagiat, grâce à un repérage des similarités entre les phrases.

271. Ivan BRUHA et A. Fazel FAMILI, *Postprocessing in Machine Learning and Data Mining*, New York, déc. 2000, doi : 10.1145/380995.381059, p. 1.

272. *Ibid.*, p. 2.

273. Dans le domaine des données, le bruit désigne un surplus d'informations inutiles dans le cadre des objectifs déterminés.

## 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

---

Il est maintenant intéressant de voir dans quelle mesure le protocole d'annotation et l'établissement de la vérité terrain ont servi la reconnaissance automatique d'images et de textes dans les livres d'heures. Contrairement à la reconnaissance de l'écriture imprimée (*Optical Character Recognition*) où les modèles peuvent être harmonisés, la reconnaissance d'écriture manuscrite (*Handwritten Text Recognition*) représente un enjeu en ce qu'elle implique des modèles spécifiques au type de documents traités. Cela est d'autant plus vrai pour les écritures médiévales, dans une langue éloignée de la nôtre, et offrant des spécificités en terme de formes des lettres et d'usage d'abréviations<sup>274</sup>. Le type d'écriture peut également varié selon l'époque et la zone de provenance du manuscrit. Il faut donc constituer un corpus d'écritures relativement homogènes pour donner plus de pertinence au modèle d'entraînement<sup>275</sup>.

### 3.2.1 Analyser la structure des pages

Avant de s'intéresser plus précisément à la reconnaissance textuelle, il est important d'entraîner le modèle sur la structuration des pages, afin qu'il distingue celles qui contiennent du texte à analyser de celles qui n'en contiennent pas<sup>276</sup>. Quel est l'état de l'art des systèmes de détections de lignes de texte, de zones dans une page de manuscrit ou d'incunable et de leur typologie ? Pour pouvoir procéder automatiquement sur un large panel de collections de livres d'heures hétérogènes, il est indispensable de créer un jeu de données annoté de pages représentatives de la variété des typologies et des présentations<sup>277</sup>. On peut ici rappeler que le corpus dépasse les 500 manuscrits numérisés, disponibles pour la plupart via la BVMM<sup>278</sup> et Gallica, tous deux sous la responsabilité de l'IRHT et de la BnF. Le tableau ci-dessous permet d'avoir une idée de l'ampleur et de la provenance des données à annoter automatiquement :

---

274. Béatrice DAILLE, Amir HAZEM, Christopher KERMORVANT, Martin MAARAND, Marie-Laurence BONHOMME, D. STUTZMANN, Jacob CURRIE et Christine JACQUIN, « Transcription automatique et segmentation thématique de livres d'heures manuscrits », *TAL*, 60–3 (2019), p. 13–36, p. 21.

275. Que l'on songe à la classification de Lief tinck-Gumbert-Derolez, qui distingue plusieurs variantes au sein des écritures gothiques, de la *textualis* à l'*hybrida* en passant par la *cursiva*. Cf. D. STUTZMANN, *Les écritures gothiques livresques : classification de Lief tinck-Gumbert-Derolez*, fr, URL : <https://oriflamms.hypotheses.org/quest-ce-que-la-paleographie/les-ecritures-gothiques-livresques-classification-de-lief tinck-gumbert-derolez> (visité le 29/02/2020).

276. Cette étape d'analyse de la mise en page est communément appelée *Document Layout Analysis*.

277. Mélodie BOILLET, M.L. BONHOMME, D. STUTZMANN et C. KERMORVANT, « HORAE : an annotated dataset of books of hours », dans *The 5th International Workshop on Historical Document Imaging and Processing*, Sydney, 2019 (2019 International Conference on Document Analysis and Recognition (ICDAR)), p. 7-12, DOI : 10.1145/3352631.3352633, p. 7.

278. Bibliothèque Virtuelle des Manuscrits Médiévaux.

PROVIDER	CITY	MSS	IMAGES
ugent.be	Gent	1	142
<b>BVMM</b> <a href="https://bvmm.irht.cnrs.fr">https://bvmm.irht.cnrs.fr</a>		<b>275</b>	<b>41902</b>
	≤ 8 MSS	114	
	Angers	21	
	Autun	12	
	Auxerre	10	
	Beaune	15	
	Chantilly	30	
	Nantes	18	
	Paris	17	
	Rennes	23	
	Toulouse	15	
<b>Gallica</b> <a href="https://gallica.bnf.fr">https://gallica.bnf.fr</a>	<b>Paris, BNF</b>	<b>183</b>	<b>52923</b>
	– Arsenal	38	
	– Manuscrits	145	
Harvard	Cambridge	32	8530
ubc.ca	Vancouver	1	224
stanford.edu	Baltimore	6	2842
wdl.org	Baltimore	2	664
<b>Total</b>		<b>500</b>	<b>107,227</b>

FIGURE 3.4 – Nombre et provenance des images et manuscrits du jeu de données. Cf. BOILLET (Mélodie), BONHOMME (Marie-Laurence), STUTZMANN (Dominique) et KERMORVANT (Christopher), « HORAE : an annotated dataset of books of hours », dans *The 5th International Workshop on Historical Document Imaging and Processing*, Sydney, 2019 (2019 International Conference on Document Analysis and Recognition (IC-DAR)), p. 7-12, DOI : 10.1145/3352631.3352633, p. 8

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

---

L'objectif est donc de traiter ce jeu de données grâce au *machine learning*, d'où le besoin de pages annotées par des experts pour entraîner le modèle. Des images de reliures ou de folios blancs ont également été incluses dans le corpus afin de faire détecter à la machine ce qui contient du texte de ce qui n'en contient pas. Pour rappel, le *machine learning* désigne un système permettant de découvrir des répétitions, des similarités dans un ou plusieurs flux de données, et de construire des prédictions à partir de statistiques. Les premiers algorithmes d'apprentissage automatique ont été conçus dès les années 1950<sup>279</sup>.

Le *machine learning* représente alors une branche de l'intelligence artificielle, mais ne peut être confondu avec cette dernière. L'intelligence artificielle désigne en effet ce qui permet à la machine d'apprendre et de comprendre des concepts en faisant des liens entre les données, en s'inspirant du système neuronal humain. L'intelligence artificielle s'appuie donc ici sur des statistiques et la création de probabilité à partir d'un jeu de données précis, afin qu'à partir d'une forme, le modèle soit capable de déduire de quelle section de page, lettre, mot, ou phrase il peut s'agir. Quant au *deep learning*, il participe à l'intelligence artificielle en ce qu'il permet à la machine d'apprendre par elle-même au fur et à mesure qu'on lui fournit des données de qualité, sans avoir besoin d'être entraînée, mais il ne recouvre pas l'ensemble des techniques inhérentes à l'intelligence artificielle<sup>280</sup>.

En ce qui concerne la classification automatique des images de livres d'heures, elles ont été classées à partir des classes définies ci-dessous à partir d'un réseau de neurones d'apprentissage profond<sup>281</sup> :

- *page*<sup>282</sup>.
- *decorated\_border* et *illustrated\_border* pour les bordures ; la première catégorie concerne les bordures avec des ornements sans signification particulière (par exemple les motifs floraux ou végétaux), tandis que la deuxième se réfère aux illustrations signifiantes (scènes inspirées du texte biblique par exemple).
- *miniature*.
- *text* pour les zones de textes sur la page, donc si une page est divisée en deux colonnes, deux zones de texte sont annotées.
- *border\_text* pour le texte en marge dans les bordures.
- Les initiales peuvent être annotées avec trois classes, *decorated\_initial* pour les initiales ornées, *simple\_initial* pour les initiales sans décoration particulière, seulement un peu plus épaisses que le reste du texte, *historiated\_initial* pour

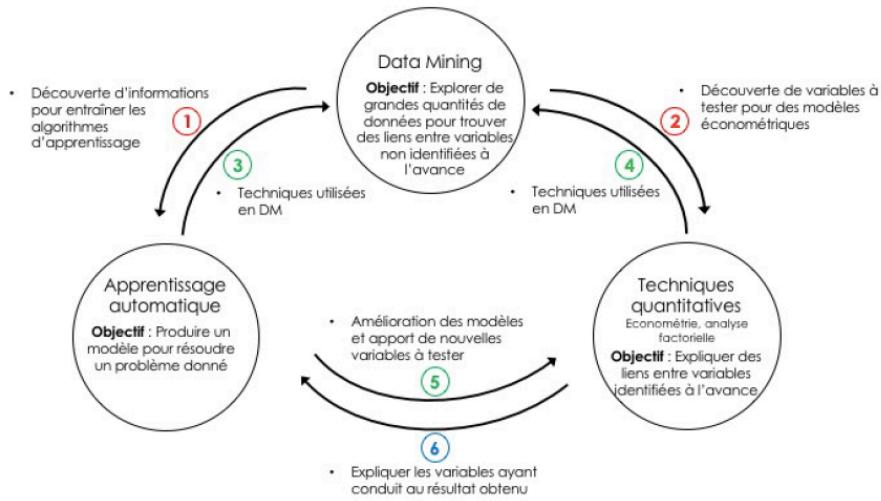
---

279. Le plus célèbre est le Perceptron, mis au point par Frank Rosenblatt en 1957. Il s'agit d'un classifier linéaire qui consiste à classer ensemble des échantillons qui ont des propriétés communes, et de séparer ceux qui se différencient.

280. Thomas MEIMOUN, URL : <https://www.quantmetry.com/intelligence-artificielle-data-science-automatisation-machine-learning-suffisait/> (visité le 14/07/2020).

281. Les classes de subdivision d'une page sont documentées dans le cadre du projet à l'adresse suivante : [https://gitlab.com/mlbonhomme/arkindex\\_annotation/-/tree/master/HORAE](https://gitlab.com/mlbonhomme/arkindex_annotation/-/tree/master/HORAE).

282. Si une image présente une double page, il est important d'annoter chaque page individuellement.



Farchy, Denis, 2020

FIGURE 3.5 – Schéma des liens entre les diverses approches d'analyse des données. Cf. BENSAMOUN (Alexandra) et FARCHY (Joëlle), *MISSION INTELLIGENCE ARTIFICIELLE ET CULTURE, Rapport final*, rapp. tech., Conseil supérieur de la propriété littéraire et artistique (CSPLA), 2020, p. 13

les initiales historiées.

- Les autres éléments décoratifs sont répartis entre *line\_filler*, soit les fins de ligne ornementées, ou bien *ornament* pour les décos à l'intérieur du texte qui ne sont pas des miniatures ; il peut s'agir des notations musicales.
- la classe *text\_line* définit les zones de texte ligne par ligne.



FIGURE 3.6 – Exemple d'initiales à annoter avec la classe *historiated\_initial*

Cette méthodologie appartient donc au domaine du *deep learning*. Sous-domaine du *machine learning*, il est constitué de réseaux de neurones convolutionnels profonds, à l'image du fonctionnement du cerveau humain<sup>283</sup>. Les successions de neurones prennent

283. Edouard OYALLON, *Analyzing and Introducing Structures in Deep Convolutional Neural Net-*

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE



FIGURE 3.7 – Exemple d'initiales à annoter avec la classe *decorated\_initial*

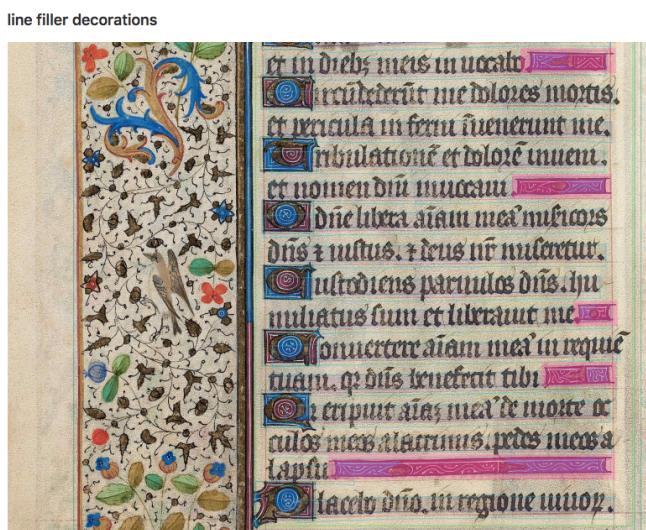


FIGURE 3.8 – Exemple d'élément décoratif avec la classe *line\_filler*

en entrée les sorties des couches de neurones précédentes, afin que la machine apprenne d'elle-même. Les premières approches en *deep learning* datent des années 1980, avec la mise au point d'un type d'algorithme appelé *Convolutional neural network*<sup>284</sup>. Pour être efficace, l'apprentissage profond doit donc fonctionner avec un volume de données bien plus considérable que les réseaux de *machine learning*, ce qui lui permet de réaliser des tâches plus complexes et plus précises.

Une fois les pages classifiées automatiquement, la deuxième étape consiste à les regrouper selon leurs ressemblances. L'objectif est ainsi de détecter les pages qui sortent du lot, qui présentent un agencement suffisamment rare pour former un groupe à part<sup>285</sup>. À partir de cette détection, un échantillon de pages à annoter a été sélectionné afin de

works, Theses, Paris Sciences et Lettres, 2017, URL : <https://hal.archives-ouvertes.fr/tel-02353134> (visité le 20/08/2020), p. 68-69.

284. *Ibid.*, p. 1-3.

285. M. BOILLET, M.L. BONHOMME, D. STUTZMANN, *et al.*, « HORAE : an annotated dataset of books of hours »..., p. 8.

représenter dans l'apprentissage la diversité des mises en page. Parmi les 600 images de l'échantillon, 141 correspondent aux dispositions les plus fréquentes, et 459 aux dispositions les plus rares et variées<sup>286</sup>.

Le processus d'annotation s'est ensuite effectué grâce au logiciel Transkribus, développé dans le cadre du projet européen READ<sup>287</sup>. Si une partie des classes doit être annotée manuellement afin d'établir une vérité terrain, les *text lines* et les *text regions* sont annotées automatiquement par Transkribus, bien qu'une vérification s'impose. En effet, le logiciel peut être sensible à la couleur de l'encre et donc mal distinguer les lignes si cette couleur est semblable à celle du parchemin. La vérité terrain est ici déterminée grâce à un corpus inter-annotateur de 10 images, où chaque annotateur travaille sur le même corpus, afin de mesurer les éventuels écarts et désaccords dans l'annotation, mais aussi de voir ce qui fait consensus.

Au vu de la variété des classes et des pages, il a fallu choisir un outil d'analyse automatique flexible mais aussi capable d'uniformisation. C'est alors le réseau *dhSegment* qui a été adopté, car cet outil a fait preuve de résultats satisfaisants pour le traitement de documents historiques. Il présente notamment les avantages de travailler avec une petite quantité de données d'entraînement, et de réaliser des opérations variées, comme l'extraction de lignes et l'analyse de la mise en page<sup>288</sup>. En s'appuyant sur les réseaux de neurones, *dhSegment* propose des solutions pour découper une page de manuscrit, extraire des enluminures du texte, localiser le texte dans une image numérisée, entre autres choses. Ces problématiques étaient traitées à l'aide de méthodes de segmentation disparates et hétérogènes, s'appliquant à des typologies de documents précises<sup>289</sup>.

Utilisant l'outil d'apprentissage automatique *TensorFlow*, notamment employé pour la détection d'objets, la segmentation se fait en deux étapes. La première s'appuie sur les réseaux de neurones convolutionnels qui prennent une image du document en entrée et rendent en sortie une cartographie des probabilités des caractéristiques prédictives attribuées à chaque pixel. La deuxième étape consiste à transformer les prédictions en la sortie souhaitée pour la tâche à effectuer<sup>290</sup>.

L'intérêt de cette méthode est de limiter les tâches de la deuxième étape relative au post-traitement à de simples opérations de prédiction. La première consiste à établir des critères selon les classes à trouver, afin que les prédictions divisent le corpus selon les classes à trouver. La deuxième opération relève de considérations morphologiques afin d'analyser les structures géométriques du document. Les deux dernières étapes recouvrent

---

286. *Ibid.*

287. Cf. <https://read.transkribus.eu/e-learning/>.

288. *Ibid.*, p. 9.

289. *dhSegment : A generic deep-learning approach for document segmentation*, 2019, DOI : 10.1109/ICFHR-2018.2018.00011, p. 1.

290. *Ibid.*, p. 1-2.

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

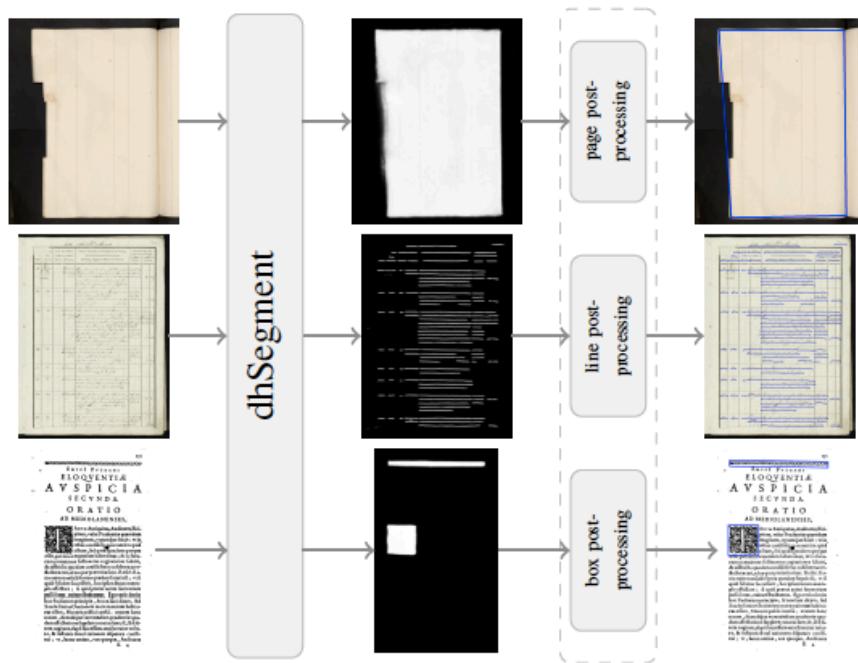


FIGURE 3.9 – La méthode de segmentation *dhSegment* développée par Sofia Ares Oliveira et Benoît Seguin. Cf. *dhSegment : A generic deep-learning approach for document segmentation*, 2019, DOI : 10.1109/ICFHR-2018.2018.00011, p. 2

l’analyse des composants connectés, afin de filtrer ceux qui resteraient après les deux premières étapes, puis la vectorisation, technique pour que plusieurs opérations soient traitées simultanément, afin que les zones détectées soient rassemblées en un ensemble de coordonnées<sup>291</sup>. En adaptant ces critères au cas souhaité, on peut entraîner le corpus. Les résultats de la première expérimentation du *dhSegment* témoignent de la possibilité pour un outil construit à partir de l’architecture générique du *deep learning* de s’appliquer à des tâches de segmentation spécifiques tout en utilisant un processus standardisé. Cela signifie que des bouts de programmation peuvent être entraînés par des non-spécialistes, et être efficaces pour de documents divers et variés<sup>292</sup>.

Dans le cadre du jeu de données composé de livres d’heures, les résultats de cette méthode sont globalement satisfaisants, mais peuvent être améliorés. En effet, l’indice d’*Intersection-over-Union*, soit le pourcentage de recouvrement de la zone établie par la vérité terrain par la zone de prédiction détectée automatiquement est entre 80% et 90% pour les lignes de texte et entre 60% et 80% pour l’agencement de la page<sup>293</sup>. Les données bénéficieraient donc à être traitées avec un réseau de neurones plus complexe.

291. *Ibid.*, p. 2.

292. *Ibid.*, p. 6.

293. Le pourcentage varie selon la taille de la page d’entrée, cf.Id., « HORAE : an annotated dataset of books of hours »..., p. 10.

Une des prochaines étapes concerne la reconnaissance des couleurs. Il faudrait alors entraîner le modèle avec les class « *red* », « *blue* », « *gold* » et « *green* » et analyser la variation de pixels au sein d'un même manuscrit. Cela serait intéressant pour avoir par exemple une vision du pourcentage de rubriques dans les livres d'heures. Toutefois, ce processus exclut les manuscrits numérisés en noir et blanc.

Une des parties richement colorées est justement celle formée du calendrier.

### 3.2.2 Le cas spécifique des calendriers

Au cours de la réflexion sur la classification des pages, la classe *Calendar* a été ajoutée pour les calendriers. En effet, si l'on annote ligne par ligne la structure tabulaire des calendriers, on perd l'intelligence des données. Les pages contenant des calendriers répondent ainsi à une structuration spécifique et à une typologie précise, et leur analyse dans le cadre des livres d'heures est précieuse pour comprendre les types d'usages et de dévotions auxquels nous sommes confrontés. Un calendrier peut ainsi mentionner des fêtes et des saints donnant des indications à la fois sur la période dans laquelle il a été composé (un saint peut avoir été récemment canonisé), comme sur sa destination (avec la mention d'un saint local par exemple)<sup>294</sup>. Riche en informations, les calendriers contiennent généralement les éléments suivants :

- les calendes indiquent le premier jour de chaque mois ;
- les nones indiquent le 5 du mois, sauf pour les mois de mars, mai, juillet et octobre où elles indiquent le septième jour du mois ;
- les ides indiquent les 13 du mois, sauf pour mois de mars, mai, juillet et octobre où elles sont placées au 15 du mois ;
- le nombre d'or, sous la forme de chiffre romain inscrit à gauche en face de certains jours<sup>295</sup> ;
- la *littera dominicalis*, la lettre indiquant le dimanche, renouvelée tous les ans<sup>296</sup> ;
- le *terminus*, soit la dernière date possible pour une fête mobile ;
- l'indication *claves pasche*, qui détermine le jour à partir duquel on commence le décompte pour la date de Pâques, dépendant elle-même du cycle métonique.

On peut également y trouver d'autres indications d'ordre astronomique, avec les nombres d'heures de jour et de nuit dans un mois, la mention des équinoxes ou des embolismes<sup>297</sup> mais aussi l'indication de jours considérés comme néfastes pour certaines

---

294. T. HEIKKILA et T. ROOS, « Quantitative methods for the analysis of medieval calendars »..., p. 767.

295. Ce chiffre indique les nouvelles lunes pour le calcul du cycle métonique, soit une période d'environ 19 ans où les cycles lunaires et solaires s'alignent.

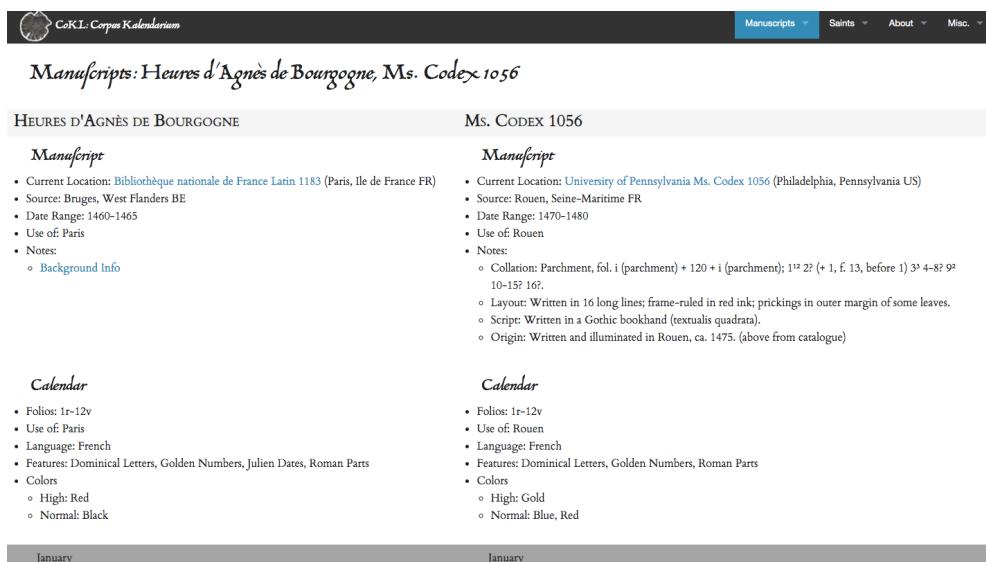
296. Chaque jour de la semaine est ainsi associé à une lettre, de A à G

297. Il s'agit du nombre de jours de différence entre l'année solaire et l'année lunaire.

## 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

activités<sup>298</sup>, les signes du zodiaque ou des indications sur les activités du mois. L'analyse d'un grand nombre de calendriers, de leurs points communs ou de leur différences selon les époques, permet ainsi de mieux visualiser l'émergence de nouveaux cultes comme leur propagation<sup>299</sup>.

À cet égard, une base de données relationnelle d'après les travaux de Aaron Macks rassemble un corpus de calendriers de livres d'heures. Il s'agit du site *CoKL : Corpus Kalendarium*<sup>300</sup>. Il est ainsi possible d'y faire des recherches par noms de saints ou par dates de fêtes et de voir dans quels livres d'heures ils sont mentionnés. On peut également comparer les calendriers de divers manuscrits en mettant en regard les usages, les langues, la foliation, ou encore les transcriptions de calendriers.



The screenshot shows the CoKL website interface. At the top, there's a navigation bar with links for Manuscripts, Saints, About, and Misc. Below the navigation, there are two main sections: 'HEURES D'AGNÈS DE BOURGOGNE' and 'Ms. CODEX 1056'. Each section has a 'Manuscript' tab and a 'Calendar' tab. The 'Manuscript' tab for HEURES D'AGNÈS DE BOURGOGNE lists: Current Location: Bibliotheque nationale de France Latin 1183 (Paris, Ile de France FR); Source: Bruges, West Flanders BE; Date Range: 1460-1465; Use of: Paris; Notes: Background Info. The 'Manuscript' tab for Ms. CODEX 1056 lists: Current Location: University of Pennsylvania Ms. Codex 1056 (Philadelphia, Pennsylvania US); Source: Rouen, Seine-Maritime FR; Date Range: 1470-1480; Use of: Rouen; Notes: Collation: Parchment, fol. i (parchment) + 120 + i (parchment); 1<sup>er</sup> 2<sup>e</sup> (+ 1, f. 13, before 1) 3<sup>er</sup> 4-8? 9<sup>e</sup> 10-15? 16<sup>e</sup>. Layout: Written in 16 long lines; frame-ruled in red ink; prickings in outer margin of some leaves. Script: Written in a Gothic bookhand (textualis quadrata). Origin: Written and illuminated in Rouen, ca. 1475. (above from catalogue). The 'Calendar' tabs for both manuscripts show January pages with specific folio ranges and layout details.

Les pages contenant un calendrier mériteraient donc à être détectées avec leur spécificité, sous la forme d'une structuration tabulaire elle-même divisée en zones de texte. La détection de la structuration des pages et leur classification est ainsi un préliminaire indispensable à la reconnaissance automatique de caractères.

### 3.2.3 Analyser le contenu textuel

Le projet HORAE met en avant des méthodes et traitements automatiques spécifiquement adaptés à la structure des livres d'heures. Ces méthodes utilisent l'approche de segmentation semi-supervisée afin de mieux retrouver leur structure malgré le bruit engendré par la reconnaissance d'écriture<sup>301</sup>.

298. Ce jours peuvent être signalés par un « D » à droite dans la marge.

299. *Ibid.*, p. 768.

300. Cf. <http://www.cokldb.org/>.

301. B. DAILLE, A. HAZEM, C. KERMORVANT, *et al.*, « Transcription automatique et segmentation thématique de livres d'heures manuscrits »..., p. 13.

FIGURE 3.10 – Exemple de visualisation comparative de calendriers d’après la base de données relationnelle *CoKL : Corpus Kalendarium*.

Une fois les lignes de textes localisées dans chaque image contenant une classe de type « page », elles sont extraites afin que le système de reconnaissance d’écriture soit appliqué sur chaque imagette de la ligne. La reconnaissance d’écriture comprend alors deux étapes<sup>302</sup> :

1. l’application d’un modèle optique qui reconnaît des caractères, fragments de caractères ou de mots ;
2. l’application d’un modèle de langue qui détermine les séquences de caractères et de mots les plus vraisemblables.

Pour la reconnaissance des caractères dans les livres d’heures, les membres du projet ont choisi d’utiliser la librairie logicielle KALDI<sup>303</sup>. Grâce à la généralisation de l’utilisation de réseaux de neurones profonds, l’outil, initialement prévu pour la reconnaissance de la parole, est aisément adapatable à la reconnaissance d’écriture<sup>304</sup>. Grâce à Kaldi Horae Recognizer, on est capable de calculer le vocabulaire commun à la même section de plusieurs manuscrits. Ainsi, le 23 juillet dernier a été confirmé le fait que 40% à 48% du vocabulaire des 247 lignes transcrives de la prière *Obsecro te* issue d’un corpus de dix manuscrits est commun à la Bible.

Le système repose sur la combinaison de réseaux de neurones profonds d’une part, et de modèles de Markov cachés d’autre part. Si le fonctionnement des réseaux de neurones rejoint ce qui a été fait pour la détection des zones dans une image, un Modèle de Markov

302. *Ibid.*, p. 22.

303. *Kaldi Speech Recognition Toolkit*. La documentation et le code source sont disponibles à l’adresse suivante : <https://github.com/kaldi-asr/kaldi>.

304. *Ibid.*, p. 23.

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

---

Caché (MMC) est un modèle statistique avec des paramètres inconnus. Il est un des premiers systèmes capables de transcrire complètement une image de mots ou de ligne<sup>305</sup>. À partir d'un état donné, le système calcule des probabilités de transition et de valeur de sortie. Ainsi, le modèle optique, capable de reconnaître des formes et d'en déduire des caractères, est composé de plusieurs couches de réseaux de neurones convolutionnés ainsi que de couches TDNN<sup>306</sup> pour modéliser les caractères en contexte. Les prédictions sont ensuite utilisées par le modèle HMM qui modélise les mots comme des séquences de caractères. Les séquences de mots sont alors modélisées par un modèle de langue statistique de type n-gramme<sup>307</sup>. Un n-gramme est une sous-séquence de  $n$  éléments, qui prend la forme de caractères ou de mots, afin de définir l'ensemble des enchaînements de caractères ou de mots possibles<sup>308</sup>. À partir d'un corpus d'apprentissage, le modèle cherche les probabilités que la prochaine lettre soit telle lettre avec un historique de taille  $n$ . Toutefois, ce modèle présente le risque de mener à des sur-représentations ou à des sous-représentations, car plus le pas  $n$  est grand, plus le nombre d'éléments est grand, et plus on risque d'obtenir des enchaînements rares<sup>309</sup>.

Le modèle a été entraîné à partir des corpus de manuscrits médiévaux formés dans le cadre de précédents projets, notamment ORIFLAMMS<sup>310</sup> et ECMEN<sup>311</sup>, constituant des corpus en latin et en ancien français. Dans le cadre d'HORAE, le modèle a été entraîné sur 247 lignes transcrrites manuellement issues des *Obsecro te* de huit livres d'heures du corpus cible. Si les taux d'erreurs mots (WER pour *Word Error Rate*<sup>312</sup>) et les taux d'erreurs caractères (CER pour *Character Error Rate*<sup>313</sup>) restent assez élevés, ils n'empêchent pas l'identification des textes<sup>314</sup>.

En effet, la reconnaissance des écritures manuscrites sert dans le projet HORAE l'identification automatique des textes. Cette identification se fait par la segmentation de textes, qui recouvre trois tâches distinctes<sup>315</sup> :

- la segmentation thématique, où les segments sont homogènes ;

305. Luc MIOULET, *Reconnaissance de l'écriture manuscrite avec des réseaux récurrents*, thèse de doct., Université de Rouen, 2015, URL : <https://hal.archives-ouvertes.fr/tel-01301728> (visité le 28/07/2020), p. 66.

306. *Time Delay Neural Networks*.

307. B. DAILLE, A. HAZEM, C. KERMORVANT, *et al.*, « Transcription automatique et segmentation thématique de livres d'heures manuscrits »..., p. 24.

308. L. MIOULET, *Reconnaissance de l'écriture manuscrite avec des réseaux récurrents...*, p. 66.

309. *Ibid.*, p. 65.

310. *Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts*.

311. *Écriture Médiévale et outils Numériques*.

312. Cette métrique permet de calculer le taux d'erreurs par mot, c'est-à-dire qu'elle indique le taux de mots mal reconnus en se calquant sur un texte de référence.

313. Cette métrique renvoie au taux d'erreurs par caractère, calculer sur le même principe que le WER.

314. B. DAILLE, A. HAZEM, C. KERMORVANT, *et al.*, « Transcription automatique et segmentation thématique de livres d'heures manuscrits »..., p. 25.

315. *Ibid.*, p. 26.

- l'identification thématique, où des thèmes sont assignés aux segments ;
- le suivi thématique, où sont établies des relations entre les thèmes des segments, notamment de nature hiérarchique.

Pour segmenter un texte et détecter les changements de thèmes, les principales approches utilisées reposent sur l'analyse lexicale ou le calcul de la cohésion lexicale<sup>316</sup>. Toutefois, jusqu'à présent, les segmentations de textes linéaires ou hiérarchiques se sont appliquées à des textes scientifiques, narratifs ou à des dialogues écrits et retranscrits. Il s'agit ici de segmenter un texte manuscrit dont la reconnaissance de l'écriture a présenté un certain taux d'erreurs<sup>317</sup>. C'est donc une approche semi-supervisée qui a été privilégiée, fondée sur une représentation des parties des livres d'heures par prolongement de mots.

L'approche semi-supervisée est une classe de techniques de l'apprentissage automatique qui s'appuie sur des données annotées et non annotées. Il constitue donc un entre-deux entre l'apprentissage supervisé, qui ne peut se faire qu'à partir d'un ensemble de données annotées, ce qui demande un travail humain indispensable pour préparer les données, et l'apprentissage non-supervisé, ne se nourrissant que de données non annotées, ce qui implique que la machine détecte par elle-même les structures sous-jacentes des données d'entrée. La particularité du projet HORAE est, non pas d'utiliser des similarités lexicales entre deux blocs d'un même document, mais entre le document et une base de références externes contenant les textes préalablement annotés. Il s'agit donc d'aligner les textes de référence des livres d'heures et les textes transcrits, découpés arbitrairement en blocs distincts<sup>318</sup>.

Par rapport à l'état de l'art dans les méthodes de segmentation, l'approche proposée est une de celles qui donnent les meilleurs résultats. En terme d'analyse lexicale, le LS2N nous apprend par exemple que 45% à 50% du contenu de la Bible est présent dans les 10 livres d'heures analysés issus du corpus inter-annotateur, sachant que, en moyenne, un tiers du livre d'heures est composé de psaumes. Cette approche semi-supervisée produit donc des résultats encourageants pour les deux premiers niveaux hiérarchiques, même à partir d'une transcription imparfaite.

Lors des réunions entre les partenaires, des pistes sont abordées pour améliorer l'HTR. Par exemple, si les lignes sont globalement courtes dans les livres d'heures<sup>319</sup>, cela s'explique par sa position de manuscrit de luxe pour les gens non-spécialistes de la lecture. Les scribes évitent donc de couper les mots. Toutefois, si la situation se rencontre, il faut penser à fusionner les fins de lignes et les débuts de lignes pour améliorer la reconnaissance de vocabulaire.

Que nous apporte, d'un point de vue historique et anthropologique, la reconnaissance

---

316. *Ibid.*, p. 27.

317. *Ibid.*, p. 28.

318. *Ibid.*, p. 29.

319. On estime la longueur des lignes de livres d'heures à moins de sept mots par ligne en moyenne.

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

---

	Niveaux de segmentation					
	Niveau 1		Niveau 2		Niveaux 1 et 2	
Approche	P <sub>k</sub>	WD	P <sub>k</sub>	WD	P <sub>k</sub>	WD
TextTiling	66,9	99,9	48,1	60,4	46,0	57,5
C99	68,7	96,8	60,0	676	59,2	66,1
U00	23,6	39,5	38,0	39,4	35,6	38,7
MinCut	40,9	49,2	48,4	52,1	45,2	48,7
HierBays	14,2	25,3	36,7	39,9	32,9	38,5
TopicTiling	60,3	87,0	42,0	48,3	42,0	47,4
Approche proposée	27,2	33,5	29,9	31,4	31,4	32,6

FIGURE 3.11 – Analyse des différentes méthodes de segmentation pour les niveaux 1 et 2 de deux livres d'heures, dont l'un a une transcription imparfaite. P<sub>k</sub> et WD indiquent les taux d'erreurs. Cf. DAILLE (Béatrice), HAZEM (Amir), KERMORVANT (Christopher), MAARAND (Martin), BONHOMME (Marie-Laurence), STUTZMANN (Dominique), CURRIE (Jacob) et JACQUIN (Christine), « Transcription automatique et segmentation thématique de livres d'heures manuscrits », *TAL*, 60–3 (2019), p. 13–36, p. 32

automatique d'écritures et de textes d'un large corpus de livres d'heures ?

#### 3.2.4 Des données structurées en masse : quand le numérique sert les humanités

Il s'agit de contextualiser et de saisir le caractère heuristique de l'utilisation des technologies numériques dans l'étude des livres d'heures, et au-delà, des pratiques dévotionnelles intimes des populations occidentales du Moyen Âge tardif. En effet, la constitution d'un corpus de livres d'heures pour l'annotation atteignant 794 livres, numérisés et segmentés, apporte des possibilités d'analyses et d'interprétations nouvelles. Ces livres sont conservés en France, en Suisse, en Allemagne, aux États-Unis, au Canada, en Grande-Bretagne, au Vatican ou en Belgique, dans des bibliothèques, des services d'archives ou encore des musées.

L'analyse textuelle d'un grand nombre de livres d'heures permet en premier lieu de cerner leur composition, avec les grandes tendances que l'on retrouve quasi systématiquement et les variations, qui peuvent témoigner d'usages et de dévotions particuliers, propres à un temps donné, à une région géographique ou tout simplement à un intérêt personnel pour tel ordre religieux, tel saint, etc. Si chaque livre d'heures correspond à l'usage d'un diocèse, il s'inspire des grandes dévotions catholiques, celles de la Vierge

Marie, de la Croix, du Saint-Esprit, ainsi que du culte des saints et des morts<sup>320</sup>. Plus la liste de manuscrits et incunables analysés automatiquement s'agrandit, plus la liste des classes au sein des différents niveaux hiérarchiques de l'annotation s'affine. Cela est particulièrement visible pour la liste des suffrages, dont on compte jusqu'à présent environ 191 classes différentes<sup>321</sup>.

Les suffrages sont composés d'une antienne, d'un verset et d'une oraison. Ils sont récités après les vêpres ou les laudes, en l'honneur de Dieu ou de saints. De manière générale, ils débutent par la sainte Trinité, puis viennent dans un ordre hiérarchisé la sainte Vierge, saint Michel, saint Jean-Baptiste, les apôtres, les martyrs, les confesseurs et les saintes. On peut ainsi y détecter les saints les plus invoqués au Moyen Âge, particulièrement contre les maladies d'après Victor Leroquais. Si leur ordre d'apparition par catégories est commun, le choix des saints célébrés varie. Cet indice de dévotions particulières interroge sur la ou les personne(s) à l'origine du choix des saints vénérés : s'agit-il de l'éditeur ? Du copiste ? D'une entente entre le transcriveur et le miniaturiste ? Du chef d'atelier ? Du destinataire du livre d'heures<sup>322</sup> ?

Toujours d'après le chanoine Leroquais, l'une des parties les plus révélatrices du livre d'heures sont les prières<sup>323</sup>. En effet, d'ordre privé et extra-liturgique, elle est « celle qui a jailli spontanément de l'âme populaire, qui a traduit à un moment donné ses besoins et ses aspirations »<sup>324</sup>. Elles représentent une communication immédiate avec Dieu ou les saints. Du point de vue de leur analyse, il serait intéressant de constater la part des prières en latin et en langues vernaculaires, ainsi que leur date approximative selon leurs formules. Si la plupart sont d'auteurs anonymes, les auteurs mentionnés font souvent plus figures d'autorité que de figures historiques véridiques<sup>325</sup>. Au-delà du potentiel choix du destinataire, la comparaison entre ces prières, leur fréquence au sein des livres d'heures, permet de mesurer des tendances au sein de la piété populaire<sup>326</sup>.

La question des usages liturgiques est particulièrement complexe dans les livres d'heures. On trouve par exemple plus de 200 versions différentes du petit office de la Vierge dans les livres d'heures conservés à la BnF et observés par Leroquais<sup>327</sup>. Contrairement aux autres livres liturgiques, l'usage des offices ne reflète pas nécessairement le lieu de production et de destination, d'autant plus que les usages de Rome et de Paris sont les plus répandus. Les usages présents dans les livres d'heures, parfois différents d'une pièce à l'autre au sein d'un même manuscrit, peuvent s'inscrire dans des usages d'origine monas-

---

320. V. LEROQUAIS, *Les livres d'heures manuscrits de la Bibliothèque nationale....*, p. VII-IX.

321. Une liste de ces classes propres aux suffrages est disponible en annexes section C.1.

322. *Ibid.*, p. XXI-XXII.

323. Les listes des messes et offices votifs découverts dans les livres d'heures du corpus sont disponibles en annexes, section C.1.

324. *Ibid.*, p. XXIX.

325. *Ibid.*

326. *Ibid.*, p. XXX.

327. *Ibid.*, p. XXXVI.

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

---

tique, d'ordres religieux réguliers et séculiers. Parmi ceux observés par Victor Leroquais se trouvent :

- l'ordre de Cluny, ordre monastique fondé en 910 ;
- les Chartreux, ordre monastique fondé en 1082 par Bruno de Cologne ;
- les Cisterciens, ordre monastique fondé en 1098 par Robert de Molesme ;
- les Guillemites, ordre monastique fondé au XII<sup>e</sup> siècle par Guillaume de Maléval ;
- les Prémontrés, ordre de chanoines réguliers fondé au XII<sup>e</sup> siècle par saint Norbert ;
- Les Franciscains, ordre mendiant mineur fondé en 1210 par François d'Assise ;
- Les Dominicains, ordre mendiant fondé en 1215 par Dominique Guzman ;
- les Ermites de saint Augustin, ordre mendiant fondé en 1243 ;
- les Célestins, ordre bénédictin fondé en 1248 par Pierre de Morrone, futur Pape Célestin V ;

Les usages de ces différents ordres se distinguent par des célébrations importantes de saints particuliers, notamment celui portant le nom du fondateur, étant parfois devenu lui-même un saint - ce qui par ailleurs est un bon indicateur pour la datation du manuscrit, selon la date de canonisation du fondateur, mais aussi pour les jours de dédicaces des édifices religieux ou de la réception de reliques, la *translatio*<sup>328</sup>.

Toutefois, l'utilisation de données entièrement numériques peut aussi limiter l'étude de certains aspects, notamment codicologiques. Par exemple, si l'on peut voir les réclames<sup>329</sup> sur les pages numérisées, il est plus délicat de faire de la collation<sup>330</sup>, car cela demande une approche physique pour mieux se rendre compte de certains paramètres.

En effet, dans le cas des livres d'heures, une approche anthropologique qui étudie le livre avant tout sous l'angle de son statut matériel d'objet peut se révéler particulièrement pertinente, d'autant plus pour ce type d'objet relevant du quotidien et de l'intimité. Restitué dans son contexte, l'objet apparaît comme un moyen d'influence des pensées et des actions, notamment dans l'expression de pratiques dévotionnelles et de cultes<sup>331</sup>. Les livres d'heures constituent par ailleurs des objets en mouvement et en métamorphose, qu'ils soient remaniés pour l'ajout d'un saint récemment canonisé, ou bien amputés de

---

328. Ces réflexions sont inspirées d'une formation continue de l'École des Chartes, « Les livres liturgiques manuscrits et imprimés : principes de catalogage », donnée par Laura Albiero du 2 au 4 septembre 2020.

329. « Indication des premiers mots de la page suivante inscrite au bas d'une page, le plus souvent à la jonction entre deux cahiers, permettant de contrôler la bonne succession des feuillets ou cahiers. ». Institut de recherche et d'histoire des textes, *Codicologia*, 2011, URL : <http://codicologia.irht.cnrs.fr/>.

330. Technique consistant à calculer le nombre de feuillets ou de cahiers que comporte un volume.

331. Michèle COQUET, « Alfred Gell, Art and Agency. An Anthropological Theory », *L'Homme*, 157 (2007), p. 261-263, URL : <http://journals.openedition.org/lhomme/5658> (visité le 20/06/2020), p. 262.

feuillets enluminés, comme nous l'avons vu dans les travaux de Kathryn M. Rudy<sup>332</sup>.

Un autre aspect des livres d'heures pourrait être étudié, si les avancées en reconnaissance automatique de caractères le permet : la musique. Si les livres liturgiques sont indissociables du chant, et donc riches en notations musicales, il faut toutefois rappeler que cela est plus rare dans les livres d'heures<sup>333</sup>. L'identification de traditions mélodiques peut ainsi être un indice de plus pour identifier la diffusion d'usages particuliers, tout comme la notation musicale peut indiquer une datation et une provenance géographique plus précise du livre<sup>334</sup>.

L'usage du numérique dans les SHS est de plus en plus prégnant. Xavier Darcos, chancelier de l'Institut de France, parle de « déplacement des frontières »<sup>335</sup>, à la fois car le numérique encourage la transdisciplinarité, mais aussi parce que le double enjeu des découvertes scientifiques et de l'amélioration des possibilités techniques se nouent. Il renouvelle également l'accès de tous au patrimoine. Il est vrai que la recherche, l'histoire médiévale en particulier, a toujours eu des rapports étroits avec l'ordinateur, que l'on songe à la revue *Le Médiéviste et l'Ordinateur*, née en 1979. De plus, le « tournant numérique » concerne l'ensemble des activités humaines, et pas que les humanités<sup>336</sup>. Le projet HORAE s'inscrit donc dans cette mouvance qui fait du numérique à la fois un instrument de recherche, un outil de communication et un objet de recherche. L'utilisation de plus en plus sophistiquée du numérique dans les humanités a donné lieu au terme anglo-saxon « *digital humanities* », terme popularisé avec la parution du livre *A Companion to Digital Humanities*. D'après le *Manifeste des digital humanities*, rédigé en 2010 à Paris, cosigné par plus de 250 chercheurs et 10 institutions,

[...] les *digital humanities* concernent l'ensemble des Sciences humaines et sociales, des Arts et des Lettres. Les *digital humanities* ne font pas table rase du passé. Elles s'appuient, au contraire, sur l'ensemble des paradigmes, savoir-faire et connaissances propres à ces disciplines, tout en mobilisant les outils et les perspectives singulières du champ du numérique. Les *digital humanities* désignent une transdiscipline, porteuse des méthodes, des dispositifs et des perspectives heuristiques liés au numérique dans le domaine des sciences humaines et sociales<sup>337</sup>.

---

332. K. M. RUDY, *Piety in Pieces, How Medieval Readers Customized their Manuscripts...*

333. Giacomo BAROFFIO, « Testo e musica nei libri d'ore », *Rivista Italiana di musicologia*, XXXIV (2011), p. 19-87, URL : [https://www.academia.edu/39811280/Testo\\_e\\_musica\\_nei\\_libri\\_dore](https://www.academia.edu/39811280/Testo_e_musica_nei_libri_dore) (visité le 24/07/2020), p. 57.

334. Parmi les différents types de notations, on distingue notamment la notation neumatique, dont une des plus anciennes apparaît dès le IX<sup>e</sup> siècle, *in campo aperto* ou sur ligne, et la notation carrée, se développant principalement au XII<sup>e</sup> siècle en France.

335. Marin DACOS et Pierre MOUNIER, *Humanités numériques État des lieux et positionnement de la recherche française dans le contexte international*, Paris, 2014, p. 3.

336. *Ibid.*, p. 5.

337. *Ibid.*, p. 7-8.

### 3.2. LE PROTOCOLE D'ANNOTATION POUR GUIDER L'APPRENTISSAGE MACHINE

---

Pourtant, ce terme ne fait pas consensus. Certains acteurs du numérique au service du patrimoine culturel, comme Gautier Poupeau, ingénieur de la donnée à l'INA et chargé de cours dans le master « Technologies numériques appliquées à l'histoire », stipule que la nécessaire digitalisation des SHS rend le terme « Humanités numériques » obsolète. Il compare la situation avec la physique ou la biologie, disciplines pour lesquelles l'usage du numérique n'a pas créé un mouvement revendicatif de l'usage du numérique dans les sciences. En faire une discipline à part restreindrait la place du numérique au lieu d'en faire une composante naturelle et intégrée au sein des différentes disciplines des sciences humaines et sociales<sup>338</sup>.

Dans un projet comme celui d'HORAE, où le champ des humanités numériques s'inscrit pleinement dans la recherche historique, chercheurs, ingénieurs, techniciens, *data scientists*, s'unissent pour donner du sens à un flot de données, en l'occurrence les livres d'heures numérisés. En effet, l'approfondissement de leur étude, la possibilité de les lire avec des transcriptions, leur mise en relation et l'enrichissement des métadonnées les rendent bien plus visibles et accessibles qu'une simple numérisation de leur contenu<sup>339</sup>.

Si le numérique a ses limites, et ne peut pas restituer l'émotion d'un contact physique humain avec l'objet étudié, les possibilités offertes par le *machine learning* dans le traitement de données nombreuses est un gage de renouvellement de notre compréhension de la ré-appropriation des usages liturgiques dans les livres d'heures, ainsi que des réseaux de transmissions des livres, et donc des idées, croyances et cultes qu'ils diffusent.

---

338. Gautier POUPEAU, *Repenser la place du numérique dans les SHS*, 2019, URL : <http://www.lespetitescases.net/repenser-la-place-du-numerique-dans-les-shs> (visité le 22/03/2020).

339. Sur les questions d'intelligibilité des données accessibles sur le web et le développement d'une compréhension partagée, cf. Thomas BARTSCHERER et Roderick COOVER, *Switching Codes, thinking through digital technology in the humanities and the arts*, 2011, p. 38-60.



# Conclusion

Ainsi, les trois grandes étapes du stage exposées dans ce mémoire permettent de dresser un bilan des apports, limites et promesses offertes par le numérique dans l'étude des lives d'heures.

Le travail de structuration des métadonnées, plus précisément du catalogue de notices de livres d'heures élaboré par Victor Leroquais<sup>340</sup>, permet la création d'une édition numérique normalisée et interopérable, selon les standards définis par TEI All, afin de faciliter l'échange et l'analyse d'informations. Il est désormais plus facile de répertorier et de comparer la présence de certains textes dans les livres d'heures conservés à la Bibliothèque nationale, de détecter leur structure, la place qu'occupent certaines pièces liturgiques, leurs usages. Le catalogue sous forme numérique donne une meilleure visibilité du contenu et des aspects codicologiques des livres d'heures tout en fournissant un apport pour les données numérisées amassées dans le cadre du projet. L'utilisation de langages de programmation participe à l'automatisation du processus d'encodage sur des centaines de notices, et constitue évidemment un gain de temps considérable.

Toutefois, l'automatisation reste perfectible, d'où l'intérêt de réfléchir à différentes possibilités, Les langages Python, XSLT ou l'exemple de GROBID pour l'apprentissage machine en l'occurrence, afin de choisir le programme le plus efficace. Différentes solutions peuvent parfois se compléter, comme l'utilisation des possibilités offertes par le logiciel Word, XSLT et XQuery dans notre cas. La mise en place du programme le plus efficace peut certes être chronophage, mais peut aussi s'avérer être un gain de temps pour la suite et éviter certains rattrapages manuels dans la structuration des métadonnées. Il peut être difficile de cerner ce qui est le plus efficace dans l'instant et ce qui le sera pour la suite. De plus, le numérique ne peut pas tout. L'étape de vérification de l'intégrité des données reste indispensable, afin de s'assurer de l'objectif principal, l'établissement de données de qualité pour la recherche en sciences humaines et sociales.

La modélisation et l'import des données du cursus dans une base de données relationnelle s'avèrent essentiels pour la gestion et la visibilité d'un grand nombre de données reliées entre elles. Leur implémentation dans une base de données, dans le respect de leur cohérence, permet ainsi de mieux les visualiser et de produire des outils d'analyse. On

---

340. V. LEROQUAIS, *Les livres d'heures manuscrits de la Bibliothèque nationale....*

peut penser à la génération de cartes sur la circulation des usages, des lieux de conservation des manuscrits du corpus, de graphes de liens entre les pièces et sections du livre d'heures selon les usages, etc. Les bases de données, relationnelles entre autres, sont un outil indispensable pour la manipulation et la visualisation des données.

L'étape de modélisation est alors cruciale pour comprendre l'objet à étudier, n'oublier aucun de ses aspects. Cette étape ne peut être négligée pour donner une forme complète et cohérente aux données dans la base. La difficulté est donc de créer un modèle reflétant la complexité de la réalité tout en pouvant être implémenté et mis à jour dans la base de données. Il faut ici souligner l'importance d'un logiciel ergonomique répondant aux besoins des différents projets de recherche, ainsi qu'un dialogue constant avec les techniciens en charge de la conception des bases de données utilisées. Dans le cadre d'HORAE, Heurist, base développée pour la TGIR<sup>341</sup> Huma-Num, a ainsi pour objectif de donner accès à une interface simple d'utilisation sans avoir au préalable de grandes notions de programmation.

L'utilisation du *machine learning* pour la reconnaissance automatique de caractères, d'écritures puis de textes témoigne de la coopération entre différents acteurs qui s'enrichissent mutuellement dans un projet en Humanités numériques. Si le numérique a des limites, ce pan du projet montre toutes ses capacités heuristiques en terme de renouvellement des conclusions historiques. L'établissement d'une vérité terrain pour la détection et la segmentation automatique permet d'enrichir constamment les classes servant à l'apprentissage machine, comme en témoigne la liste des suffrages.

L'utilisation du *deep learning* pour la reconnaissance d'écritures puis de textes offre des perspectives prometteuses dans l'analyse d'un nombre important de livres d'heures, afin de repérer les éventuelles variantes, les tendances et les exceptions dans la structure des livres d'heures. Les promesses offertes par l'utilisation de ces technologies laissent songer à leur appropriation pour d'autres aspects de l'étude des livres d'heures, tels que les couleurs, les calendriers, les analyses lexicales, la détection de plagiat, de points communs avec d'autres textes religieux de référence. Les technologies développées dans le cadre d'HORAE peuvent ainsi devenir des outils essentiels à d'autres projets.

*In fine*, les analyses découlant de l'apprentissage automatique permettront peut-être de saisir une dévotion plus intime, voire de noter des aspects qui confirment ou infirment l'idée d'une dévotion féminine ou masculine. L'utilisation de toutes les possibilités offertes par le numérique pour l'étude des livres d'heures permet du moins certainement de saisir au plus proche une part de la foi des êtres de la fin du Moyen Âge, en écho aux lacunes qu'avait soulignées le chanoine Victor Leroquais.

---

341. Très Grande Infrastructure de Recherche.

## **Annexes**



# **Annexe A**

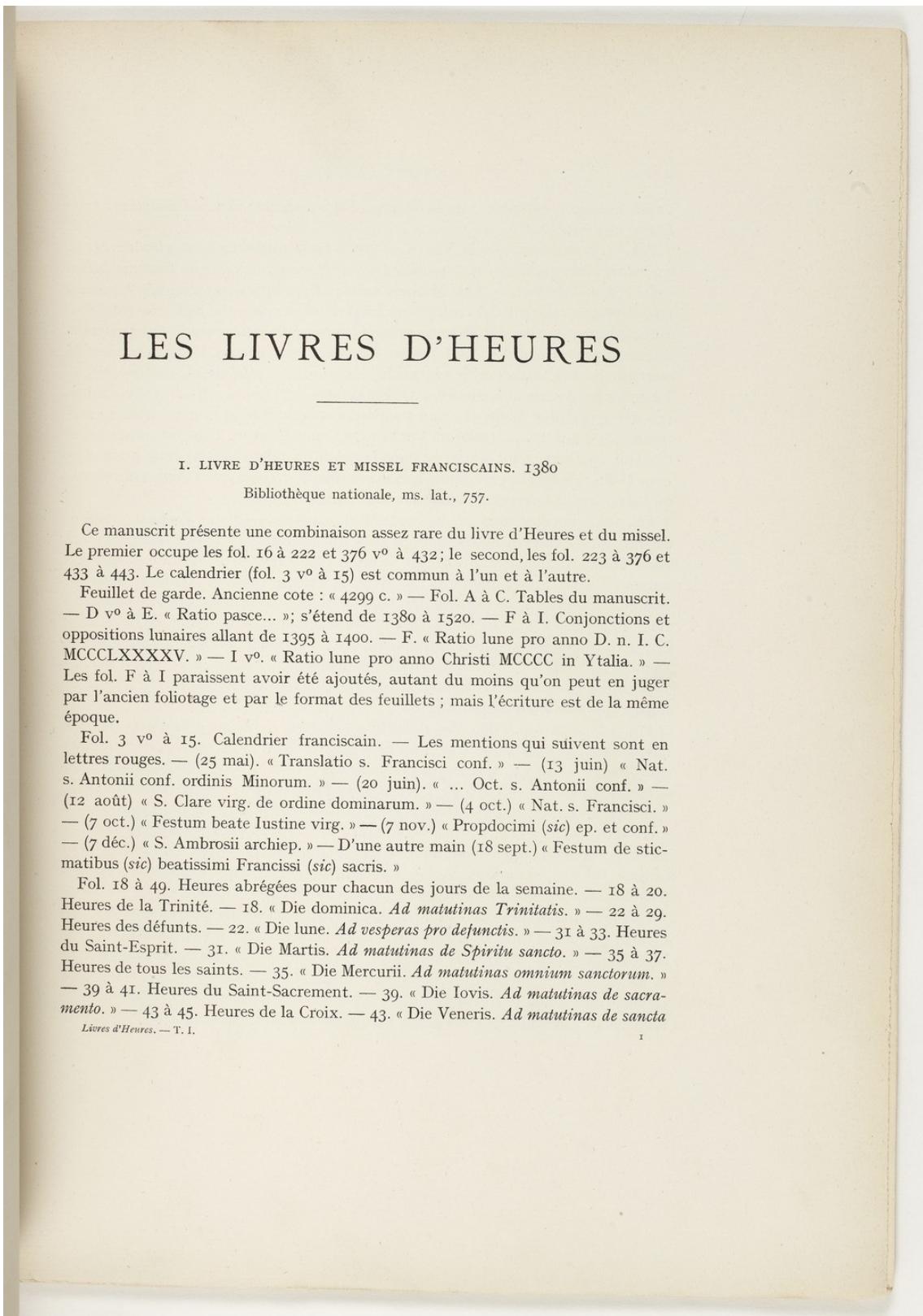
## **Structuration semi-automatisée d'un catalogue de notices**

### **A.1 Structurer des notices de livres d'heures**

#### **A.1.1 Documents sources**

Ci-dessous ont été reproduites les trois notices qui ont servi de tests et de modèles pour le projet de structuration des métadonnées fournies par Victor Leroquais. Le terme de « Document sources » est mis au pluriel car le document de départ se trouve être les notices établies par Leroquais, qui ont été numérisées et qui sont disponibles sur le site Gallica, mais le document de travail servant à la transformation est le document océrisé. Or, ce document est lui-même une transformation du document de départ. La version papier et la version océrisée ont donc été mises en regard.

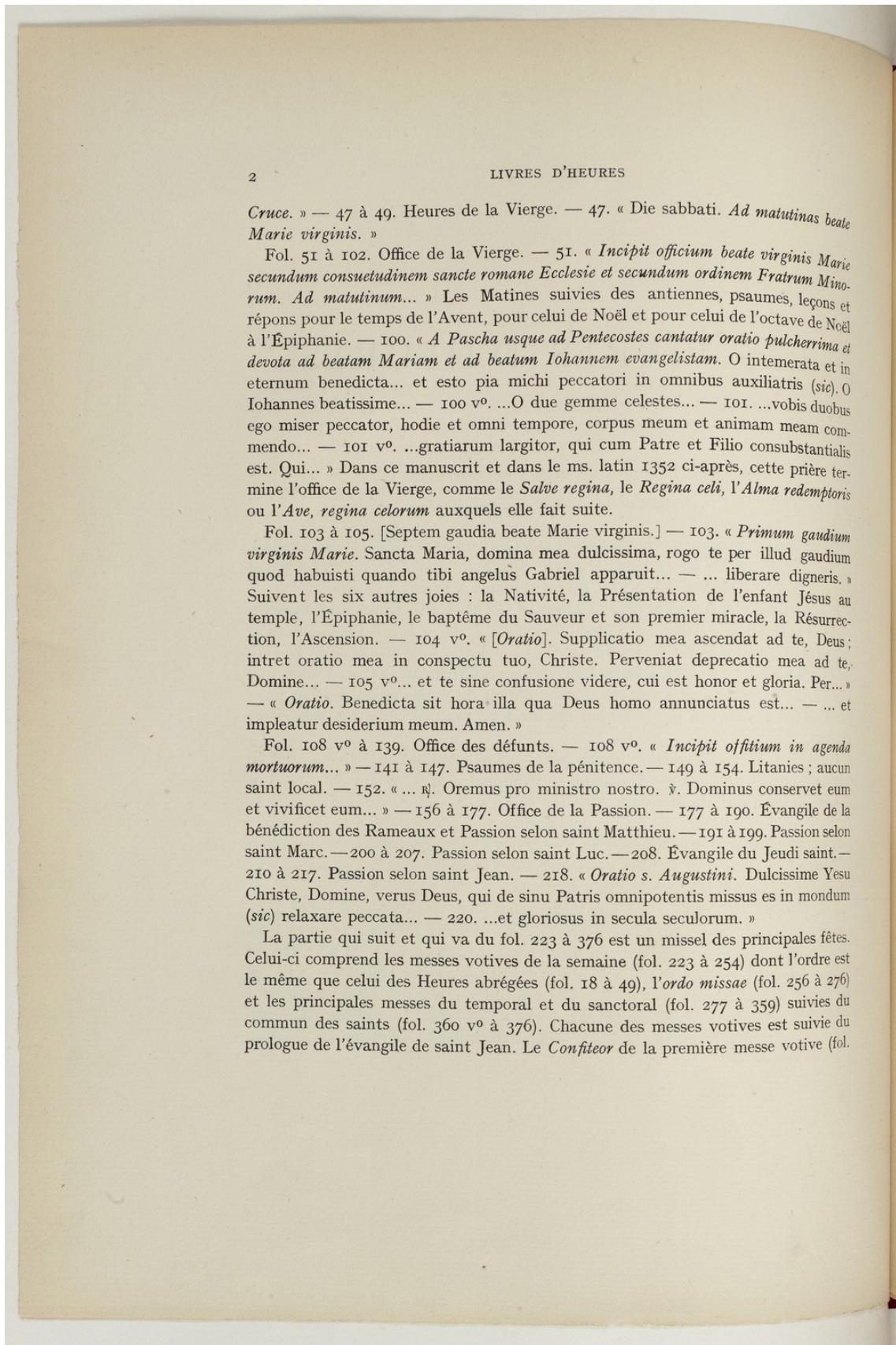
#### **Notices papiers**



Source gallica.bnf.fr / Bibliothèque nationale de France

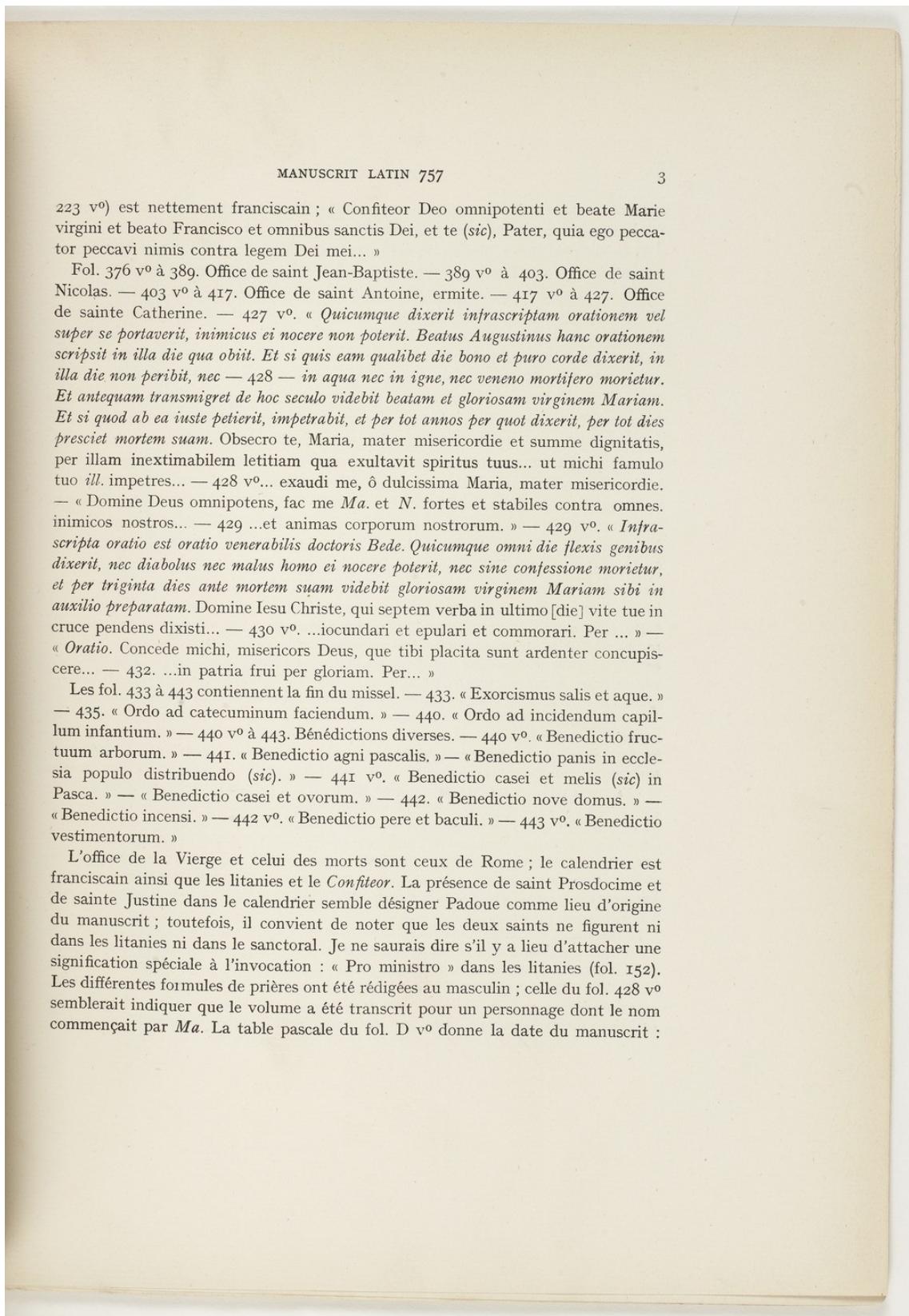
FIGURE A.1 – Notice 1 numérisée, première page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 1

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES



Source gallica.bnf.fr / Bibliothèque nationale de France

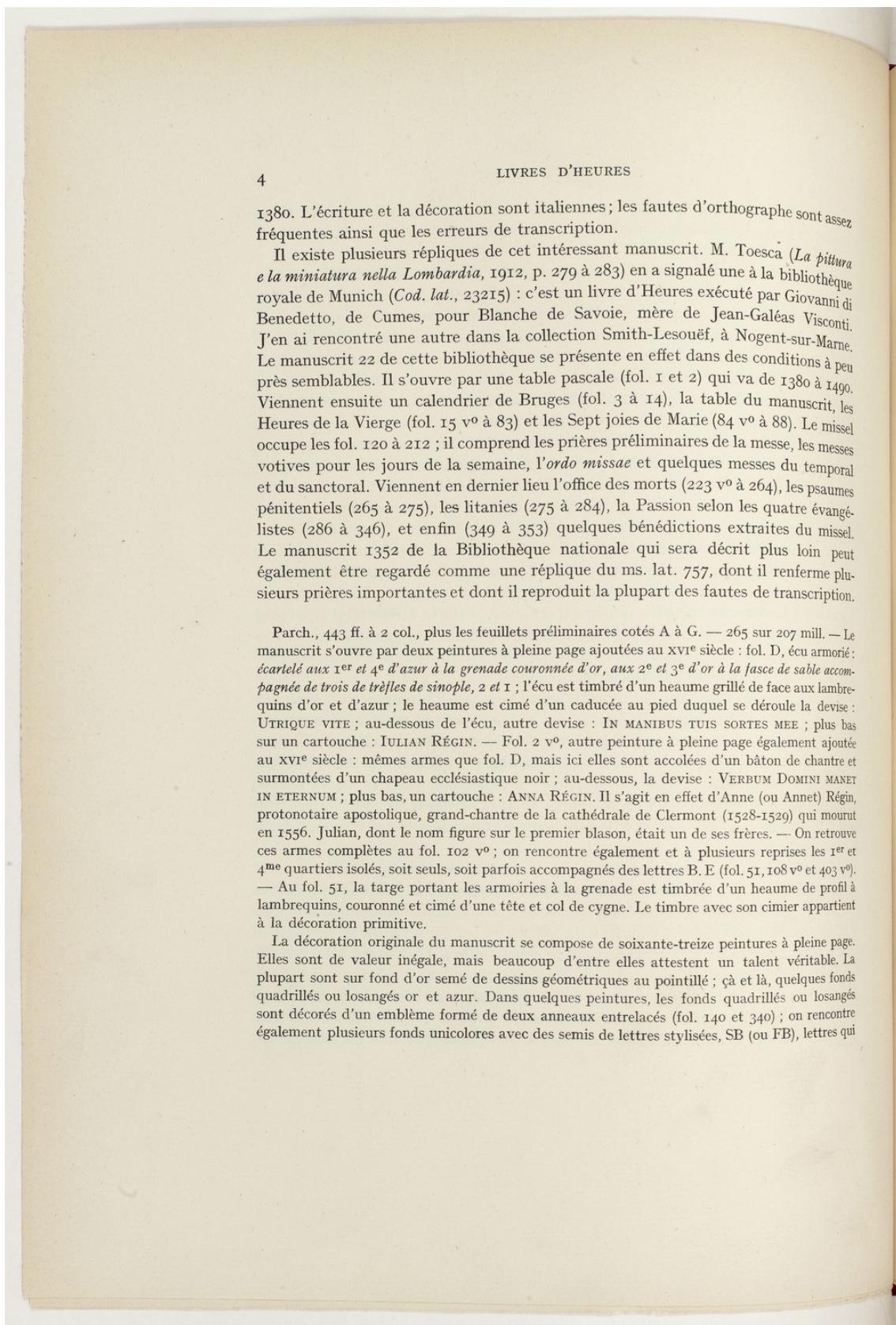
FIGURE A.2 – Notice 1 numérisée, deuxième page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 2



Source gallica.bnf.fr / Bibliothèque nationale de France

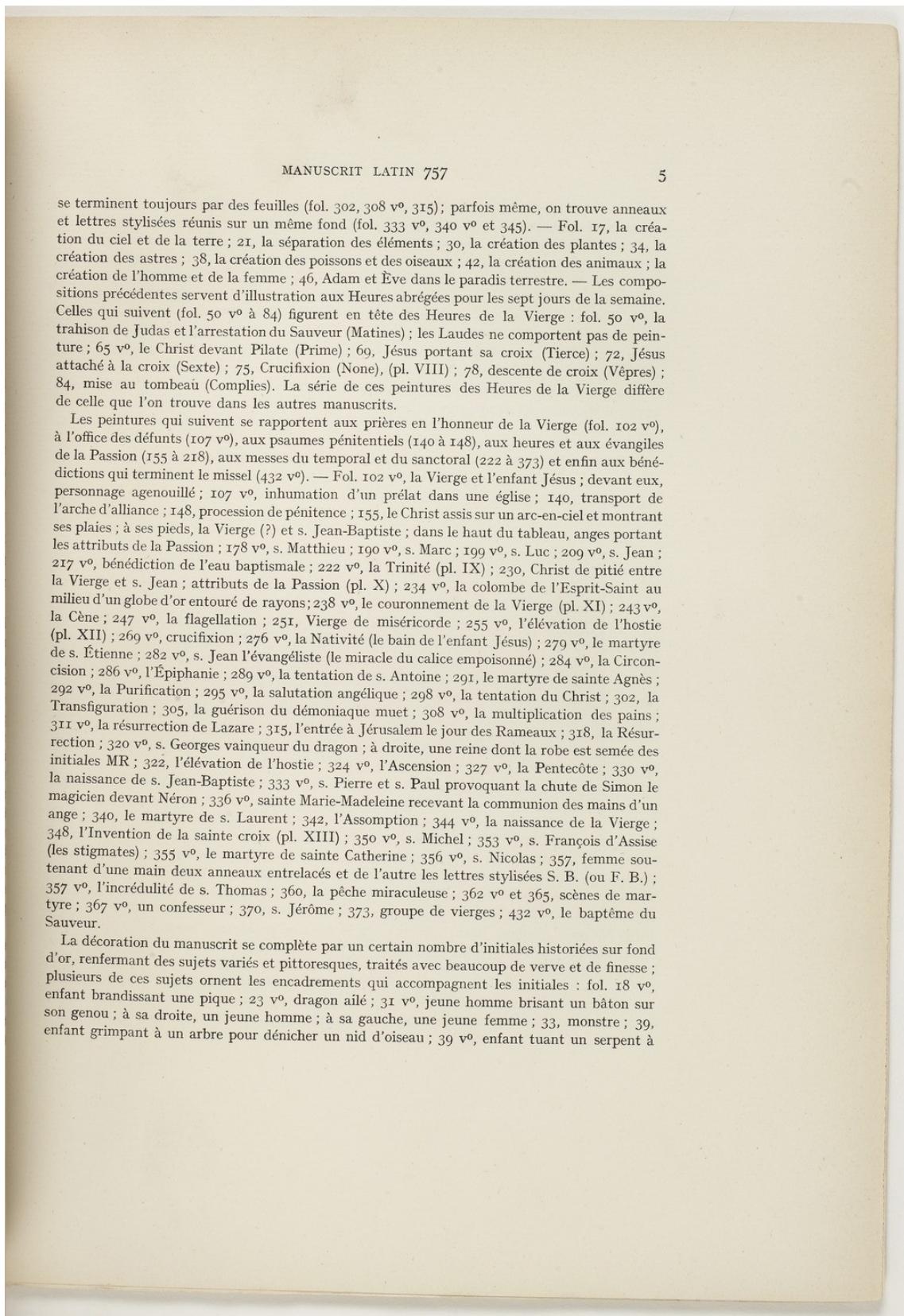
FIGURE A.3 – Notice 1 numérisée, troisième page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 3

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES



Source gallica.bnf.fr / Bibliothèque nationale de France

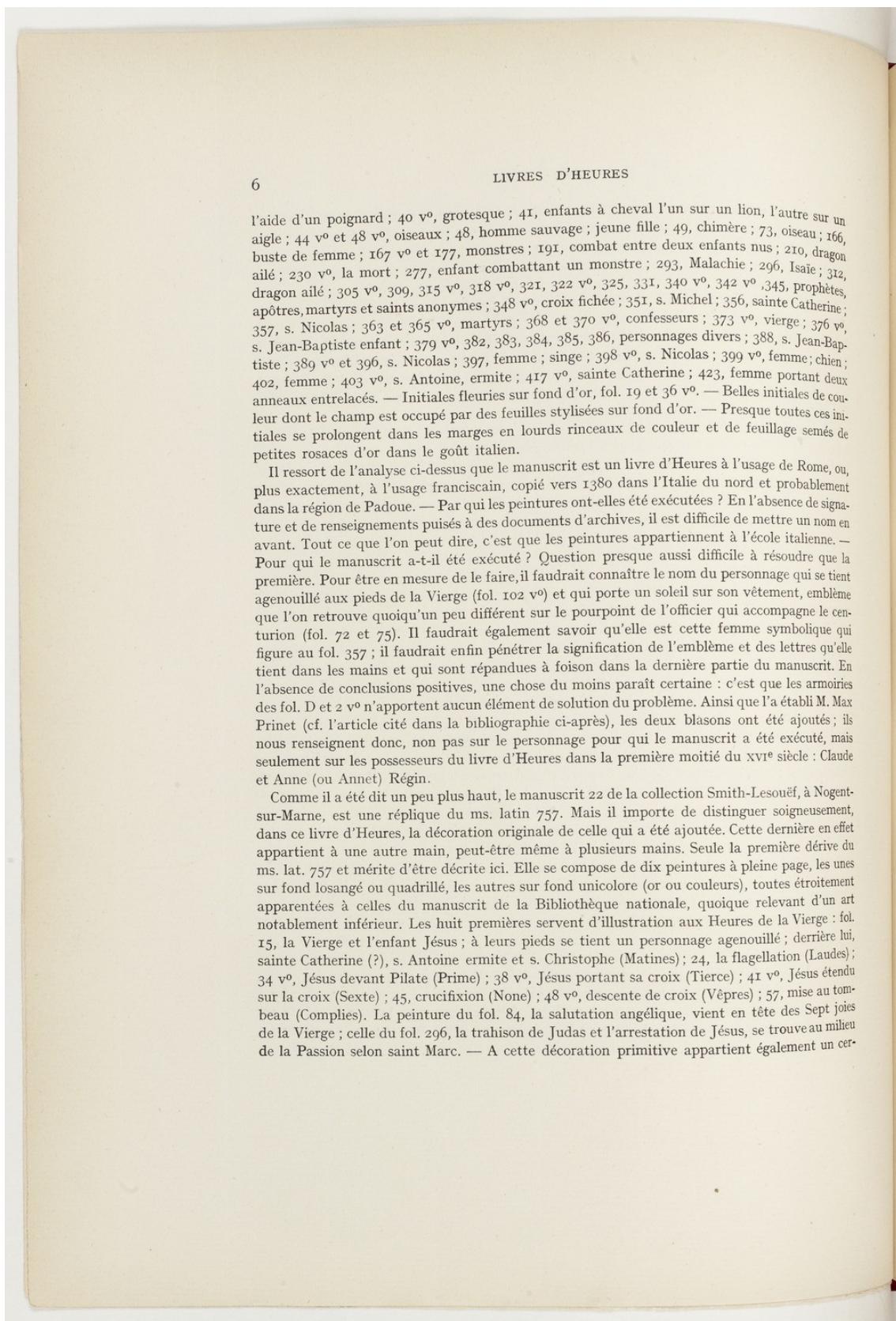
FIGURE A.4 – Notice 1 numérisée, quatrième page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 4



Source gallica.bnf.fr / Bibliothèque nationale de France

FIGURE A.5 – Notice 1 numérisée, cinquième page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 5

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES



Source gallica.bnf.fr / Bibliothèque nationale de France

FIGURE A.6 – Notice 1 numérisée, sixième page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 6

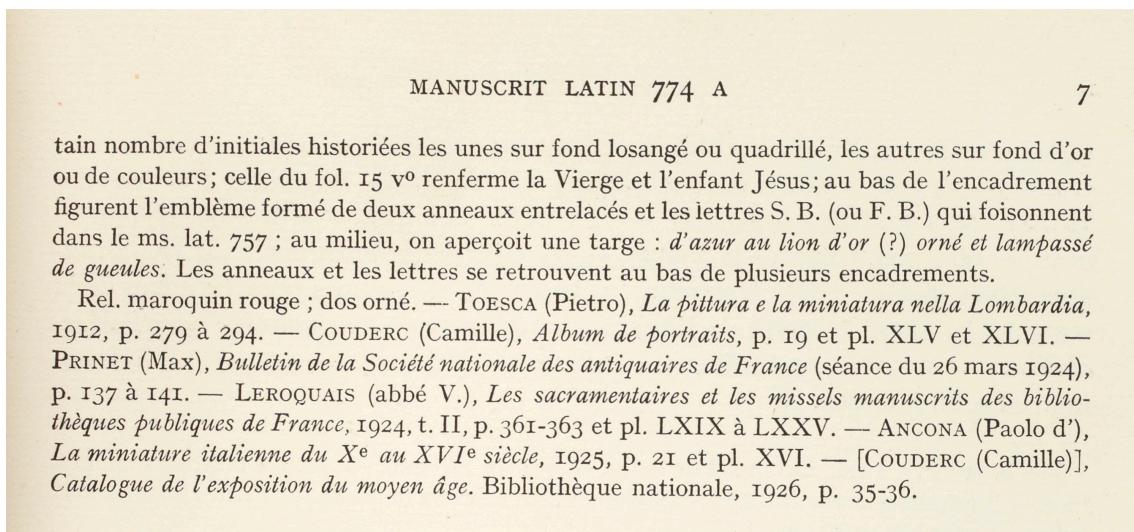
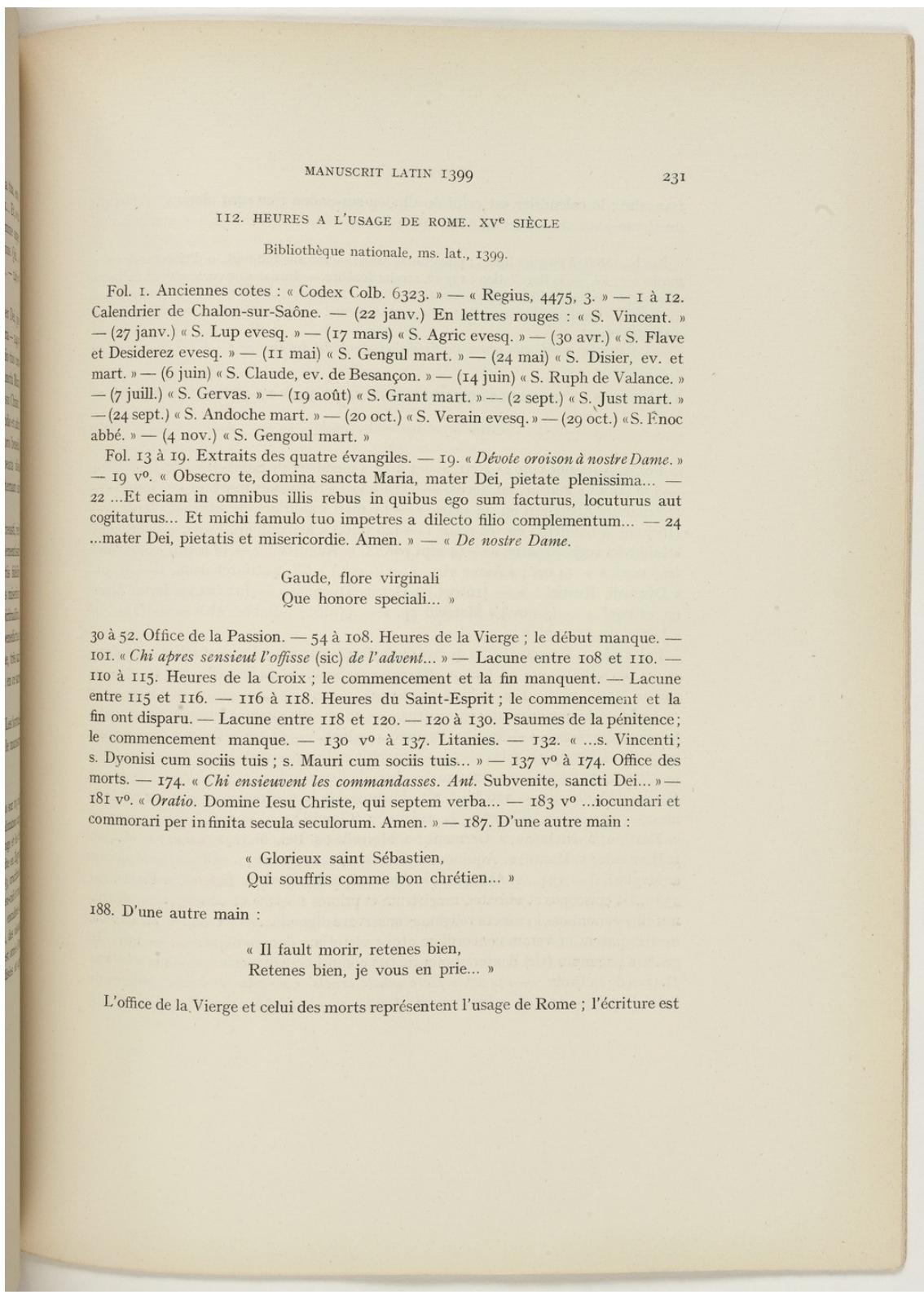


FIGURE A.7 – Notice 1 numérisée, septième page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 7

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES



Source gallica.bnf.fr / Bibliothèque nationale de France

FIGURE A.8 – Notice 112 numérisée, première page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 231

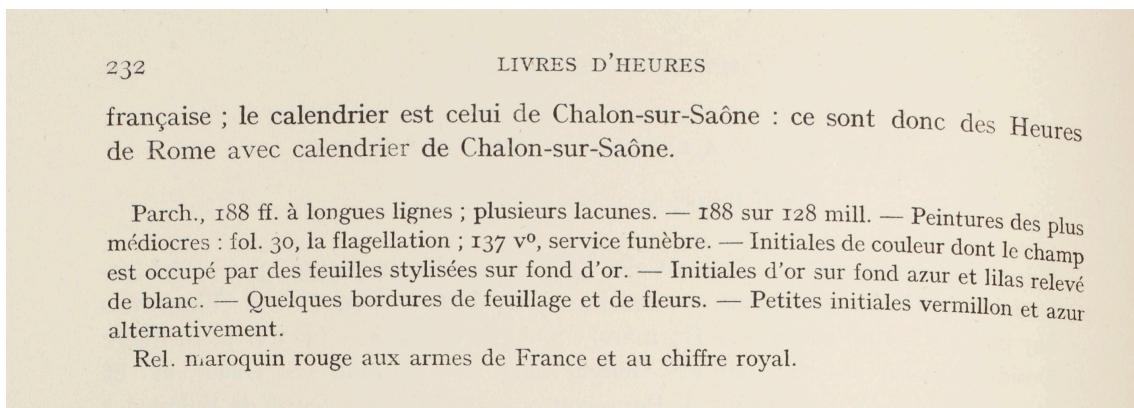


FIGURE A.9 – Notice 112 numérisée, deuxième page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 232

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

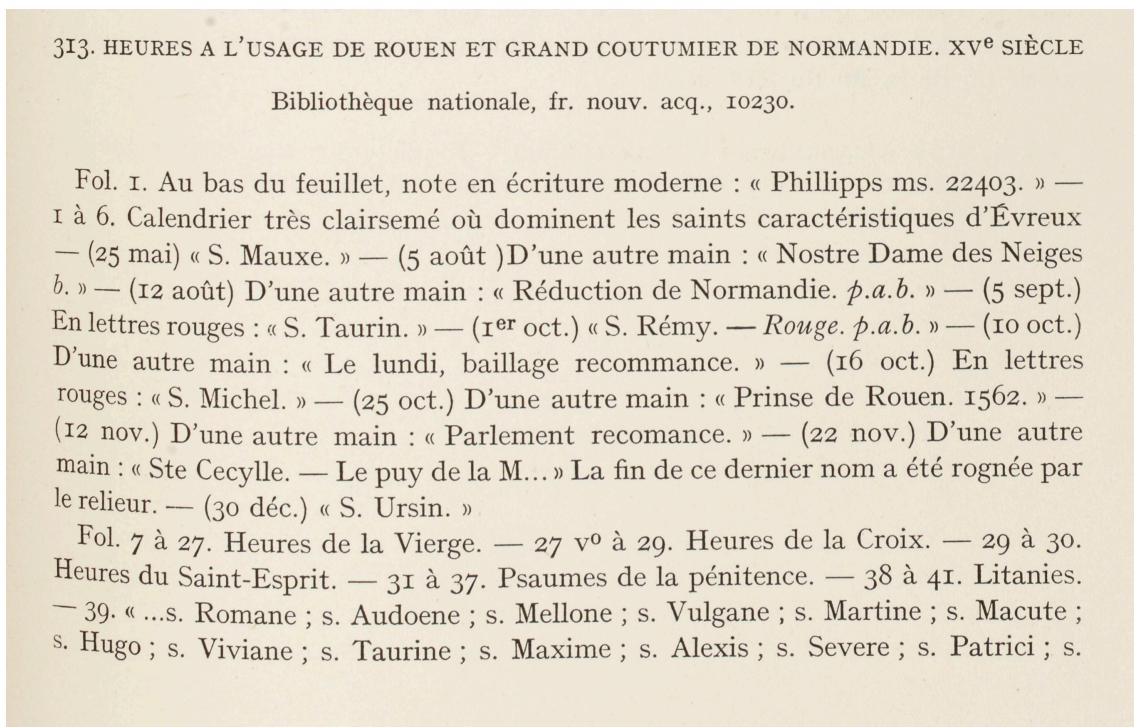
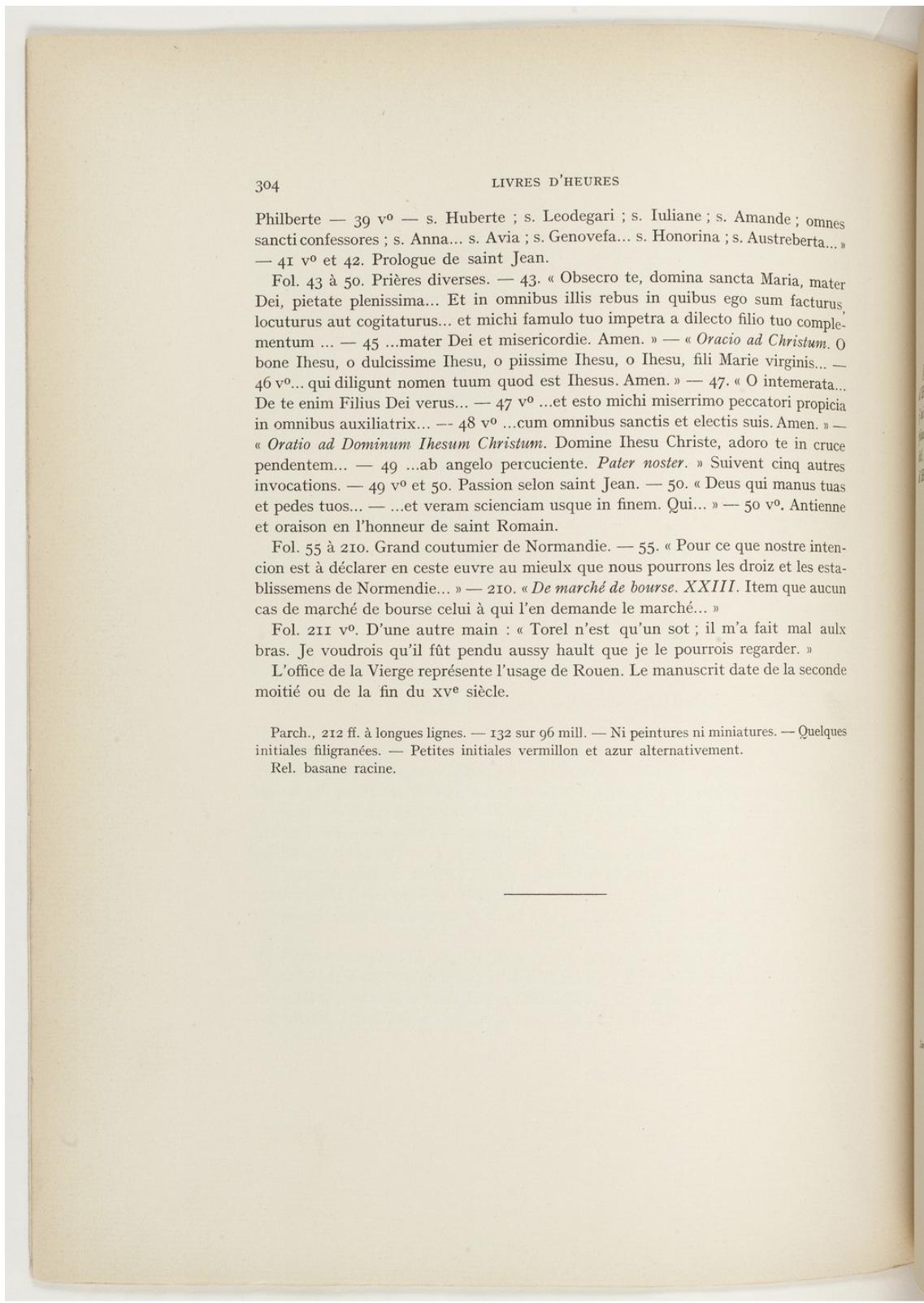


FIGURE A.10 – Notice 313 numérisée, première page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 304



Source gallica.bnf.fr / Bibliothèque nationale de France

FIGURE A.11 – Notice 313 numérisée, deuxième page. Cf. LEROQUAIS (Victor), *Les livres d'heures manuscrits de la Bibliothèque nationale*. 3 t., Paris, 1927, p. 305

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

### A.1.2 Notices océrisées

1. LIVRE D'HEURES ET MISSEL FRANCISCAINS. 1380  
Bibliothèque nationale, ms. lat., 757.

Ce manuscrit présente une combinaison assez rare du livre d'Heures et du missel. Le premier occupe les fol. 16 à 222 et 376 v° à 432; le second, les fol. 223 à 376 et 433 à 443. Le calendrier (fol. 3 v° à 15) est commun à l'un et à l'autre.

Feuillet de garde. Ancienne cote : « 4299 c. » — Fol. A à C. Tables du manuscrit.

— D v° à E. « Ratio pasce... », s'étend de 1380 à 1520. — F à I. Conjonctions et oppositions lunaires allant de 1395 à 1400. — F. « Ratio lune pro anno D. n. I. C. MCCCLXXXV. » — I v°. « Ratio lune pro anno Christi MCCCC in Ytalia. » — Les fol. F à I paraissent avoir été ajoutés, autant du moins qu'on peut en juger par l'ancien foliotage et par le format des feuilles; mais l'écriture est de la même époque.

Fol. 3 v° à 15. Calendrier franciscain. — Les mentions qui suivent sont en lettres rouges. — (25 mai). « Translatio s. Francisci conf. » — (13 juin) « Nat. s. Antonii conf. ordinis Minorum. » — (20 juin). « ... Oct. s. Antonii conf. » — (12 août) « S. Clare virg. de ordine dominarum. » — (4 oct.) « Nat. s. Francisci. » — (7 oct.) « Festum beate lustine virg. » — (7 nov.) « Propdocimi (*sic*) ep. et conf. » — (7 déc.) « S. Ambrosii archiep. » — D'une autre main (18 sept.) « Festum de stic- matibus (*sic*) beatissimi Francissi (*sic*) sacris. »

Fol. 18 à 49. Heures abrégées pour chacun des jours de la semaine. — 18 à 20. Heures de la Trinité. — 18. « Die dominica. *Ad matutinas Trinitatis.* » — 22 à 29. Heures des défunt. — 22. « Die lune. *Ad vespertas pro defunctis.* » — 31 à 33. Heures du Saint-Esprit. — 31. « Die Martis. *Ad matutinas de Spiritu sancto.* » — 35 à 37. Heures de tous les saints. — 35. « Die Mercurii. *Ad matutinas omnium sanctorum.* »

— 39 à 41. Heures du Saint-Sacrement. — 39. « Die Iovis. *Ad matutinas de sacramento.* » - 43 à 45. Heures de la Croix. — 43. « Die Veneris. *Ad matutinas de sancta*

*Livres d'Heures. — T. I.*

FIGURE A.12 – Notice 1 océrisée, première page

*Cruice.* » — 47 à 49. Heures de la Vierge. — 47. « *Die sabbati. Ad matutinas beate Marie virginis.* »

Fol. 51 à 102. Office de la Vierge. — 51. « *Incipit officium beate virginis Marie secundum consuetudinem sancte romane Ecclesie et secundum ordinem Fratrum Minorum. Ad matutinum...* » Les Matines suivies des antennes, psaumes, leçons et répons pour le temps de l'Avent, pour celui de Noël et pour celui de l'octave de Noël à l'Épiphanie. — 100. « *A Pascha usque ad Pentecostes cantatur oratio pulcherrima et devota ad beatam Mariam et ad beatum Iohannem evangelistam. O intermerita et in eternum benedicta... et esto pia michi peccatori in omnibus auxiliatrix (stc). O Iohannes beatissime... — 100 v°. ...O due gemme celestes... — 101. ...vobis duobus ego miser peccator, hodie et omni tempore, corpus meum et animam meam commendo... — 101 v°. ...gratiarum largitor, qui cum Patre et Filio consubstantialis est. Qui... » Dans ce manuscrit et dans le ms. latin 1352 ci-après, cette prière termine l'office de la Vierge, comme le *Salve regina*, le *Regina celi*, *l'Alma redemptoris* ou *l'Ave, regina celorum*, auxquels elle fait suite.*

Fol. 103 à 105. [Septem gaudia beate Marie virginis.] — 103. « *Primum gaudium virginis Marie. Sancta Maria, domina mea dulcissima, rogo te per illud gaudium quod habuisti quando tibi angelus Gabriel apparuit... — ... liberare digneris.* » Suivent les six autres joies : la Nativité, la Présentation de l'enfant Jésus au temple, l'Épiphanie, le baptême du Sauveur et son premier miracle, la Résurrection, l'Ascension. — 104 v°. « *[Oratio]. Supplicatio mea ascendat ad te, Deus: intret oratio mea in conspectu tuo, Christe. Perveniat deprecatio mea ad te. Domine... — 105 v°... et te sine confusione videre, cui est honor et gloria. Per...» — « *Oratio. Benedicta sit hora illa qua Deus homo annunciatus est... — ... et impleatur desiderium meum. Amen.* »*

Fol. 108 v° à 139. Office des défunt. — 108 v°. « *Incipit officium in agenda mortuorum... — 141 à 147. Psaumes de la pénitence — 149 à 154. Litanies ; aucun saint local. — 152. u 8J. Oramus pro ministro nostro, £. Dominus conservet eum et vivificet eum... — 156 à 177. Office de la Passion. — 177 à 190. Évangile de la bénédiction des Rameaux et Passion selon saint Matthieu. — 191 à 199. Passion selon saint Marc. — 200 à 207. Passion selon saint Luc. — 208. Évangile du Jeudi saint — 210 à 217. Passion selon saint Jean. — 218. « *Oratio s. Augustini. Dulcissime Yesu Christe, Domine, verus Deus, qui de sinu Patris omnipotentis missus es in mondum (sic) relaxare peccata... — 220. ...et gloriosus in secula seculorum.* »*

La partie qui suit et qui va du fol. 223 à 376 est un missel des principales fêtes. Celui-ci comprend les messes votives de la semaine (fol. 223 à 254) dont l'ordre est le même que celui des Heures abrégées (fol. 18 à 49), *l'ordo missae* (fol. 256 à 276) et les principales messes du temporal et du sanctoral (fol. 277 à 359) suivies du commun des saints (fol. 360 v° à 376). Chacune des messes votives est suivie du prologue de l'évangile de saint Jean. Le *Confiteor* de la première messe votive (fol.

FIGURE A.13 – Notice 1 océrisée, deuxième page

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

223 v°) est nettement franciscain ; « *Confiteor Deo omnipotenti et beate Marie virginis et beato Francisco et omnibus sanctis Dei, et te (sic), Pater, quia ego peccator peccavi nimis contra legem Dei mei... »*

Fol. 376 v° à 389. Office de saint Jean-Baptiste. — 389 v° à 403. Office de saint Nicolas. — 403 v° à 417. Office de saint Antoine, ermite. — 417 v° à 427. Office de sainte Catherine. — 427 v°. « *Quicumque dixerit infrascriptam orationem vel super se portaverit, inimicus ei nocere non poterit. Beatus Augustinus hanc orationem scripsit in illa die qua obiit. Et si quis eam qualibet die bono et puro corde dixerit, in illa die non peribit, nec — 428 — in aqua nec in igne, nec veneno mortifero morietur. Et antequam transmigret de hoc seculo videbit beatam et gloriosam virginem Mariam. Et si quod ab ea iuste petierit, impetrabit, et per tot annos per quot dixerit, per tot dies prescribet mortem suam. Obsecro te, Maria, mater misericordie et summe dignitatis, per illam inextirpabilem letitiam qua exultavit spiritus tuus... ut michi famulo tuo ill. impetres... — 428 v°... exaudi me, O dulcissima Maria, mater misericordie.*

— « *Domine Deus omnipotens, fac me Ma. et 2V. fortes et stabiles contra omnes, inimicos nostros... — 429 ...et animas corporum nostrorum. » — 429 v°. « Infrascripta oratio est oratio venerabilis doctoris Bede. Quicumque omni die flexis genibus dixerit, nec diabolus nec malus homo ei nocere poterit, nec sine confessione morietur, et per triginta dies ante mortem suam videbit gloriosam virginem Mariam sibi in auxilio preparatam. Domine Iesu Christe, qui septem verba in ultimo [die] vite tue in cruce pendens dixisti... — 430 v°. ...iocundari et epulari et commorari. Per ... » — « *Oratio. Concede michi, misericors Deus, que tibi placita sunt ardenter concupiscere... — 432. ...in patria frui per gloriam. Per... »**

Les fol. 433 à 443 contiennent la fin du missel. — 433. « *Exorcismus salis et aquae. »* — 435- « *Ordo ad catecumenum faciendum. » — 440. « *Ordo ad incidendum capillum infantium. » — 440 v° à 443. Bénédictions diverses. — 440 v°. « *Benedictio fructuum arborum. » — 441. « *Benedictio agni pascalis. » — « *Benedictio panis in ecclesia populo distribuendo (sic). » — 441 v°. « *Benedictio casei et melis (sic) in Pasca. » — « *Benedictio casei et ovorum. » — 442. « *Benedictio nove dormus. » — « *Benedictio incensi. » — 442 v°. « *Benedictio pere et baculi. » — 443 v°. « *Benedictio vestimentorum. »***********

L'office de la Vierge et celui des morts sont ceux de Rome ; le calendrier est franciscain ainsi que les litanies et le *Confiteor*. La présence de saint Prosdocime, et de sainte Justine dans le calendrier semble désigner Padoue comme lieu d'origine du manuscrit ; toutefois, il convient de noter que les deux saints ne figurent ni dans les litanies ni dans le sanctoral. Je ne saurais dire s'il y a lieu d'attacher une signification spéciale à l'invocation : « *Pro ministro* » dans les litanies (fol. 152). Les différentes formules de prières ont été rédigées au masculin ; celle du fol. 428 v° semblerait indiquer que le volume a été transcrit pour un personnage dont le nom commençait par *Ma*. La table pascale du fol. D v° donne la date du manuscrit :

FIGURE A.14 – Notice 1 océrisée, troisième page

1380. L'écriture et la décoration sont italiennes; les fautes d'orthographe sont assez fréquentes ainsi que les erreurs de transcription.

Il existe plusieurs répliques de cet intéressant manuscrit. M. Toesca (*La pittura e la miniatura nella Lombardia*, 1912, p. 279 à 283) en a signalé une à la bibliothèque royale de Munich (*Cod. lat.*, 23215) : c'est un livre d'Heures exécuté par Giovanni di Benedetto, de Cumes, pour Blanche de Savoie, mère de Jean-Gaiéas Visconti. J'en ai rencontré une autre dans la collection Smith-Lesouëf, à Nogent-sur-Marne. Le manuscrit 22 de cette bibliothèque se présente en effet dans des conditions à peu près semblables. Il s'ouvre par une table pascale (fol. 1 et 2) qui va de 1380 à 1490. Viennent ensuite un calendrier de Bruges (fol. 3 à 14), la table du manuscrit, les Heures de la Vierge (fol. 15 v° à 83) et les Sept joies de Marie (84 v° à 88). Le missel occupe les fol. 120 à 212 ; il comprend les prières préliminaires de la messe, les messes votives pour les jours de la semaine, *Yor do missae* et quelques messes du temporal et du sanctoral. Viennent en dernier lieu l'office des morts (223 v° à 264), les psaumes pénitentiels (265 à 275), les litanies (275 à 284), la Passion selon les quatre évangélistes (286 à 346), et enfin (349 à 353) quelques bénédictions extraites du missel. Le manuscrit 1352 de la Bibliothèque nationale qui sera décrit plus loin peut également être regardé comme une réplique du ms. lat. 757, dont il renferme plusieurs prières importantes et dont il reproduit la plupart des fautes de transcription.

Parch., 443 ff. à 2 col., plus les feuillets préliminaires cotés A à G. — 265 sur 207 mill. — Le manuscrit s'ouvre par deux peintures à pleine page ajoutées au XVI<sup>e</sup> siècle : fol. D, écu armorié: écartelé aux 1<sup>e</sup> et 4<sup>e</sup> d'azur à la grenade couronnée d'or, aux 2<sup>e</sup> et 3<sup>e</sup> d'or à la fasce de sable accompagnée de trois de trèfles de sinople, 2 éli ; Vécu est timbré d'un heaume grillé de face aux lambrequins d'or et d'azur ; le heaume est cimé d'un caducée au pied duquel se déroule la devise : UTRIQUE VITE ; au-dessous de l'écu, autre devise : IN MANIBUS TUIS SORTES MEE ; plus bas sur un cartouche : JULIAN REGIN. — Fol. 2 v°, autre peinture à pleine page également ajoutée au XVI<sup>e</sup> siècle : mêmes armes que fol. D, mais ici elles sont accolées d'un bâton de chantre et surmontées d'un chapeau ecclésiastique noir ; au-dessous, la devise : VERBUM DOMINI MANET IN ETERNUM ; plus bas, un cartouche : ANNA REGIN. Il s'agit en effet d'Anne (ou Annet) Regin, protonotaire apostolique, grand-chantre de la cathédrale de Clermont (1528-1529) qui mourut en 1556. Julian, dont le nom figure sur le premier blason, était un de ses frères. — On retrouve ces armes complètes au fol. 102 v° ; on rencontre également et à plusieurs reprises les 1<sup>e</sup> et 4<sup>e</sup> quartiers isolés, soit seuls, soit parfois accompagnés des lettres B. E (fol. 51, 108 v° et 403 V°). — Au fol. 51, la targe portant les armoiries à la grenade est timbrée d'un heaume de profil à lambrequins, couronné et cimé d'une tête et col de cygne. Le timbre avec son cimier appartient à la décoration primitive.

La décoration originale du manuscrit se compose de soixante-treize peintures à pleine page. Elles sont de valeur inégale, mais beaucoup d'entre elles attestent un talent véritable. La plupart sont sur fond d'or semé de dessins géométriques au pointillé ; çà et là, quelques fonds quadrillés ou losangés or et azur. Dans quelques peintures, les fonds quadrillés ou losangés sont décorés d'un emblème formé de deux anneaux entrelacés (fol. 140 et 340) ; on rencontre également plusieurs fonds unicolores avec des semis de lettres stylisées, SB (ou FB), lettres qui

FIGURE A.15 – Notice 1 océrisée, quatrième page

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

se terminent toujours par des feuilles (fol. 302, 308 v°, 315); parfois même, on trouve anneaux et lettres stylisées réunis sur un même fond (fol. 333 v°, 340 v° et 345). — Fol. 17, la création du ciel et de la terre ; 21, la séparation des éléments ; 30, la création des plantes ; 34, la création des astres ; 38, la création des poissons et des oiseaux ; 42, la création des animaux ; la création de l'homme et de la femme ; 46, Adam et Ève dans le paradis terrestre. — Les compositions précédentes servent d'illustration aux Heures abrégées pour les sept jours de la semaine. Celles qui suivent (fol. 50 v° à 84) figurent en tête des Heures de la Vierge : fol. 50 v°, la trahison de Judas et l'arrestation du Sauveur (Matines) ; les Laudes ne comportent pas de peinture ; 65 v°, le Christ devant Pilate (Prime) ; 69, Jésus portant sa croix (Tierce) ; 72, Jésus attaché à la croix (Sexte) ; 75, Crucifixion (None), (pl. VIII) ; 78, descente de croix (Vépres) ; 84, mise au tombeau (Complies). La série de ces peintures des Heures de la Vierge diffère de celle que l'on trouve dans les autres manuscrits.

Les peintures qui suivent se rapportent aux prières en l'honneur de la Vierge (fol. 102 v°), à l'office des défunt (107 v°), aux psaumes pénitentiels (140 à 148), aux heures et aux évangiles de la Passion (155 à 218), aux messes du temporal et du sanctoral (222 à 373) et enfin aux bénédicitions qui ferment le missel (432 v°). — Fol. 102 v°, la Vierge et l'enfant Jésus ; devant eux, personnage agenouillé ; 107 v°, inhumation d'un prélat dans une église ; 140, transport de l'arche d'alliance ; 148, procession de pénitence ; 155, le Christ assis sur un arc-en-ciel et montrant ses plaies ; à ses pieds, la Vierge (?) et s. Jean-Baptiste ; dans le haut du tableau, anges portant les attributs de la Passion ; 178 v°, s. Matthieu ; 190 v°, s. Marc ; 199 v°, s. Luc ; 209 v°, s. Jean ; 217 v°, bénédiction de l'eau baptismale ; 222 v°, la Trinité (pl. IX) ; 230, Christ de pitié entre la Vierge et s. Jean ; attributs de la Passion (pl. X) ; 234 v°, la colombe de l'Esprit-Saint au milieu d'un globe d'or entouré de rayons ; 238 v°, le couronnement de la Vierge (pl. XI) ; 243 v°, la Cène ; 247 v°, la flagellation ; 251, Vierge de miséricorde ; 255 v°, l'élévation de l'hostie (pl. XII) ; 269 v°, crucifixion ; 276 v°, la Nativité (le bain de l'enfant Jésus) ; 279 v°, le martyre de s. Étienne ; 282 v°, s. Jean l'évangéliste (le miracle du calice empoisonné) ; 284 v°, la Circoncision ; 286 v°, l'Épiphanie ; 289 v°, la tentation de s. Antoine ; 291, le martyre de sainte Agnès ; 292 v°, la Purification ; 295 v°, la salutation angélique ; 298 v°, la tentation du Christ ; 302, la Transfiguration ; 305, la guérison du démoniaque muet ; 308 v°, la multiplication des pains ; 311 v°, la résurrection de Lazare ; 315, l'entrée à Jérusalem le jour des Rameaux ; 318, la Résurrection ; 320 v°, s. Georges vainqueur du dragon ; à droite, une reine dont la robe est semée des initiales MR ; 322, l'élévation de l'hostie ; 324 v°, l'Ascension ; 327 v°, la Pentecôte ; 330 v°, la naissance de s. Jean-Baptiste ; 333 v°, s. Pierre et s. Paul provoquant la chute de Simon le magicien devant Néron ; 336 v°, sainte Marie-Madeleine recevant la communion des mains d'un ange ; 340, le martyre de s. Laurent ; 342, l'Assomption ; 344 v°, la naissance de la Vierge ; 348, l'Invention de la sainte croix (pl. XIII) ; 350 v°, s. Michel ; 353 v°, s. François d'Assise (les stigmates) ; 355 v°, le martyre de sainte Catherine ; 356 v°, s. Nicolas ; 357, femme soutenant d'une main deux anneaux entrelacés et de l'autre les lettres stylisées S. B. (ou F. B.) ; 357 v°, l'incredulité de s. Thomas ; 360, la pêche miraculeuse ; 362 v° et 365, scènes de martyre ; 367 v°, un confesseur ; 370, s. Jérôme ; 373, groupe de vierges ; 432 v°, le baptême du Sauveur.

La décoration du manuscrit se complète par un certain nombre d'initiales historiées sur fond d'or, renfermant des sujets variés et pittoresques, traités avec beaucoup de verve et de finesse ; plusieurs de ces sujets ornent les encadrements qui accompagnent les initiales : fol. 18 v°, enfant brandissant une pique ; 23 v°, dragon ailé ; 31 v°, jeune homme brisant un bâton sur son genou ; à sa droite, un jeune homme ; à sa gauche, une jeune femme ; 33, monstre ; 39, enfant grimpant à un arbre pour dénicher un nid d'oiseau ; 39 v°, enfant tuant un serpent à

FIGURE A.16 – Notice 1 océrisée, cinquième page

l'aide d'un poignard ; 40 v°, grotesque ; 41, enfants à cheval l'un sur un lion, l'autre sur un aigle ; 44 v° et 48 v°, oiseaux ; 48, homme sauvage ; jeune fille ; 49, chimère ; 73, oiseau ; 166, buste de femme ; 167 v° et 177, monstres ; 191, combat entre deux enfants nus ; 210, dragon ailé ; 230 v°, la mort ; 277, enfant combattant un monstre ; 293, Malachie ; 296, Isaïe ; 312, dragon ailé ; 305 v°, 309, 315 v°, 318 v°>, 321, 322 v°>, 325, 331, 340 v°, 342 v°, 345, prophètes, apôtres, martyrs et saints anonymes ; 348 v°, croix fichée ; 351, s. Michel ; 356, sainte Catherine ; 357, s. Nicolas ; 363 et 365 v°, martyrs ; 368 et 370 v°, confesseurs ; 373 v°, vierge ; 376 v°, s. Jean-Baptiste enfant ; 379 v°, 382, 383, 384, 385, 386, personnages divers ; 388, s. Jean-Baptiste ; 389 v° et 396, s. Nicolas ; 397, femme ; singe ; 398 v°, s. Nicolas ; 399 v°, femme ; chien ; 402, femme ; 403 v°, s. Antoine, ermite ; 417 v°, sainte Catherine ; 423, femme portant deux anneaux entrelacés. — Initiales fleuries sur fond d'or, fol. 19 et 36 v°. — Belles initiales de couleur dont le champ est occupé par des feuilles stylisées sur fond d'or. — Presque toutes ces initiales se prolongent dans les marges en lourds rinceaux de couleur et de feuillage semés de petites rosaces d'or dans le goût italien.

Il ressort de l'analyse ci-dessus que le manuscrit est un livre d'Heures à l'usage de Rome, ou, plus exactement, à l'usage franciscain, copié vers 1380 dans l'Italie du nord et probablement dans la région de Padoue. — Par qui les peintures ont-elles été exécutées ? En l'absence de signature et de renseignements puisés à des documents d'archives, il est difficile de mettre un nom en avant. Tout ce que l'on peut dire, c'est que les peintures appartiennent à l'école italienne. — Pour qui le manuscrit a-t-il été exécuté ? Question presque aussi difficile à résoudre que la première. Pour être en mesure de le faire, il faudrait connaître le nom du personnage qui se tient agenouillé aux pieds de la Vierge (fol. 102 v°) et qui porte un soleil sur son vêtement, emblème que l'on retrouve quoiqu'un peu différent sur le pourpoint de l'officier qui accompagne le centurion (fol. 72 et 75). Il faudrait également savoir qu'elle est cette femme symbolique qui figure au fol. 357 ; il faudrait enfin pénétrer la signification de l'emblème et des lettres quelle tient dans les mains et qui sont répandues à foison dans la dernière partie du manuscrit. En l'absence de conclusions positives, une chose du moins paraît certaine ; c'est que les armoiries des fol. D et 2 v° n'apportent aucun élément de solution du problème. Ainsi que l'a établi M. Max Prinet (cf. l'article cité dans la bibliographie ci-après), les deux blasons ont été ajoutés ; ils nous renseignent donc, non pas sur le personnage pour qui le manuscrit a été exécuté, mais seulement sur les possesseurs du livre d'Heures dans la première moitié du xvi<sup>e</sup> siècle : Claude et Anne (ou Annet) Régis.

Comme il a été dit un peu plus haut, le manuscrit 22 de la collection Smith-Lesouëf, à Nogent-sur-Marne, est une réplique du ms. latin 757. Mais il importe de distinguer soigneusement, dans ce livre d'Heures, la décoration originale de celle qui a été ajoutée. Cette dernière en effet appartient à une autre main, peut-être même à plusieurs mains. Seule la première dérive du ms. lat. 757 et mérite d'être décrite ici. Elle se compose de dix peintures à pleine page, les unes sur fond losangé ou quadrillé, les autres sur fond unicolore (or ou couleurs), toutes étroitement apparentées à celles du manuscrit de la Bibliothèque nationale, quoique relevant d'un art notamment inférieur. Les huit premières servent d'illustration aux Heures de la Vierge : fol. 15, la Vierge et l'enfant Jésus ; à leurs pieds se tient un personnage agenouillé ; derrière lui, sainte Catherine (?), s. Antoine ermite et s. Christophe (Matines) ; 24, la flagellation (Laudes) ; 34 v°, Jésus devant Pilate (Prime) ; 38 v°, Jésus portant sa croix (Tierce) ; 41 v°, Jésus étendu sur la croix (Sexte) ; 45, crucifixion (None) ; 48 v°, descente de croix (Vêpres) ; 57, mise au tombeau (Complies). La peinture du fol. 84, la salutation angélique, vient en tête des Sept joies de la Vierge ; celle du fol. 296, la trahison de Judas et l'arrestation de Jésus, se trouve au milieu de la Passion selon saint Marc. — A cette décoration primitive appartient également un cer-

FIGURE A.17 – Notice 1 océrisée, sixième page

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

tain nombre d'initiales historiées les unes sur fond losangé ou quadrille, les autres sur fond d'or ou de couleurs; celle du fol. 15 v° renferme la Vierge et l'enfant Jésus; au bas de l'encadrement figurent l'emblème formé de deux anneaux entrelacés et les lettres S. B. (ou F. B.) qui foisonnent dans le ms. lat. 757 ; au milieu, on aperçoit une targe : *d'azur au lion d'or (?) orné et lampassé de gueules*. Les anneaux et les lettres se retrouvent au bas de plusieurs encadrements.

Rel. maroquin rouge ; dos orné. — TOESCA (Pietro), *La pittura e la miniatura nella Lombardia*, 1912, p. 279 à 294. — COUDERC (Camille), *Album de portraits*, p. 19 et pl. XLV et XLVI. — PRINET (Max), *Bulletin de la Société nationale des antiquaires de France* (séance du 26 mars 1924), p. 137 à 141. — LEROQUAIS (abbé V.), *Les sacramentaires et les missels manuscrits des bibliothèques publiques de France*, 1924, t. II, p. 361-363 et pl. LXIX à LXXV. — ANCONA (Paolo d'), *La miniature italienne du X<sup>e</sup> au XVI<sup>e</sup> siècle*, 1925, p. 21 et pl. XVI. — [COUDERC (Camille)], *Catalogue de l'exposition du moyen âge*. Bibliothèque nationale, 1926, p. 35-36.

FIGURE A.18 – Notice 1 océrisée, septième page

II2. HEURES A L'USAGE DE ROME. XV<sup>e</sup> SIÈCLE

Bibliothèque nationale, ms. lat. 1399.

Fol. i. Anciennes cotes : « Codex Colb. 6323. » — « Regius, 4475, 3. » — 1 à 12. Calendrier de Chalon-sur-Saône. — (22 janv.) En lettres rouges : « S. Vincent. » — (27 janv.) « S. Lup evesq. » — (17 mars) « S. Agric evesq. » — (30 avr.) « S. Flave et Desiderez evesq. » — (11 mai) « S. Gengul mart. » — (24 mai) « S. Disier, ev. et mart. » — (6 juin) « S. Claude, ev. de Besançon. » — (14 juin) « S. Ruph de Valance. » — (7 juill.) « S. Gervas. » — (19 août) « S. Grant mart. » — (2 sept.) « S. Just mart. » abbé. » — (4 nov.) « S. Gengoul mart. »

Fol. 13 à 19. Extraits des quatre évangiles. — 19. « *Dévote oraison à nostre Dame.* » — 19 v°. « Obsecro te, domina sancta Maria, mater Dei, pietate plenissima... — 22 ...Et eciam in omnibus illis rebus in quibus ego sum facturus, locuturus aut cogitaturus... Et michi famulo tuo impetres a dilecto filio complementum... 24 ...mater Dei, pietatis et misericordie. Arnen. » — « *De nostre Dame.* »

Gaudie, flore virginali Que  
honore speciali... »

30 à 52. Office de la Passion. — 54 à 108. Heures de la Vierge ; le début manque. — 101. « Chi apres sensicut l'offisse (sic) de V advent.. » — Lacune entre 108 et 110.

110 à 115. Heures de la Croix ; le commencement et la fin manquent. — Lacune entre 115 et 116. — 116 à 118. Heures du Saint-Esprit; le commencement et la fin ont disparu. — Lacune entre 118 et 120. — 120 à 130. Psaumes de la pénitence; le commencement manque. — 130 v° à 137. Litanies. — 132. « ...s. Vincenti; s. Dyonisi cum sociis tuis ; s. Mauri cum sociis tuis... » — 137 v° à 174. Office des morts. — 174. « *Chi ensieuvent les commandasses.* Ant. Subvenite, sancti Dei... » — 181 v°. « *Oratio.* Domine Iesu Christe, qui septem verba... — 183 v° ...iocundari et commorari per infinita seculorum. Arnen. » — 187. D'une autre main :

« Glorieux saint Sébastien,  
Qui souffris comme bon chrétien... »

188. D'une autre main :

« Il fault morir, retenes bien, Retenes  
bien, je vous en prie... »

L'office de la Vierge et celui des morts représentent l'usage de Rome ; l'écriture est

FIGURE A.19 – Notice 112 océrisée, première page

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

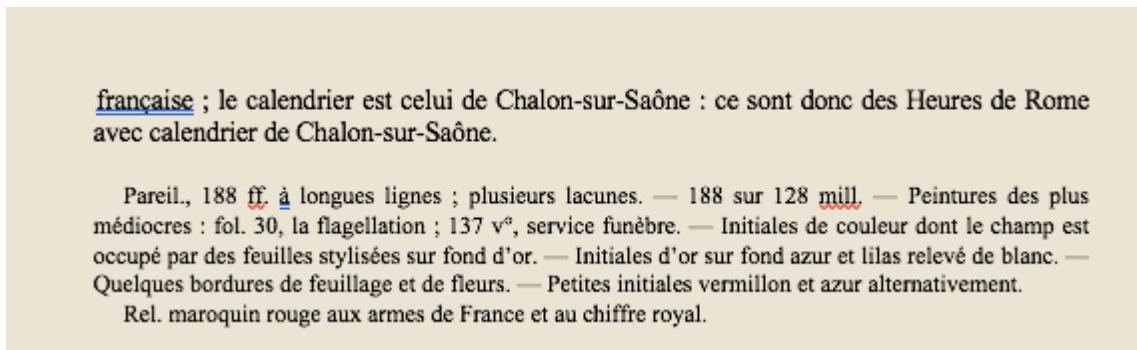


FIGURE A.20 – Notice 112 océrisée, deuxième page

313. HEURES A L'USAGE DE ROUEN ET GRAND COUTUMIER DE NORMANDIE. XV<sup>e</sup> SIÈCLE  
Bibliothèque nationale, fr. nouv. acq., 10230.

Fol. i. Au bas du feuillett, note en écriture moderne : « Phillipps ms. 22403. » — 1 à 6. Calendrier très clairsemé où dominent les saints caractéristiques d'Évreux — (25 mai) « S. Mauxe. » — (5 août) I'une autre main : « Nostre Dame des Neiges b. » — (12 août) D'une autre main : « Réduction de Normandie, p.a.b. » — (5 sept.) En lettres rouges : « S. Taurin. » — (1<sup>er</sup> oct.) « S. Rémy. — *Bouge, p.a.b.* » — (10 oct.) D'une autre main : « Le lundi, baillage recommanc. » — (16 oct.) En lettres rouges : « S. Michel. » — (25 oct.) D'une autre main : « Prinse de Rouen. 1562. » — (12 nov.) D'une autre main : « Parlement recommanc. » — (22 nov.) D'une autre main : « Ste Cecylle. — Le puy de la M... » La fin de ce dernier nom a été rognée par le relieur. — (30 déc.) « S. Ursin. »

Fol. 7 à 27. Heures de la Vierge. — 27 v° à 29. Heures de la Croix. — 29 à 30. Heures du Saint-Esprit. — 31 à 37. Psaumes de la pénitence. — 38 à 41. Litanies.

— 39. « ...s. Romane ; s. Audioene ; s. Mellone ; s. Vulgane ; s. Martine ; s. Macute ; s. Hugo ; s. Viviane ; s. Taurine ; s. Maxime ; s. Alexis ; s. Severe ; s. Patrici ; s.

FIGURE A.21 – Notice 313 océrisée, première page

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

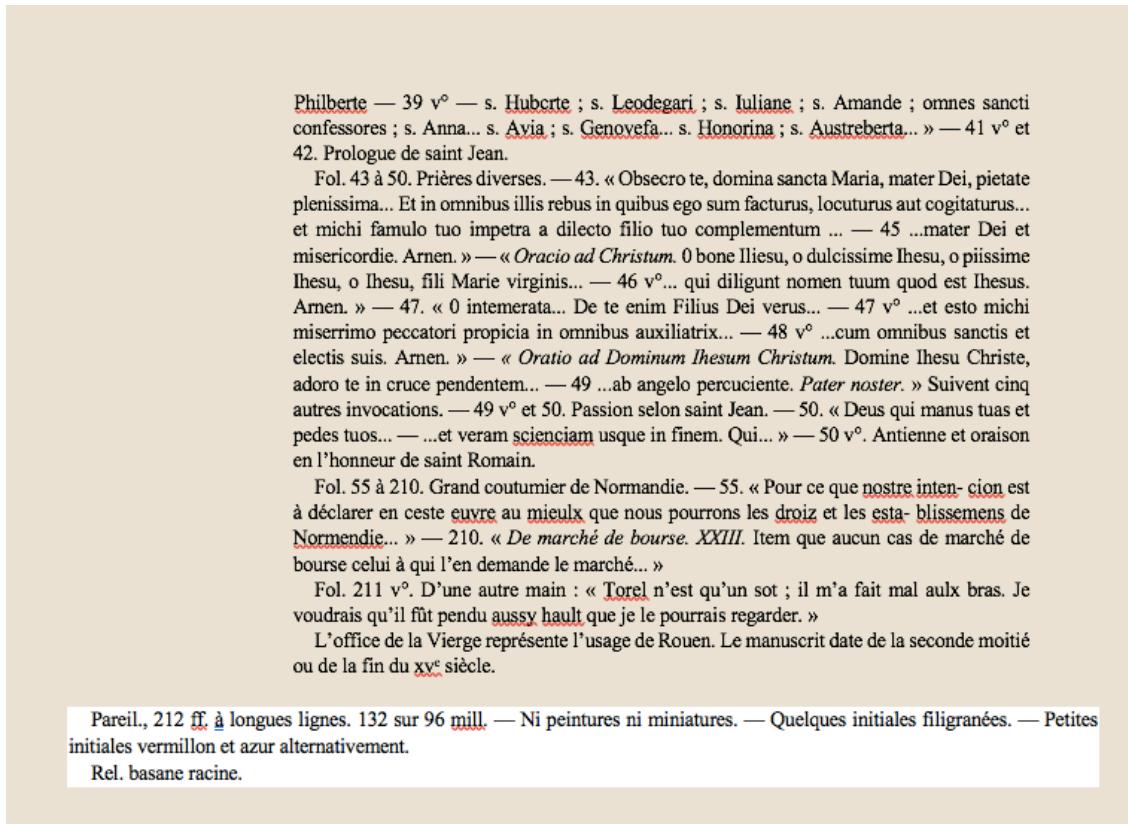


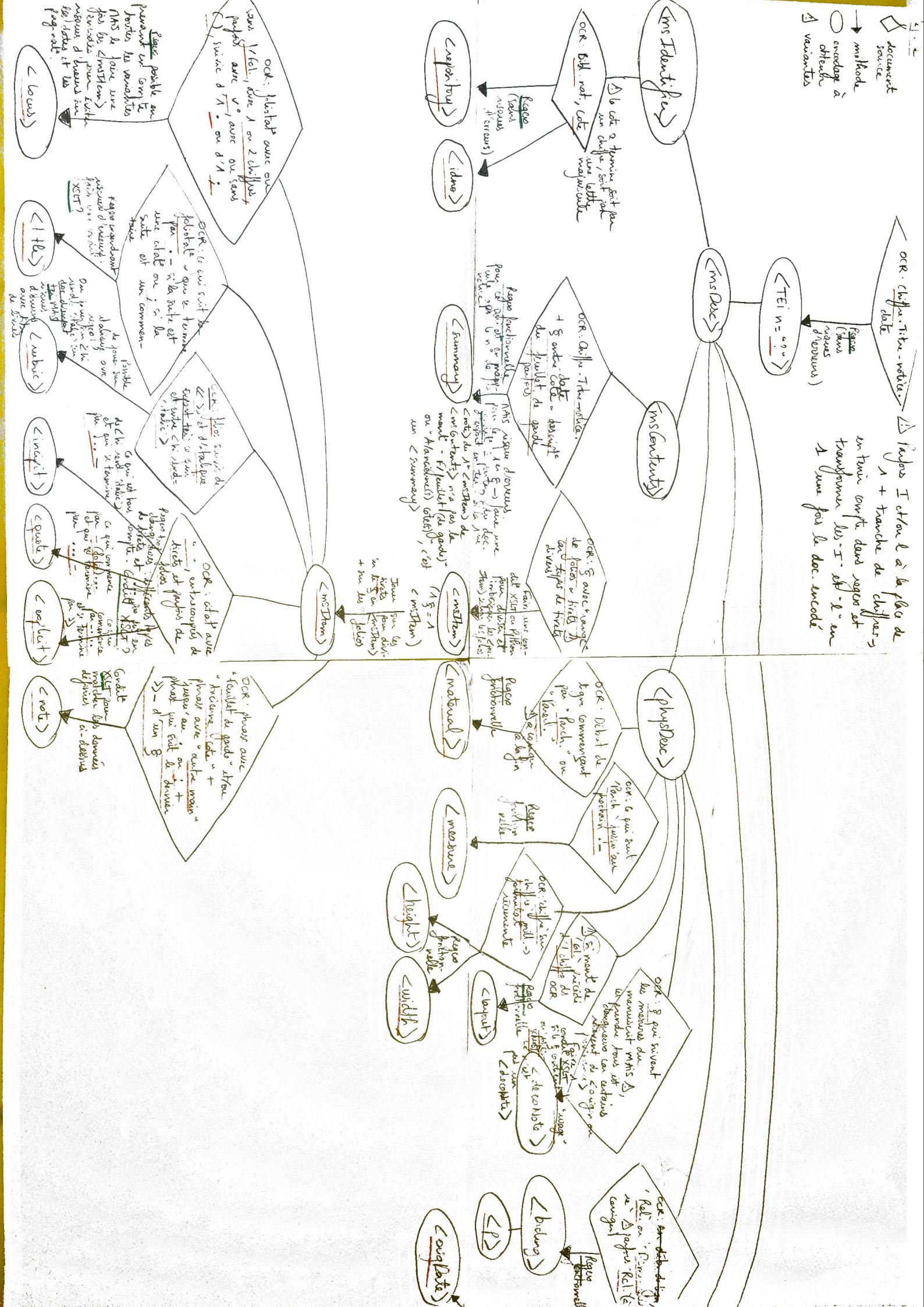
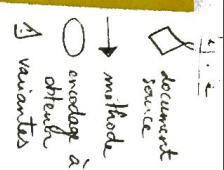
FIGURE A.22 – Notice 313 océrisée, deuxième page

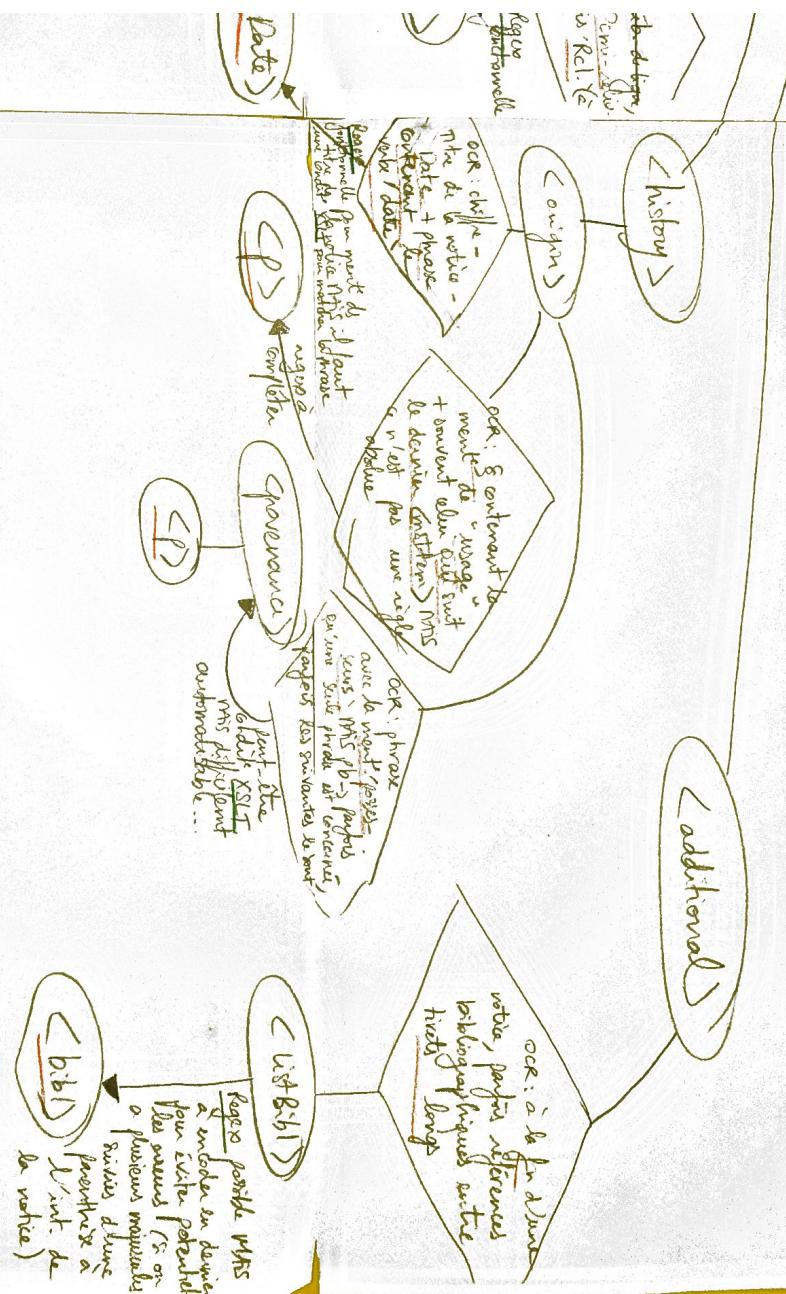
### A.1.3 Analyses du document à encoder

La phase d'observation des documents à encoder le plus automatiquement possible s'est concrétisée par la formalisation d'arbres de décision, censés refléter d'une part la structure des notices et d'autre part leurs irrégularités. Face à ces cas concrets s'offre un ensemble de choix et de décisions à prendre selon le document de sortie *in fine* souhaité. L'arbre de décision a pour intérêt de souligner les grandes étapes qui doivent apparaître dans le code de transformation.

#### Arbre de décision : première version

La version ici proposée a pour principal atout de mettre en valeur la structure finale souhaitée en partant du document d'origine. Il permet donc de visualiser, selon les balises TEI, où se trouve les informations à encoder dans le document source, et donc d'avoir conscience des informations qui se trouvent à divers endroits du document d'origine mais qui sont ensuite regroupées sous une même balise. Si cette situation reste relativement rare, l'information relevant de la balise <summary> illustre ce cas. Désignant ce qui concerne le contenu général du manuscrit, l'information se situe systématiquement dans le titre de la notice en majuscule, mais aussi parfois dans un premier paragraphe. Toutefois, cette version a pour défaut de ne pas assez représenter la structure des notices rédigées par Leroquais.





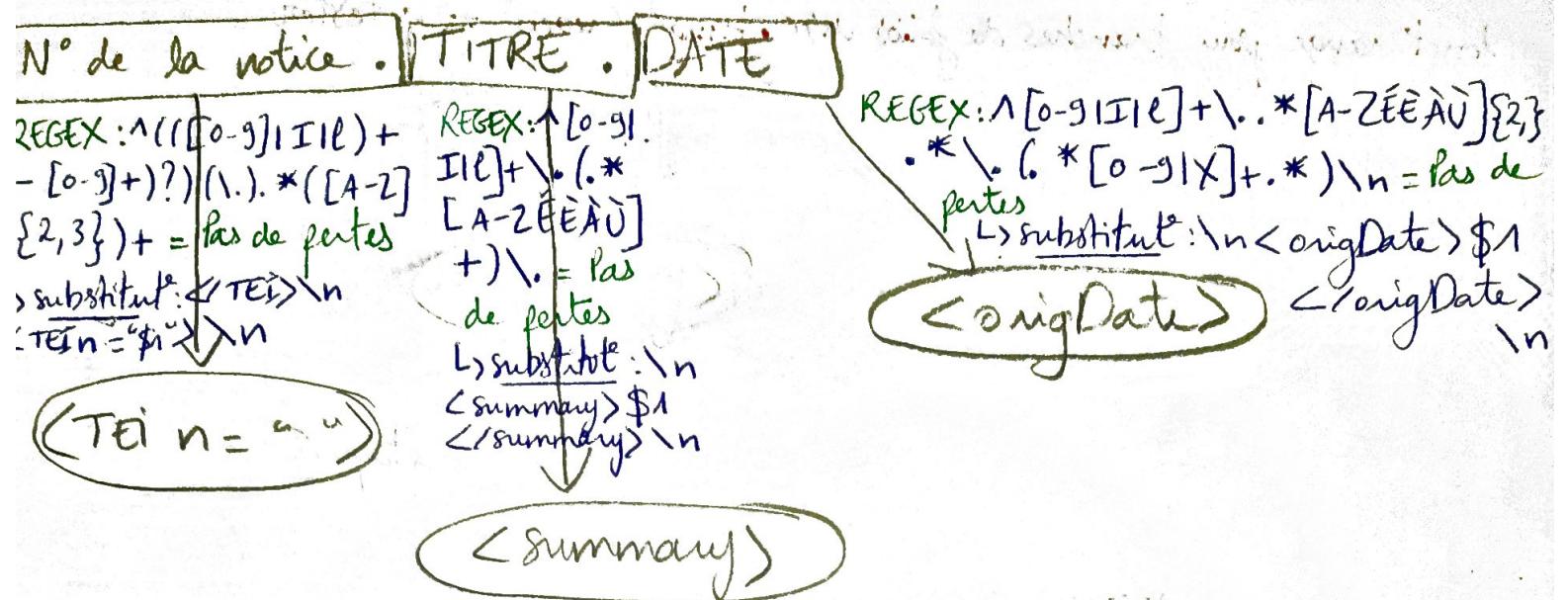
## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

### **Arbre de décision : deuxième version**

Contrairement à la première, cette deuxième version se veut plus proche du document d'origine à encoder, tout en soulignant les solutions, difficultés et interrogations qui ont émergées autour de la possibilité d'une automatisation de l'encodage.

# Arbre Notice papier



Bibliothèque nationale, cote.

REGEX: ^((Bibliothèque nationale), )\{1\} = Pas de pertes  
 > substitut: \n <repository> \$2  
 <repository> \n

REGEX: ^((Bibliothèque nationale), \{1\}) (. \* [0-9+ | A-Z] ? | .) \$ \n = Pas de pertes  
 > substitut: \n <idno> \$3  
 <idno> \n

**<repository>**

Si **1<sup>er</sup> §** ne débute pas par un folio ou la mention d'un feuillet de garde (fol. 1 feuillett)

Condit Python ou XSLT: Si il y a un § apr. la cote et avant la description du feuillet de garde et/ou de l'ancienne cote (à tester)

→ **<summary>**

§ avec folios, citat et tirets longs:

→ Majorer la § → tiret: long

→ Regex trop dangereuses

Condit en Python ou XSLT: jouer sur les §, les tirets et les n° de folios (à tester)

↓ sauf si le tiret long est suivie d'une ( (y calendrier)

**<msItem>**

→ si dans un même §, les n° de folios qui suivent sont compris dans la tranche de folios qui précède (n<sup>e</sup> à n<sup>e</sup>) → créer un <msItem> jusqu'au prochain

titre long suivi d'un n° de folio > au dernier de la tran-  
che définie, ou bien = avec v°. Condition XSLT ou Python avec  
fonct range pour tranches de folios avec .format (regex) → à tester

↳  $\boxed{(\text{f/fol.}) \text{n}^{\circ} (\text{à } \text{n}^{\circ}) (\text{v}^{\circ})}$ ; ou ; (A les n°  
de folios sont parfois encadrés par des ( ))

↓  
à faire apr. avoir enco-  
dé les autres éléments  
du <constItem>  
étape ⑥

↳  $n^{\circ} \text{ de folio(s)} + \boxed{\text{titre de la partie}} . -$

REGEX:  $\begin{array}{l} - ? ((\text{à } \text{l et } \text{l}^{\wedge} \ll \cdot \ast)) [0-9] \\ A-21 \text{v}^{\circ} ] \{1,3\} \backslash . ) ([A-2\ddot{\text{E}}\ddot{\text{E}}\ddot{\text{A}}\ddot{\text{U}}]^+ \\ [a-2\ddot{\text{E}}\ddot{\text{E}}\ddot{\text{A}}\ddot{\text{U}}]^+ (. \ast ?) [; \text{N} \cdot \text{N}]) \\ \Rightarrow \text{pas de parties MAIS cela ne matche pas les rubriques qui sont officielles de titre} \end{array}$

étape ①

↳  $\boxed{n^{\circ} \text{ de folio(s)} + \ll \text{italique} \gg}$

Comment jouer sur  
italique  
(seule différence  
avec incipit)?

étape ②

<rubric>

↳  $\boxed{(\text{n}^{\circ} \text{ de folio(s)}) + (\ll \text{) testé + \{ : \} \gg} . -}$

étape ③

<incipit> ?

$(\text{n}^{\circ} \text{ de folio(s)}) \dots \text{ testé } \dots . -$

étape ④

<quote> ?

↳  $\boxed{- (\text{n}^{\circ}) \dots + \text{testé} + \{ \cdot \} \gg}$

Excepté que l'id n'y a pas de différences  
mais entre les 3 types de citat.  
REGEX:  $\begin{array}{l} (- ?) ((0-9) A-21 \text{v}^{\circ}) \{1,3\} \backslash - \\ ? \ll ((\cdot \ast ?) \{1\} \cdot ) \{3\} ? ! (\backslash \cdot \cdot \cdot ) \\ (\gg ?) \{1\} \cdot \cdot \cdot \{3\} \text{ citat } \text{perdues} \\ (= \text{signe de testé } \text{matchées } \text{perdues}) \end{array}$

à faire apr. avoir délimité  
les <constItem> pour éviter les erreurs

Etape 5

③

< explicit > ?

Etape 1

Ancienne côte

au · ou au

→ d'un > / Etape 3 / Rest du calendrier de · — à >

avec jours entre { } :

Credit XSLT ou Python (parment du calendrier)

car regex trop dangereuses (peuvent être qu'il ne faut) : si un titre a la ment "calendrier" suivi d'un · d'au plus espaces et jusqu'à un saut de ligne prendre ce qui suit le 1- (Avec first) jusqu'au dernier > suivi de fol / chiffre / Maj.

< note >

REGEX (jeu feuillets de garde x anc côte)  
(C.\*?)(Feuillet(s)?|Feuillet(s)?+(de garde))  
l'anciennes(s)?|Anciennes(s)? côte)+(\*?)  
(\-[0-9]{1,2}A-Z){1,2} Mais trouver  
moyen d'exclure la phrase qui contient  
"Parch/Pareil"

↳ substitut: \n<note>\$1</note>\n

\*

§ suivant ceux sur le contenu : ment de l'usage, souvent de l'office et du calendrier

↓ Etape 1

< origin >

REGEX: ((.+)((usage|calendrier|office)+(.+?))\n=> pas de pertes mais  
risque de perdre que il ne faut (élevant de décompte) notammen

↳ substitut: \n<p>\$1</p>\n

↳ phrases avec la ment de date

→ < origDate >

Etape 2  
REGEX: \.\s+([A-ZÉÉÀÀÙÙ]) (.+?) date(.+?)\.\.  
=> risque de ne pas tout matcher  
↳ substitut: \n<origDate>\$1<origDate>\n

Credit Python ou XSLT:抓取 phrases contenant vb "date".

Saut de ligne (double) pris Parch./Pareil.

REGEX: \n\s\*(Parch\.|Pareil\.), => pas de pertes

↳ substitut: \n<material>\$1</material>\n

→ < material >

nbre de folios

(nbre de col.) jusqu'au · —

REGEX: ([0-9]+ col\.-)

=> pas de pertes

↳ substitut: ([0-9]+ col\.-)

< layout >

REGEX: \n^ (Parch\.|Pareil\.), \s ([0-9]+ \s) \. (.+?) (col\.-)? (.+?) \. )

=> pas de pertes

↳ substitut: \n<measure>\$2</measure>\n

< measure >

✓ chiffre sur chiffre mill.

(4)



REGEX:  $([0-9]+)$  sur  $([0-9]+)$  mill  
=> pas de pertes  
↳ substitut: \n < height > \$1 < height >  
\n < width > \$2 < width >

— phrase(s) ou § avec ment de l'"peinture",  
1° "décorat", 2° "initiales", 3° "miniatures" → entre  
les dimens° du manuscrit et la reliure.

+ On dit Python / XSLT (regressions trop dangereuses): à chaque notice, matcher ce qui est  
entre mill. — et Rel. / Rcl. / Demi-reliure.  
étape ①

< decoNote >

étape ② => si "usage" et/ou "possesseurs"

REGEX: \n (.\*)? (usage | possesseur) (.\*) \n => pas  
de pertes pour ce qui a été  
matché pour < decoNote >

< origin >

+ On dit Python ou XSLT pour  
extraire la phrase et les suivantes  
contenant "possesseur".

< p >

< provenance >

< CPS >

§ commençant par l'" Rel. / Rcl. / Demi-reliure " jusqu'  
au .(→)

< binding >

REGEX: ((Rel|.Rcl|.Demi)-  
reliure)(.\*?)(.)(( |-) [A-ZÉ  
ÄÀÙ])? => pas de pertes  
↳ substitut: \n < p > \$1 < p > \n

Si — Non ( jusqu'à chiffre ou MAISULT. 5)

étape ① REGEX: (- )?((\D)?[A-ZÉÉÀÙ]{2,3}\.(.|\\*\?))(-|\n) => pas de pertes pour biblio. à la fin de la notice MAIS ↗ si dans la notice Maj. suivie de ( et de lettres dans la notice, ça marche aussi. Ne pas prendre en compte ce qui a déjà été encodé?

étape ② REGEX: ((\D)?[A-ZÉÉÀÙ]{2,3}\.(.|\\*\?))\.+|-+|\n) => Sépare les réf à partir des tirets MAIS ↗ ne marche pas la dernière réf.  
Corriger les erreurs d'OCR rencontrées à la fin de l'encodage :

- Pareil. → Panch.
- Armen → Armen
- Rcl. → Rel.
- ds @n de <TTI> → remplacer "I" et/ou "l" par 1

\* <note> au sein d'un <msItem> :

étape ④

- fin d'un § après >> -

? Regles trop dangereuse car \n.\*>>\s+ - \s+.\*) peut matcher la bonne <note> mais aussi d'autres fins de § construites de la même manière (citat ou titre d'un <msItem>)

étape ⑤

- § au milieu de 2 autres sans aucun  
de folios suivis d'un titre ↴ aux sauts de lignes dans un  
m§ crées par l'OCR

Condition XSLT ou Python : regles (>>\s+\n?|\n.\*?\n\*\(.\*\|\n\*\.\*\)\n fol | [0-9A-Z]) trop dangereuse car matche ④ qu'il ne faut. Bien  
matcher un § en entier dans la condition.

étape ②

- Pour la ment de "L/lacune"  
entre - et . - / ou st de ligne

REGEX : ((lacune|lacune) entre(.\*)?)\.\{1\}(\s+|-)? => pas de  
pertes

↳ substitutP = \n<note>\$1</note>\n

étape ⑥

- Pour la ment d'une autre main  
précédée de - ou de . ou de ) et suivie de ( ou de :

Regles délicate : (\-|([0-9A-Z]\.))|([a-zA-Z]\))\{1\}+(.\*)? une  
autre main (.\*)?)) (\(1:)) => Pas de pertes mais cela matche parfois  
trop. Match juste pour ". D'une autre main : "

#### A.1.4 Définition du document cible

Personnalisation du fichier ODD relatif au msDesc

```
<!--Checking module msdescription-->
<classRef key="att.msClass"/>
<classRef key="model.physDescPart"/>
<moduleRef key="msdescription"
            include="msDesc catchwords dimensions dim height
                     width locus material origDate origPlace
                     signatures watermark msIdentifier repository
                     altIdentifier colophon explicit finalRubric
                     incipit msContents msItem rubric summary
                     physDesc objectDesc supportDesc support
                     collation foliation condition layoutDesc
                     layout handDesc scriptDesc musicNotation
                     decoDesc decoNote bindingDesc binding
                     history origin provenance additional
                     adminInfo recordHist source surrogates msPart"/>
<elementSpec ident="msDesc" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
        <attDef ident="type" mode="delete"/>
        <!-- Suppression de l'attribut "status" qui apparaît
             systématiquement sur l'élément <msDesc>
             avec la valeur "draft". -->
        <attDef ident="status" mode="delete" />
    </attList>
</elementSpec>
<elementSpec ident="catchwords" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<attDef ident="facs" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="dimensions" mode="change">
    <!-- Définition d'une séquence d'éléments
        à l'intérieur de l'élément "dimensions". -->
    <content>
        <sequence>
            <elementRef key="height" minOccurs="1" maxOccurs="1"/>
            <elementRef key="width" minOccurs="1" maxOccurs="1"/>
            <elementRef key="dim" minOccurs="0" maxOccurs="1"/>
        </sequence>
    </content>
    <attList>
        <!-- Ajout de l'attribut "type" avec les valeurs
            suggérées "justification" et "leaf". -->
        <attDef ident="type" mode="change">
            <valList mode="add">
                <valItem ident="justification"/>
                <valItem ident="leaf"/>
            </valList>
        </attDef>
        <!-- Ajout de l'attribut "scope" avec la valeur
            suggérée "all". -->
        <attDef ident="scope" mode="change">
            <valList mode="add">
                <valItem ident="all"/>
            </valList>
        </attDef>
        <!-- Ajout de l'attribut "unit" avec la valeur
            "mm" pour normaliser l'unité de mesure. -->
        <attDef ident="unit" mode="change">
            <valList mode="add">
                <valItem ident="mm"/>
            </valList>
        </attDef>
```

```

<!-- Ajout de l'attribut "corresp". -->
<attDef ident="corresp" mode="change"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
<attDef ident="quantity" mode="delete"/>
<attDef ident="atLeast" mode="delete"/>
<attDef ident="atMost" mode="delete"/>
<attDef ident="min" mode="delete"/>
<attDef ident="max" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="dim" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
<attDef ident="type" mode="delete"/>
<attDef ident="unit" mode="delete"/>
<attDef ident="quantity" mode="delete"/>
<attDef ident="scope" mode="delete"/>
<attDef ident="atLeast" mode="delete"/>
<attDef ident="atMost" mode="delete"/>
<attDef ident="min" mode="delete"/>
<attDef ident="max" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="height" mode="change">
<attList>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
<attDef ident="unit" mode="delete"/>
<!-- Ajout de l'attribut "quantity" dans l'élément
"height" qui a pour type de données un nombre. -->
<attDef ident="quantity" mode="change">
    <datatype>
        <dataRef key="teidata.numeric"/>
    </datatype>
</attDef>
</attList>
</elementSpec>
<elementSpec ident="width" mode="change">
    <attList>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
        <attDef ident="unit" mode="delete"/>
        <!-- Ajout de l'attribut "quantity" dans l'élément
        "width" qui a pour type de données un nombre. -->
        <attDef ident="quantity" mode="change">
            <datatype>
                <dataRef key="teidata.numeric"/>
            </datatype>
        </attDef>
    </attList>
</elementSpec>
<elementSpec ident="locus" mode="change">
    <attList>
        <attDef ident="scheme" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
```

```

<attDef ident="source" mode="delete"/>
<attDef ident="target" mode="delete"/>
<attDef ident="type" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="material" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
<attDef ident="ref" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="origDate" mode="change">
<attList>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="unit" mode="delete"/>
<attDef ident="quantity" mode="delete"/>
<attDef ident="scope" mode="delete"/>
<attDef ident="atLeast" mode="delete"/>
<attDef ident="atMost" mode="delete"/>
<attDef ident="min" mode="delete"/>
<attDef ident="max" mode="delete"/>
<attDef ident="type" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="origPlace" mode="change">
<attList>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="role" mode="delete"/>

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<attDef ident="notBefore" mode="delete"/>
<attDef ident="notAfter" mode="delete"/>
<attDef ident="from" mode="delete"/>
<attDef ident="to" mode="delete"/>
<attDef ident="type" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="signatures" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="facis" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="watermark" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="facis" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="msIdentifier" mode="change">
<!-- Définition d'un enchaînement d'éléments
à l'intérieur de l'élément &lt;elementSpec&gt; : la séquence
comprend au minimum l'élément &lt;idno&gt;. --&gt;
&lt;content&gt;
&lt;sequence preserveOrder="true"&gt;
&lt;elementRef key="settlement" minOccurs="0" maxOccurs="1"/&gt;
&lt;elementRef key="repository" minOccurs="0" maxOccurs="1"/&gt;
&lt;elementRef key="idno" minOccurs="1" maxOccurs="1"/&gt;
&lt;elementRef key="altIdentifier" minOccurs="0" maxOccurs="1"/&gt;</pre>
```

```

        </sequence>
    </content>
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="repository" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
        <attDef ident="role" mode="delete"/>
        <attDef ident="ref" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="altIdentifier" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
        <attDef ident="type" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="colophon" mode="change">
    <attList>

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="explicit" mode="change">
<attList>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
<attDef ident="type" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="finalRubric" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
<attDef ident="type" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="incipit" mode="change">
<attList>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
```

```

<attDef ident="source" mode="delete"/>
<attDef ident="type" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="msContents" mode="change">
    <!-- Définition d'une séquence d'éléments à l'intérieur
        de l'élément <msContents>, qui est parfois vide.
        Il est préférable d'avoir un élément "msItem" par œuvre
        au sein du manuscrit, et de ne pas les multiplier au sein
        de l'élément qui prend son sens dans l'individualité
        de ce qu'il décrit. -->
    <content>
        <sequence preserveOrder="true">
            <elementRef key="summary" minOccurs="0" maxOccurs="1" />
            <elementRef key="textLang" minOccurs="0" maxOccurs="1" />
            <elementRef key="msItem" minOccurs="0" maxOccurs="unbounded" />
        </sequence>
    </content>
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
        <attDef ident="class" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="msItem" mode="change">
    <constraintSpec ident="class" scheme="isoschematron">
        <!-- Restriction des valeurs de l'attribut "class" aux "xml:id"
            déclarés dans les éléments <category>.
            La valeur de l'attribut "class" renvoie en effet aux catégories
            de livres définies dans l'élément "category".
            Message d'alerte si l'attribut "class" n'apparaît pas. -->
        <constraint>
            <!--<rule xmlns="http://purl.oclc.org/dsdl/schematron"
                context="//tei:msItem" role="warning">-->

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<assert xmlns="http://purl.oclc.org/dsdl/schematron"
        test="substring-after(@class, '#') =
ancestor::tei:teiCorpus/tei:teiHeader
/tei:encodingDesc/tei:classDecl/tei:taxonomy
/tei:category/@xml:id">
    Please select a valid class</assert>
<!--</rule>-->
</constraint>
</constraintSpec>
<!-- Définition d'une séquence d'éléments dans l'élément &lt;msItem&gt;.
Aucun des éléments ci-dessous n'apparaissent systématiquement.
Par ailleurs, on peut retrouver plusieurs fois la séquence
dans un même &lt;msItem&gt;. Le titre et l'auteur peuvent apparaître
plusieurs fois pour signaler d'éventuels traductions
dans plusieurs langues. --&gt;
&lt;content&gt;
    &lt;sequence preserveOrder="true"&gt;
        &lt;elementRef key="locus" minOccurs="1"
maxOccurs="1"/&gt;
        &lt;elementRef key="author" minOccurs="0"
maxOccurs="unbounded"/&gt;
        &lt;elementRef key="title" minOccurs="0"
maxOccurs="unbounded"/&gt;
        &lt;elementRef key="rubric" minOccurs="0"
maxOccurs="1"/&gt;
        &lt;elementRef key="incipit" minOccurs="0"
maxOccurs="1"/&gt;
        &lt;elementRef key="quote" minOccurs="0"
maxOccurs="unbounded"/&gt;
        &lt;elementRef key="explicit" minOccurs="0"
maxOccurs="1"/&gt;
        &lt;elementRef key="colophon" minOccurs="0"
maxOccurs="1"/&gt;
        &lt;elementRef key="finalRubric" minOccurs="0"
maxOccurs="1"/&gt;
        &lt;elementRef key="note" minOccurs="0"
maxOccurs="unbounded"/&gt;
    &lt;/sequence&gt;
&lt;/content&gt;</pre>
```

```

<attList>
    <attDef ident="class" mode="change" usage="rec"/>
    <attDef ident="corresp" mode="delete"/>
    <attDef ident="sameAs" mode="delete"/>
    <attDef ident="ana" mode="delete"/>
    <attDef ident="fac" mode="delete"/>
    <attDef ident="cert" mode="delete"/>
    <attDef ident="resp" mode="delete"/>
    <attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="rubric" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
        <attDef ident="type" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="summary" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="physDesc" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="objectDesc" mode="change">
<!-- Définition d'une séquence d'éléments dans
l'élément &lt;objectDesc&gt;. Les éléments &lt;supportDesc&gt; et
&lt;layoutDesc&gt; apparaissent systématiquement une fois. --&gt;
&lt;content&gt;
&lt;sequence preserveOrder="true"&gt;
&lt;elementRef key="supportDesc" minOccurs="1"
maxOccurs="1"/&gt;
&lt;elementRef key="layoutDesc" minOccurs="1"
maxOccurs="1"/&gt;
&lt;/sequence&gt;
&lt;/content&gt;
&lt;attList&gt;
&lt;attDef ident="corresp" mode="delete"/&gt;
&lt;attDef ident="sameAs" mode="delete"/&gt;
&lt;attDef ident="ana" mode="delete"/&gt;
&lt;attDef ident="fac" mode="delete"/&gt;
&lt;attDef ident="cert" mode="delete"/&gt;
&lt;attDef ident="resp" mode="delete"/&gt;
&lt;attDef ident="source" mode="delete"/&gt;
<!-- Ajout de l'attribut "form" rendu obligatoire,
qui peut prendre la valeur "codex". --&gt;
&lt;attDef ident="form" mode="change" usage="req"&gt;
&lt;valList&gt;
&lt;valItem ident="codex" mode="add"/&gt;
&lt;/valList&gt;
&lt;/attDef&gt;
&lt;/attList&gt;
&lt;/elementSpec&gt;
&lt;elementSpec ident="supportDesc" mode="change"&gt;
<!-- Définition d'une séquence d'éléments imbriqués
dans la balise &lt;supportDesc&gt; : support, extent,
folitaion et collation. --&gt;</pre>
```

```

<content>
  <sequence>
    <elementRef key="support" mode="change"/>
    <elementRef key="extent" mode="change"/>
    <elementRef key="foliation" mode="change"/>
    <elementRef key="collation" mode="change"/>
  </sequence>
</content>
<attList>
  <attDef ident="corresp" mode="delete"/>
  <attDef ident="sameAs" mode="delete"/>
  <attDef ident="ana" mode="delete"/>
  <attDef ident="facis" mode="delete"/>
  <attDef ident="cert" mode="delete"/>
  <attDef ident="resp" mode="delete"/>
  <attDef ident="source" mode="delete"/>
  <!-- Attribut "material" dans l'élément "supportDesc"
  pour des questions de normalisation, avec les valeurs adéquates
  conformes aux teiguidelines. -->
  <attDef ident="material" mode="change" usage="req">
    <valList>
      <valItem ident="parch" mode="add"/>
      <valItem ident="paper" mode="add"/>
      <valItem ident="mixed" mode="add"/>
    </valList>
  </attDef>
</attList>
</elementSpec>
<elementSpec ident="support" mode="change">
  <!-- Élément "material" à l'intérieur de
  l'élément "support" rendu obligatoire. -->
  <content>
    <sequence>
      <elementRef key="material" minOccurs="1"
                  maxOccurs="1"/>
    </sequence>
  </content>
  <attList>
    <attDef ident="corresp" mode="delete"/>

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="collation" mode="change">
    <!-- L'élément "collation" peut contenir
        la liste d'éléments suivants.
        Attention à bien mettre les bonnes informations
        dans les éléments correspondants
        pour une meilleure structuration des données. -->
<content>
    <sequence>
        <elementRef key="desc" minOccurs="0" maxOccurs="1"/>
        <elementRef key="formula" minOccurs="0" maxOccurs="1"/>
        <elementRef key="p" minOccurs="0" maxOccurs="unbounded"/>
        <elementRef key="q" minOccurs="0" maxOccurs="unbounded"/>
        <elementRef key="signatures" minOccurs="0" maxOccurs="1"/>
        <elementRef key="catchwords" minOccurs="0" maxOccurs="1"/>
    </sequence>
</content>
<attList>
    <attDef ident="corresp" mode="delete"/>
    <attDef ident="sameAs" mode="delete"/>
    <attDef ident="ana" mode="delete"/>
    <attDef ident="fac" mode="delete"/>
    <attDef ident="cert" mode="delete"/>
    <attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="foliation" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
```

```

<attDef ident="cert" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="condition" mode="change">
  <attList>
    <attDef ident="corresp" mode="delete"/>
    <attDef ident="sameAs" mode="delete"/>
    <attDef ident="ana" mode="delete"/>
    <attDef ident="fac" mode="delete"/>
    <attDef ident="cert" mode="delete"/>
    <attDef ident="resp" mode="delete"/>
    <attDef ident="source" mode="delete"/>
  </attList>
</elementSpec>
<elementSpec ident="layoutDesc" mode="change">
  <!-- Apparition au moins une fois
  de l'élément "layout" dans l'élément "layoutDesc". -->
  <content>
    <sequence>
      <elementRef key="layout" minOccurs="1" maxOccurs="unbounded"/>
    </sequence>
  </content>
  <attList>
    <attDef ident="sameAs" mode="delete"/>
    <attDef ident="ana" mode="delete"/>
    <attDef ident="fac" mode="delete"/>
    <attDef ident="cert" mode="delete"/>
    <attDef ident="resp" mode="delete"/>
    <attDef ident="source" mode="delete"/>
  </attList>
</elementSpec>
<elementSpec ident="layout" mode="change">
  <!-- Ajout de l'élément "desc" qui peut apparaître
  plusieurs fois ou ne pas apparaître au sein
  de l'élément "layout". -->
  <content>
    <sequence>
      <elementRef key="desc" minOccurs="0"

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
        maxOccurs="unbounded" />
    </sequence>
</content>
<!-- Ajout des attributs "columns" et "writtenlines"
sur l'élément "layout", attributs
qui ont tous deux des valeurs numériques. --&gt;
&lt;attList&gt;
    &lt;attDef ident="columns" mode="change"/&gt;
    &lt;datatype&gt;
        &lt;dataRef key="teidata.numeric"/&gt;
    &lt;/datatype&gt;
    &lt;attDef ident="writtenlines" mode="change"/&gt;
    &lt;datatype&gt;
        &lt;dataRef key="teidata.numeric"/&gt;
    &lt;/datatype&gt;
    &lt;attDef ident="streams" mode="delete"/&gt;
    &lt;attDef ident="sameAs" mode="delete"/&gt;
    &lt;attDef ident="ana" mode="delete"/&gt;
    &lt;attDef ident="fac" mode="delete"/&gt;
    &lt;attDef ident="cert" mode="delete"/&gt;
    &lt;attDef ident="resp" mode="delete"/&gt;
    &lt;attDef ident="source" mode="delete"/&gt;
&lt;/attList&gt;
&lt;/elementSpec&gt;
&lt;elementSpec ident="handDesc" mode="change"&gt;
    &lt;attList&gt;
        &lt;attDef ident="hands" mode="delete"/&gt;
        &lt;attDef ident="corresp" mode="delete"/&gt;
        &lt;attDef ident="sameAs" mode="delete"/&gt;
        &lt;attDef ident="ana" mode="delete"/&gt;
        &lt;attDef ident="fac" mode="delete"/&gt;
        &lt;attDef ident="cert" mode="delete"/&gt;
        &lt;attDef ident="resp" mode="delete"/&gt;
        &lt;attDef ident="source" mode="delete"/&gt;
    &lt;/attList&gt;
&lt;/elementSpec&gt;
&lt;elementSpec ident="scriptDesc" mode="change"&gt;
    &lt;attList&gt;
        &lt;attDef ident="sameAs" mode="delete"/&gt;</pre>
```

```

<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="musicNotation" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="decoDesc" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="decoNote" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="source" mode="delete"/>
<attDef ident="type" mode="delete"/>
</attList>

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
</elementSpec>
<elementSpec ident="bindingDesc" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="binding" mode="change">
    <attList>
        <attDef ident="contemporary" mode="delete"/>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="source" mode="delete"/>
        <attDef ident="when" mode="delete"/>
        <attDef ident="notBefore" mode="delete"/>
        <attDef ident="notAfter" mode="delete"/>
        <attDef ident="from" mode="delete"/>
        <attDef ident="to" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="history" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
    </attList>
</elementSpec>
```

```

<elementSpec ident="origin" mode="change">
    <attList>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="facis" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
        <attDef ident="evidence" mode="delete"/>
        <attDef ident="when" mode="delete"/>
        <attDef ident="notBefore" mode="delete"/>
        <attDef ident="notAfter" mode="delete"/>
        <attDef ident="from" mode="delete"/>
        <attDef ident="to" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="provenance" mode="change">
    <constraintSpec ident="provStruct" scheme="isoschematron">
        <constraint>
<!-- &lt;rule xmlns="http://purl.oclc.org/dsdl/schematron" context="//tei:provenance" role="warning"&gt; --&gt;
            &lt;assert xmlns="http://purl.oclc.org/dsdl/schematron"
                test = "(child::tei:orgName
                    or child::tei:placeName
                    or child::tei:persName)
                    and child::tei:p"&gt;
                Provenance must contain
                (orgName or placeName or persName)
                and (p)&lt;/assert&gt;
            &lt!--&lt;/rule&gt;--&gt;
        &lt;/constraint&gt;
    &lt;/constraintSpec&gt;
    &lt;attList&gt;
        &lt;attDef ident="corresp" mode="delete"/&gt;
        &lt;attDef ident="sameAs" mode="delete"/&gt;
        &lt;attDef ident="ana" mode="delete"/&gt;
        &lt;attDef ident="facis" mode="delete"/&gt;
        &lt;attDef ident="cert" mode="delete"/&gt;
        &lt;attDef ident="source" mode="delete"/&gt;
</pre>

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<attDef ident="from" mode="delete"/>
<attDef ident="to" mode="delete"/>
<attDef ident="type" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="additional" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="adminInfo" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="recordHist" mode="change">
<attList>
<attDef ident="corresp" mode="delete"/>
<attDef ident="sameAs" mode="delete"/>
<attDef ident="ana" mode="delete"/>
<attDef ident="fac" mode="delete"/>
<attDef ident="cert" mode="delete"/>
<attDef ident="resp" mode="delete"/>
<attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="source" mode="change">
```

```

<attList>
    <attDef ident="corresp" mode="delete"/>
    <attDef ident="sameAs" mode="delete"/>
    <attDef ident="ana" mode="delete"/>
    <attDef ident="fac" mode="delete"/>
    <attDef ident="cert" mode="delete"/>
    <attDef ident="source" mode="delete"/>
</attList>
</elementSpec>
<elementSpec ident="surrogates" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
    </attList>
</elementSpec>
<elementSpec ident="msPart" mode="change">
    <attList>
        <attDef ident="corresp" mode="delete"/>
        <attDef ident="sameAs" mode="delete"/>
        <attDef ident="ana" mode="delete"/>
        <attDef ident="fac" mode="delete"/>
        <attDef ident="cert" mode="delete"/>
        <attDef ident="resp" mode="delete"/>
        <attDef ident="source" mode="delete"/>
        <attDef ident="type" mode="delete"/>
    </attList>
</elementSpec>

```

## Établissement de notices modèles

Les trois notices ici encodées correspondent à celles numérisées et océrisées dans la section A.1.1 des annexes<sup>342</sup>.

---

<sup>342.</sup> Les notices encodées ont ici été abrégées par souci d'économie de papier, mais elles sont disponibles dans leur intégralité dans les livrables techniques.

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Les livres d'heures manuscrits de la Bibliothèque
          nationale / Abbé V.
          Leroquais</title>
        <author>Leroquais, Victor (1875-1946)</author>
        <respStmt>
          <resp>Encodage réalisé pour l'IRHT dans le cadre du
            projet HORAE, structurant sémantiquement le travail
            intellectuel de Victor Leroquais avec l'aide des
            recommandations de la TEI.</resp>
        <persName>
          <forename>Gwenaëlle</forename>
          <surname>Patat</surname>
        </persName>
      </respStmt>
    </titleStmt>
    <editionStmt>
      <edition>
        <date>2020-05-07</date>
      </edition>
    </editionStmt>
    <publicationStmt>
      <p>IRHT, mai 2020.</p>
    </publicationStmt>
    <sourceDesc>
      <biblFull>
        <titleStmt>
          <title>Les livres d'heures manuscrits
            de la Bibliothèque nationale / Abbé V.
            Leroquais</title>
          <author>
            <forename>Victor</forename>
            <surname>Leroquais</surname>
          </author>
        </titleStmt>
        <publicationStmt>
```

```
<publisher>
    <date>1927</date>
</publisher>
<pubPlace>Paris</pubPlace>
<availability>
    <licence>Domaine Public</licence>
</availability>
</publicationStmt>
</biblFull>
</sourceDesc>
</fileDesc>
</teiHeader>
<!-- Ajout d'un @n qui reprend le numéro de la notice.
Notice 1, p. 1-7 de l'OCR. --&gt;
&lt;TEI n="1"&gt;
&lt;teiHeader&gt;
&lt;fileDesc&gt;
    &lt;titleStmt&gt;
        &lt;title/&gt;
    &lt;/titleStmt&gt;
    &lt;publicationStmt&gt;
        &lt;p&gt;cf. supra&lt;/p&gt;
    &lt;/publicationStmt&gt;
    &lt;sourceDesc&gt;
        &lt;listWit&gt;
            &lt;witness&gt;
                &lt;msDesc&gt;
                    &lt;msIdentifier&gt;
                        &lt;repository&gt;Bibliothèque
                        nationale&lt;/repository&gt;
                        &lt;idno&gt;ms.lat. 757&lt;/idno&gt;
                    &lt;/msIdentifier&gt;
                    &lt;msContents&gt;
                        &lt;summary&gt;LIVRE D'HEURES ET
                        MISSEL FRANCISCAINS. Ce manuscrit
                        présente une combinaison assez rare
                        du livre d'Heures et du missel. Le premier
                        occupe les fol. 16 à 222 et
                        376 v° à 432; le second, les fol. 223 à</pre>
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

376 et 433 à 443.

Le calendrier(fol. 3 v° à 15) est commun  
à l'un et à l'autre.</summary>

```
<msItem>
    <note>Feuillet de garde.
    Ancienne cote : « 4299 c. »</note>
</msItem>
<msItem>
    <locus from="Ar" to="Cr">A à C</locus>
    <title>Tables du manuscrit</title>
</msItem>
<msItem>
    <locus from="Dv" to="Er">D à E</locus>
    <incipit>« Ratio pasce... »</incipit>
    <note>s'étend de 1380 à 1520.</note>
</msItem>
<msItem>
    <locus from="Fr" to="Ir">Fr à Ir</locus>
    <title>Conjonctions et oppositions
    lunaires allant de 1395 à 1400</title>
    <incipit><locus n="Fr">F</locus>Ratio
    lune pro anno D. n. I. C.
    MCCCLXXXXV.</incipit>
    <quote><locus n="Iv">I v°</locus>.
    Ratio lune pro anno Christi MCCCC
    in Ytalia.</quote>
    <note>Les fol. F à I paraissent avoir
    été ajoutés, autant du moins
    qu'on peut en juger par l'ancien
    foliotage et par le format des
    feuillets ; mais l'écriture est
    de la même époque.</note>
</msItem>
<msItem>
    <locus from="3v" to="15r">3 à 15</locus>
    <title>Calendrier franciscain</title>
    <note>Les mentions qui suivent sont en
    lettres rouges. - (25 mai). «
    Translatio s. Francisci conf. » - (13 juin)
```

```
    « Nat. s. Antonii conf.  
    ordinis Minorum. » - (20 juin). « ... Oct.  
    s. Antonii conf. » - (12  
    août) « S. Clare virg. de ordine dominarum. »  
    - (4 oct.) « Nat. s.  
    Francisci. »</note>  
</msItem>  
<msItem>  
    <locus from="18r" to="49r">18 à 49</locus>  
    <title>Heures abrégées pour chacun des jours  
    de la semaine.</title>  
<msItem>  
    <locus n="18r" to="20r">18 à 20</locus>  
    <title>Heures de la Trinité.</title>  
    <rubric><locus n="18r">f.18r</locus>Die  
    dominica. Ad matutinas Trinitatis.</rubric>  
</msItem>  
<msItem>  
    <locus from="22r" to="29r">22 à 29</locus>  
    <title>Heures des défunts.</title>  
    <rubric><locus n="22r">22r</locus>Die  
    lune. Ad vesperas pro  
    defunctis.</rubric>  
</msItem>  
<msItem>  
    <locus from="31r" to="33r">31 à 33</locus>  
    <title>Heures du Saint-Esprit.</title>  
    <rubric><locus n="31r">31r</locus> Die  
    Martis. Ad matutinas de  
    Spiritu sancto.</rubric>  
</msItem>  
<msItem>  
    <locus from="35r" to="37r">35 à 37</locus>  
    <title>Heures de tous les saints.</title>  
    <rubric><locus n="35r">35r</locus> Die  
    Mercurii. Ad matutinas  
    omnium sanctorum.</rubric>  
</msItem>  
<msItem>
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<locus from="39r" to="41r">39 à 41</locus>
<title>Heures du Saint-Sacrement.</title>
<rubric><locus>39r</locus> Die Iovis.
Ad matutinas de
    sacra¬mento.</rubric>
</msItem>
<msItem>
    <locus from="43r" to="45r">43 à 45</locus>
    <title>Heures de la Croix.</title>
    <rubric><locus n="43r">43r</locus>
        Die Veneris. Ad matutinas de
        sanctaCruce.</rubric>
    </msItem>
    <msItem>
        <locus from="47r" to="49r">47 à 49</locus>
        <title>Heures de la Vierge.</title>
        <rubric><locus n="47r">47r</locus> Die
            sabbati. Ad matutinas beate
            Marie virginis.</rubric>
        </msItem>
    </msItem>
    <msItem>
        <locus from="51r" to="102r">Fol. 51 à 102</locus>
        <title>Office de la Vierge.</title>
        <rubric><locus n="51r">51</locus>Incipit
            officium beate virginis Marie
            secundum consuetudinem sancte romane Ecclesie
            et secundum ordinem
            Fratrum Minorum. Ad matutinum...</rubric>
        <note>Les Matines suivies des antennes, psaumes,
            leçons et répons pour le temps de l'Avent,
            pour celui de Noël et pour celui de
            l'octave de Noël à l'Épiphanie.</note>
        <msItem>
            <locus n="100r">100</locus>
            <rubric>A Pascha usque ad Pentecostes
                cantatur oratio pulcherrima
                et devota ad beatam Mariam et ad beatum
                Iohannem
```

```
evangelistam.</rubric>
<incipit>O intemerata et in eternum benedicta...
et esto pia michi
peccatori in omnibus auxiliatrixis (stc).O
Iohannes
beatissime...</incipit>
<quote><locus n="100v">100 v°.</locus> ...
O due gemme
celestes...</quote>
<quote><locus n="101r">101.</locus> ...
vobis duobus ego miser
peccator, hodie et omni tempore,
corpus meum et animam meam
com-mendo...</quote>
<quote><locus n="101v">101 v°.</locus>
...gratiarum largitor, qui
cum Patre et Filio consubstantialis est.
Qui..</quote>
<note>Dans ce manuscrit et dans le ms. latin
1352 ci-après, cette prière termine l'office
de la Vierge, comme le Salve regina, le
Regina celi, l'Alma redemptoris ou l'Ave,
regina celorum auxquels elle fait suite.</note>
</msItem>
</msItem>
<msItem>
<locus from="103r" to="105r">103 à 105</locus>
<title>Septem gaudia beate Marie virginis.</title>
<msItem>
<rubric><locus n="103r">f.103</locus> Primum
gaudium virginis
Marie.</rubric>
<incipit>Sancta Maria, domina mea dulcissima,
rogo te per illud
gaudium quod habuisti quando tibi angelus
Gabriel
apparuit...</incipit>
<explicit>... liberare digneris.</explicit>
<note>Suivent les six autres joies : la Nativité,
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

la Présentation  
de l'enfant Jésus au temple, l'Épiphanie,  
le baptême du Sauveur  
et son premier miracle, la Résurrection,  
l'Ascension.</note>

</msItem>

<msItem>

<locus n="104v">104 v°</locus>

<title>Oratio</title>

<incipit>Supplicatio mea ascendat ad te,  
Deus: intret oratio mea in  
conspectu tuo, Christe. Perveniat  
deprecationis mea ad te.  
Domine...</incipit>

<explicit><locus n="105v">105 v°</locus>...  
et te sine confusione  
videre, cui est honor et gloria. Per...</explicit>

</msItem>

<msItem>

<title>Oratio</title>

<incipit>Benedicta sit hora illa qua Deus  
homo annunciatus  
est...</incipit>

<explicit>... et impleatur desiderium meum.  
Amen.</explicit>

</msItem>

</msItem>

<msItem>

<locus from="108v" to="103r">108 v° à 139</locus>

<title>Office des défunts.</title>

<rubric><locus n="118v">f.108 v°</locus>

Incipit officium in agenda  
mortuorum...</rubric>

</msItem>

<msItem>

<locus from="141r" to="147r">141 à 147</locus>

<title>Psaumes de la pénitence.</title>

</msItem>

```
<msItem>
    <locus from="149r" to="154r">149 à 154</locus>
    <title>Litanies</title>
    <note>aucun saint local</note>
    <explicit><locus n="152r">f.152</locus>...R/
        Oremus pro ministro
        nostro, V/ Dominus conservet eum et vivificet
        eum...</explicit>
</msItem>
<msItem>
    <locus from="156r" to="177r">156 à 177</locus>
    <title>Office de la Passion.</title>
</msItem>
<msItem>
    <locus from="177r" to="190r">177 à 190</locus>
    <title>Évangile de la bénédiction des Rameaux
    et Passion selon saint
    Matthieu.</title>
</msItem>
<msItem>
    <locus from="191r" to="199r">191 à 199</locus>
    <title>Passion selon saint Marc.</title>
</msItem>
<msItem>
    <locus from="200r" to="207r">200 à 207</locus>
    <title>Passion selon saint Luc.</title>
</msItem>
<msItem>
    <locus n="208r">208</locus>
    <title>Évangile du Jeudi saint.</title>
</msItem>
<msItem>
    <locus from="210r" to="217r">210 à 217</locus>
    <title>Passion selon saint Jean.</title>
</msItem>
<msItem>
    <locus n="218r">218</locus>
    <rubric><title>Oratio s. Augustini.</title>
    </rubric>
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<incipit>Dulcissime Yesu Christe, Domine, verus  
Deus, qui de sinu  
    Patris omnipotentis missus es in mondum  
        (sic) relaxare  
            peccata...</incipit>  
<explicit><locus n="220r">220r</locus>...et  
    gloriosus in secula  
        seculorum.</explicit>  
</msItem>  
<!-- Choix ici de mettre l'ensemble des informations  
dans une balise <note> car la structuration du  
paragraphe diffère des autres et  
la structuration d'informations n'est alors pas  
automatisable. -->  
<msItem>  
    <note>La partie qui suit et qui va du fol.  
223 à 376 est un missel des  
    principales fêtes. [...]  
    Pater, quia ego peccator peccavi nimis  
    contra legem Dei mei...  
    ></note>  
</msItem>  
<msItem>  
    <locus from="326v" to="389r">376 v° à 389</locus>  
    <title>Office de saint Jean-Baptiste.</title>  
</msItem>  
<msItem>  
    <locus from="389v" to="403r">389 v° à 403</locus>  
    <title>Office de saint Nicolas.</title>  
</msItem>  
<msItem>  
    <locus from="403v" to="417r">403 v° à 417</locus>  
    <title>Office de saint Antoine, ermite.</title>  
</msItem>  
<msItem>  
    <locus from="417v" to="427r">417 v° à 427</locus>  
    <title>Office de sainte Catherine.</title>  
</msItem>  
<msItem>
```

```
<locus n="427v">427 v°</locus>
<rubric>Quicumque dixerit infrascriptam
[...]
mortem suam.</rubric>
<incipit>Obsecro te, Maria, mater misericordie et
summe dignitatis,
per illam inextimabilem letitiam qua exultavit
spiritus tuus... ut
michi famulo tuo ill. impetres... </incipit>
<explicit><locus n="428v">428 v°</locus>... exaudi
me, O dulcissima
Maria, mater misericordie.</explicit>
</msItem>
<msItem>
<incipit><locus n="428v">428 v°</locus>Domine Deus
omnipotens, fac me
Ma. et N. fortes et stabiles contra omnes,
inimicos
nostros...</incipit>
<explicit><locus n="429r">429</locus> ...et animas
corporum
nostrorum.</explicit>
</msItem>
<msItem>
<rubric><locus n="429v">429 v°</locus>Infrascripta
[...]
preparatam.</rubric>
<incipit>Domine Iesu Christe, qui septem verba
in ultimo [die] vite
tue in cruce pendens dixisti... </incipit>
<explicit><locus n="430v">430v v°</locus> ...
iocundari et epulari et
commorari. Per ...</explicit>
</msItem>
<msItem>
<rubric><title>Oratio.</title></rubric>
<incipit>Concede michi, misericors Deus, que
tibi placita sunt
ardenter concupis-cere...</incipit>
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<explicit><locus n="432r">432</locus> ... in  
patria frui per gloriam.  
Per...</explicit>  
</msItem>  
<msItem>  
    <note>Les fol. 433 à 443 contiennent la  
    fin du missel.</note>  
    <msItem>  
        <rubric><locus n="433r">433</locus>  
        <title>Exorcismus salis et  
            aque.</title></rubric>  
    </msItem>  
    <msItem>  
        <rubric><locus n="435r">435</locus>  
        <title>Ordo ad catecumenum  
            faciendum.</title></rubric>  
    </msItem>  
    <msItem>  
        <rubric><locus n="440r">440</locus>  
        <title>Ordo ad incidendum  
            capil-lum infantium.</title></rubric>  
    </msItem>  
    <msItem>  
        <locus from="440v" to="443r">440 v° à  
        443</locus>  
        <title>Bénédictions diverses.</title>  
        <msItem>  
            <rubric><locus n="440v">440 v°</locus>  
            Benedictio fructuum  
            arborum.</rubric>  
        </msItem>  
        <msItem>  
            <rubric><locus n="441r">441</locus>  
            Benedictio agni  
            pascalis.</rubric>  
        </msItem>  
        <msItem>  
            <rubric>Benedictio panis in ecclesia  
            populo distribuendo
```

```
(sic).</rubric>
</msItem>
<msItem>
<rubric><locus n="441v">441 v°</locus>
Benedictio casei et melis
(sic) in Pasca.</rubric>
</msItem>
<msItem>
<rubric>Benedictio casei et ovorum.</rubric>
</msItem>
<msItem>
<rubric><locus n="442r">442</locus>
Benedictio nove
domus.</rubric>
</msItem>
<msItem>
<rubric>Benedictio incensi.</rubric>
</msItem>
<msItem>
<rubric><locus n="442v">442 v°
</locus>Benedictio pere et
baculi.</rubric>
</msItem>
<msItem>
<rubric><locus n="443v">443 v°
</locus>Benedictio
vestimentorum.</rubric>
</msItem>
</msItem>
</msItem>
</msContents>
<physDesc>
<objectDesc>
<supportDesc material="parch">
<support>
<material>Parch.</material>
</support>
<extent>
<measure type="composition">
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
unit="leaf" quantity="443">443 ff. à
2 col., plus les feuillets
préliminaires cotés A à
G.</measure>
<dimensions type="leaf" unit="mm">
    <height quantity="265">265</height>
    <width quantity="207">207</width>
</dimensions>
</extent>
</supportDesc>
<layoutDesc>
    <layout columns="2">2 col.</layout>
</layoutDesc>
</objectDesc>
<scriptDesc>
    <summary>L'écriture et la décoration sont
italiennes; les fautes
d'orthographe sont assez fréquentes
ainsi que les erreurs de
transcription.</summary>
    <scriptNote/>
</scriptDesc>
<decoDesc>
    <decoNote>
        <p>Le manuscrit s'ouvre par deux peintures
[...]
        à la décoration primitive.</p>
        <p>La décoration originale du manuscrit
[...]
        autres manuscrits.</p>
        <p>Les peintures qui suivent se rapportent
[...]
        Sauveur.</p>
        <p>La décoration du manuscrit se complète
[...]
        goût italien.</p>
        <p>Comme il a été dit un peu plus haut,
[...]
        encadrements.</p>
```

```

        </decoNote>
    </decoDesc>
    <bindingDesc>
        <binding>
            <p>Rel. maroquin rouge ; dos orné.</p>
        </binding>
    </bindingDesc>
</physDesc>
<history>
    <origin>
        <origDate>1380</origDate>
        <p>L'office de la Vierge et celui des morts
        sont ceux de Rome ; [...]
            personnage dont le nom commençait par Ma.</p>
        <p>Il existe plusieurs répliques de cet intéressant
        manuscrit. [...]
            fautes de transcription.</p>
        <origDate>La table pascale du fol. D v°
        donne la date du manuscrit :
            1380.</origDate>
        <p>Il ressort de l'analyse ci-dessus que
        [...]
            aucun élément de solution du problème.</p>
    </origin>
<!-- Division du paragraphe entre &lt;origin&gt;
et &lt;provenance&gt; difficile à automatiser,
à moins de matcher sur des mots-clés
("usage", "possesseurs" etc.). --&gt;
    &lt;provenance&gt;
        &lt;p&gt;Ainsi que l'a établi M. Max Prinet
        (cf. l'article cité dans la
            bibliographie ci-après), les deux
            blasons
            ont été ajoutés; ils nous
            renseignent donc, non pas sur le personnage
            pour qui le manuscrit a
            été exécuté, mais seulement sur les
            possesseurs du livre d'Heures
            dans la première moitié du XVIe siècle
</pre>

```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
: Claude et Anne (ou Annet)
Régine.</p>
</provenance>
</history>
<additional>
  <listBibl>
    <bibl>TOESCA (Pietro), La pittura e
      la miniatura nella Lombardia,
      1912, p. 279 à 294.</bibl>
    <bibl>COUDERC (Camille), Album de portraits,
      p. 19 et pl. XLV et
      XLVI.</bibl>
    <bibl>PRINET (Max), Bulletin de la Société
      nationale des antiquaires
      de France (séance du 26 mars 1924),
      p. 137 à 141.</bibl>
    <bibl>LEROQUAIS (abbé V.), Les
      sacramentaires et les missels
      manuscrits des biblio-thèques
      publiques de France, 1924, t. II, p.
      361-363 et pl. LXIX à LXXV.</bibl>
    <bibl> ANCONA (Paolo d'), La miniature
      italienne du Xe au XVIe siècle,
      1925, p. 21 et pl. XVI.</bibl>
    <bibl>[COUDERC (Camille)], Catalogue
      de l'exposition du moyen âge.
      Bibliothèque nationale, 1926, p. 35-36.</bibl>
  </listBibl>
</additional>
</msDesc>
</witness>
</listWit>
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
  <body>
    <p/>
  </body>
```

```
</text>
</TEI>
<!-- Notice 2, p. 232-233 de l'OCR. -->
<TEI n="112">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title/>
      </titleStmt>
      <publicationStmt>
        <p>cf. supra</p>
      </publicationStmt>
      <sourceDesc>
        <listWit>
          <witness>
            <msDesc>
              <msIdentifier>
                <repository>Bibliothèque nationale</repository>
                <idno>ms. lat., 1399.</idno>
              </msIdentifier>
              <msContents>
                <summary>HEURES A L'USAGE DE ROME.</summary>
                <msItem>
                  <locus n="1r">1</locus>
                  <note>Anciennes cotes. « Codex Colb.  
6323. Regius, 4475, 3.»</note>
                </msItem>
                <msItem>
                  <locus from="1r" to="12r">1 à 12</locus>
                  <title>Calendrier de Chalon-sur-Saône.</title>
                  <note>(22 janv.) En lettres rouges :  
« S. Vincent. » [...]  
Gengoul mart. »</note>
                </msItem>
                <msItem>
                  <locus from="13r" to="19r">13 à 19</locus>
                  <title>Extraits des quatre évangiles.</title>
                </msItem>
                <msItem>
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<locus n="19r">19.</locus>
<rubric>Dévote oraison à nostre Dame.</rubric>
<incipit><locus n="19v">19 v°.</locus>
Obsecro te, domina sancta
    Maria, mater Dei, pietate plenissima...</incipit>
<quote><locus n="22r">22.</locus> ...Et
    eciam in omnibus illis rebus
        in quibus ego sum facturus, locuturus
        aut cogitaturus... Et michi
        famulo tuo impetres a dilecto filio
        complementum...</quote>
<explicit>
    <locus n="24r">24.</locus>...mater
    Dei, pietatis et misericordie.
    Amen.</explicit>
</msItem>
<msItem>
    <rubric>De nostre Dame.</rubric>
    <incipit>Gaude, flore virginali Que
        honore speciali...</incipit>
    </msItem>
    <msItem>
        <locus from="30r" to="52r">30 à 52</locus>
        <title>Office de la Passion.</title>
    </msItem>
    <msItem>
        <locus from="54r" to="108r">54 à 108</locus>
        <title>Heures de la Vierge</title>
        <quote><locus n="101r">101.</locus>Chi
            apres sensicut l'offisse (sic)
            de Vadvent...</quote>
        <note>le début manque.</note>
    </msItem>
    <msItem>
        <note>Lacune entre 108 et 110.</note>
    </msItem>
    <msItem>
        <locus from="110r" to="115r">110 à 115</locus>
        <title>Heures de la Croix</title>
```

```
        <note>le commencement et la fin manquent.</note>
    </msItem>
    <msItem>
        <note>Lacune entre 115 et 116.</note>
    </msItem>
    <msItem>
        <locus from="116r" to="118r">116 à 118</locus>
        <title>Heures du Saint-Esprit</title>
        <note>le commencement et la fin ont disparu.</note>
    </msItem>
    <msItem>
        <note>Lacune entre 118 et 120.</note>
    </msItem>
    <msItem>
        <locus from="120r" to="130r">120 à 130</locus>
        <title>Psaumes de la pénitence</title>
        <note>le commencement manque.</note>
    </msItem>
    <msItem>
        <locus from="130v" to="137r">130 v° à 137</locus>
        <title>Litanyes.</title>
        <quote><locus n="132r">132</locus>... s.
            Vincenti; s. Dyonisi cum
            sociis tuis ; s. Mauri cum sociis tuis...</quote>
    </msItem>
    <msItem>
        <locus from="1737v" to="174r">137 v° à 174</locus>
        <title>Office des morts.</title>
    </msItem>
    <msItem>
        <locus n="174r">174</locus>
        <rubric>Chi ensieuvent les commandasses.</rubric>
        <incipit>Ant. Subvenite, sancti Dei...</incipit>
    </msItem>
    <msItem>
        <locus n="181v">181 v°</locus>
        <rubric><title>Oratio.</title></rubric>
        <incipit>Domine Iesu Christe,
            qui septem verba...</incipit>
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<explicit><locus n="183v">183 v°</locus>...
.iocundari et commorari per
infinita secula seculorum. Amen.</explicit>
</msItem>
<msItem>
<locus n="187r">187</locus>
<note>D'une autre main :</note>
<incipit>Glorieux saint Sébastien,
Qui souffris comme bon
chrétien...</incipit>
</msItem>
<msItem>
<locus n="188r">188</locus>
<note>D'une autre main :</note>
<incipit>Il fault morir, retenes bien,
Retenes bien, je vous en
prie...</incipit>
</msItem>
</msContents>
<physDesc>
<objectDesc>
<supportDesc material="parch">
<support>
<material>Parch.</material>
</support>
<extent>
<measure type="composition" unit="leaf"
quantity="188">188 ff. à
longues lignes ; plusieurs lacunes.</measure>
<dimensions type="leaf" unit="mm">
<height quantity="188">188</height>
<width quantity="128">128</width>
</dimensions>
</extent>
</supportDesc>
</objectDesc>
<decoDesc>
<!-- Les paragraphes dans &lt;decoNote&gt; ont été
déterminés grâce aux tirets longs. --&gt;</pre>
```

```
<decoNote>
    <p>Peintures des plus médiocres : fol. 30,
    la flagellation ; 137
        v°, service funèbre.</p>
    <p>Initiales de couleur dont le champ est
    occupé par des feuilles
        stylisées sur fond d'or.</p>
    <p>Initiales d'or sur fond azur et lilas
    relevé de blanc.</p>
    <p>Quelques bordures de feuillage et de fleurs.</p>
    <p>Petites initiales vermillon et azur
    alternativement.</p>
</decoNote>
</decoDesc>
<bindingDesc>
    <binding>
        <p>Rel. maroquin rouge aux armes de France
        et au chiffre royal.</p>
    </binding>
</bindingDesc>
</physDesc>
<history>
    <origin>
        <origDate>XVe SIÈCLE</origDate>
        <p>L'office de la Vierge et celui des morts
        représentent l'usage de
            Rome ; l'écriture est française ;
            le calendrier est celui de
            Chalon-sur-Saône : ce sont donc
            des Heures de Rome avec calendrier
            de Chalon-sur-Saône.</p>
    </origin>
</history>
</msDesc>
</witness>
</listWit>
</sourceDesc>
</fileDesc>
</teiHeader>
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<text>
  <body>
    <p/>
  </body>
</text>
</TEI>
<!-- Notice 3, p. 675-676 de l'OCR. --&gt;
&lt;TEI n="313"&gt;
  &lt;teiHeader&gt;
    &lt;fileDesc&gt;
      &lt;titleStmt&gt;
        &lt;title/&gt;
      &lt;/titleStmt&gt;
      &lt;publicationStmt&gt;
        &lt;p&gt;cf. supra&lt;/p&gt;
      &lt;/publicationStmt&gt;
      &lt;sourceDesc&gt;
        &lt;listWit&gt;
          &lt;witness&gt;
            &lt;msDesc&gt;
              &lt;msIdentifier&gt;
                &lt;repository&gt;Bibliothèque nationale&lt;/repository&gt;
                &lt;idno&gt;fr. nouv. acq., 10230&lt;/idno&gt;
              &lt;/msIdentifier&gt;
              &lt;msContents&gt;
                &lt;summary&gt;HEURES A L'USAGE DE ROUEN ET
                GRAND COUTUMIER DE
                NORMANDIE.&lt;/summary&gt;
              &lt;msItem&gt;
                &lt;locus n="1r"&gt;1&lt;/locus&gt;
                &lt;note&gt;Au bas du feuillett, note en écriture
                moderne : « Phillipps ms.
                22403. »&lt;/note&gt;
              &lt;/msItem&gt;
              &lt;msItem&gt;
                &lt;locus from="1r" to="6r"&gt;1 à 6&lt;/locus&gt;
                &lt;title&gt;Calendrier très clairsemé où
                dominent les saints
                caractéristiques d'Évreux&lt;/title&gt;
              &lt;/msItem&gt;
            &lt;/msContents&gt;
          &lt;/witness&gt;
        &lt;/listWit&gt;
      &lt;/sourceDesc&gt;
    &lt;/fileDesc&gt;
  &lt;/teiHeader&gt;
  &lt;body&gt;
    &lt;p/&gt;
  &lt;/body&gt;
&lt;/TEI&gt;</pre>
```

```
<note>(25 mai) « S. Mause. » - (5 août )  
[...]  
- (30 déc.) « S. Ursin. »</note>  
</msItem>  
<msItem>  
    <locus from="7r" to="27r">7 à 27</locus>  
    <title>Heures de la Vierge.</title>  
</msItem>  
<msItem>  
    <locus from="27v" to="29r">27 v° à 29</locus>  
    <title>Heures de la Croix.</title>  
</msItem>  
<msItem>  
    <locus from="29r" to="30r">29 à 30</locus>  
    <title>Heures du Saint-Esprit.</title>  
</msItem>  
<msItem>  
    <locus from="31r" to="37r">31 à 37</locus>  
    <title>Psaumes de la pénitence.</title>  
</msItem>  
<msItem>  
    <locus from="38r" to="41r">38 à 41</locus>  
    <title>Litaines.</title>  
    <quote><locus n="39r">39</locus>... s. Romane ;  
    s. Audoene ; s. Mellone  
    ; s. Vulgane ; s. Martine ; s. Macute ; s.  
    Hugo ; s. Viviane ; s.  
    Taurine ; s. Maxime ; s. Alexis ; s. Severe  
    ; s. Patrici ;  
    s. Philberte</quote>  
    <quote><locus n="39v">39 v°</locus>s. Hubcrte ;  
    s. Leodegari ; s.  
    Iuliane ; s. Amande ; omnes sancti confessores  
    ; s. Anna... s. Avia  
    ; s. Genovefa... s. Honorina ; s. Austreberta.  
    ..</quote>  
</msItem>  
<msItem>  
    <locus from="41v" to="42r">41 v° et 42</locus>
```

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
<title>Prologue de saint Jean.</title>
</msItem>
<!-- On peut ici subdiviser chaque prière en un
<msItem> à l'intérieur du <msItem>
sur les prières diverses, mais cela serait
difficile à automatiser,
car les tirets longs séparent souvent des
citations, et un paragraphe correspond en
général à un <msItem>.
Cela peut être une opération à compléter
à la main.-->
<msItem>
<locus from="43r" to="50r">43 à 50</locus>
<title>Prières diverses.</title>
<msItem>
<incipit><locus n="43r">43.</locus>Obsecro
te, domina sancta Maria,
mater Dei, pietate plenissima... Et in
omnibus illis rebus in
quibus ego sum facturus, locuturus aut
cogitaturus... et michi
famulo tuo impetra a dilecto filio tuo
complementum ... </incipit>
<explicit><locus n="45r">45.</locus> ...
mater Dei et misericordie.
Amen.</explicit>
</msItem>
<msItem>
<rubicid>Oracio ad Christum.</rubicid>
<incipit>O bone Iliesu, o dulcissime Ihesu,
o piissime Ihesu, o
Ihesu, fili Marie virginis...</incipit>
<explicit><locus n="46v">46 v°</locus>...
qui diligunt nomen tuum
quod est Ihesus. Amen.</explicit>
</msItem>
<msItem>
<incipit><locus n="47r">47.</locus>0
intemerata... De te enim
```

Filius Dei verus... </incipit>  
<quote><locus n="47v">47 v°</locus> ...  
.et esto michi miserrimo  
peccatori propicia in omnibus auxiliatrix  
...</quote>  
<explicit><locus n="48v">48 v°</locus> ...  
.cum omnibus sanctis et  
electis suis. Amen.</explicit>  
</msItem>  
<msItem>  
<r rubric>Oratio ad  
Dominum Ihesum Christum.</r rubric>  
<incipit>Domine Ihesu Christe, adoro te in cruce  
pendentem...</incipit>  
<explicit><locus n="49r">49</locus> ...ab  
angelo percuciente. Pater  
noster.</explicit>  
<note>Suivent cinq autres invocations.</note>  
</msItem>  
<msItem>  
<locus from="49v" to="50r">49 v° et 50.</locus>  
<title>Passion selon saint Jean.</title>  
<incipit><locus n="50r">50.</locus>Deus  
qui manus tuas et pedes  
tuos... - ...et veram scienciam usque  
in finem. Qui...</incipit>  
</msItem>  
</msItem>  
<msItem>  
<locus n="50v">50 v°</locus>  
<title>Antienne et oraison en l'honneur de  
saint Romain.</title>  
</msItem>  
<msItem>  
<locus from="55r" to="210r">Fol. 55 à 210</locus>  
<title>Grand coutumier de Normandie.</title>  
<incipit><locus n="55r">55</locus>Pour ce que  
nostre intencion est à  
déclarer en ceste euvre au mieulx que nous

## A.1. STRUCTURER DES NOTICES DE LIVRES D'HEURES

---

```
        pourrons les droiz et
        les esta- blissemens de Normendie...</incipit>
</msItem>
<msItem>
    <locus n="210r">210.</locus>
    <rubric>De marché de bourse. XXIII.</rubric>
    <explicit>Item que aucun cas de marché de
    bourse celui à qui l'en
        demande le marché...</explicit>
</msItem>
<msItem>
    <locus n="211v">211 v°</locus>
    <incipit><note>D'une autre main :</note>Torel
    n'est qu'un sot ; il m'a
        fait mal aulx bras. Je voudrais qu'il fût
        pendu aussy hault que je
        le pourrais regarder.</incipit>
</msItem>
</msContents>
<physDesc>
    <objectDesc>
        <supportDesc material="parch">
            <support>
                <material>Parch.</material>
            </support>
            <extent>
                <measure type="composition" unit="leaf"
                    quantity="212">212 ff. à
                    longues lignes.</measure>
                <!-- Les informations à mettre dans
                    <dimensions> seront à diviser dans un
                    deuxième temps
                    en <height> et <width> en ne gardant
                    que les chiffres. -->
                <dimensions type="leaf" unit="mm">
                    <height quantity="132">132</height>
                    <width quantity="96">96</width>
                </dimensions>
            </extent>
        </objectDesc>
    </physDesc>
```

```
</supportDesc>
</objectDesc>
<decoDesc>
  <decoNote>
    <p>Ni peintures ni miniatures.</p>
    <p>Quelques initiales filigranées.</p>
    <p>Petites initiales vermillon et azur
    alternativement.</p></decoNote>
  </decoDesc>
  <bindingDesc>
    <binding>
      <p>Rel. basane racine.</p>
    </binding>
  </bindingDesc>
</physDesc>
<history>
  <origin>
    <origDate>XVe SIÈCLE</origDate>
    <p>L'office de la Vierge représente
    l'usage de Rouen.</p>
    <origDate>Le manuscrit date de la seconde moitié
    ou de la fin du xve
    siècle.</origDate>
  </origin>
  </history>
</msDesc>
</witness>
</listWit>
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
  <body>
    <p/>
  </body>
</text>
</TEI>
</teiCorpus>
```

## A.2 Documents utiles à la transformation

### A.2.1 Documents d'entrée

Ajout de styles dans le document Word contenant les notices océrisées : l'exemple de la première notice

**LES LIVRES D'HEURES**

**1. LIVRE D'HEURES ET MISSEL FRANCISCAINS. 1380**

Bibliothèque nationale, ms. lat., 757

Ce manuscrit présente une combinaison assez rare du livre d'Heures et du missel. Le premier occupe les fol. 16 à 222 et 376 v° à 432; le second, les fol. 223 à 376 et 433 à 443. Le calendrier (fol. 3 v° à 15) est commun à l'un et à l'autre.

Feuillet de garde. Ancienne cote : « 4299 c. » — Fol. A à C. Tables du manuscrit.

— D v° à E. « Ratio pasce... »; s'étend de 1380 à 1520. — F à I. Conjonctions et oppositions lunaires allant de 1395 à 1400. — F. « Ratio lune pro anno D. n. I. C. MCCCLXXXV. » — I v°. « Ratio lune pro anno Christi MCCCC in Ytalia. » — Les fol. F à I paraissent avoir été ajoutés, autant du moins qu'on peut en juger par l'ancien foliotage et par le format des feuilles; mais l'écriture est de la même époque.

Fol. 3 v° à 15. Calendrier franciscain. — Les mentions qui suivent sont en lettres rouges. — (25 mai). « Translatio s. Francisci conf. » — (13 juin) « Nat. s. Antonii conf. ordinis Minorum. » — (20 juin). « ... Oct. s. Antonii conf. » — (12 août) « S. Clare virg. de ordine dominarum. » — (4 oct.) « Nat. s. Francisci. »

— (7 oct.) « Festum beate Iustine virg. » — (7 nov.) « Propdocimi (*sic*) ep. et conf. »

— (7 déc.) « S. Ambrosii archiep. » — D'une autre main (18 sept.) « Festum de stigmatibus (*sic*) beatissimi Francissi (*sic*) sacrис. »

Fol. 18 à 49. Heures abrégées pour chacun des jours de la semaine. — 18 à 20. Heures de la Trinité. — 18. « Die dominica. *Ad matutinas Trinitatis.* » — 22 à 29. Heures des défunts. — 22. « Die lune. *Ad vesperas pro defunctis.* » — 31 à 33. Heures du Saint-Esprit. — 31. « Die Martis. *Ad matutinas de Spiritu sancto.* » — 35 à 37. Heures de tous les saints. — 35. « Die Mercurii. *Ad matutinas omnium sanctorum.* »

— 39 à 41. Heures du Saint-Sacrement. — 39. « Die Iovis. *Ad matutinas de sacramento.* » - 43 à 45. Heures de la Croix. — 43. « Die Veneris. *Ad matutinas de sancta*

FIGURE A.23 – Notice 1 avec styles Word, première page

*Cruce.* » — 47 à 49. Heures de la Vierge. — 47. « Die sabbati. *Ad matutinas beate Marie virginis.*

»

Fol. 51 à 102. Office de la Vierge. — 51. « *Incipit officium beate virginis Marie secundum consuetudinem sancte romane Ecclesie et secundum ordinem Fratrum Minorum. Ad matutinum...* » Les Matines suivies des antiennes, psaumes, leçons et répons pour le temps de l'Avent, pour celui de Noël et pour celui de l'octave de Noël à l'Épiphanie. — 100. « *A Pascha usque ad Pentecostes cantatur oratio pulcherrima et devota ad beatam Mariam et ad beatum Iohannem evangelistam. O intemerata et in eternum benedicta... et esto pia michi peccatori in omnibus auxiliatrix (stc). O Iohannes beatissime... — 100 v°. ...O due gemme celestes... — 101. ...vobis duobus ego miser peccator, hodie et omni tempore, corpus meum et animam meam commendo... — 101 v°. ...gratiarum largitor, qui cum Patre et Filio consubstantialis est. Qui... » Dans ce manuscrit et dans le ms. latin 1352 ci-après, cette prière termine l'office de la Vierge, comme le *Salve regina*, le *Regina celi*, l'*Alma redemptoris* ou l'*Ave, regina celorum* auxquels elle fait suite.*

Fol. 103 à 105. [Septem gaudia beate Marie virginis.] — 103. « *Primum gaudium virginis Marie. Sancta Maria, domina mea dulcissima, rogo te per illud gaudium quod habuisti quando tibi angelus Gabriel apparuit... — ... liberare digneris.* » Suivent les six autres joies : la Nativité, la Présentation de l'enfant Jésus au temple, l'Épiphanie, le baptême du Sauveur et son premier miracle, la Résurrection, l'Ascension. — 104 v°. « *[Oratio]. Supplicatio mea ascendat ad te, Deus: intret oratio mea in conspectu tuo, Christe. Perveniat deprecatione mea ad te. Domine... — 105 v°... et te sine confusione videre, cui est honor et gloria. Per...» — « *Oratio. Benedic sit hora illa qua Deus homo annunciatus est... — ... et impleatur desiderium meum. Amen.* »*

Fol. 108 v° à 139. Office des défuns. — 108 v°. « *Incipit officium in agenda mortuorum... » — 141 à 147. Psalms de la pénitence — 149 à 154. Litanies ; aucun saint local. — 152. a RJ. Oremus pro ministro nostro, £. Dominus conservet eum et vivificet eum... » — 156 à 177. Office de la Passion. — 177 à 190. Évangile de la bénédiction des Rameaux et Passion selon saint Matthieu. — 191 à 199. Passion selon saint Marc. — 200 à 207. Passion selon saint Luc. — 208. Évangile du Jeudi saint — 210 à 217. Passion selon saint Jean. — 218. « *Oratio s. Augustini. Dulcissime Yesu Christe, Domine, verus Deus, qui de sinu Patris omnipotentis missus es in mundum (sic) relaxare peccata... — 220. ...et gloriosus in secula seculorum.* »*

La partie qui suit et qui va du fol. 223 à 376 est un missel des principales fêtes. Celui-ci comprend les messes votives de la semaine (fol. 223 à 254) dont l'ordre est le même que celui des Heures abrégées (fol. 18 à 49), l'*ordo missae* (fol. 256 à 276) et les principales messes du temporal et du sanctoral (fol. 277 à 359) suivies du commun des saints (fol. 360 v° à 376). Chacune des messes votives est suivie du prologue de l'évangile de saint Jean. Le *Confiteor* de la première messe votive (fol.

FIGURE A.24 – Notice 1 avec styles Word, deuxième page

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

223 v°) est nettement franciscain ; « *Confiteor Deo omnipotenti et beate Marie virgini et beato Francisco et omnibus sanctis Dei, et te (sic), Pater, quia ego peccator peccavi nimis contra legem Dei mei... »*

Fol. 376 v° à 389. Office de saint Jean-Baptiste. — 389 v° à 403. Office de saint Nicolas. — 403 v° à 417. Office de saint Antoine, ermite. — 417 v° à 427. Office de sainte Catherine. — 427 v°. « *Quicumque dixerit infrascriptam orationem vel super se portaverit, inimicus ei nocere non poterit. Beatus Augustinus hanc orationem scripsit in illa die qua obiit. Et si quis eam qualibet die bono et puro corde dixerit, in illa die non peribit, nec — 428 — in aqua nec in igne, nec veneno mortifero morietur. Et antequam transmigret de hoc seculo videbit beatam et gloriosam virginem Mariam. Et si quod ab ea iuste petierit, impetrabit, et per tot annos per quot dixerit, per tot dies presciet mortem suam. Obsecro te, Maria, mater misericordie et summe dignitatis, per illam inextimabilem letitiam qua exultavit spiritus tuus... ut michi famulo tuo ill. impetres... — 428 v°... exaudi me, 0 dulcissima Maria, mater misericordie.*

— « Domine Deus omnipotens, fac me *Ma.* et 2V. fortes et stabiles contra omnes, inimicos nostros... — 429 ...et animas corporum nostrorum. » — 429 v°. « *Infrascripta oratio est oratio venerabilis doctoris Bede. Quicumque omni die flexis genibus dixerit, nec diabolus nec malus homo ei nocere poterit, nec sine confessione morietur, et per triginta dies ante mortem suam videbit gloriosam virginem Mariam sibi in auxilio preparatam.* Domine Iesu Christe, qui septem verba in ultimo [die] vite tue in cruce pendens dixisti... — 430 v°. ...iocundari et epulari et commorari. Per ... » — « *Oratio.* Concede michi, misericors Deus, que tibi placita sunt ardenter concupiscere... — 432. ...in patria frui per gloriam. Per... »

Les fol. 433 à 443 contiennent la fin du missel. — 433. « *Exorcismus salis et aque.* » — 435- « *Ordo ad catecumenum faciendum.* » — 440. « *Ordo ad incidendum capillum infantium.* » — 440 v° à 443. Bénédictions diverses. — 440 v°. « *Benedictio fructuum arborum.* » — 441. « *Benedictio agni pascalis.* » — « *Benedictio panis in ecclesia populo distribuendo (sic).* » — 441 v°. « *Benedictio casei et melis (sic) in Pasca.* » — « *Benedictio casei et ovorum.* » — 442. « *Benedictio nove domus.* » — « *Benedictio incensi.* » — 442 v°. « *Benedictio pere et baculi.* » — 443 v°. « *Benedictio vestimentorum.* »

L'office de la Vierge et celui des morts sont ceux de Rome ; le calendrier est franciscain ainsi que les litanies et le *Confiteor*. La présence de saint Prodocime et de sainte Justine dans le calendrier semble désigner Padoue comme lieu d'origine du manuscrit ; toutefois, il convient de noter que les deux saints ne figurent ni dans les litanies ni dans le sanctoral. Je ne saurais dire s'il y a lieu d'attacher une signification spéciale à l'invocation : « *Pro ministro* » dans les litanies (fol. 152). Les différentes formules de prières ont été rédigées au masculin ; celle du fol. 428 v° semblerait indiquer que le volume a été transcrit pour un personnage dont le nom commençait par *Ma*. La table pascale du fol. D v° donne la date du manuscrit :

FIGURE A.25 – Notice 1 avec styles Word, troisième page

1380. L'écriture et la décoration sont italiennes; les fautes d'orthographe sont assez fréquentes ainsi que les erreurs de transcription.

Il existe plusieurs répliques de cet intéressant manuscrit. M. Toesca (*La pittura e la miniatura nella Lombardia*, 1912, p. 279 à 283) en a signalé une à la bibliothèque royale de Munich (*Cod. lat.*, 23215) : c'est un livre d'Heures exécuté par Giovanni di Benedetto, de Cumes, pour Blanche de Savoie, mère de Jean-Gaiéas Visconti. J'en ai rencontré une autre dans la collection Smith-Lesouëf, à Nogent-sur-Marne. Le manuscrit 22 de cette bibliothèque se présente en effet dans des conditions à peu près semblables. Il s'ouvre par une table pascale (fol. 1 et 2) qui va de 1380 à 1490. Viennent ensuite un calendrier de Bruges (fol. 3 à 14), la table du manuscrit, les Heures de la Vierge (fol. 15 v° à 83) et les Sept joies de Marie (84 v° à 88). Le missel occupe les fol. 120 à 212 ; il comprend les prières préliminaires de la messe, les messes votives pour les jours de la semaine, *Yor do missae* et quelques messes du temporal et du sanctoral. Viennent en dernier lieu l'office des morts (223 v° à 264), les psaumes pénitentiels (265 à 275), les litanies (275 à 284), la Passion selon les quatre évangélistes (286 à 346), et enfin (349 à 353) quelques bénédictions extraites du missel. Le manuscrit 1352 de la Bibliothèque nationale qui sera décrit plus loin peut également être regardé comme une réplique du ms. lat. 757, dont il renferme plusieurs prières importantes et dont il reproduit la plupart des fautes de transcription.

Parch., 443 ff. à 2 col., plus les feuillets préliminaires cotés A à G. — 265 sur 207 mill. —

Le manuscrit s'ouvre par deux peintures à pleine page ajoutées au XVI<sup>e</sup> siècle : fol. D, écu armorié : écartelé aux 1<sup>e</sup> et 4<sup>e</sup> d'azur à la grenade couronnée d'or, aux 2<sup>e</sup> et 3<sup>e</sup> d'or à la fasce de sable accompagnée de trois de trèfles de sinople, 2 éli ; Vécu est timbré d'un heaume grillé de face aux lambrequins d'or et d'azur ; le heaume est cimé d'un caducée au pied duquel se déroule la devise : UTRIQUE VITE ; au-dessous de l'écu, autre devise : IN MANIBUS TUIS SORTES MEE ; plus bas sur un cartouche : IULIAN REGIN. — Fol. 2 v°, autre peinture à pleine page également ajoutée au XVI<sup>e</sup> siècle : memes armes que fol. D, mais ici elles sont accolées d'un bâton de chantre et surmontées d'un chapeau ecclésiastique noir ; au-dessous, la devise : VERBUM DOMINI MANET IN ETERNUM ; plus bas, un cartouche : ANNA REGIN. Il s'agit en effet d'Anne (ou Annet) Regin, protonotaire apostolique, grand-chantre de la cathédrale de Clermont (1528-1529) qui mourut en 1556. Julian, dont le nom figure sur le premier blason, était un de ses frères. — On retrouve ces armes complètes au fol. 102 v° ; on rencontre également et à plusieurs reprises les 1<sup>e</sup> et 4<sup>me</sup> quartiers isolés, soit seuls, soit parfois accompagnés des lettres B. E (fol. 51, 108 v° et 403 V°). — Au fol. 51, la targe portant les armoiries à la grenade est timbrée d'un heaume de profil à lambrequins, couronné et cimé d'une tête et col de cygne. Le timbre avec son cimier appartient à la décoration primitive.

La décoration originale du manuscrit se compose de soixante-treize peintures à pleine page. Elles sont de valeur inégale, mais beaucoup d'entre elles attestent un talent véritable. La plupart sont sur fond d'or semé de dessins géométriques au pointillé ; ça et là, quelques fonds quadrillés ou losangés or et azur. Dans quelques peintures, les fonds quadrillés ou losanges sont décorés d'un emblème formé de deux anneaux entrelacés (fol. 140 et 340) ; on rencontre également plusieurs fonds unicolores avec des semis de lettres stylisées, SB (ou FB), lettres qui

FIGURE A.26 – Notice 1 avec styles Word, quatrième page

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

se terminent toujours par des feuilles (fol. 302, 308 v°, 315) ; parfois même, on trouve anneaux et lettres stylisées réunis sur un même fond (fol. 333 v°, 340 v° et 345). — Fol. 17, la création du ciel et de la terre ; 21, la séparation des éléments ; 30, la création des plantes ; 34, la création des astres ; 38, la création des poissons et des oiseaux ; 42, la création des animaux ; la création de l'homme et de la femme ; 46, Adam et Ève dans le paradis terrestre. — Les compositions précédentes servent d'illustration aux Heures abrégées pour les sept jours de la semaine. Celles qui suivent (fol. 50 v° à 84) figurent en tête des Heures de la Vierge : fol. 50 v°, la trahison de Judas et l'arrestation du Sauveur (Matines) ; les Laudes ne comportent pas de peinture ; 65 v°, le Christ devant Pilate (Prime) ; 69, Jésus portant sa croix (Tierce) ; 72, Jésus attaché à la croix (Sexte) ; 75, Crucifixion (None), (pl. VIII) ; 78, descente de croix (Vêpres) ; 84, mise au tombeau (Complies). La série de ces peintures des Heures de la Vierge diffère de celle que l'on trouve dans les autres manuscrits.

Les peintures qui suivent se rapportent aux prières en l'honneur de la Vierge (fol. 102 v°), à l'office des défunts (107 v°), aux psaumes pénitentiels (140 à 148), aux heures et aux évangiles de la Passion (155 à 218), aux messes du temporal et du sanctoral (222 à 373) et enfin aux bénédicitions qui terminent le missel (432 v°). — Fol. 102 v°, la Vierge et l'enfant Jésus ; devant eux, personnage agenouillé ; 107 v°, inhumation d'un prélat dans une église ; 140, transport de l'arche d'alliance ; 148, procession de pénitence ; 155, le Christ assis sur un arc-en-ciel et montrant ses plaies ; à ses pieds, la Vierge (?) et s. Jean-Baptiste ; dans le haut du tableau, anges portant les attributs de la Passion ; 178 v°, s. Matthieu ; 190 v°, s. Marc ; 199 v°, s. Luc ; 209 v°, s. Jean ; 217 v°, bénédiction de l'eau baptismale ; 222 v°, la Trinité (pl. IX) ; 230, Christ de pitié entre la Vierge et s. Jean ; attributs de la Passion (pl. X) ; 234 v°, la colombe de l'Esprit-Saint au milieu d'un globe d'or entouré de rayons ; 238 v°, le couronnement de la Vierge (pl. XI) ; 243 V0, la Cène ; 247 v°, la flagellation ; 251, Vierge de miséricorde ; 255 v°, l'élévation de l'hostie (pl. XII) ; 269 v°, crucifixion ; 276 v°, la Nativité (le bain de l'enfant Jésus) ; 279 v°, le martyre de s. Étienne ; 282 v°, s. Jean l'évangéliste (le miracle du calice empoisonné) ; 284 v°, la Circoncision ; 286 v°, l'Épiphanie ; 289 v°, la tentation de s. Antoine ; 291, le martyre de sainte Agnès ; 292 v°, la Purification ; 295 v°, la salutation angélique ; 298 v°, la tentation du Christ ; 302, la Transfiguration ; 305, la guérison du démoniaque muet ; 308 v°, la multiplication des pains ; 311 v°, la résurrection de Lazare ; 315, l'entrée à Jérusalem le jour des Rameaux ; 318, la Résurrection ; 320 v°, s. Georges vainqueur du dragon ; à droite, une reine dont la robe est semée des initiales MR ; 322, l'élévation de l'hostie ; 324 v°, l'Ascension ; 327 v°, la Pentecôte ; 330 v°, la naissance de s. Jean-Baptiste ; 333 v°, s. Pierre et s. Paul provoquant la chute de Simon le magicien devant Néron ; 336 v°, sainte Marie-Madeleine recevant la communion des mains d'un ange ; 340, le martyre de s. Laurent ; 342, l'Assomption ; 344 v°, la naissance de la Vierge ; 348, l'Invention de la sainte croix (pl. XIII) ; 350 v°, s. Michel ; 353 v°, s. François d'Assise (les stigmates) ; 355 v°, le martyre de sainte Catherine ; 356 v°, s. Nicolas ; 357, femme soutenant d'une main deux anneaux entrelacés et de l'autre les lettres stylisées S. B. (ou F. B.) ; 357 v°, l'incrédulité de s. Thomas ; 360, la pêche miraculeuse ; 362 v° et 365, scènes de martyre ; 367 v°, un confesseur ; 370, s. Jérôme ; 373, groupe de vierges ; 432 v°, le baptême du Sauveur.

La décoration du manuscrit se complète par un certain nombre d'initiales historiées sur fond d'or, renfermant des sujets variés et pittoresques, traités avec beaucoup de verve et de finesse ; plusieurs de ces sujets ornent les encadrements qui accompagnent les initiales : fol. 18 v°, enfant brandissant une pique ; 23 v°, dragon ailé ; 31 v°, jeune homme brisant un bâton sur son genou ; à sa droite, un jeune homme ; à sa gauche, une jeune femme ; 33, monstre ; 39, enfant grimpant à un arbre pour dénicher un nid d'oiseau ; 39 v°, enfant tuant un serpent à

FIGURE A.27 – Notice 1 avec styles Word, cinquième page

l'aide d'un poignard ; 40 v°, grotesque ; 41, enfants à cheval l'un sur un lion, l'autre sur un aigle ; 44 v° et 48 v°, oiseaux ; 48, homme sauvage ; jeune fille ; 49, chimère ; 73, oiseau ; 166, buste de femme ; 167 v° et 177, monstres ; 191, combat entre deux enfants nus ; 210, dragon ailé ; 230 v°, la mort ; 277, enfant combattant un monstre ; 293, Malachie ; 296, Isaïe ; 312, dragon ailé ; 305 v°, 309, 315 v°, 318 v°, 321, 322 v°, 325, 331, 340 v°, 342 v°, 345, prophètes, apôtres, martyrs et saints anonymes ; 348 v°, croix fichée ; 351, s. Michel ; 356, sainte Catherine ; 357, s. Nicolas ; 363 et 365 v°, martyrs ; 368 et 370 v°, confesseurs ; 373 v°, vierge ; 376 v°, s. Jean-Baptiste enfant ; 379 v°, 382, 383, 384, 385, 386, personnages divers ; 388, s. Jean-Baptiste ; 389 v° et 396, s. Nicolas ; 397, femme ; singe ; 398 v°, s. Nicolas ; 399 v°, femme ; chien ; 402, femme ; 403 v°, s. Antoine, ermite ; 417 v°, sainte Catherine ; 423, femme portant deux anneaux entrelacés. — Initiales fleuries sur fond d'or, fol. 19 et 36 v°. — Belles initiales de couleur dont le champ est occupé par des feuilles stylisées sur fond d'or. — Presque toutes ces initiales se prolongent dans les marges en lourds rinceaux de couleur et de feuillage semés de petites rosaces d'or dans le goût italien.

Il ressort de l'analyse ci-dessus que le manuscrit est un livre d'Heures à l'usage de Rome, ou, plus exactement, à l'usage franciscain, copié vers 1380 dans l'Italie du nord et probablement dans la région de Padoue. — Par qui les peintures ont-elles été exécutées ? En l'absence de signature et de renseignements puisés à des documents d'archives, il est difficile de mettre un nom en avant. Tout ce que l'on peut dire, c'est que les peintures appartiennent à l'école italienne. — Pour qui le manuscrit a-t-il été exécuté ? Question presque aussi difficile à résoudre que la première. Pour être en mesure de le faire, il faudrait connaître le nom du personnage qui se tient agenouillé aux pieds de la Vierge (fol. 102 v°) et qui porte un soleil sur son vêtement, emblème que l'on retrouve quoiqu'un peu différent sur le pourpoint de l'officier qui accompagne le centurion (fol. 72 et 75). Il faudrait également savoir qu'elle est cette femme symbolique qui figure au fol. 357 ; il faudrait enfin pénétrer la signification de l'emblème et des lettres quelle tient dans les mains et qui sont répandues à foison dans la dernière partie du manuscrit. En l'absence de conclusions positives, une chose du moins paraît certaine ; c'est que les armoiries des fol. D et 2 v° n'apportent aucun élément de solution du problème. Ainsi que l'a établi M. Max Prinet (cf. l'article cité dans la bibliographie ci-après), les deux blasons ont été ajoutés : ils nous renseignent donc, non pas sur le personnage pour qui le manuscrit a été exécuté, mais seulement sur les possesseurs du livre d'Heures dans la première moitié du xvi<sup>e</sup> siècle : Claude et Anne (ou Annet) Regin.

Comme il a été dit un peu plus haut, le manuscrit 22 de la collection Smith-Lesouëf, à Nogent-sur-Marne, est une réplique du ms. latin 757. Mais il importe de distinguer soigneusement, dans ce livre d'Heures, la décoration originale de celle qui a été ajoutée. Cette dernière en effet appartient à une autre main, peut-être même à plusieurs mains. Seule la première dérive du ms. lat. 757 et mérite d'être décrite ici. Elle se compose de dix peintures à pleine page, les unes sur fond losangé ou quadrillé, les autres sur fond unicolore (or ou couleurs), toutes étroitement apparentées à celles du manuscrit de la Bibliothèque nationale, quoique relevant d'un art notablement inférieur. Les huit premières servent d'illustration aux Heures de la Vierge : fol. 15, la Vierge et l'enfant Jésus ; à leurs pieds se tient un personnage agenouillé ; derrière lui, sainte Catherine (?), s. Antoine ermite et s. Christophe (Matines) ; 24, la flagellation (Laudes) ; 34 v°, Jésus devant Pilate (Prime) ; 38 v°, Jésus portant sa croix (Tierce) ; 41 v°, Jésus étendu sur la croix (Sexte) ; 45, crucifixion (None) ; 48 v°, descente de croix (Vêpres) ; 57, mise au tombeau (Complies). La peinture du fol. 84, la salutation angélique, vient en tête des Sept joies de la Vierge ; celle du fol. 296, la trahison de Judas et l'arrestation de Jésus, se trouve au milieu de la Passion selon saint Marc. — A cette décoration primitive appartient également un cer-

FIGURE A.28 – Notice 1 avec styles Word, sixième page

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

tain nombre d'initiales historiées les unes sur fond losangé ou quadrille, les autres sur fond d'or ou de couleurs; celle du fol. 15 v<sup>e</sup> renferme la Vierge et l'enfant Jésus; au bas de l'encadrement figurent l'emblème formé de deux anneaux entrelacés et les lettres S. B. (ou F. B.) qui foisonnent dans le ms. lat. 757 ; au milieu, on aperçoit une targe : *d'azur au lion d'or (?) orné et lampassé de gueules*. Les anneaux et les lettres se retrouvent au bas de plusieurs encadrements.

Rel. maroquin rouge ; dos orné. — Toesca (Pietro), La pittura e la miniatura nella Lombardia, 1912, p. 279 à 294. — Couderc (Camille), Album de portraits, p. 19 et pl. XLV et XLVI. — PRINET (Max), Bulletin de la Société nationale des antiquaires de France (séance du 26 mars 1924), p. 137 à 141. — Leroquais (abbé V.), Les sacramentaires et les missels manuscrits des bibliothèques publiques de France, 1924, t. II, p. 361-363 et pl. LXIX à LXXV. — Ancona (Paolo d'), La miniature italienne du Xe au XVI<sup>e</sup> siècle, 1925, p. 21 et pl. XVI. — [Couderc (Camille)], Catalogue de l'exposition du moyen âge. Bibliothèque nationale, 1926, p. 35-36.

FIGURE A.29 – Notice 1 avec styles Word, septième page

### Correction de la structuration du document océrisé au format XML

Dans un premier, nous avons repéré les paragraphes problématiques :

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:xs="http://www.w3.org/2001/XMLSchema" exclude-result-prefixes="xs"
  xmlns:tei="http://www.tei-c.org/ns/1.0" version="3.0"
  xpath-default-namespace="http://www.tei-c.org/ns/1.0">
  <xsl:strip-space elements="*"/>
  <xsl:output indent="yes" method="xml" encoding="UTF-8"/>

  <!-- Copy source file : Ce bout de code permet de ne pas perdre
       d'informations lors de la transformation en copiant tous les
       nœuds de document source.--&gt;
  &lt;xsl:template match="@* | node()" mode="#all"&gt;
    &lt;xsl:choose&gt;
      &lt;xsl:when test="matches(name(.),
        '^part|instant|anchored|default|full|status)$')"/&gt;
      &lt;xsl:otherwise&gt;
        &lt;xsl:copy&gt;
          &lt;xsl:apply-templates select="@*
            | node()" mode="#current"/&gt;
        &lt;/xsl:copy&gt;
      &lt;/xsl:otherwise&gt;
    &lt;/xsl:choose&gt;
  &lt;/xsl:template&gt;

  <!-- Vision des paragraphes mal-découpés relatifs aux msItem
       dans le document océrisé : ceux ne commençant pas par un F
       majuscule, un chiffre [pour ce dernier cas, ceux débutant</pre>
```

*par un chiffre ont aussi été vérifiés], ceux débutant par un guillemet ouvrant, un tiret ou trois points de suspension. -->*

```

<xsl:template match="p[@rend = 'Texte du corps (2)']">
    <xsl:choose>
        <xsl:when test="not(starts-with(., 'F'))">
            <p>
                <xsl:value-of select="normalize-space(.)"/>
            </p>
        </xsl:when>
        <xsl:when test="not(matches(., '^(\d)+'))">
            <p>
                <xsl:value-of select="normalize-space(.)"/>
            </p>
        </xsl:when>
        <xsl:when test="matches(., '^(\d)+')">
            <p>
                <xsl:value-of select="normalize-space(.)"/>
            </p>
        </xsl:when>
        <xsl:when test="starts-with(., '<')">
            <p>
                <xsl:value-of select="normalize-space(.)"/>
            </p>
        </xsl:when>
        <xsl:when test="starts-with(., '-')">
            <p>
                <xsl:value-of select="normalize-space(.)"/>
            </p>
        </xsl:when>
        <xsl:when test="starts-with(., '...')">
            <p>
                <xsl:value-of select="normalize-space(.)"/>
            </p>
        </xsl:when>
        <xsl:otherwise>
            <xsl:copy-of select="."/>
        </xsl:otherwise>
    </xsl:choose>
</xsl:template>

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
</xsl:stylesheet>
```

Puis nous avons automatisé, dans la mesure du possible, la fusion de certains paragraphes problématiques :

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
    xmlns:xs="http://www.w3.org/2001/XMLSchema" exclude-result-prefixes="xs"
    xmlns:tei="http://www.tei-c.org/ns/1.0" version="3.0"
    xpath-default-namespace="http://www.tei-c.org/ns/1.0">
    <xsl:strip-space elements="*"/>
    <xsl:output indent="yes" method="xml" encoding="UTF-8"/>

    <!-- Copy source file : Ce bout de code permet de ne pas perdre
        d'informations lors de la transformation en copiant tous les
        nœuds de document source. -->
    <xsl:template match="@* | node()" mode="#all">
        <xsl:choose>
            <xsl:when test="matches(name(.),
                '^part|instant|anchored|default|full|status)$')"/>
            <xsl:otherwise>
                <xsl:copy>
                    <xsl:apply-templates select="@*
                        | node()" mode="#current"/>
                </xsl:copy>
            </xsl:otherwise>
        </xsl:choose>
    </xsl:template>

    <!-- Fusion des paragraphes mal découpés débutant par
        un tiret, l'abréviation s. pour saint, un guillemet
        ouvrant ou trois points de suspension. -->
    <xsl:template match="p[@rend = 'Texte du corps (2)']">
        <xsl:choose>
            <xsl:when
                test="following-sibling::p[@rend =
                    'Texte du corps (2)'][1][starts-with(., '-')
                    or starts-with(., '-')]">
                <xsl:copy>
                    <xsl:attribute name="rend">
```

```

        <xsl:value-of select="@rend"/>
    </xsl:attribute>
    <xsl:apply-templates/>
    <xsl:apply-templates
        select="following-sibling::p[@rend =
        'Texte du corps (2)'][1]" mode="keep"/>
    </xsl:copy>
</xsl:when>
<xsl:when test="starts-with(., '-') or
starts-with(., '-')"/>
<xsl:otherwise>
    <xsl:copy-of select=".."/>
</xsl:otherwise>
</xsl:choose>
</xsl:template>
<xsl:template
    match="p[@rend = 'Texte du corps (2)'][starts-with
    (., '-') or starts-with(., '-')]"
    mode="keep">
    <xsl:apply-templates/>
</xsl:template>

<xsl:template match="p[@rend = 'Texte du corps (2)']">
    <xsl:choose>
        <xsl:when
            test="following-sibling::p[@rend = 'Texte du
            corps (2)][1][starts-with(., 's.')]">
            <xsl:copy>
                <xsl:attribute name="rend">
                    <xsl:value-of select="@rend"/>
                </xsl:attribute>
                <xsl:apply-templates/>
                <xsl:apply-templates
                    select="following-sibling::p[@rend =
                    'Texte du corps (2)'][1]" mode="keep"/>
            </xsl:copy>
        </xsl:when>
        <xsl:when test="starts-with(., 's.')"/>
        <xsl:otherwise>

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
        <xsl:copy-of select=". "/>
    </xsl:otherwise>
</xsl:choose>
</xsl:template>
<xsl:template match="p[@rend = 'Texte du corps (2)']
[starts-with(., 's.')] " mode="keep">
    <xsl:apply-templates/>
</xsl:template>

<xsl:template match="p[@rend = 'Texte du corps (2)']">
    <xsl:choose>
        <xsl:when
            test="following-sibling::p[@rend = 'Texte du corps
(2)'][1][starts-with(., '<')]">
            <xsl:copy>
                <xsl:attribute name="rend">
                    <xsl:value-of select="@rend"/>
                </xsl:attribute>
                <xsl:apply-templates/>
                <xsl:apply-templates
                    select="following-sibling::p[@rend =
'Texte du corps (2)'][1]" mode="keep"/>
            </xsl:copy>
        </xsl:when>
        <xsl:when test="starts-with(., '<')"/>
        <xsl:otherwise>
            <xsl:copy-of select=". "/>
        </xsl:otherwise>
    </xsl:choose>
</xsl:template>
<xsl:template match="p[@rend = 'Texte du corps (2)']
[starts-with(., '<')] " mode="keep">
    <xsl:apply-templates/>
</xsl:template>

<xsl:template match="p[@rend = 'Texte du corps (2)']">
    <xsl:choose>
        <xsl:when
            test="following-sibling::p[@rend =
```

```

'Texte du corps (2)'][1][starts-with(., '...')]">
<xsl:copy>
    <xsl:attribute name="rend">
        <xsl:value-of select="@rend"/>
    </xsl:attribute>
    <xsl:apply-templates/>
    <xsl:apply-templates
        select="following-sibling::p[@rend =
        'Texte du corps (2)'][1]" mode="keep"/>
</xsl:copy>
</xsl:when>
<xsl:when test="starts-with(., '...')"/>
<xsl:otherwise>
    <xsl:copy-of select=". "/>
</xsl:otherwise>
</xsl:choose>
</xsl:template>
<xsl:template match="p[@rend = 'Texte du corps (2)']
[starts-with(., '...')] mode="keep">
    <xsl:apply-templates/>
</xsl:template>

<xsl:template match="p[@rend = 'Calendrier']">
    <xsl:choose>
        <xsl:when
            test="following-sibling::p[@rend = 'Calendrier']
[1][starts-with(., '-') or starts-with(., '-')]">
            <xsl:copy>
                <xsl:attribute name="rend">
                    <xsl:value-of select="@rend"/>
                </xsl:attribute>
                <xsl:apply-templates/>
                <xsl:apply-templates select="following-sibling:
                    :p[@rend = 'Calendrier'][1]"
                    mode="keep"/>
            </xsl:copy>
        </xsl:when>
        <xsl:when test="starts-with(., '-') or starts-with
(., '-')"/>

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
<xsl:otherwise>
    <xsl:copy-of select=". "/>
</xsl:otherwise>
</xsl:choose>
</xsl:template>
<xsl:template match="p[@rend = 'Calendrier'][starts-with(., '-') or starts-with(., '-')]"
mode="keep">
    <xsl:apply-templates/>
</xsl:template>

<xsl:template match="p[@rend = 'Décoration']">
    <xsl:choose>
        <xsl:when
            test="following-sibling::p[@rend = 'Décoration'][1]
            [starts-with(., '-') or starts-with(., '-')]">
            <xsl:copy>
                <xsl:attribute name="rend">
                    <xsl:value-of select="@rend"/>
                </xsl:attribute>
                <xsl:apply-templates/>
                <xsl:apply-templates select="following-sibling::p
                [@rend = 'Décoration'][1]"
                mode="keep"/>
            </xsl:copy>
        </xsl:when>
        <xsl:when test="starts-with(., '-') or starts-with
        (., '-')"/>
        <xsl:otherwise>
            <xsl:copy-of select=". "/>
        </xsl:otherwise>
    </xsl:choose>
</xsl:template>
<xsl:template match="p[@rend = 'Décoration'][starts-with(., '-') or starts-with(., '-')]"
mode="keep">
    <xsl:apply-templates/>
</xsl:template>
```

```

<xsl:template match="p[@rend = 'Histoire']">
    <xsl:choose>
        <xsl:when
            test="following-sibling::p[@rend = 'Histoire'][1]
            [starts-with(., '-') or starts-with(., '-')]">
            <xsl:copy>
                <xsl:attribute name="rend">
                    <xsl:value-of select="@rend"/>
                </xsl:attribute>
                <xsl:apply-templates/>
                <xsl:apply-templates select="following-sibling::
                    p[@rend = 'Histoire'][1]"
                    mode="keep"/>
            </xsl:copy>
        </xsl:when>
        <xsl:when test="starts-with(., '-') or starts-with
            (., '-')"/>
        <xsl:otherwise>
            <xsl:copy-of select=". "/>
        </xsl:otherwise>
    </xsl:choose>
</xsl:template>
<xsl:template match="p[@rend = 'Histoire'] [starts-with(., '-')
or starts-with(., '-')]"
    mode="keep">
    <xsl:apply-templates/>
</xsl:template>

</xsl:stylesheet>

```

## A.2.2 Documents de transformation

### Code de transformation avec XSLT

Le code suivant crée l'ensemble des balises attendues excepté les informations de rubrication.

```

<?xml version="1.0" encoding="UTF-8"?>
<!-- On inclut la déclaration non préfixé de
TEI en déclaration d'espace de nom par défaut pour ne

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

*ne pas avoir à la répéter dans chaque nouvelle création de balise TEI. -->*

```
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/
Transform"
  xmlns:xs="http://www.w3.org/2001/XMLSchema" exclude-
  result-prefixes="xs tei"
  xmlns:tei="http://www.tei-c.org/ns/1.0" version="3.0"
  xmlns="http://www.tei-c.org/ns/1.0"
  xpath-default-namespace="http://www.tei-c.org/ns/1.0">
  <!-- Suppression des espaces blancs en trop pour tous
les éléments. -->
  <xsl:strip-space elements="*"/>
  <xsl:output indent="yes" method="xml" encoding="UTF-8"/>

  <!-- Création de l'élément racine et de son premier
descendant. -->
  <xsl:template match="TEI">
    <xsl:element name="teiCorpus" namespace="http:/
/www.tei-c.org/ns/1.0">
      <xsl:apply-templates/>
    </xsl:element>
  </xsl:template>

  <!-- Création du teiHeader du document, qui reprend
ici les informations du document source converti. Il
sera à modifier selon ce qui a été déterminé pour les
trois notices modèles dans le document cible. -->
  <xsl:template match="teiHeader">
    <xsl:copy-of select=". "/>
  </xsl:template>

  <!-- On ne prend pas la première <div> qui contient
uniquement le titre du recueil de notices dans le document source. -->
  <xsl:template match="div[1]">

  <!-- On applique à chaque <div> du document source, donc
à chaque notice, la strcuture TEI appropriée et définie en format cible.
Ossature d'une notice structurée. -->
  <xsl:template match="div[position() != 1]">
```

```

<xsl:element name="TEI">
    <xsl:attribute name="n">
        <xsl:apply-templates select="head/hi[@rend =
        'Numero']"/>
    </xsl:attribute>
    <xsl:element name="teiHeader">
        <xsl:element name="fileDesc">
            <xsl:element name="titleStmt">
                <xsl:element name="title"/>
            </xsl:element>
            <xsl:element name="publicationStmt">
                <xsl:element name="p">
                    <xsl:text>cf. supra</xsl:text>
                </xsl:element>
            </xsl:element>
            <xsl:element name="sourceDesc">
                <xsl:element name="listWit">
                    <xsl:element name="witness">
                        <xsl:element name="msDesc">
                            <xsl:element name="msIdentifier">
                                <xsl:element name="repository">
                                    <xsl:text>Bibliothèque
                                    nationale</xsl:text>
                                </xsl:element>
                                <xsl:apply-templates select=
                                "p[@rend = 'Cote']"/>
                                <xsl:apply-templates select=
                                "p/hi[@rend = 'Cote_Car']"/>
                            </xsl:element>
                            <xsl:element name="msContents">
                                <xsl:element name="summary">
                                    <xsl:apply-templates select=
                                    "head"/>
                                </xsl:element>
                                <xsl:apply-templates
                                    select="p[@rend = 'Calendrier']
                                    | p/hi[@rend = 'Calendrier_Car']"/>
                                <xsl:apply-templates
                                    select="p[@rend = 'Texte du corps

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
(2) ']" />
</xsl:element>
<xsl:element name="physDesc">
  <xsl:element name="objectDesc">
    <xsl:apply-templates select=
      "p[@rend = 'Codico']" />
  </xsl:element>
  <xsl:if test="p[@rend = 'Décoration'] |
  p/hi[@rend='Décoration_Car']">
    <xsl:element name="decoDesc">
      <xsl:element name="decoNote">
        <xsl:apply-templates
          select="
            p[@rend = 'Décoration'] |
            p/hi[@rend=
              'Décoration_Car']" />
      </xsl:element>
    </xsl:element>
  </xsl:if>
  <xsl:apply-templates
    select="p[@rend = 'Reliure'] |
    | p/hi[@rend = 'Reliure_Car']" /
  >
</xsl:element>
<xsl:element name="history">
  <xsl:element name="origin">
    <xsl:apply-templates
      select="head/hi[@rend =
        'Date_Manuscrit']" />
    <xsl:apply-templates select=
      "p[@rend = 'Histoire']" />
  </xsl:element>
</xsl:element>
<xsl:apply-templates
  select="p/hi[@rend =
    'Références_Bibliographie_Car']" />
</xsl:element>
</xsl:element>
</xsl:element>
```

```

        </xsl:element>
    </xsl:element>
</xsl:element>
<xsl:element name="text">
    <xsl:element name="body">
        <xsl:element name="p"/>
    </xsl:element>
</xsl:element>
</xsl:element>
</xsl:template>

<!-- On applique un template à chaque partie du document
source dont on récupère l'information et dont on modifie la
structure selon le format cible souhaité. --&gt;

&lt;xsl:template match="head/hi[@rend = 'Numero']"&gt;
    &lt;xsl:value-of select=". "/&gt;
&lt;/xsl:template&gt;

&lt;xsl:template match="p[@rend = 'Cote'] | p/hi[@rend =
'Cote_Car']"&gt;
    &lt;xsl:element name="idno"&gt;
        &lt;xsl:variable name="cote" select=". "/&gt;
        &lt;xsl:analyze-string select="$cote" regex="",
        \s*(.*(\n*).*)"&gt;
            &lt;xsl:matching-substring&gt;
                &lt;xsl:value-of select="normalize-space
(regex-group(1))"/&gt;
            &lt;/xsl:matching-substring&gt;
        &lt;/xsl:analyze-string&gt;
    &lt;/xsl:element&gt;
&lt;/xsl:template&gt;

&lt;xsl:template match="head"&gt;
    <!-- solution 1 : on prend le texte qui descend
    directement de &lt;head&gt;, cela supprime tout encodage supplémentaire --&gt;
    &lt;xsl:value-of select="normalize-space(./text()[1])"/&gt;
    <!-- solution 2: on garde l'encodage supplémentaire,
</pre>

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
mais on a un autre template qui supprime <hi/> -->
<!-- <xsl:apply-templates select=". ." mode="head"/>-->
<!-- Récupération du premier paragraphe s'il concerne
l'ensemble du contenu du mansuscrit. -->
<xsl:value-of
    select="normalize-space(following-sibling::p[@rend =
'Texte du corps (2)' ][1][not(starts-with(., 'F'))])"
/>
</xsl:template>
<!--Suite solution 2 -->
<!-- <xsl:template mode="head" match="hi"/>-->

<xsl:template match="p[@rend = 'Calendrier'] | p/hi[@rend =
'Calendrier_Car']">
    <xsl:variable name="locus_full">
        <xsl:choose>
            <xsl:when test="matches(., 'Fol.')">
                <xsl:analyze-string select=". " regex=
"^-?\s*Fol((.*?)(\n*))\.">
                    <xsl:matching-substring>
                        <xsl:value-of select="normalize-space
(regex-group(1))"/>
                    </xsl:matching-substring>
                </xsl:analyze-string>
            </xsl:when>
            <xsl:when test="matches(., 'fol.')">
                <xsl:analyze-string select=". " regex=
"^-?\s*fol((.*?)(\n*))\.">
                    <xsl:matching-substring>
                        <xsl:value-of select="normalize-space
(regex-group(1))"/>
                    </xsl:matching-substring>
                </xsl:analyze-string>
            </xsl:when>
            <xsl:when test="matches(., '\.{3}\s*(\d)+\s*
(v°)?\s*\.{3}')">
                <xsl:analyze-string select=". " regex="\.{3}
\s*((\d)+\s*(v°?)|\s*\.{3})">
                    <xsl:matching-substring>
```

```

        <xsl:value-of select="normalize-space
            (regex-group(1))"/>
    </xsl:matching-substring>
</xsl:analyze-string>
</xsl:when>
<xsl:when test="matches(., 'P')">
    <xsl:analyze-string select=". " regex="-?\s*P(
        (.*)?(\n*))\.">
        <xsl:matching-substring>
            <xsl:value-of select="normalize-space
                (regex-group(1))"/>
        </xsl:matching-substring>
    </xsl:analyze-string>
</xsl:when>
<xsl:otherwise>
    <xsl:value-of select="substring-before(., '.')"/>
</xsl:otherwise>
</xsl:choose>
</xsl:variable>
<xsl:element name="msItem">
    <xsl:element name="locus">
        <xsl:choose>
            <xsl:when test="matches($locus_full, 'à')
                or matches($locus_full, 'et')">
                <xsl:attribute name="from">
                    <xsl:value-of
                        select="normalize-space(substring-
                            before($locus_full, 'et'))"/>
                <xsl:value-of
                        select="normalize-space(substring-
                            before($locus_full, 'à'))"/>
                <xsl:if test="not(matches(substring-
                    before($locus_full, 'et|à'), 'v'))">
                    <xsl:text>r</xsl:text>
                </xsl:if>
                <xsl:if test="matches(substring-before
                    ($locus_full, 'et|à'), 'v')">
                    <xsl:value-of select="replace(., 'v°', 'v')"/>
                </xsl:if>

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
</xsl:attribute>
<xsl:attribute name="to">
    <xsl:value-of
        select="normalize-space(substring-
            after($locus_full, 'et'))"/>
    <xsl:value-of
        select="normalize-space(substring-
            after($locus_full, 'à'))"/>
    <xsl:if test="not(contains(substring-
        before($locus_full, 'et|à'), 'v'))">
        <xsl:text>r</xsl:text>
    </xsl:if>
    <xsl:if test="matches(substring-before
        ($locus_full, 'et|à'), 'v')">
        <xsl:value-of select="replace(., 'v°', 'v')"/>
    </xsl:if>
</xsl:attribute>
</xsl:when>
<xsl:otherwise>
    <xsl:attribute name="n">
        <xsl:if
            test="matches($locus_full, '\d+')
            or matches($locus_full, '[A-Z]+')">
            <xsl:value-of select="normalize-space(.)"/>
            <xsl:if test="not(matches(., 'v'))">
                <xsl:text>r</xsl:text>
            </xsl:if>
            <xsl:if test="matches(., 'v')">
                <xsl:value-of select="replace
                    (., 'v°', 'v')"/>
            </xsl:if>
        </xsl:if>
    </xsl:attribute>
</xsl:otherwise>
</xsl:choose>
<xsl:value-of select="normalize-space
    ($locus_full)"/>
</xsl:element>
<xsl:variable name="title">
```

```

<xsl:value-of select="substring-after
(., '.')"/>
</xsl:variable>
<xsl:if test="matches($title, 'Calendrier')">
    <xsl:element name="title">
        <xsl:variable name="calendrier"
select=". "/>
        <xsl:analyze-string select="$calendrier"
regex="([0-9|v°]+\\.\\s*)?(Calendrier
+(.*?)\\n*(.*?))([;|\\.|:|-])+>
        <xsl:matching-substring>
            <xsl:value-of select="normalize-
space(regex-group(2))"/>
        </xsl:matching-substring>
    </xsl:analyze-string>
    </xsl:element>
</xsl:if>
<!-- Choix de mettre tout le contenu du paragraphe
dans &lt;note&gt; pour éviter d'éventuelles pertes d'information
avec des regex.--&gt;
&lt;xsl:element name="note"&gt;
    &lt;xsl:value-of select=". "/&gt;
&lt;/xsl:element&gt;
&lt;/xsl:element&gt;
&lt;/xsl:template&gt;

&lt;xsl:template match="p[@rend = 'Codico']"&gt;
    &lt;xsl:element name="supportDesc"&gt;
        &lt;xsl:variable name="material"&gt;
            &lt;xsl:if
                test="
                    starts-with(., 'Par')
                    or starts-with(., 'Vel')
                    or starts-with(., 'Vél')"
                &gt;parch&lt;/xsl:if&gt;
            &lt;xsl:if test="starts-with(., 'Pap')"&gt;paper&lt;/xsl:if&gt;
        &lt;/xsl:variable&gt;
        &lt;xsl:attribute name="material"&gt;
            &lt;xsl:value-of select="normalize-space($material)"/&gt;
</pre>

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
</xsl:attribute>
<xsl:variable name="material_developped">
    <xsl:if test="$material = 'parch'">Parchemin</xsl:if>
    <xsl:if test="$material = 'paper'">Papier</xsl:if>
</xsl:variable>
<xsl:element name="support">
    <xsl:element name="material">
        <xsl:value-of select="normalize-space
($material_developped)"/>
    </xsl:element>
</xsl:element>
<xsl:element name="extent">
    <xsl:if test="matches(., '(\d+)\s*(f|p)')">
        <measure type="composition" unit="leaf">
            <xsl:attribute name="quantity">
                <xsl:variable name="measure" select=". "/>
                <xsl:analyze-string select="$measure"
regex="((\d+)\s*(f|p))">
                    <xsl:matching-substring>
                        <xsl:value-of select="normalize-
space(regex-group(2))"/>
                    </xsl:matching-substring>
                </xsl:analyze-string>
            </xsl:attribute>
            <xsl:variable name="measure" select=". "/>
            <xsl:analyze-string select="$measure"
regex="(\d+)\s*(f|p)">
                <xsl:matching-substring>
                    <xsl:value-of select="normalize-space
(regex-group(1))"/>
                </xsl:matching-substring>
            </xsl:analyze-string>
        </measure>
    </xsl:if>
    <xsl:if test="matches(., '(\d+)\s*sur\s*(\d+)\s*mill')">
        <dimensions scope="all" type="leaf" unit="mm">
            <xsl:element name="height">
                <xsl:attribute name="quantity">
                    <xsl:variable name="height" select=". "/>
```

```

        <xsl:analyze-string select="$height"
regex="(\d+)\s*sur\s*(\d+)\s*mill">
        <xsl:matching-substring>
            <xsl:value-of select="normalize-
space(regex-group(1))"/>
        </xsl:matching-substring>
    </xsl:analyze-string>
</xsl:attribute>
<xsl:variable name="height" select=". . ."/>
<xsl:analyze-string select="$height"
regex="(\d+)\s*sur\s*(\d+)\s*mill">
        <xsl:matching-substring>
            <xsl:value-of select="normalize-
space(regex-group(1))"/>
        </xsl:matching-substring>
    </xsl:analyze-string>
</xsl:element>
<xsl:element name="width">
    <xsl:attribute name="quantity">
        <xsl:variable name="width" select=". . ."/>
        <xsl:analyze-string select="$width"
regex="(\d+)\s*sur\s*(\d+)\s*mill">
        <xsl:matching-substring>
            <xsl:value-of select="normalize-
space(regex-group(1))"/>
        </xsl:matching-substring>
    </xsl:analyze-string>
</xsl:attribute>
<xsl:variable name="width" select=". . ."/>
<xsl:analyze-string select="$width"
regex="(\d+)\s*sur\s*(\d+)\s*mill">
        <xsl:matching-substring>
            <xsl:value-of select="normalize-
space(regex-group(1))"/>
        </xsl:matching-substring>
    </xsl:analyze-string>
</xsl:element>
</dimensions>
</xsl:if>
```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
</xsl:element>
<!-- Choix de prendre dans <condition> l'ensemble des
informations relatives à la codicologie pour éviter
d'éventuelles pertes d'informations. -->
<xsl:element name="condition">
    <xsl:value-of select="normalize-space(.)"/>
    <!-- Tentative de récupérer uniquement les informations
relatives à l'état matériel du manuscrit. -->
    <!--<xsl:choose>
        <xsl:when test="contains(., 'incomplet/mutilés/lacunes')">
            <xsl:variable name="etat" select=".//p[@rend='Codico']"/>
            <xsl:analyze-string select="$etat"
                regex="(col\./lignes\s*), /;(.*?
                (incomplet|mutilés/lacunes)+.*?\n*)\.\s+(-)?">
                <xsl:matching-substring>
                    <xsl:value-of select="regex-group(2)"/>
                </xsl:matching-substring>
            </xsl:analyze-string>
        </xsl:when>
        <xsl:otherwise>
            <xsl:value-of select=".//p[@rend='Codico']"/>
        </xsl:otherwise>
    </xsl:choose>-->
</xsl:element>
</xsl:element>
<xsl:element name="layoutDesc">
    <xsl:element name="layout">
        <xsl:attribute name="columns">
            <xsl:if test="matches(., 'col')">
                <xsl:variable name="col" select=".."/>
                <xsl:analyze-string select="$col" regex=
                    "(.*) (\d+) \s* col (.*)">
                    <xsl:matching-substring>
                        <xsl:value-of select="normalize-
                            space(regex-group(2))"/>
                    </xsl:matching-substring>
                </xsl:analyze-string>
            </xsl:if>
            <xsl:if test="not(matches(., 'col')) or
```

```

        matches(., 'longues lignes')">1</xsl:if>
    </xsl:attribute>
    <xsl:if test="matches(., 'col')">
        <xsl:variable name="col" select=".">

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```

<xsl:element name="origDate">
    <xsl:value-of select="normalize-space(.)"/>
    <!-- Tentative de matcher les mentions de
        date dans les paragraphes relatifs
        au contexte de production du manuscrit : pas
        automatisable ? Cf. problème de l'information de date
        éclatée dans plusieurs paragraphes. -->
    <!-- <xsl:if test="following::p[@rend='Histoire']
        [contains(., 'date')]">
        <xsl:variable name="date" select=". "/>
        <xsl:analyze-string select="$date"
            regex="\.(.*?)(\n*)(date)(.*?)(\n*)\.">
            <xsl:matching-substring>
                <xsl:value-of select="regex-group(1)"/>
            </xsl:matching-substring>
        </xsl:analyze-string>
    </xsl:if>-->
</xsl:element>
</xsl:template>

<xsl:template match="p[@rend = 'Histoire']">
    <xsl:element name="p">
        <xsl:value-of select="normalize-space(.)"/>
    </xsl:element>
    <!-- Tentative de prendre les phrases mentionnant les
        possesseurs des manuscrits : pas automatisable ? L'information
        peut concerner plus d'une phrase.... -->
    <!--<xsl:if test=". /following-sibling::p[@rend='Histoire']
        [contains(., 'possesseur/possesseurs')]">
        <xsl:element name="provenance" namespace="http://www.tei-c.org/ns/1.0">
        <xsl:element name="p" namespace="http://www.tei-c.org/ns/1.0">
        <xsl:variable name="provenance" select=". "/>
        <xsl:analyze-string select="$provenance"
            regex="\.(.*?\n*)(possesseur/possesseurs)(.*?\n*\.).">
            <xsl:matching-substring>
                <xsl:value-of select="regex-group(1)"/>
            </xsl:matching-substring>
        </xsl:analyze-string>
    </xsl:element>
</xsl:template>

```

```

</xsl:element>
</xsl:if>-->
</xsl:template>

<xsl:template match="p/hi[@rend = 'Références_Bibliographie_Car']">
    <xsl:element name="additional">
        <xsl:element name="listBibl">
            <xsl:variable name="biblio" select="normalize-space(.)"/>
            <xsl:for-each select="tokenize($biblio, '-')">
                <xsl:element name="bibl" namespace="http://
www.tei-c.org/ns/1.0">
                    <xsl:value-of select=".."/>
                </xsl:element>
            </xsl:for-each>
        </xsl:element>
    </xsl:element>
</xsl:template>

<xsl:variable name="italic">
    <xsl:value-of select=".//hi[@rend = 'italic']"/>
</xsl:variable>

<xsl:template match="p[@rend = 'Texte du corps (2)']">
    <xsl:variable name="p-id" select="generate-id(.)"/>
    <xsl:variable name="msItem_hypothese" select="tokenize
(., '-')"/>
    <xsl:for-each select="$msItem_hypothese">
        <xsl:variable name="locus_full">
            <xsl:choose>
                <xsl:when test="matches(., 'Fol.')">
                    <xsl:analyze-string select=".." regex=
"-?\s*Fol\.(.*?)(\n*)\.\.">
                        <xsl:matching-substring>
                            <xsl:value-of select="normalize-
space(regex-group(1))"/>
                        </xsl:matching-substring>
                    </xsl:analyze-string>
                </xsl:when>
                <xsl:when test="matches(., 'fol.')">

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
<xsl:analyze-string select=". " regex=
"-?\s*fol\.(.*?)(\n*)\.">
    <xsl:matching-substring>
        <xsl:value-of select="normalize-
            space(regex-group(1))"/>
    </xsl:matching-substring>
</xsl:analyze-string>
</xsl:when>
<xsl:when test="matches
(., '\.\.{3}\s*(\d)+\s*(v°)?\s*\.\.{3}')">
    <xsl:analyze-string select=". " regex=
"\.\.{3}\s*((\d)+\s*(v°)?)\s*\.\.{3}">
        <xsl:matching-substring>
            <xsl:value-of select=
                "normalize-space(regex-group(1))"/>
        </xsl:matching-substring>
</xsl:analyze-string>
</xsl:when>
<xsl:when test="matches(., 'P')">
    <xsl:analyze-string select=". "
        regex="-?\s*P\.(.*?)(\n*)\.">
        <xsl:matching-substring>
            <xsl:value-of select=
                "normalize-space(regex-group(1))"/>
        </xsl:matching-substring>
</xsl:analyze-string>
</xsl:when>
<xsl:otherwise>
    <xsl:value-of select=
        "substring-before(., '.')"/>
</xsl:otherwise>
</xsl:choose>
</xsl:variable>
<xsl:variable name="after_locus">
    <xsl:value-of select="substring-after
(., $locus_full)"/>
</xsl:variable>
<xsl:element name="msItem">
    <xsl:attribute select="$p-id" name="corresp"/>
```

```

<xsl:element name="locus">
    <xsl:choose>
        <xsl:when test="matches($locus_full, 'à')"
            or matches($locus_full, 'et')">
            <xsl:attribute name="from">
                <xsl:value-of
                    select="normalize-space
                        (substring-before($locus_full, 'et'))"/>
            <xsl:value-of
                    select="normalize-space
                        (substring-before($locus_full, 'à'))"/>
            <xsl:if
                    test="not(contains(substring-before
                        ($locus_full, 'et | à'), 'v°'))">
                <xsl:if test="matches($locus_full, 'Fol|fol')">
                    <xsl:text>r</xsl:text>
                </xsl:if>
            </xsl:if>
            <xsl:if
                    test="matches(substring-before
                        ($locus_full, 'et | à'), 'v°')">
                <xsl:value-of
                    select="replace(substring-before
                        ($locus_full, 'et | à'), 'v°', '')"/>
                <xsl:if test="matches($locus_full, 'Fol|fol')">
                    <xsl:text>v</xsl:text>
                </xsl:if>
            </xsl:if>
        </xsl:choose>
        <xsl:attribute name="to">
            <xsl:value-of
                select="normalize-space(substring-after
                    ($locus_full, 'et'))"/>
            <xsl:value-of
                select="normalize-space(substring-after
                    ($locus_full, 'à'))"/>
            <xsl:if
                    test="not(matches(substring-after
                        ($locus_full, 'et | à'), 'v°'))">

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
<xsl:if test="matches($locus_full, 'Fol|fol')">
    <xsl:text>r</xsl:text>
</xsl:if>
</xsl:if>
<xsl:if
    test="matches(substring-after
        ($locus_full, 'et | à'), 'v°')"
<xsl:value-of
    select="replace(substring-after
        ($locus_full, 'et | à'), 'v°', '')"/>
<xsl:if test="matches($locus_full, 'Fol|fol')">
    <xsl:text>v</xsl:text>
</xsl:if>
</xsl:if>
</xsl:attribute>
</xsl:when>
<xsl:otherwise>
    <xsl:attribute name="n">
        <xsl:if
            test="matches($locus_full, '\d+')
            or matches($locus_full, '[A-Z]+')"
<xsl:if test="not(matches($locus_full, 'v'))">
    <xsl:value-of select=
        "normalize-space($locus_full)"/>
    <xsl:text>r</xsl:text>
</xsl:if>
<xsl:if test="matches($locus_full, 'v°')"
<xsl:value-of
    select="normalize-space
        (replace($locus_full, 'v°', ''))"/>
    <xsl:text>v</xsl:text>
</xsl:if>
</xsl:if>
</xsl:attribute>
</xsl:otherwise>
</xsl:choose>
<xsl:value-of select="normalize-space($locus_full)"/>
</xsl:element>
<xsl:if
```

```

test="not(matches($after_locus, '<')) and
not(matches($after_locus, '>')) and not
(matches($after_locus, '...'))">
<xsl:element name="title">
    <xsl:analyze-string select="$after_locus"
        regex="((.*?)\n*(.*?))([;|\.]*)">
        <xsl:matching-substring>
            <xsl:value-of select="normalize-
                space(regex-group(1))"/>
        </xsl:matching-substring>
    </xsl:analyze-string>
</xsl:element>
</xsl:if>
<xsl:if test="matches($after_locus, '<')">
    <xsl:element name="incipit">
        <!-- Regex qui prend en compte
        toutes les possibilités observées :
        (([0-9/vº/A-Z]+\. \s*)?<\s*[A-ZÉÈÀÙ]{1}\{1\}(.*?)(\n*)
        (.*?)(\.\{3\})/:-)) -->
        <xsl:analyze-string select="$after_locus"
            regex="<\s*((.*?)(\n*)(.*?)\.\{3\})">
            <xsl:matching-substring>
                <xsl:value-of select=
                    "normalize-space(regex-group(1))"/>
            </xsl:matching-substring>
        </xsl:analyze-string>
    </xsl:element>
</xsl:if>
<xsl:if
    test="matches($after_locus, '...') and
not(matches($after_locus, '<')) and
not(matches($after_locus, '>'))">
    <xsl:element name="quote">
        <!-- Regex qui prend en compte toutes
        les possibilités observées :
        (([0-9/vº/A-Z]+\. \s*)?(\.\{3\})?((.*?)(\n*)
        (.*?))\.\{3\}) -->
        <xsl:analyze-string select="$after_locus"
            regex="\.\{3\}(.*?)(\n*)(.*?\.\{3\})">

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```

<xsl:matching-substring>
    <xsl:value-of select="normalize-
        space(regex-group(0))"/>
</xsl:matching-substring>
</xsl:analyze-string>
</xsl:element>
</xsl:if>
<xsl:if test="matches($after_locus, '»')">
    <xsl:element name='explicit'>
        <!-- Regex qui prend en compte toutes
            les possibilités observées : (([0-9/vº/A-Z]
            +\.\s*)?\.\{\{3\}\}(.*)?(\n*)(.*?)(\.\.\{\{3\}\}))» -->
        <xsl:analyze-string select=". " regex=
            "((\.\{\{3\}\})?(.*?)(\n*)(.*?)\.)»">
            <xsl:matching-substring>
                <xsl:value-of select="normalize-
                    space(regex-group(1))"/>
            </xsl:matching-substring>
        </xsl:analyze-string>
    </xsl:element>
</xsl:if>
<xsl:element name="note">
    <xsl:value-of select="normalize-space(.)"/>
</xsl:element>
</xsl:element>
</xsl:for-each>
</xsl:template>
</xsl:stylesheet>

```

Le code suivant récupère toutes les informations en italique dans le document source :

```

<?xml version="1.0" encoding="UTF-8"?>
<!-- On inclut la déclaration non prefixé de TEI en déclaration
d'espace de nom par défaut pour ne pas avoir à la répéter
dans chaque nouvelle création de balise TEI. -->
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
    xmlns:xs="http://www.w3.org/2001/XMLSchema" exclude-result-
    prefixes="xs tei"
    xmlns:tei="http://www.tei-c.org/ns/1.0" version="3.0"
    xmlns="http://www.tei-c.org/ns/1.0"

```

```

xpath-default-namespace="http://www.tei-c.org/ns/1.0">
<xsl:strip-space elements="*"/>
<xsl:output indent="yes" method="xml" encoding="UTF-8"/>

<xsl:template match="TEI">
    <xsl:element name="teiCorpus">
        <xsl:apply-templates/>
    </xsl:element>
</xsl:template>

<xsl:template match="teiHeader">
    <xsl:copy-of select=". "/>
</xsl:template>

<!-- On ne prend pas la première <div> qui contient
uniquement le titre du recueil de notices. -->
<xsl:template match="div[1]"/>

<!-- On applique à chaque <div> du document source,
donc à chaque notice, la strcuture TEI appropriée et
définie en format cible.
Ossature d'une notice structurée. -->
<xsl:template match="div[position() != 1]">
    <xsl:element name="TEI">
        <xsl:attribute name="n">
            <xsl:apply-templates select="head/hi[@rend =
'Numero']"/>
        </xsl:attribute>
        <xsl:element name="teiHeader">
            <xsl:element name="fileDesc">
                <xsl:element name="titleStmt">
                    <xsl:element name="title"/>
                </xsl:element>
                <xsl:element name="publicationStmt">
                    <xsl:element name="p">
                        <xsl:text>cf. supra</xsl:text>
                    </xsl:element>
                </xsl:element>
                <xsl:element name="sourceDesc">

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
<xsl:element name="listWit">
  <xsl:element name="witness">
    <xsl:element name="msDesc">
      <xsl:element name="msIdentifier">
        <xsl:element name="repository">
          <xsl:text>Bibliothèque
          nationale</xsl:text>
        </xsl:element>
      <xsl:apply-templates select="p
[@rend = 'Cote']"/>
      <xsl:apply-templates select="p/hi
[@rend = 'Cote_Car']"/>
    </xsl:element>
    <xsl:element name="msContents">
      <xsl:element name="summary">
        <xsl:apply-templates select="head"/>
      </xsl:element>
      <xsl:apply-templates
select="p[@rend = 'Calendrier']
| p/hi[@rend = 'Calendrier_Car']"/>
      <xsl:apply-templates
select="p[@rend = 'Texte du corps
(2)']"/>
    </xsl:element>
    <xsl:element name="physDesc">
      <xsl:element name="objectDesc">
        <xsl:apply-templates select="p
[@rend = 'Codico']"/>
      </xsl:element>
      <xsl:if test="p[@rend = 'Décoration'] |
p/hi[@rend='Décoration_Car']">
        <xsl:element name="decoDesc">
          <xsl:element name="decoNote">
            <xsl:apply-templates
select="p[@rend =
'Décoration'] |
p/hi[@rend='Décoration_Car']"/>
          </xsl:element>
        </xsl:element>
      </xsl:if>
    </xsl:element>
  </xsl:element>
</xsl:element>
```

```

        </xsl:if>
        <xsl:apply-templates
            select="p[@rend = 'Reliure'
            | p/hi[@rend = 'Reliure_Car']]"
        />
    </xsl:element>
    <xsl:element name="history">
        <xsl:element name="origin">
            <xsl:apply-templates
                select="head/hi
                    [@rend = 'Date_Manuscrit']"/>
            <xsl:apply-templates
                select="p[@rend = 'Histoire']"/>
        </xsl:element>
    </xsl:element>
    <xsl:apply-templates
        select="p/hi[@rend =
            'Références_Bibliographie_Car']"/>
    </xsl:element>
    </xsl:element>
    </xsl:element>
    </xsl:element>
    </xsl:element>
    </xsl:element>
    </xsl:element>
    <xsl:element name="text">
        <xsl:element name="body">
            <xsl:element name="p"/>
        </xsl:element>
    </xsl:element>
    </xsl:element>
</xsl:template>

<!-- On applique un template à chaque partie du document
source dont on récupère l'information et dont on modifie la
structure selon le format cible souhaité. --&gt;

&lt;xsl:template match="head/hi[@rend = 'Numéro']"&gt;
    &lt;xsl:value-of select=". "/&gt;
&lt;/xsl:template&gt;
</pre>

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
<!-- Récupération des <rubric> au sein d'un paragraphe. -->
<xsl:template match="p[@rend = 'Texte du corps (2)']">
    <xsl:variable name="p-id" select="generate-id(.)"/>
    <xsl:element name="msItem">
        <xsl:attribute select="$p-id" name="corresp"/>
        <xsl:for-each select=".">
            <xsl:for-each select=".//hi[@rend='italic']">
                <xsl:element name="rubric">
                    <xsl:value-of select="."/>
                </xsl:element>
            </xsl:for-each>
        </xsl:for-each>
        <xsl:element name="note">
            <xsl:value-of select="."/>
        </xsl:element>
    </xsl:element>
</xsl:template>
</xsl:stylesheet>
```

### Code de transformation avec Python

Code inachevé pour tester les possibilités offertes par Python.

```
import re
import docx2txt

notices = docx2txt.process('/Users/gwenaellepatat/Desktop/
Stage_TNAH/MémoireHORAE/Catalogue_VL/noticesTestsocr.docx', 'r')
manuscrits=str(re.split(r'(\W+)', notices))

#paragraphe = re.findall('.*\n.*\n', notices)
#num_notice = re.search('^(([0-9]/I/l)+(-[0-9]+)?)(\.\.\.([A-Z]{2,3})+)', notices)
#titre_notice = re.search('([0-9]/I/l)+\.(.*[A-ZÉÈÀÙ]+)\.', notices)
#print(titre_notice.group(1))

#Capture des titres de notices dans l'ordre stockées
#dans un dictionnaire
```

```

livres_heures={}
for char in range(1, 320):
    try:
        titre=re.search(r"\n"+'((1|I){1,2})?' + str(char) +
        "\.\" + '.*?[A-ZÉÈÙÀ]{2,}.*?\n', notices).group(0)
        livres_heures[char]=titre
        #print(titre)
    except: #Exceptions utiles pour détecter les possibles
        #dénominations irrégulières
        #print(char)
        continue

#Utilisation des valeurs du dictionnaire, soit les titres de notices,
#pour diviser le texte et associer le contenu à chaque titre
regexPattern = '|'.join(map(re.escape, livres_heures.values()))
contenu=re.split(regexPattern, notices)

#Zip de la liste pour créer une liste de tuples contenant les titres
#et leur contenu
structure_notices = list(zip(livres_heures.values(), contenu[1:]))

#for char in structure_notices:
#    print('Titre : %s \n Contenu : %s \n\n' % (char[0], char[1]))

#Capture des descriptions matérielles des manuscrits
for index, char in enumerate(structure_notices):
    try:
        physDesc=re.search(r"(Pareil\.|Parch\.){1}
        ((.*?)\n)*(Rel\.|Rcl\.|Demi\|-reliure){1}(.*?)" + '(-)?',
        char[1]).group(0)
        # char[1] est le contenu dans notre tuple
        structure_notices[index].append(physDesc)
        #Modification de la liste principale
        #print(physDesc)
    except :
        #print(index)
        continue

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
#Capture de la bibliographie
for index, char in enumerate(structure_notices):
    try:
        additional=re.search('^(Rel\.|Rcl\.|Demi\~-reliure){1}
        (.*)\.\s+\-\s+([A-ZÉÈÀÙ]{1,}.*\n', char[1]).group(3)
        # char[1] est le contenu dans notre tuple
        structure_notices[index].append(additional)
        #Modification de la liste principale
        #print(additional)

    except :
        print(index)
        continue

#Capture de l'historique du manuscrit :
#Regex avec faible risque d'erreurs
for index, char in enumerate(structure_notices):
    try:
        history=re.search(r"(\n.*?(usage|possesseur)+.*" + '(\n.*')
        [r"^\Pareil\.|\Parch\.|\Rel\.|\Rcl\.|
        Demi\~-reliure"](.?))' + "(Pareil\.|\Parch\.|
        |\Rel\.|\Rcl\.|Demi\~-reliure)?", char[1]).group(1)
        # char[1] est le contenu dans notre tuple
        structure_notices[index].append(additional)
        #Modification de la liste principale
        #print(history)

    except :
        print(index)
        continue
```

### A.2.3 Documents de sortie

#### Résultats des transformations avec XSLT

Le résultat suivant provient de la fusion des transformations en deux passes : la première pour récupérer toutes les informations sauf l’italique pour les `<rubric>` et les `<finalRubric>`, la deuxième pour récupérer l’italique dans les `<msItem>` divisés sur les paragraphes. A été reproduite ici la notice 313 qui figure parmi les notices modèles<sup>343</sup>.

---

<sup>343.</sup> Certains passages de la notice sont abrégés par souci d’économie de papier, mais l’ensemble des notices encodées après transformation sont disponibles dans les livrables techniques.

```

<TEI n="313">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title/>
      </titleStmt>
      <publicationStmt>
        <p>cf. supra</p>
      </publicationStmt>
      <sourceDesc>
        <listWit>
          <witness>
            <msDesc>
              <msIdentifier>
                <repository>Bibliothèque nationale</repository>
                <idno>fr. nouv. acq., 10230.</idno>
              </msIdentifier>
              <msContents xmlns:tei="http://www.tei-c.org/ns/1.0">
                <summary>HEURES A L'USAGE DE ROUEN ET GRAND
                COUTUMIER DE NORMANDIE.</summary>
                <msItem corresp="d1e15396">
                  <locus n="">i</locus>
                  <incipit/>
                  <explicit/>
                  <note>Fol. i. Au bas du feuillet, note en
                  écriture moderne : « Phillipps ms.
                  22403. »</note>
                </msItem>
                <msItem>
                  <locus from="1r" to="6r">1 à 6</locus>
                  <title>Calendrier très clairsemé où dominent
                  les saints caractéristiques
                  d'Évreux</title>
                  <note>1 à 6. Calendrier très clairsemé où
                  [...]
                  « S. Ursin. »</note>
                </msItem>
                <msItem corresp="d1e15412">
                  <locus from="7r" to="27r">7 à 27</locus>

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
<title>Heures de la Vierge</title>
<note>Fol. 7 à 27. Heures de la Vierge.</note>
</msItem>
<msItem corresp="d1e15412">
<locus from="27v" to="29r">27 v° à 29</locus>
<title>Heures de la Croix</title>
<note>27 v° à 29. Heures de la Croix.</note>
</msItem>
<msItem corresp="d1e15412">
<locus from="29r" to="30r">29 à 30</locus>
<title>Heures du Saint-Esprit</title>
<note>29 à 30. Heures du Saint-Esprit.</note>
</msItem>
<msItem corresp="d1e15412">
<locus from="31r" to="37r">31 à 37</locus>
<title>Psaumes de la pénitence</title>
<note>31 à 37. Psaumes de la pénitence.</note>
</msItem>
<msItem corresp="d1e15412">
<locus from="38r" to="41r">38 à 41</locus>
<title>Litanies</title>
<note>38 à 41. Litanies.</note>
<quote><locus n="39r">39</locus>...s. Romane ;
s. Audioene ; s. Mellone ; s. Vulgane ; s. Martine ;
s. Macute ; s. Hugo ; s. Viviane ; s. Taurine ;
s. Maxime ; s. Alexis ; s.
Severe ; s. Patrici ; s. Philberte</quote>
<quote><locus n="39v">39 v°</locus> s. Huberte ;
s. Leodegari ; s. Iuliane ; s. Amande ; omnes sancti
confessores ; s. Anna... s. Avia ; s. Genovefa...
s. Honorina ; s.
Austreberta... </quote>
</msItem>
</msItem>
<msItem corresp="d1e15412">
<locus from="41v" to="42r">41 v° et 42</locus>
<title>Prologue de saint Jean</title>
<note>41 v° et 42. Prologue de saint Jean.</note>
</msItem>
```

```

<msItem corresp="d1e15414">
    <locus from="43r" to="50r">43 à 50</locus>
    <title>Prières diverses</title>
    <note>Fol. 43 à 50. Prières diverses.</note>
<msItem corresp="d1e15414">
    <locus n="43r">43</locus>
    <incipit>Obsecro te, domina sancta Maria, mater
    Dei, pietate
        plenissima...</incipit>
    <note>43. « Obsecro te, domina sancta Maria, mater
    Dei, pietate plenissima... Et
        in omnibus illis rebus in quibus ego sum facturus,
        locuturus aut
        cogitaturus... et michi famulo tuo impetra a dilecto
        filio tuo complementum
        ...</note>
    <explicit><locus n="45r">45</locus> ...mater Dei et
    misericordie. Amen. »</explicit>
</msItem>
<msItem corresp="d1e15414">
    <locus n="45r">45</locus>
    <rubric>Oracio ad Christum</rubric>
    <incipit>O bone Ihesu, o dulcissime Ihesu, o piissime
    Ihesu,
        o Ihesu, fili Marie virginis...</incipit>
    <explicit><locus n="46v">46v</locus>... qui diligunt
    nomen tuum quod est Ihesus. Amen. »</explicit>
</msItem>
<msItem corresp="d1e15414">
    <locus n="47r">47</locus>
    <incipit>O intemerata...</incipit>
    <quote>... De te enim Filius Dei verus...</quote>
    <quote><locus n="47v">47 v°</locus>...et esto michi
    miserrimo peccatori propicia in omnibus
        auxiliatrix...</quote>
    <explicit><locus n="48v">48 v°</locus>cum omnibus
    sanctis et electis suis. Amen. »</explicit>
</msItem>
<msItem corresp="d1e15414">

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
<locus n="48v">48 v°</locus>
<rubric>« Oratio ad Dominum Ihesum Christum.</rubric>
<quote>Domine Ihesu Christe, adoro te in
    cruce pendentem...</quote>
<explicit><locus n="49r">49</locus>ab angelo
percuciente</explicit>
<finalRubric>Pater noster</finalRubric>
<note>Suivent cinq autres
    invocations.</note>
</msItem>
<msItem corresp="d1e15414">
    <locus from="49v" to="50r">49 v° et 50</locus>
    <title>Passion selon saint Jean</title>
    <note>49 v° et 50. Passion selon saint Jean.</note>
</msItem>
<msItem corresp="d1e15414">
    <locus n="50r">50</locus>
    <incipit>Deus qui manus tuas et pedes tuos...
    </incipit>
    <quote>...et veram scienciam usque in finem.
    Qui... »</quote>
    <note>50. « Deus qui manus tuas et pedes
    tuos...</note>
</msItem>
<msItem corresp="d1e15414">
    <locus n="50v">50 v°</locus>
    <title>Antienne et oraison en l'honneur de
    saint Romain</title>
    <note>50 v°. Antienne et oraison en l'honneur
    de saint Romain.</note>
</msItem>
</msItem>
<msItem corresp="d1e15425">
    <locus from="55r" to="210r">55 à 210</locus>
    <title>Grand coutumier de Normandie</title>
    <note>Fol. 55 à 210. Grand coutumier de
    Normandie.</note>
    <rubric>De marché de bourse. XXIII.</rubric>
</msItem>
```

```

<msItem corresp="d1e15425">
    <locus n="55r">55</locus>
    <incipit>Pour ce que nostre intencion est à
    déclarer en ceste euvre au mieulx
        que nous pourrons les droiz et les esta-
        blissemens de Normendie...</incipit>
    <explicit/>
    <note>55. « Pour ce que nostre intencion est
    à déclarer en ceste euvre au mieulx
        que nous pourrons les droiz et les esta-
        blissemens de Normendie... »</note>
    <rubric>De marché de bourse. XXIII.</rubric>
</msItem>
<msItem corresp="d1e15425">
    <locus n="210r">210</locus>
    <incipit>De marché de bourse. XXIII. Item que
    aucun cas de marché de bourse
        celui à qui l'en demande le marché...</incipit>
    <explicit/>
    <note>210. « De marché de bourse. XXIII. Item
    que aucun cas de marché de bourse
        celui à qui l'en demande le marché... »</note>
    <rubric>De marché de bourse. XXIII.</rubric>
</msItem>
<msItem corresp="d1e15430">
    <locus n="211v">211 v°</locus>
    <incipit/>
    <explicit/>
    <note>Fol. 211 v°. D'une autre main : « Torel
    n'est qu'un sot ; il m'a fait mal
    aux bras. Je voudrais qu'il fût pendu aussy
    hault que je le pourrais
    regarder. »</note>
</msItem>
</msContents>
<physDesc xmlns:tei="http://www.tei-c.org/ns/1.0">
    <objectDesc>
        <supportDesc material="parch">
            <support>

```

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

```
<material>Parchemin</material>
</support>
<extent>
  <measure type="composition" unit="leaf"
    quantity="212">212</measure>
  <dimensions scope="all" type="leaf" unit="mm">
    <height quantity="132">132</height>
    <width quantity="96">96</width>
  </dimensions>
</extent>
<condition>Parch., 212 ff. à longues lignes.
  132 sur 96 mill. -</condition>
</supportDesc>
<layoutDesc>
  <layout columns="1">1 col.</layout>
</layoutDesc>
</objectDesc>
<decoDesc>
  <decoNote>
    <p>Ni peintures ni miniatures. - Quelques
      initiales filigranées. - Petites
      initiales vermillon et azur alternativement.</p>
  </decoNote>
</decoDesc>
<bindingDesc>
  <binding>
    <p>Rel. basane racine.</p>
  </binding>
</bindingDesc>
</physDesc>
<history xmlns:tei="http://www.tei-c.org/ns/1.0">
  <origin>
    <origDate>XVe SIÈCLE</origDate>
    <p>L'office de la Vierge représente l'usage de
      Rouen. Le manuscrit date de la
      seconde moitié ou de la fin du xve siècle.</p>
  </origin>
</history>
</msDesc>
```

```
</witness>
</listWit>
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
  <body>
    <p/>
  </body>
</text>
</TEI>
```

Nous voyons qu'une reprise à la main est indispensable pour la structuration des <msItem>. Il faut être particulièrement attentif aux différents types de citations, aux erreurs de regex pour les titres et les folios, à la restitution des <rubric> et <finalRubric> à leur juste place, et enfin, aux éventuelles imbrications de sections et pièces liturgiques. Il est également important de vérifier les possibles répétitions d'informations dans les <msItem> de données relatives aux calendriers (à remplacer s'ils n'apparaissent pas dans l'ordre de déroulé du livre d'heures) et aux résumés du contenu (<summary>).

## Résultats des transformations avec Python

Ce premier résultat montre la séparation des notices entre elles, toujours à partir des trois notices modèles sélectionnées au départ (notice 1, 112 et 313), puis de leur titre et de leur contenu<sup>344</sup> :

```
<TEI> Titre :
1. LIVRE D'HEURES ET MISSEL FRANCISCAINS. 1380
```

```
Content :
Bibliothèque nationale, ms. lat., 757.
```

```
[...]
```

Rel. maroquin rouge ; dos orné. - Toesca (Pietro), La pittura e la miniatura nella Lombardia, 1912, p. 279 à 294. - Couderc (Camille), Album de portraits, p. 19 et pl. XLV et XLVI. - PRINET (Max), Bulletin de la Société nationale des antiquaires de France (séance du 26 mars 1924), p. 137 à 141. - Leroquais

---

344. Les résultats ci-dessous, le découpage du texte comme la liste de tuples, ont été abrégés dans le texte pour des soucis de lisibilité et d'économie de papier, mais ils sont disponibles dans leur intégralité dans les livrables techniques.

## A.2. DOCUMENTS UTILES À LA TRANSFORMATION

---

(abbé V.), Les sacramentaires et les missels manuscrits des bibliothèques publiques de France, 1924, t. II, p. 361-363 et pl. LXIX à LXXV. - Ancona (Paolo d'), La miniature italienne du Xe au XVIe siècle, 1925, p. 21 et pl. XVI. - [Couderc (Camille)], Catalogue de l'exposition du moyen âge. Bibliothèque nationale, 1926, p. 35-36.

</TEI>

<TEI> Titre :  
II2. HEURES A L'USAGE DE ROME. XVe SIÈCLE

Content :  
Bibliothèque nationale, ms. lat., 1399.

[...]

Rel. maroquin rouge aux armes de France et au chiffre royal.  
</TEI>

<TEI> Titre :  
313. HEURES A L'USAGE DE ROUEN ET GRAND COUTUMIER DE NORMANDIE. XVe SIÈCLE

Content :  
Bibliothèque nationale, fr. nouv. acq., 10230.

[...]

Rel. basane racine. </TEI>

On retrouve ici la liste de tuples obtenue à partir des trois notices modèles :

```
[('\'n1. LIVRE D'HEURES ET MISSEL FRANCISCAINS. 1380\n', "\nBibliothèque nationale,\nms. lat., 757. [...] - [Couderc (Camille)],\nCatalogue de l'exposition du moyen âge. Bibliothèque nationale, 1926,\np. 35-36.\n\n"), ('\'nII2.\tHEURES A L'USAGE DE ROME. XVe SIÈCLE\n', "\nBibliothèque nationale, ms. lat., 1399. [...] \n\nRel. maroquin rouge aux armes de\nau chiffre royal.\n"), ('\'n313. HEURES A L'USAGE DE ROUEN ET GRAND COUTUMIER\nDE NORMANDIE. XVe SIÈCLE\n', '\nBibliothèque nationale, fr. nouv. acq., 10230.\n [...] \n\nRel. basane racine.')]
```

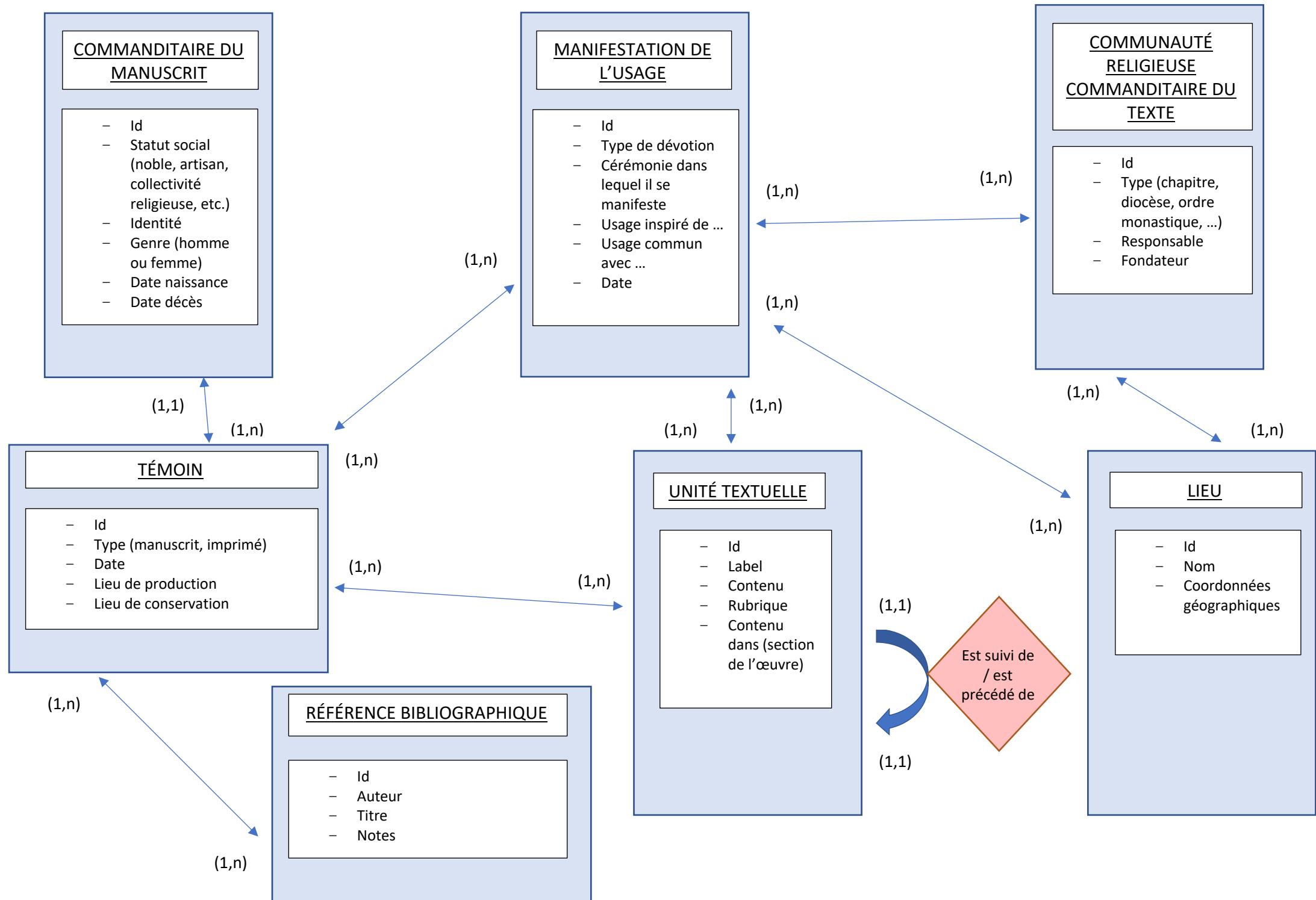


## **Annexe B**

# **Une base de données pour les sciences humaines et sociales : l'exemple d'Heurist**

### **B.1 Documents de modélisation**

#### **B.1.1 Idée de modèle conceptuel**



## B.1. DOCUMENTS DE MODÉLISATION

### B.1.2 Idée de modèle relationnel logique

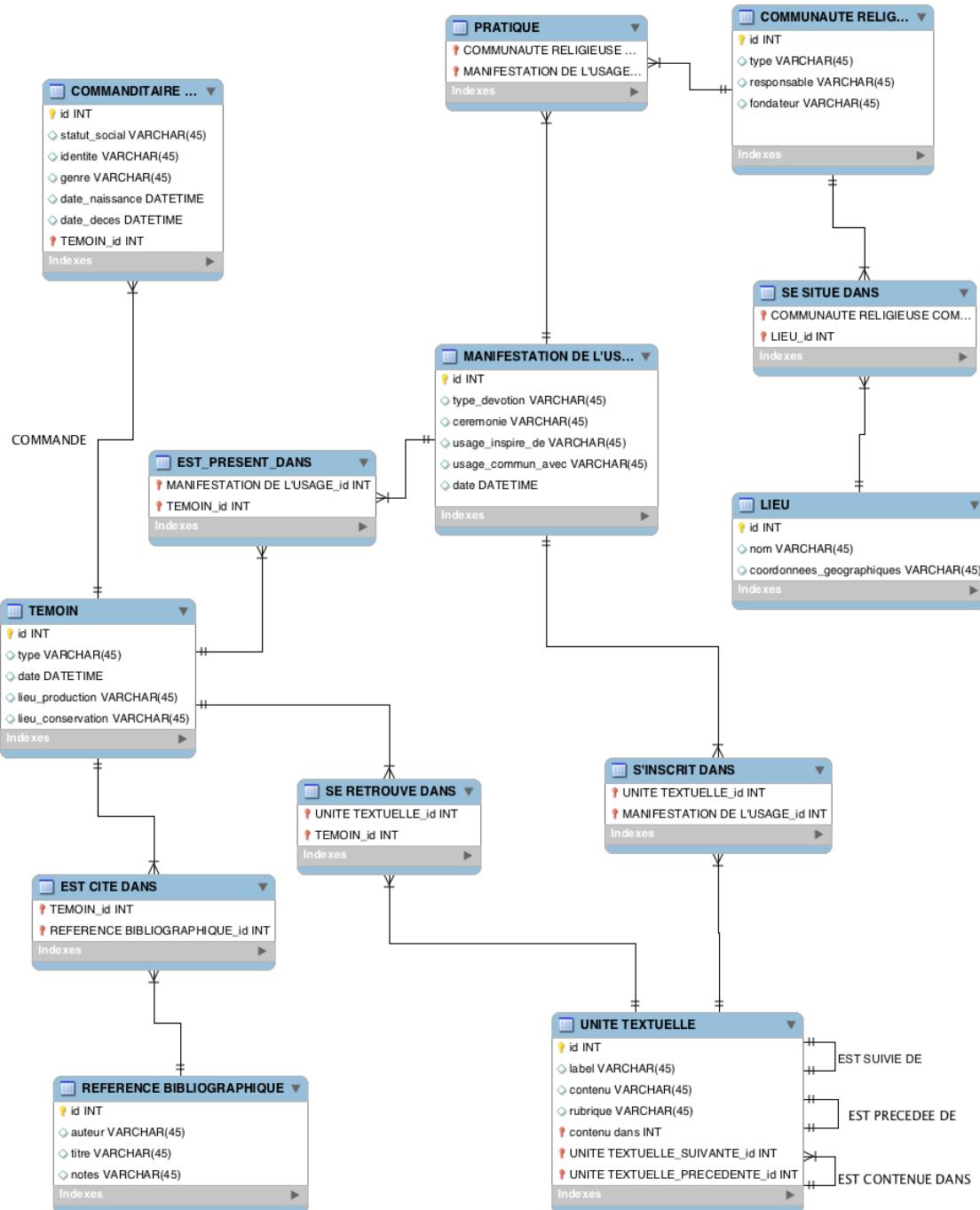
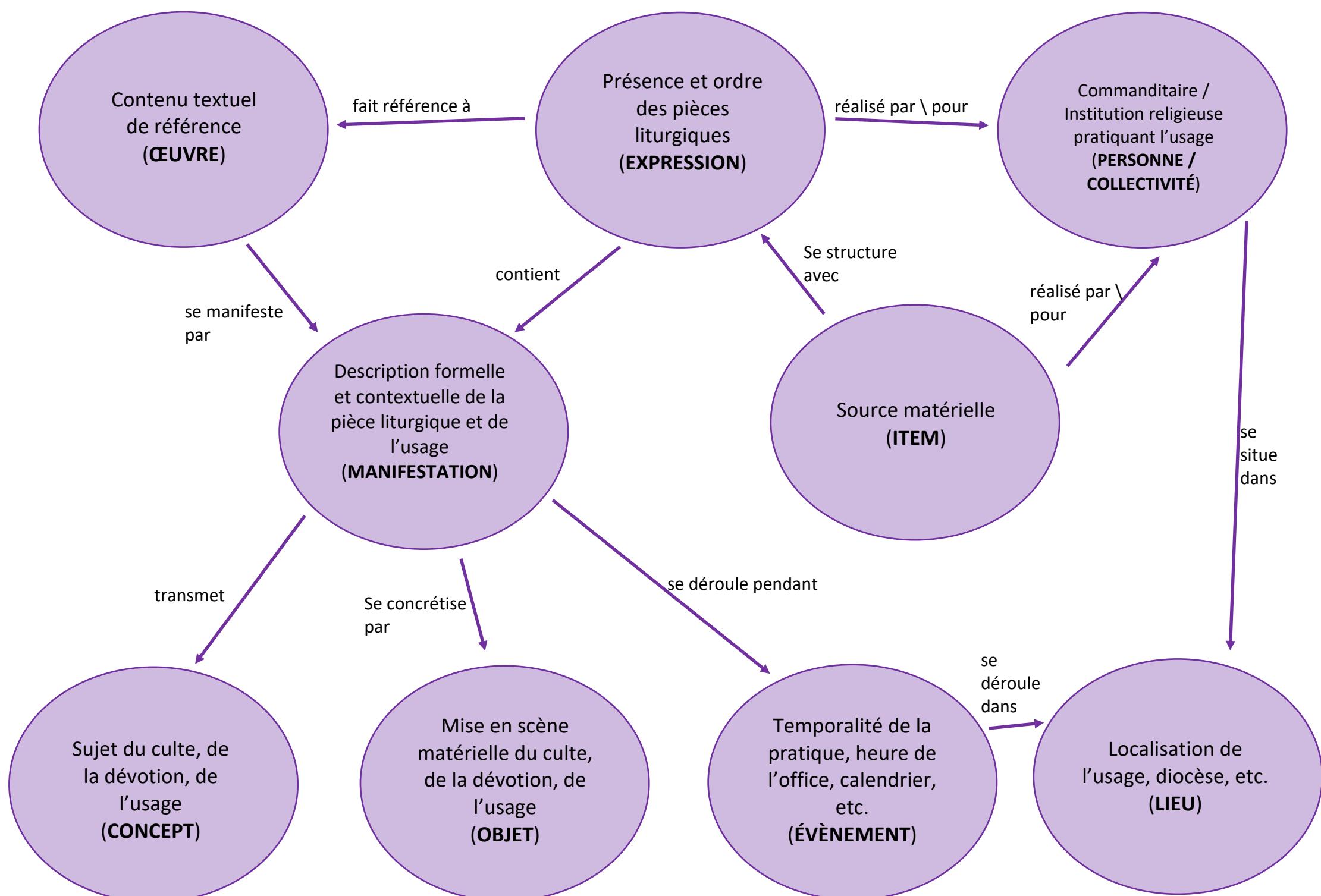


FIGURE B.1 – Modèle relationnel logique des usages dans les livres d'heures

### B.1.3 Modèle conceptuel inspiré des standards RDF et FRBR



## B.2 Documents pour l'import des données

### B.2.1 Définition du format cible pour l'import en XML

Le document ci-dessous représente la structure correcte pour importer les données en XML. On y trouve en commentaires des rappels du modèle fourni par Heurist pour la construction des enregistrements selon le type d'entité.

```
<?xml version="1.0" encoding="UTF-8"?>
<hml xmlns="http://heuristnetwork.org" xmlns:xsi="http://www.w3.org/2001
/XMLSchema-instance"
      xsi:schemaLocation="http://heuristnetwork.org/reference/
schema_hml.xsd">
<database id="0">stutzmann_horae</database>

<!-- Problème sur les RECORDS-IDENTIFIERS
qui n'avaient pas été générés préalablement
dans la base de données. --&gt;

&lt;records&gt;

    &lt;!-- Template pour Use :
    &lt;record&gt;
        &lt;!-- Specify the entity identifier in
            the source database (numeric or alphanumeric)
            if entity may be the target of a record
            pointer field, including the target record
            pointer of a relationship record.--&gt;
        &lt;id&gt;RECORD-IDENTIFIER&lt;/id&gt;
        &lt;!-- type specifies the record (entity)
            type of the record --&gt;
        &lt;type conceptID="0000-92"&gt;Use&lt;/type&gt;
        &lt;url&gt;URL&lt;/url&gt;
        &lt;notes/&gt;
        &lt;detail conceptID="2-9" name="Date"&gt;DATE
        &lt;/detail&gt;
        &lt;detail conceptID="2-21" name="Organisation" isRecordPointer="true"&gt;
            RECORD_REFERENCE
        &lt;/detail&gt;
        &lt;detail conceptID="0000-1011" name="bibl" isRecordPointer="true"&gt;</pre>
```

## B.2. DOCUMENTS POUR L'IMPORT DES DONNÉES

---

```
RECORD_REFERENCE
</detail>
<detail conceptID="0000-1012" name="Note">
MEMO_TEXT</detail>
<detail conceptID="0000-1187" name="Office">
TEXT</detail>
<detail conceptID="0000-1191" name="Cursus" isRecordPointer="true">
RECORD_REFERENCE
</detail>
<detail conceptID="0000-1194" name="Source" isRecordPointer="true">
RECORD_REFERENCE
</detail>
</record>
-->

<!-- 3 essais de type "Use" : Agen, Aix,
Amiens-->

<record>
<id>H-ID-600000</id>
<type conceptId="0000-92">Use</type>
<detail conceptID="2-1" name="Title">
Agen (Use) [2]</detail>
<detail conceptID="2-21" name="Organisation" isRecordPointer="true">
402892</detail>
</record>
<record>
<id>H-ID-600001</id>
<type conceptId="0000-92">Use</type>
<detail conceptID="2-1" name="Title">
Aix-en-Provence (Use) [2]</detail>
<detail conceptID="2-21" name="Organisation" isRecordPointer="true">
402960</detail>
</record>
<record>
<id>H-ID-600002</id>
<type conceptId="0000-92">Use</type>
<detail conceptID="2-1" name="Title">
Amiens (Use) [2]</detail>
```

```
<detail conceptID="2-21" name="Organisation" isRecordPointer="true">  
H-ID-10004</detail>  
</record>  
  
<!-- Template pour UseItem :  
<record>  
  <!-- Specify the entity identifier in  
      the source database (numeric or alphanumeric)  
      if entity may be the target of a record  
      pointer field, including the target record  
      pointer of a relationship record.-->  
  <id>RECORD-IDENTIFIER</id>  
  <!-- type specifies the record (entity)  
      type of the record -->  
  <type conceptID="0000-93">useItem</type>  
  <url>URL</url>  
  <notes/>  
  <detail conceptID="0000-1012" name="Note">  
MEMO_TEXT</detail>  
  <detail conceptID="0000-1134" name="Work" isRecordPointer="true">  
RECORD_REFERENCE  
  </detail>  
  <detail conceptID="0000-1188" name="Work (Cursus)">TEXT</detail>  
  <detail conceptID="0000-1192" name="Sequence">  
TEXT</detail>  
</record>-->  
  
<!--Quel texte mettre dans "Sequence" ? -->  
  
<!-- 6 essais de type useItem pour pouvoir  
faire des relations de succession et de hiérarchie -->  
  
<record>  
  <id>H-ID-700000</id>  
  <type conceptID="0000-93">useItem</type>  
  <detail conceptID="0000-1134" name="Work" isRecordPointer="true">  
H-ID-395039</detail>  
  <detail conceptID="0000-1188" name="Work  
(Cursus)">Quem terra pontus sidera colunt
```

## B.2. DOCUMENTS POUR L'IMPORT DES DONNÉES

---

```
adorant praedicant trinam regentem  
machinam claustrum Mariae bajulat ;  
Cui luna sol  
et omnia deserviunt per tempora  
perfusa caeli gratia gestant puellae  
viscera ; Beata  
mater munere cuius supernus artifex  
mundum pugillo continens ventris sub arca  
clausus est ; Beata caeli nuntio fecunda  
sancto spiritu desideratus gentibus cuius  
per alvum fusus est ; Jesu tibi sit gloria  
qui natus es de virgine cum patre et almo  
spiritu in sempiterna saecula ; Amen [Hymn] (useItem)</detail>  
</record>  
<record>  
    <id>H-ID-700001</id>  
    <type conceptID="0000-93">useItem</type>  
    <detail conceptID="0000-1134" name="Work" isRecordPointer="true">  
        H-ID-398123</detail>  
        <detail conceptID="0000-1188" name="Work  
(Cursus)">Domine, Dominus noster, quam  
            admirabile est nomen tuum in universa  
            terra !quoniam elevata est magnificentia  
            tua super caelos. Ex ore infantium et  
            lactentium perfecisti laudem propter  
            inimicos  
            tuos, ut destruas inimicum et ultorem.  
            Quoniam videbo caelos tuos, opera  
            digitorum tuorum, lunam et stellas quae  
            tu fundasti.  
            Quid est homo, quod memor es ejus ?aut  
            filius hominis, quoniam visitas eum ?  
            Minuisti eum paulominus ab angelis ;  
            gloria et  
            honore coronasti eum ; et constituisti  
            eum super opera manuum tuarum. Omnia  
            subjecisti sub pedibus ejus, oves et  
            boves universas, insuper et pecora  
            campi, volucres caeli, et pisces
```

```
    maris qui perambulant semitas maris.  
    Domine, Dominus  
    noster, quam admirabile est nomen tuum  
    in universa terra ! [Psalm]  
    (useItem)</detail>  
  
<!-- Possible de faire des liens vers  
d'autres entités (comme Use ou les relations) ?  
    <detail conceptID="0000-20" name="Use" isRecordPointer="true">  
        H-ID-600001  
    </detail>  
    <detail conceptID="0000-102"  
        name="Record Relationship">  
        H-ID-800002</detail>-->  
  
</record>  
<record>  
    <id>H-ID-700002</id>  
    <type conceptID="0000-93">useItem</type>  
    <detail conceptID="0000-1188" name="Work  
(Cursus)">O admirabile commercium creator  
        generis humani animatum corpus sumens  
        de virgine nasci dignatus est et procedens  
        homo sine semine largitus est nobis  
        suam deitatem [Antiphon]  
        (useItem)</detail>  
    <detail conceptID="0000-1134"  
        name="Work" isRecordPointer="true">  
        H-ID-399221</detail>  
  
<!-- <detail conceptID="0000-20"  
        name="Use" isRecordPointer="true"  
        >H-ID-600002</detail>  
    <detail conceptID="0000-102"  
        name="Record Relationship">  
        H-ID-800003</detail>-->  
  
</record>  
<record>
```

## B.2. DOCUMENTS POUR L'IMPORT DES DONNÉES

---

```
<id>H-ID-700003</id>
<type conceptID="0000-93">useItem</type>
<detail conceptID="0000-1188"
name="Work (Cursus)">Matins
(useItem)</detail>
<detail conceptID="0000-1134"
name="Work" isRecordPointer="true">
H-ID-394820</detail>

<!-- <detail conceptID="0000-20"
name="Use" isRecordPointer="true">
H-ID-600000</detail>
<detail conceptID="0000-102"
name="Record Relationship">
H-ID-800001</detail>-->

</record>
<record>
<id>H-ID-700004</id>
<type conceptID="0000-93">
useItem</type>
<detail conceptID="0000-1188"
name="Work (Cursus)">Benedicta
tu in mulieribus et
    benedictus fructus ventris
    tui [Antiphon] (useItem)</detail>
<detail conceptID="0000-1134"
name="Work" isRecordPointer="true">
H-ID-395935</detail>

<!--<detail conceptID="0000-20"
name="Use" isRecordPointer="true">
H-ID-600001</detail>
<detail conceptID="0000-102"
name="Record Relationship">
H-ID-800002</detail> -->

</record>
<record>
```

```
<id>H-ID-700005</id>
<type conceptID="0000-93">
useItem</type>
<detail conceptID="0000-1188"
name="Work (Cursus)">Hours of the Virgin
(useItem)</detail>
<detail conceptID="0000-1134"
name="Work" isRecordPointer="true">
H-ID-394805</detail>
</record>

<!-- <detail conceptID="0000-20"
name="Use" isRecordPointer="true">
H-ID-600002</detail>
<detail conceptID="0000-102"
name="Record Relationship">
H-ID-800003</detail>-->

<!-- Template pour Relationship
<record>
<!--\-- Specify the entity
identifier in the source database
(numeric or alphanumeric) if entity
may be the target of a record pointer
field, including the target record pointer
of a relationship record.--\-->
<id>RECORD-IDENTIFIER</id>
<!--\-- type specifies the record (entity)
type of the record --\-->
<type conceptID="2-1">Record
relationship</type>
<url>URL</url>
<notes/>
<detail conceptID="2-1" name="Title for relationship">TEXT</detail>
<detail conceptID="2-3" name="Short description">MEMO_TEXT</detail>
<detail conceptID="2-5" name="Target
record" isRecordPointer="true">
RECORD_REFERENCE</detail>
<detail conceptID="2-6" name="Relationship
```

## B.2. DOCUMENTS POUR L'IMPORT DES DONNÉES

---

```
type" termID="VALUE"/>
<detail conceptID="2-7" name="Source
record" isRecordPointer="true">
RECORD_REFERENCE</detail>
<detail conceptID="2-8" name=
"Interpretation / commentary &gt;">
isRecordPointer="true">
RECORD_REFERENCE</detail>
<detail conceptID="2-10" name=
"Start date/time">DATE</detail>
<detail conceptID="2-11" name=
"End date/time">DATE</detail>
<detail conceptID="1125-101"
name="Certainty" termID="VALUE"/>
<detail conceptID="0000-1121"
name="Source">TEXT</detail>
<detail conceptID="0000-1135"
name="Variant use">TEXT</detail>
</record>-->

<!-- 3 essais de type relations/clés étrangères -->

<record>
<!-- Matins : Contains :
Quem terra pontus... -->
<id>H-ID-800001</id>
<type conceptID="2-1">
Record relationship</type>
<detail conceptID="2-1"
name="Title for relationship">
Contains | Matins (useItem)
&lt;-&gt; Quem terra
pontus sidera colunt
adorant praedicant trinam
regentem machinam claustrum
Mariae bajulat ; Cui luna sol
et omnia deserviunt per tempora
perfusa caeli gratia gestant
puellae viscera ; Beata
```

```
mater munere cuius supernus
artifex mundum pugillo continens
ventris sub arca
clausus est ; Beata caeli nuntio
fecunda sancto spiritu desideratus
gentibus cuius
per alvum fusus est ; Jesu tibi sit
gloria qui natus es de virgine cum
patre et almo
spiritu in sempiterna saecula ;
Amen [Hymn] (useItem)</detail>
<detail conceptID="2-7" name="Source
record" isRecordPointer="true">
H-ID-700003</detail>
<detail conceptID="2-6" name=
"Relationship type" termID="3262"
termConceptID="2-3262"
ParentTerm="Overlap">
Contains</detail>
<detail conceptID="2-5"
name="Target record"
isRecordPointer="true">
H-ID-700000</detail>
</record>
<record>
<!-- Domine, Dominus noster, ... :
ImmediatelyFollows : Benedicta tu... --&gt;
&lt;id&gt;H-ID-800002&lt;/id&gt;
&lt;type conceptID="2-1"&gt;Record
relationship&lt;/type&gt;
&lt;detail conceptID="2-1" name=
"Title for relationship"&gt;
ImmediatelyFollows | Domine,
Dominus noster,quam admirabile
est nomen tuum in
universa terra !quoniam elevata
est magnificentia tua super caelos.
Ex ore infantium
et lactentium perfecisti laudem</pre>
```

## B.2. DOCUMENTS POUR L'IMPORT DES DONNÉES

---

propter inimicos tuos, ut destruas  
inimicum et  
ultorem. Quoniam videobo caelos tuos,  
opera digitorum tuorum, lunam et stellas  
quae tu  
fundasti. Quid est homo, quod memor es  
eius ? aut filius hominis, quoniam  
visitas  
eum ? Minuisti eum paulominus ab  
angelis ; gloria et honore coronasti  
eum ; et  
constituisti eum super opera manuum  
tuarum. Omnia subjecisti sub pedibus  
eius, oves  
et boves universas, insuper et pecora  
campi, volucres caeli, et pisces maris qui  
perambulant semitas maris. Domine, Dominus  
noster, quam admirabile est nomen tuum in  
universa terra ! [Psalm] (useItem)  
&lt;-&gt; Benedicta tu in  
mulieribus et  
benedictus fructus ventris tui [Antiphon] (useItem) </detail>  
<detail conceptID="2-7" name="Source record" isRecordPointer="true">  
H-ID-700001</detail>  
<detail conceptID="2-6" name="Relationship type" termID="3266"  
termConceptID="2-3266"  
ParentTerm="Sequence">ImmediatelyFollows  
</detail>  
<detail conceptID="2-5" name="Target record" isRecordPointer="true"  
>H-ID-700004</detail>  
</record>  
<record>  
<!-- Hours of the Virgin : Contains :<br/>O admirabile commercium... -->  
<id>H-ID-800003</id>  
<type conceptID="2-1">Record  
relationship</type>  
<detail conceptID="2-1" name="Title for relationship">Contains |  
Hours of the

```
Virgin (useItem) &lt;-&gt; 0 admirabile
commercium creator generis humani
animatum corpus sumens de virgine
nasci dignatus
est et procedens homo sine semine
largitus est nobis suam deitatem [Antiphon]
(useItem)</detail>
<detail conceptID="2-7" name="Source
record" isRecordPointer="true">
H-ID-700005</detail>
<detail conceptID="2-6" name=
"Relationship type" termID="3262"
termConceptID="2-3262"
ParentTerm="Overlap">
Contains</detail>
<detail conceptID="2-5" name=
"Target record" isRecordPointer=
"true">H-ID-700002</detail>
</record>
</records>
</hml>
```

## B.2.2 Codes python pour générer le document d'import

Un premier code avait commencé à être élaboré dans l'optique d'obtenir un document conforme à la première cible définie :

```
from lxml import etree
import xml.etree.ElementTree as ET
from bs4 import BeautifulSoup
import untangle
from datetime import datetime
import csv
import urllib

#Paramétrage de l'URL dans la balise citeAs
def param_URL_citeAs(recId, db):
"""

Génère automatiquement l'URL associé à l'enregistrement de la donnée
dans laquelle elle se trouve
```

## B.2. DOCUMENTS POUR L'IMPORT DES DONNÉES

---

```
:param recId: id correspondant au record dans lequel on se trouve
:type recId: str
:param db: nom de la base de données
:type db: str
:returns: URL associée au bon record
:rtype: str
"""

param = urllib.parse.urlencode({'recId': recId, 'db': db})
root_url = "https://heurist.huma-num.fr:443/heurist/"
url_finale = root_url + '?' + param
return url_finale

#Récupération des données depuis le document d'export XML de la base Heurist
exportXML = untangle.parse('/Users/gwenaellepatat/Desktop/Stage_TNAH
/MémoireHORAE/BaseHeurist/Import_données/
Export_stutzmann_horae_20200424152411.xml')

#Element racine
hml= etree.Element("hml")

#Sous-éléments de la balise hml
database = etree.SubElement(hml, "database")
query = etree.SubElement(hml, "query")
datestamp = etree.SubElement(hml, "dateStamp")
resultcount = etree.SubElement(hml, "resultCount")
records = etree.SubElement(hml, "records")
recordcount = etree.SubElement(hml, "recordCount")

#Sous-élément de la balise records
#Multiplier le nombre de balise record en fonction du nombre de
données déclaré et ajouté depuis le CSV
with open ("/Users/gwenaellepatat/Desktop/Stage_TNAH/MémoireHORAE
/BaseHeurist/Données/UseItem_Test_LL.csv") as csvfile:
    donnees_Heurist = csv.reader(csvfile, delimiter=';', quotechar='''')
    #id_record = 599999
    for id_record, row in enumerate(donnees_Heurist, 600000) :
        #id_record += 1
        record = etree.SubElement(records, "record")
```

```
#Sous-éléments de la balise record
id_node = etree.SubElement(record, "id")
id_node.text = str(id_record)

type_node = etree.SubElement(record, "type")
#Valeurs d'attributs de la balise type dépendant de la base Heurist
type_node.set("id", exportXML.hml.records.record[0].type["id"])
#Récupération du nom de l'entité dans l'élément type depuis le CSV
type_node.text = str(row[1])

citeas = etree.SubElement(record, "citeAs")
citeas.text = param_URL_citeAs(str(id_record), "stutzmann_horae")

title = etree.SubElement(record, "title")
title.text = str(row[10])

added = etree.SubElement(record, "added")
added.text = str(datetime.now())

modified = etree.SubElement(record, "modified")
modified.text = str(datetime.now())

workGroup = etree.SubElement(record, "workGroup")
workGroup.text = "public"

detail = etree.SubElement(record, "detail")

# Attribut xmlns : espace de nom vers la page d'accueil de la base Heurist,
#Lien vers le schema hml sur Heurist où sont stockées les données :
#à faire à la main
#Espace de nom pour le schéma de référence d'XML :
#à faire à la main
hml.set("xmlns", exportXML.hml["xmlns"])
#hml.set("xmlns:xsi", "http://www.w3.org/2001/XMLSchema-instance")
#hml.set("xsi:schemaLocation", exportXML.hml["xsi:schemaLocation"])

#Attribut et texte relatifs à la balise database :
database.text = str(exportXML.hml.database.cdata)
```

## B.2. DOCUMENTS POUR L'IMPORT DES DONNÉES

---

```
database.set("id", exportXML.hml.database["id"])

#Attributs relatifs à la balise query :
query.set("q", exportXML.hml.query["q"])
query.set("db", exportXML.hml.query["db"])
query.set("depth", exportXML.hml.query["depth"])

#Texte de l'élément dateStamp qui correspond au jour et à l'heure
d'export de la base
datestamp.text = str(datetime.now())

print(etree.tostring(hml, xml_declaration=True, encoding="UTF-8", pretty_print=True))
tree = ET.ElementTree(hml)
tree.write("ImportUseHeurist.xml".encode('utf8'))
```

Le premier document cible ne correspondant finalement pas à la structure correcte pour l'implémentation des données dans Heurist, nous l'avons ajusté, mais nous ne pouvions générer toutes les balises <detail> nécessaires d'une part car cela demandait de chercher l'information dans plusieurs documents csv qui n'étaient peut-être plus à jour au moment de la reprise du code, d'autre part à cause du problème de récupération d'identifiants générés par Heurist, comme expliqué plus haut.

```
from lxml import etree
import xml.etree.ElementTree as ET
from bs4 import BeautifulSoup
import untangle
from datetime import datetime
import csv

#Récupération des données depuis le document d'export XML de la base Heurist
exportXML = untangle.parse('/Users/gwenaellepatat/Desktop/Stage_TNAH
/MémoireHORAE/BaseHeurist/Import_données/
Export_stutzmann_horae_20200424152411.xml')

#Element racine
hml= etree.Element("hml")

#Sous-éléments de la balise hml
database = etree.SubElement(hml, "database")
records = etree.SubElement(hml, "records")
```

```
#Sous-élément de la balise records
#Multiplier le nombre de balise record en fonction
#du nombre de données déclaré et ajouté depuis le CSV
with open ("/Users/gwenaellepatat/Desktop/Stage_TNAH/
MémoireHORAE/BaseHeurist/Données/UseItem_Test_LL.csv")
as csvfile:
    donnees_Heurist = csv.reader(csvfile, delimiter=';',
                                quotechar='''')
    #id_record = 599999
    for id_record, row in enumerate(donnees_Heurist,
                                    600000) :
        #id_record += 1
        record = etree.SubElement(records, "record")
        #Sous-éléments de la balise record
        id_node = etree.SubElement(record, "id")
        id_node.text = "H-ID-" + str(id_record)

        type_node = etree.SubElement(record, "type")
        #Valeurs d'attributs de la balise type dépendant
        #de la base Heurist
        type_node.set("conceptID", exportXML.hml.records.record[0].type["conceptID"])
        #Récupération du nom de l'entité dans l'élément
        #type depuis le CSV
        for line in csvfile :
            type_node.text = str(row[1])

with open ("/Users/gwenaellepatat/Desktop/Stage_TNAH/MémoireHORAE
/BaseHeurist/Données/UseItem_Test_LL.csv") as csvfile:
    donnees_Heurist = csv.reader(csvfile, delimiter=';',
                                quotechar='''')
    #id_record = 599999
    for id_record, row in enumerate(donnees_Heurist, 600000) :

        detail1 = etree.SubElement(record, "detail")
        if type_node.text == "Use" :
            detail1.set("conceptID", "2-1")
            detail1.set("name", "Title")
```

## B.2. DOCUMENTS POUR L'IMPORT DES DONNÉES

---

```
detail1.text = str(row[1]) + " (Use) [2]"

#Difficile de récupérer depuis le document
#d'export la clé étrangère correspondant à
#la bonne organisation,
#car il faudrait pouvoir une condition
#d'égalité entre le nom de l'usage et celui de l'organisation.

"""

detail2.set("conceptID", "2-21")
detail2.set("name", "Organisation")
detail2.set("isRecordPointer", "True")
detail2.text = str(exportXML.hml.records.record.id)

"""

# Attribut xmlns : espace de nom vers la page d'accueil
#de la base Heurist,
#Lien vers le schema hml sur Heurist où sont stockées les
#données : à faire à la main
#Espace de nom pour le schéma de référence d'XML : à faire
#à la main
hml.set("xmlns", exportXML.hml["xmlns"])
#hml.set("xmlns:xsi", "http://www.w3.org/2001/XMLSchema-instance")
#hml.set("xsi:schemaLocation", exportXML.hml["xsi:schemaLocation"])

#Attribut et texte relatifs à la balise database :
database.text = str(exportXML.hml.database.cdata)
database.set("id", exportXML.hml.database["id"])

#Ajout de l'élément detail autant de fois que l'utilisateur
#enrichi les informations relatives à l'entité en question :
#à tester
#for ajout_detail in record :
^I#if element.tag == 'detail':
^I^I#detail.append(record)

print(etree.tostring(hml, xml_declaration=True, encoding="UTF-8", pretty_print=True))
```

```
tree = ET.ElementTree(hml)
tree.write("ImportUseHeurist.xml".encode('utf8'))
```

# Annexe C

## Annotation et *Machine learning*

### C.1 Le travail d'annotation

Nous rappelons ici les listes de classes répertoriées qui se sont agrandies, et qui témoignent de dévotions personnelles. Il s'agit des suffrages, des offices votifs et des messes votives.

#### C.1.1 Liste des suffrages

- Achatius
- Adrianus
- Agatha
- Agneta
- Ambrosius
- Andreas
- Angeli
- Anna
- Antonius
- Antonius de Padua
- Apollonia
- Apostle (common)
- Armagilus
- Assumptio Mariae
- Athanasius
- Audoenus
- Augustinus
- Avertinus
- Avia
- Barbara
- Barnabas
- Bartholomeus
- Benedictus
- Benignus
- Bernardinus Senensis
- Bernardus
- Blasius
- Carolus
- Catharina
- Catharina de Senis
- Cecilia
- Cesarius
- Christophorus
- Christophorus et Cucuphas
- Christus
- Clara
- Clarus
- Claudius
- Corpus Christi
- Cosmas et Damianus

- Crux Christi
- Cucuphas
- Decem milia martyres
- Demetrius
- Deus pater
- Domicianus et Rogacianus
- Dominicus
- Drogonus
- Dyonisius
- Egidius
- Eligius
- Elisabeth
- Eugenius
- Eustachius
- Eutropius
- Expectatio Mariae
- Exuperus
- Felix et Regula
- Fiacrus
- Firminus
- Florencia
- Florentius
- Franciscus
- Gabriel
- Gallus
- Gatianus
- Gemma
- Genovefa
- Georgius
- Geraldus
- Germanus
- Geronimus
- Gertrudis
- Godelena
- Godo
- Gregorius
- Guillelmus
- Guingaloeus
- Guislenus
- Helena
- Hoyldis
- Hubertus
- Hylarius
- Hyppolitus
- Ildephonsus
- Innocentius
- Ivo
- Jacobus
- Joachim et Anna
- Job
- Johannes Baptista
- Johannes evangelista
- Joseph
- Julianus
- Justus
- Launus
- Laurentius
- Lazarus
- Leochadia
- Leonardus
- Lucas
- Lucia
- Ludovicus de Marsilia
- Ludovicus rex Francie
- Lupus Senonensis
- Lupus Trecensis
- Magnobodus
- Marcellus
- Marcialis
- Marculphus
- Marcus
- Margareta
- Maria Egyptica
- Maria Magdalena
- Martha
- Martinus
- Martyr (common)
- Martyres

## C.1. LE TRAVAIL D'ANNOTATION

---

- Mastidia
- Matheus
- Mathias
- Maturinus
- Maudetus
- Mauricius
- Maurus
- Mello
- Michael
- Neomadia
- Nichasius
- Nicolaus
- Omnes angeli
- Omnes apostoli
- Omnes sancti
- Omnes virgines
- Onuphrius
- Opportuna
- Osmana
- Othmarus
- Patroclus
- Paulus
- Paulus Leonensis
- Pax
- Peregrinus
- Petrus
- Petrus et Paulus
- Petrus Luxemburgensis
- Petrus martyr
- Philippus
- Philippus et Jacobus
- Potentianus
- Purificatio Mariae
- Quatuor evangelistae
- Quintinus
- Radegundis
- Raphael
- Restituta
- Robertus
- Roch
- Romanus
- Rosa
- Sancta Facies Christi
- Sanguis Christi
- Sapientia
- Savinianus
- Sebastianus
- Simon et Judas
- Spinea Corona Christi
- Spiritus Sanctus
- Stephanus
- Susanna
- Symphorianus
- Theobaldus
- Thomas
- Thomas a Becket
- Thomas de Aquino
- Tres Magi
- Tres Mariae
- Trinitas
- Tugdualdus
- Undecim millium virginum
- Urbanus
- Ursula
- Valerius
- Vedastus
- Venissa
- Veronica
- Victor
- Vincentius
- Virgin (common)
- Virgo Maria
- Waldetrudis

### C.1.2 Liste des offices votifs

Cette liste concerne les offices hors de ceux les plus courants que sont l'office de la Vierge, de la Croix, de l'Esprit Saint et des Morts.

- *Common office of the confessors*
- *Hours of Our Lady of Pity*
- *Hours of St Ambrose*
- *Hours of St Anne*
- *Hours of St Anthony the Great*
- *Hours of St Apollonia*
- *Hours of St Barbara*
- *Hours of St Benedict*
- *Hours of St Bernardino*
- *Hours of St Catherine*
- *Hours of St Christopher*
- *Hours of St Denis*
- *Hours of St Fabian and St Sebastian*
- *Hours of St Fiacre*
- *Hours of St Humbert*
- *Hours of St John the Baptist*
- *Hours of St John the Evangelist*
- *Hours of St Joseph*
- *Hours of St Louis of France*
- *Hours of St Margaret*
- *Hours of St Martha*
- *Hours of St Martin*
- *Hours of St Mary Magdalene*
- *Hours of St Maurus*
- *Hours of St Nicolas*
- *Hours of Sts Simon and Jude*
- *Hours of the angels*
- *Hours of the Annunciation*
- *Hours of the Assumption*
- *Hours of the Compassion of the Virgin*
- *Hours of the Conception*
- *Hours of the Cross (long)*
- *Hours of the Dead*
- *Hours of the Eleven Thousand Virgins*
- *Hours of the guardian angel*
- *Hours of the Holy Name of Jesus*
- *Hours of the Holy Sacrament*
- *Hours of the Holy Spirit (long)*
- *Hours of the Messiah*
- *Hours of the Nativity of the Virgin*
- *Hours of the Passion*
- *Hours of the Presentation of the Virgin*
- *Hours of the Sorrows of the Virgin*
- *Hours of the Three Marys*
- *Hours of the Translation of St George*
- *Hours of the Translation of St Gerard*
- *Hours of the Trinity*
- *Hours of the Visitation*

### C.1.3 Liste des messes votives

Les messes votives ne sont pas liées à un temps de l'année liturgique mais aux besoins du célébrant qui la récite. Elles servent à prévenir contre les malheurs et maux du quotidien, et concernent par exemple les maladies des animaux, ou les épidémies.

- *Mass*
- *Mass of All Saints*

## C.2. LE TRAVAIL D'ALIGNEMENT

---

- *Mass of St Catherine*
- *Mass of St Francis*
- *Mass of St Mary Magdalene*
- *Mass of the angels*
- *Mass of the Cross*
- *Mass of the Dead*
- *Mass of the Five Holy Wounds*
- *Mass of the Holy Spirit*
- *Mass of the Trinity*
- *Mass of the Twelve Apostles*
- *Mass of the Virgin*

## C.2 Le travail d'alignement

Vous trouverez ci-dessous les textes de référence des sept psaumes de la Pénitence, soit les psaumes numéros 6, 32, 37, 50, 101, 129 et 142, établis à partir de la transcription du projet « *Clementine Vulgate* » qui a contribué à mettre en ligne les Psautiers de la Vulgate depuis 2005, grâce à des contributions coopératives<sup>345</sup>.

Domine ne in furore tuo arguas me neque in ira tua corripias me Miserere  
mei Domine quoniam infirmus sum sana me Domine quoniam conturbata sunt  
ossa mea Et anima mea turbata est valde et tu Domine usquequo ? Convertere  
Domine et eripe animam meam salvum me fac propter misericordiam tuam.  
Quoniam non est in morte qui memor sit tui in inferno autem quis confitebitur  
tibi ? Laboravi in gemitu meo lavabo per singulas noctes lectum meum lacri-  
mis meis stratum meum rigabo Turbatus est a furore oculus meus inveteravi  
inter omnes inimicos meos Discedite a me omnes qui operamini iniquitatem  
quoniam exaudivit Dominus vocem fletus mei. Exaudivit Dominus deprecatio-  
nem meam Dominus orationem meam suscepit. Erubescant et conturbentur  
vehementer omnes inimici mei convertantur et erubescant valde velociter.

Beati quorum remisso sunt iniquitates et quorum tecta sunt peccata. Beatus vir cui non imputabit Dominus peccatum nec est in spiritu eius dolus. Quoniam tacui inveteraverunt ossa mea dum clamarem tota die. Quoniam die ac nocte gravata est super me manus tua conversus sum in erumna mea dum configitur spina. Delictum meum cognitum tibi feci et iniustitiam meam non abscondi. Dixi confitebor adversum me iniustitiam meam Domino et tu remisisti iniquitatem peccati mei. Pro hac orabit ad te omnis sanctus in tempore opportuno. Verumtamen in diluvio aquarum multarum ad eum non approxi-

---

345. Cf. <http://vulsearch.sourceforge.net/html/Ps.html>.

mabunt. Tu es refugium meum a tribulatione que circumdedit me exultatio mea erue me a circumdantibus me. Intellectum tibi dabo et instruam te in via hac qua gradieris firmabo super te oculos meos. Nolite fieri sicut equus et mulus quibus non est intellectus. In camo et freno maxillas eorum constringe qui non approximant ad te. Multa flagella peccatoris sperantem autem in Domino misericordia circumdabit. Letamini in Domino et exultate iusti et gloriamini omnes recti corde.

Domine ne in furore tuo arguas me neque in ira tua corripias me. Quoniam sagitte tue infixae sunt mihi et confirmasti super me manum tuam. Non est sanitas in carne mea a facie ire tue non est pax ossibus meis a facie peccatorum meorum. Quoniam iniquitates mee supergressae sunt caput meum et sicut onus grave gravatae sunt super me. Putruerunt et corrupte sunt cicatrices mee a facie insipientie mee. Miser factus sum et curvatus sum usque in finem tota die contristatus ingrediebar. Quoniam lumbi mei impleti sunt illusionibus et non est sanitas in carne mea. Afflictus sum et humiliatus sum nimis rugiebam a gemitu cordis mei. Domine ante te omne desiderium meum et gemitus meus a te non est absconditus. Cor meum conturbatum est, dereliquit me virtus mea et lumen oculorum meorum et ipsum non est mecum. Amici mei et proximi mei adversum me appropinquaverunt et steterunt. Et qui iuxta me erant de longe steterunt et vim faciebant qui querebant animam meam. Et qui inquirebant mala mihi locuti sunt vanitates et dolos tota die meditabantur. Ego autem tamquam surdus non audiebam et sicut mutus non aperiens os suum. Et factus sum sicut homo non audiens et non habens in ore suo redargutiones. Quoniam in te Domine speravi tu exaudies me Domine Deus meus. Quia dixi nequando supergaudeant mihi inimici mei et dum commoventur pedes mei super me magna locuti sunt. Quoniam ego in flagella paratus et dolor meus in conspectu meo semper. Quoniam iniquitatem meam annuntiabo et cogitabo pro peccato meo. Inimici autem mei vivunt et confirmati sunt super me et multiplicati sunt qui oderunt me inique. Qui retribuunt mala pro bonis detrahebant mihi quoniam sequebar bonitatem. Ne derelinquas me Domine Deus meus ne discesseris a me. Intende in adiutorium meum Domine Deus salutis mee.

Miserere mei Deus secundum magnam misericordiam tuam. Et secundum multitudinem miserationum tuarum dele iniquitatem meam. Amplius lava me ab iniquitate mea et a peccato meo munda me. Quoniam iniquitatem meam ego cognosco et peccatum meum contra me est semper. Tibi soli peccavi et malum coram te feci ut iustificeris in sermonibus tuis et vincas cum iudicaris. Ecce enim in iniquitatibus conceptus sum et in peccatis concepit me mater mea. Ecce enim veritatem dilexisti incerta et occulta sapientie tue manifestasti

## C.2. LE TRAVAIL D'ALIGNEMENT

---

mihi. Asperges me hyssopo et mundabor lavabis me et super nivem dealbabor. Auditui meo dabis gaudium et letitiam et exultabunt ossa humiliata Averte faciem tuam a peccatis meis et omnes iniurias meas dele. Cor mundum crea in me Deus et spiritum rectum innova in visceribus meis. Ne proiicias me a facie tua et spiritum sanctum tuum ne auferas a me. Redde mihi letitiam salutaris tui et spiritu principali confirma me. Docebo iniquos vias tuas et impii ad te convertentur. Libera me de sanguinibus Deus Deus salutis mee et exultabit lingua mea iustitiam tuam. Domine labia mea aperies et os meum annuntiabit laudem tuam. Quoniam si voluisses sacrificium dedissem utique holocaustis non delectaberis. Sacrificium Deo spiritus contribulatus cor contritum et humiliatum Deus non despicies. Benigne fac Domine in bona voluntate tua Sion ut edificantur muri Hierusalem. Tunc acceptabis sacrificium iustitie oblationes et holocausta tunc imponent super altare tuum vitulos.

Domine exaudi orationem meam et clamor meus ad te veniat. Non avertas faciem tuam a me in quacumque die tribulor inclina ad me aurem tuam. In quacumque die invocavero te velociter exaudi me. Quia defecerunt sicut fumus dies mei et ossa mea sicut cremium aruerunt. Percussus sum ut fenum et aruit cor meum quia oblitus sum comedere panem meum. A voce gemitus mei adhesit os meum carni mee. Similis factus sum pelicano solitudinis factus sum sicut nycticorax in domicilio. Vigilavi et factus sum sicut passer solitarius in tecto. Tota die exprobrabant mihi inimici mei et qui laudabant me adversum me iurabant. Quia cinerem tamquam panem manducabam et poculum meum cum fletu miscebam. A facie ire indignationis tue quia elevans allisisti me. Dies mei sicut umbra declinaverunt et ego sicut fenum arui. Tu autem Domine in eternum perennes et memoriale tuum in generatione et generationem. Tu exsurgens misereberis Sion quia tempus miserendi eius quia venit tempus. Quoniam placuerunt servis tuis lapides eius et terre eius miserebuntur. Et timebunt gentes nomen tuum Domine et omnes reges terre gloriam tuam. Quia edificavit Dominus Sion et videbitur in gloria sua. Respxit in orationem humilium et non sprevit precem eorum. Scribantur hec in generatione altera et populus qui creabitur laudabit Dominum. Quia prospexit de excelso sancto suo Dominus de celo in terram aspexit. Ut audiret gemitus compeditorum ut solveret filios interemptorum. Ut annuntient in Sion nomen Domini et laudem eius in Hierusalem. In conveniendo populos in unum et reges ut serviant Dominum. Respondit ei in via virtutis sue paucitatem dierum meorum nuntia mihi. Ne revokes me in dimidio dierum meorum in generatione et generationem anni tui. Initio tu Domine terram fundasti et opera manuum tuarum sunt celi. Ipsi peribunt tu autem perennes et omnes sicut vestimentum veterascent. Et sicut opertorium mutabis eos et mutabuntur tu autem idem ipse es et anni tui

non deficient. Filii servorum tuorum habitabunt et semen eorum in seculum dirigetur.

De profundis clamavi ad te Domine Domine exaudi vocem meam. Fiant aures tue intendentes in vocem deprecationis mee. Si iniuriae observaveris Domine Domine quis sustinebit ? Quia apud te propitiatio est et propter legem tuam sustinui te Domine. Sustinuit anima mea in verbo eius speravit anima mea in Domino. A custodia matutina usque ad noctem speret Israel in Domino. Quia apud Dominum misericordia et copiosa apud eum redemptio. Et ipse redimet Israel ex omnibus iniuriatibus eius.

Domine exaudi orationem meam auribus percipe obsecrationem meam in veritate tua exaudi me in tua iustitia. Et non intres in iudicium cum servo tuo quia non iustificabitur in conspectu tuo omnis vivens. Quia persecutus est inimicus animam meam humiliavit in terra vitam meam. Collocavit me in obscuris sicut mortuos seculi et anxiatus est super me spiritus meus in me turbatum est cor meum. Memor fui dierum antiquorum meditatus sum in omnibus operibus tuis in factis manuum tuarum meditabar. Expandi manus meas ad te anima mea sicut terra sine aqua tibi. Velociter exaudi me Domine defecit spiritus meus. Non avertas faciem tuam a me et similis ero descendentibus in lacum. Auditam fac mihi mane misericordiam tuam quia in te speravi. Notam fac mihi viam in qua ambulem quia ad te levavi animam meam. Eripe me de inimicis meis Domine ad te confugi doce me facere voluntatem tuam quia Deus meus es tu. Spiritus tuus bonus deducet me in terram rectam : propter nomen tuum Domine vivificabis me in equitate tua. Educes de tribulatione animam meam et in misericordia tua disperdes inimicos meos. Et perdes omnes qui tribulant animam meam quoniam ego servus tuus sum.

### C.3 Le choix de l'interface : le développement d'Ar-kindex

### C.3. LE CHOIX DE L'INTERFACE : LE DÉVELOPPEMENT D'ARKINDEX

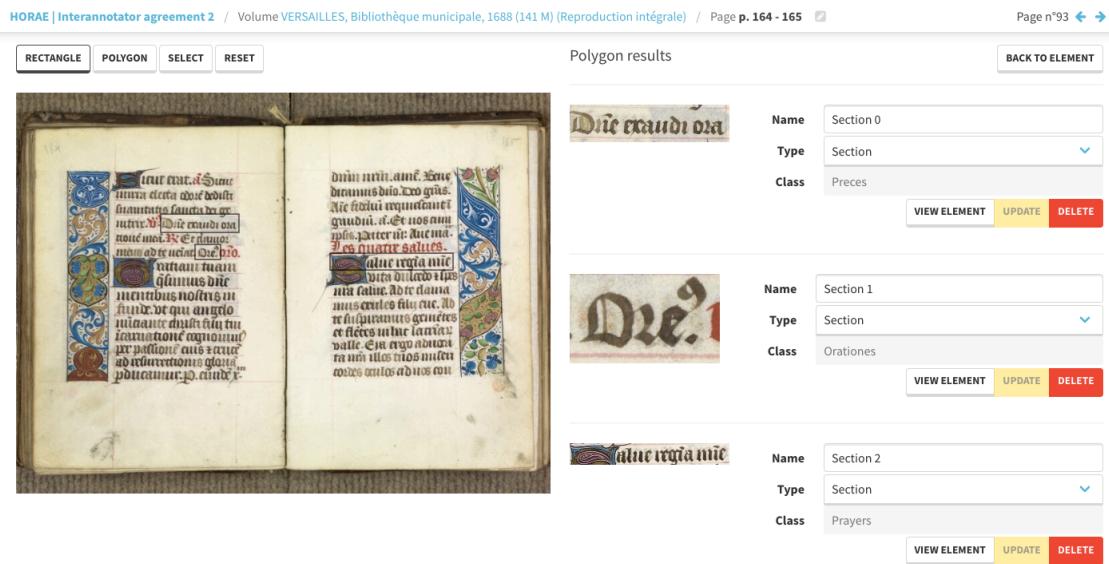


FIGURE C.1 – Illustration du travail d'annotation dans Arkindex

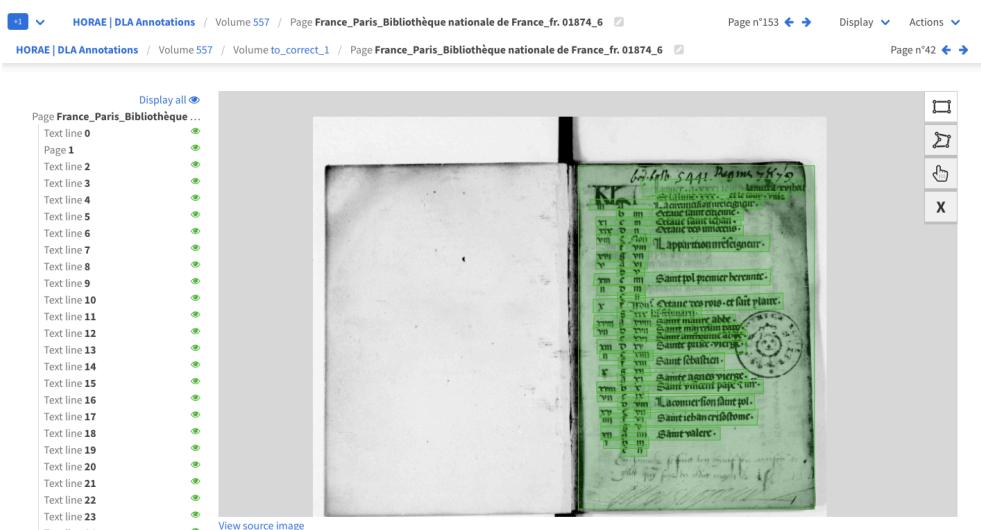


FIGURE C.2 – Le travail de segmentation d'une page dans Arkindex



# Table des figures

1.1 Extraction des termes clés avec GROBID . . . . .	31
2.1 Modèle relationnel logique des usages dans les livres d'heures . . . . .	40
2.2 Test d'implémentation n°1 . . . . .	42
2.3 Test d'implémentation n°2 . . . . .	42
2.4 Test d'implémentation n°3 . . . . .	43
2.5 Test d'implémentation n°4 . . . . .	44
2.6 Modélisation conceptuelle des usages dans les livres d'heures inspirée du modèle FRBR . . . . .	47
2.7 Modèle d'implémentation de l'entité Use . . . . .	50
2.8 Modèle d'implémentation de l'entité UseItem . . . . .	51
2.9 Modèle d'implémentation de l'entité Work . . . . .	52
2.10 Test d'import en XML de l'entité Use . . . . .	63
2.11 Test d'import en XML de l'entité UseItem . . . . .	63
2.12 Test d'import en XML de l'entité Work . . . . .	64
2.13 Cartographie des lieux de conservation des manuscrits étudiés . . . . .	75
2.14 Cartographie d'une organisation identifiée dans les usages des témoins : l'exemple d'Angers . . . . .	76
2.15 Extrait de graphe sur les relations entre les différentes sections d'un livre d'heures . . . . .	79
3.1 Exemple des différentes classes de segmentation d'une page de livre d'heures sur l'interface Arkindex . . . . .	89
3.2 L'imbrication du <i>deep learning</i> , du <i>machine learning</i> et de l'intelligence artificielle. Cf. MARKETING (Search Engine), <i>Quelles sont les différences entre le Deep learning et le Machine learning ?</i> , 2020, URL : <a href="https://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/deep-learning-vs-machine-learning/">https://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/deep-learning-vs-machine-learning/</a> (visité le 28/08/2020) . . . . .	90
3.3 Exemple d'un alignement d'un extrait de psaumes pénitentiels sur Transkribus, extrait du manuscrit ms. lat. 01403, f.78r, conservé à la Bibliothèque nationale de France. . . . .	96

3.4	Nombre et provenance des images et manuscrits du jeu de données. Cf. BOILLET (Mélodie), BONHOMME (Marie-Laurence), STUTZMANN (Dominique) et KERMORVANT (Christopher), « HORAE : an annotated dataset of books of hours », dans <i>The 5th International Workshop on Historical Document Imaging and Processing</i> , Sydney, 2019 (2019 International Conference on Document Analysis and Recognition (ICDAR)), p. 7-12, DOI : 10.1145/3352631.3352633, p. 8 . . . . .	98
3.5	Schéma des liens entre les diverses approches d'analyse des données. Cf. BENSAMOUN (Alexandra) et FARCHY (Joëlle), <i>MISSION INTELLIGENCE ARTIFICIELLE ET CULTURE, Rapport final</i> , rapp. tech., Conseil supérieur de la propriété littéraire et artistique (CSPLA), 2020, p. 13 . . . . .	100
3.6	Exemple d'initiales à annoter avec la classe <i>historiated_initial</i> . . . . .	100
3.7	Exemple d'initiales à annoter avec la classe <i>decorated_initial</i> . . . . .	101
3.8	Exemple d'élément décoratif avec la classe <i>line_filler</i> . . . . .	101
3.9	La méthode de segmentation <i>dhSegment</i> développée par Sofia Ares Oliveira et Benoît Seguin. Cf. <i>dhSegment : A generic deep-learning approach for document segmentation</i> , 2019, DOI : 10.1109/ICFHR-2018.2018.00011, p. 2103	
3.10	Exemple de visualisation comparative de calendriers d'après la base de données relationnelle <i>CoKL : Corpus Kalendarium</i> . . . . .	106
3.11	Analyse des différentes méthodes de segmentation pour les niveaux 1 et 2 de deux livres d'heures, dont l'un a une transcription imparfaite. P <sub>k</sub> et WD indiquent les taux d'erreurs. Cf. DAILLE (Béatrice), HAZEM (Amir), KERMORVANT (Christopher), MAARAND (Martin), BONHOMME (Marie-Laurence), STUTZMANN (Dominique), CURRIE (Jacob) et JACQUIN (Christine), « Transcription automatique et segmentation thématique de livres d'heures manuscrits », <i>TAL</i> , 60–3 (2019), p. 13-36, p. 32 . . . . .	109
A.1	Notice 1 numérisée, première page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 1 . .	120
A.2	Notice 1 numérisée, deuxième page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 2 . .	121
A.3	Notice 1 numérisée, troisième page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 3 . .	122
A.4	Notice 1 numérisée, quatrième page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 4 . .	123
A.5	Notice 1 numérisée, cinquième page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 5 . .	124

## TABLE DES FIGURES

---

A.6 Notice 1 numérisée, sixième page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 6 . . .	125
A.7 Notice 1 numérisée, septième page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 7 . . .	126
A.8 Notice 112 numérisée, première page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 231 . . . . .	127
A.9 Notice 112 numérisée, deuxième page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 232 . . . . .	128
A.10 Notice 313 numérisée, première page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 304 . . . . .	129
A.11 Notice 313 numérisée, deuxième page. Cf. LEROQUAIS (Victor), <i>Les livres d'heures manuscrits de la Bibliothèque nationale</i> . 3 t., Paris, 1927, p. 305 . . . . .	130
A.12 Notice 1 océrisée, première page . . . . .	131
A.13 Notice 1 océrisée, deuxième page . . . . .	132
A.14 Notice 1 océrisée, troisième page . . . . .	133
A.15 Notice 1 océrisée, quatrième page . . . . .	134
A.16 Notice 1 océrisée, cinquième page . . . . .	135
A.17 Notice 1 océrisée, sixième page . . . . .	136
A.18 Notice 1 océrisée, septième page . . . . .	137
A.19 Notice 112 océrisée, première page . . . . .	138
A.20 Notice 112 océrisée, deuxième page . . . . .	139
A.21 Notice 313 océrisée, première page . . . . .	140
A.22 Notice 313 océrisée, deuxième page . . . . .	141
A.23 Notice 1 avec styles Word, première page . . . . .	199
A.24 Notice 1 avec styles Word, deuxième page . . . . .	200
A.25 Notice 1 avec styles Word, troisième page . . . . .	201
A.26 Notice 1 avec styles Word, quatrième page . . . . .	202
A.27 Notice 1 avec styles Word, cinquième page . . . . .	203
A.28 Notice 1 avec styles Word, sixième page . . . . .	204
A.29 Notice 1 avec styles Word, septième page . . . . .	205
B.1 Modèle relationnel logique des usages dans les livres d'heures . . . . .	249
C.1 Illustration du travail d'annotation dans Arkindex . . . . .	277
C.2 Le travail de segmentation d'une page dans Arkindex . . . . .	277

*TABLE DES FIGURES*

---

# Table des matières

Résumé	iii
Remerciements	v
Introduction	xiii
<b>1 Structuration semi-automatique des métadonnées : allier quantité et qualité</b>	<b>1</b>
1.1 Analyse du document source : des données semi-structurées . . . . .	1
1.1.1 Le travail minutieux d'un chanoine passionné . . . . .	2
1.1.2 Des notices papiers aux notices océrisées . . . . .	4
1.2 Définir le document cible : trouver l'équilibre entre contrainte et adaptation . . . . .	7
1.2.1 Normaliser la description des manuscrits . . . . .	8
1.2.2 Des données structurées pour mieux exploiter les informations . . . . .	9
1.3 Encoder les métadonnées : une opération entièrement automatisable ? . . . . .	18
1.3.1 Structurer les notices avec XSLT . . . . .	19
1.3.2 Structurer les notices avec Python . . . . .	26
1.3.3 Les possibilités offertes par l'apprentissage machine . . . . .	30
<b>2 De l'importance de la modélisation des données</b>	<b>35</b>
2.1 Modéliser les usages : restituer une réalité complexe . . . . .	35
2.1.1 Qu'est-ce que l'usage liturgique ? . . . . .	35
2.1.2 S'inspirer de modèles standardisés . . . . .	45
2.2 L'impossible import des données structurées en xml : bilan d'une tentative avortée et solutions alternatives . . . . .	53
2.2.1 Définir la structure du format cible en XML . . . . .	53
2.2.2 Structurer les données : du CSV à l'XML en passant par Python . . . . .	66
2.2.3 Raisons du choix d'importer les données depuis le format CSV . . . . .	68
2.3 Possibilités en visualisation des données . . . . .	71
2.3.1 Géolocaliser les manuscrits et les usages . . . . .	72
2.3.2 Ce que nous disent les textes sur les dévotions... . . . . .	77

---

*TABLE DES MATIÈRES*

---

<b>3 L'annotation pour l'apprentissage machine : ce que le numérique apporte à l'analyse des sources</b>	<b>81</b>
3.1 Gestion et management d'un projet en humanités numériques . . . . .	81
3.1.1 La collaboration IRHT, Teklia et LS2N . . . . .	82
3.1.2 Arkindex : une interface d'annotation en construction . . . . .	88
3.2 Le protocole d'annotation pour guider l'apprentissage machine . . . . .	90
3.2.1 Analyser la structure des pages . . . . .	97
3.2.2 Le cas spécifique des calendriers . . . . .	104
3.2.3 Analyser le contenu textuel . . . . .	105
3.2.4 Des données structurées en masse : quand le numérique sert les humanités . . . . .	109
<b>Conclusion</b>	<b>115</b>
<b>Annexes</b>	<b>119</b>
<b>A Structuration semi-automatisée d'un catalogue de notices</b>	<b>119</b>
A.1 Structurer des notices de livres d'heures . . . . .	119
A.1.1 Documents sources . . . . .	119
A.1.2 Notices océrisées . . . . .	131
A.1.3 Analyses du document à encoder . . . . .	142
A.1.4 Définition du document cible . . . . .	152
A.2 Documents utiles à la transformation . . . . .	199
A.2.1 Documents d'entrée . . . . .	199
A.2.2 Documents de transformation . . . . .	212
A.2.3 Documents de sortie . . . . .	237
<b>B Une base de données pour les sciences humaines et sociales : l'exemple d'Heurist</b>	<b>247</b>
B.1 Documents de modélisation . . . . .	247
B.1.1 Idée de modèle conceptuel . . . . .	247
B.1.2 Idée de modèle relationnel logique . . . . .	249
B.1.3 Modèle conceptuel inspiré des standards RDF et FRBR . . . . .	250
B.2 Documents pour l'import des données . . . . .	252
B.2.1 Définition du format cible pour l'import en XML . . . . .	252
B.2.2 Codes python pour générer le document d'import . . . . .	262
<b>C Annotation et <i>Machine learning</i></b>	<b>269</b>
C.1 Le travail d'annotation . . . . .	269
C.1.1 Liste des suffrages . . . . .	269

## TABLE DES MATIÈRES

---

C.1.2 Liste des offices votifs . . . . .	272
C.1.3 Liste des messes votives . . . . .	272
C.2 Le travail d'alignement . . . . .	273
C.3 Le choix de l'interface : le développement d'Arkindex . . . . .	276
<b>Table des figures</b>	<b>279</b>
<b>Table des matières</b>	<b>283</b>