Modélisation des sources anciennes et édition numérique

Pierre-Yves Buard

Université de Caen Basse-Normandie **École Doctorale SIMEM**

Modélisation des sources anciennes et édition numérique

Thèse de doctorat

présentée et soutenue publiquement le 4 mai 2015

en vue de l'obtention du

Doctorat de l'Université Caen Basse-Normandie

spécialité Informatique et Application

par

Pierre-Yves Buard

Composition du jury

Rapporteurs :	Anne-Marie Turcan-Verkerk,	
	Directeur d'études	École Pratique des Hautes Études, Paris
	Thomas Lebarbé,	
	Professeur des Universités	Université Stendhal – Grenoble 3
Examinateurs :	Catherine Faron-Zucker,	
	Maître de conférences	Université Nice Sophia Antipolis
	Florence Sèdes,	
	Professeur des Universités	Université Toulouse III Paul Sabatier
Directeurs :	Hervé Le Crosnier,	
	Maître de Conférences	
	Habilité à Diriger des Recherches	Université de Caen Basse-Normandie
	Catherine Jacquemard,	
	Professeur des Universités	Université de Caen Basse-Normandie

Mis en page avec la classe thloria.

Remerciements

Je remercie Stéphanie, pour son soutien sans faille, ses encouragements constants et son écoute attentive.

Je remercie Catherine Jacquemard et Hervé Le Crosnier d'avoir accepté de diriger et d'accompagner ce travail pendant ces cinq années. Ce travail n'aurait jamais vu le jour sans leur encadrement et leurs qualités humaines.

Je remercie Anne-Marie Turcan-Verkerk et Thomas Lebarbé d'avoir accepté d'évaluer mon travail.

Je remercie Catherine Faron-Zucker et Florence Sèdes pour leur participation au jury.

Enfin, je tiens à remercier mes collègues de la MRSH, du CERTIC, du GREYC et du CRAHAM.

À ma famille

Table des matières

Partie II Métiers et pratiques

Table des figures

Introduction

Problématique

L'accès aux textes anciens et aux connaissances qu'ils apportent est un enjeu culturel central. Mais la nature même de ces sources historiques pose un certain nombre de difficultés dans le cadre des projets de mise à disposition de ces textes.

Ce travail de doctorat se focalise sur les logiques à l'œuvre dans les opérations de numérisation, d'étude et de diffusion des sources anciennes. De nombreux acteurs et objets différents sont mobilisés dans ce domaine. Nous définissons donc dans ce mémoire les notions d'édition matérielle, d'éditeurs scientifiques, de sources anciennes, etc. autrement dit, l'ensemble des acteurs, les objets qu'ils manipulent et les interactions entre ces objets. Nous insistons sur l'importance de la dimension numérique dans le développement actuel de la compréhension de ces objets et de ces activités. Les principes et modèles que nous proposons dans cette étude sont en grande partie mis en œuvre dans le cadre de différents projets d'édition, de catalogues ou d'inventaires en ligne soutenus par le Pôle Document numérique de la Maison de la Recherche en Sciences Humaines de Caen.

Notre approche trouve son origine dans l'activité éditoriale que nous avons exercée pendant plusieurs années aux Presses universitaires de Caen et au cours de laquelle nous avons été confronté à un certain nombre de questions et problèmes auxquels cette thèse propose d'apporter des réponses. Cette approche éditoriale conditionne donc en grande partie la nature de notre analyse. Nous nous intéressons aux modes de circulation des textes et aux méthodes mises en œuvre pour l'assurer.

Il s'agit pour nous d'étudier l'impact du numérique sur l'ensemble des activités concernées par la diffusion de sources anciennes. Mais bien entendu, tous les métiers concernés ont déjà développé une pratique numérique dans le cadre de leur activité qu'il s'agit de prendre en compte.

Ainsi notre travail s'appuie nécessairement sur l'étude des pratiques de la conservation, de l'étude et de l'édition des sources anciennes pour développer et proposer un modèle d'organisation du travail. Nous retrouvons ici la nécessité déjà soulignée par Yves Jeanneret et Emmanuel Souchier :

> [...] dès lors qu'on choisit de penser cette réalité en termes éditoriaux, il devient indispensable d'aborder les objets et les pratiques qui leur sont liées à partir d'un point de vue particulier, celui de l'invention des formes écrites, de l'imposition de ces formes, de la façon dont elles se disséminent et, ce faisant, de la facon dont elles encadrent la circulation des textes eux-mêmes [Jeanneret et Souchier, 2005].

L'intégration des cultures métier au cœur de notre réflexion, la convergence numérique et la présence constante du réseau nous poussent à interroger la notion même de document numérique pour l'intégrer dans une conceptualisation plus vaste et plus générique intégrant les caractéristiques de fluidité du texte. La notion de document doit en effet se doubler d'une notion de « flux de texte » qui permet de penser autrement l'édition de sources anciennes et plus largement la versatilité des textes en fonction de l'approche spécifique qui est déterminée par la recherche en sciences humaines. Nous nous interrogeons sur la capacité à sémantiser cette approche par les flux.

Notre travail s'inscrit également dans le cadre du développement des *humanités numériques* auquel il participe et par lequel il est influencé. En particulier du point de vue de l'expérimentation et de la pratique qui tiennent une place centrale dans ce mouvement en pleine expansion actuellement :

> [...] *those working in the digital humanities have long held the view that application is as important as theory* [Schreibman *et al.*, 2004].

Ainsi, l'expérimentation, considérée comme la mise à l'épreuve d'un principe ou d'un modèle, tient une place centrale dans l'ensemble de notre étude. Le travail sur l'édition des *Manuscrits du Mont Saint Michel* et de l'*Hortus Sanitatis* servent de support principal au raisonnement. La modélisation des divers participants à l'activité (archivistes, chercheurs, éditeurs) permet de mesurer la complexité de l'édition de sources anciennes, et de mieux placer la convergence numérique comme élément moteur d'une nouvelle conception de cette activité.

Au carrefour de l'informatique, des humanités et des sciences de l'information et de la communication, cette thèse de doctorat veut ouvrir un aspect déterminant des humanités numériques, au-delà de l'usage des statistiques et de la fouille de données. La réflexion sur les pratiques de recherche dans le domaine de l'édition de sources anciennes et leur possible intégration dans l'espace numérique est le guide de ce travail de doctorat.

Organisation du mémoire

Ce mémoire est organisé en quatre parties.

La première se concentre sur la présentation du contexte général de notre réflexion du point de vue des rapports du numérique et des humanités. Il s'agit en

particulier d'étudier comment le numérique, en investissant l'ensemble des activités dans un mouvement de convergence, a impacté les métiers de la conservation, de l'étude et de l'édition de sources anciennes. Une attention particulière est accordée à la recherche en sciences humaines et sociales : quel changement cela provoque-t-il ? Comment les chercheurs s'adaptent (ou non) à ces évolutions ? Cette première partie présente aussi le positionnement spécifique de notre travail dans le champ large et mal délimité des *humanités numériques*.

Une fois ce contexte posé, la seconde partie se focalise sur l'étude des pratiques et des métiers concernés par le champ que nous nous sommes fixé. Les activités des archivistes et conservateurs, des chercheurs et des éditeurs font ainsi l'objet d'une analyse détaillée. Pour chacun de ces métiers nous nous concentrons en particulier sur les objets manipulés et sur le mode d'intégration de la dimension numérique aux activités. Autrement dit, cette seconde partie est consacrée à la manière dont ces métiers se sont adaptés à la convergence numérique.

La troisième partie constitue le cœur de notre contribution. En effet, à l'issue de notre travail d'observation et d'analyse réalisé en majorité de manière empirique dans le cadre de notre activité professionnelle au cours des quinze dernières années, nous proposons dans cette troisième partie une solution d'organisation générale du travail fondée sur un modèle formel dépassant les limites imposées par la logique de document numérique. Nous explorons aussi dans cette partie les perspectives ouvertes par ce modèle en terme de recherche pour les humanités.

La quatrième partie de notre étude est consacrée à la présentation de plusieurs expérimentations et validations de notre modèle et des solutions d'organisation préconisées. Il s'agit de deux éditions de sources publiées aux Presses universitaires de Caen entre 2009 et 2013. La première est celle des *Manuscrits du Mont Saint-Michel*, diffusée sur trois supports (papier, cédérom et en ligne), composée de deux tomes : *Les chroniques latines du Mont Saint-Michel* [Bouet et Desbordes, 2009] et *Le Roman du Mont-Michel* [Bougy, 2009]. La seconde est l'édition du *Tractatus de piscibus* (traité sur les poissons) de l'*Hortus Sanitatis* [Jacquemard *et al.*, 2013]. Nous avons retenu ces deux éditions parmi d'autres exemples mobilisables pour valider notre modèle car elles présentent plusieurs qualités intéressantes pour cette étude. Tout d'abord toutes les deux mobilisent plusieurs supports de diffusion, et c'est un point important qui est largement abordé ici, car la capacité à articuler de multiples

supports de lecture est un enjeu central dans le monde de l'édition. De plus, ces deux expérimentations correspondent l'une aux premières expérimentations, l'autre aux plus récentes. Elles permettent donc à elles deux d'illustrer le modèle que nous proposons et qu'elles ont elles-mêmes fortement contribué à construire à différents niveaux.

Nous concluons cette étude par un bilan de l'ensemble des travaux réalisés et par une présentation des recherches en cours et à venir.

Première partie

Contexte : Numérique et Humanités

Introduction

De nombreux acteurs et objets différents sont mobilisés dans le domaine de l'édition de sources anciennes. Dans les pages qui suivent, l'objectif est de présenter très précisément l'ensemble de ces éléments dans le contexte que l'on nomme aujourd'hui habituellement la *convergence numérique*. En effet, toute notre réflexion sera menée dans ce cadre qu'il nous faut donc ici préciser.

La figure 1 propose une représentation schématique de la situation centrée autour de l'éditeur, ou plus précisément, de la maison d'édition. La situation de cet acteur dans le cycle de production des formes de diffusion permet en effet d'appréhender l'ensemble des problèmes. Il intervient en fait en bout de chaîne et doit en conséquence trouver des solutions à mettre en œuvre dans l'ensemble du cycle de production pour permettre la circulation des textes.

L'éditeur doit, à partir d'une grande variété de documents initiaux, assurer l'ensemble des opérations permettant la diffusion de leurs contenus, quelle que soit leur nature, à destination des publics visés par le projet. Son matériau de base peut se composer de photographies, de tableaux, de textes, de vidéos ou de sons. Tous ces documents peuvent être analogiques ou numériques. Son rôle consiste aujourd'hui à numériser ces documents, s'ils ne le sont pas nativement, dans le respect des standards des métiers concernés. À l'issue de ce travail d'acquisition, il dispose d'un ensemble documentaire cohérent qui entrera dans un processus éditorial à même de remplir plusieurs objectifs.

L'objectif le plus évident réside dans la diffusion des contenus et leur mise à disposition d'un ou plusieurs publics. Aujourd'hui encore, malgré le développement extrêmement rapide des nouveaux supports de lecture comme les liseuses, le papier reste un support de diffusion majeur. Ce support peut d'ailleurs faire l'objet de plusieurs opérations de diffusion : une première publication en grand format, puis quelque temps plus tard une seconde version papier au format poche, il s'agit ici d'un modèle de production tout à fait habituel.

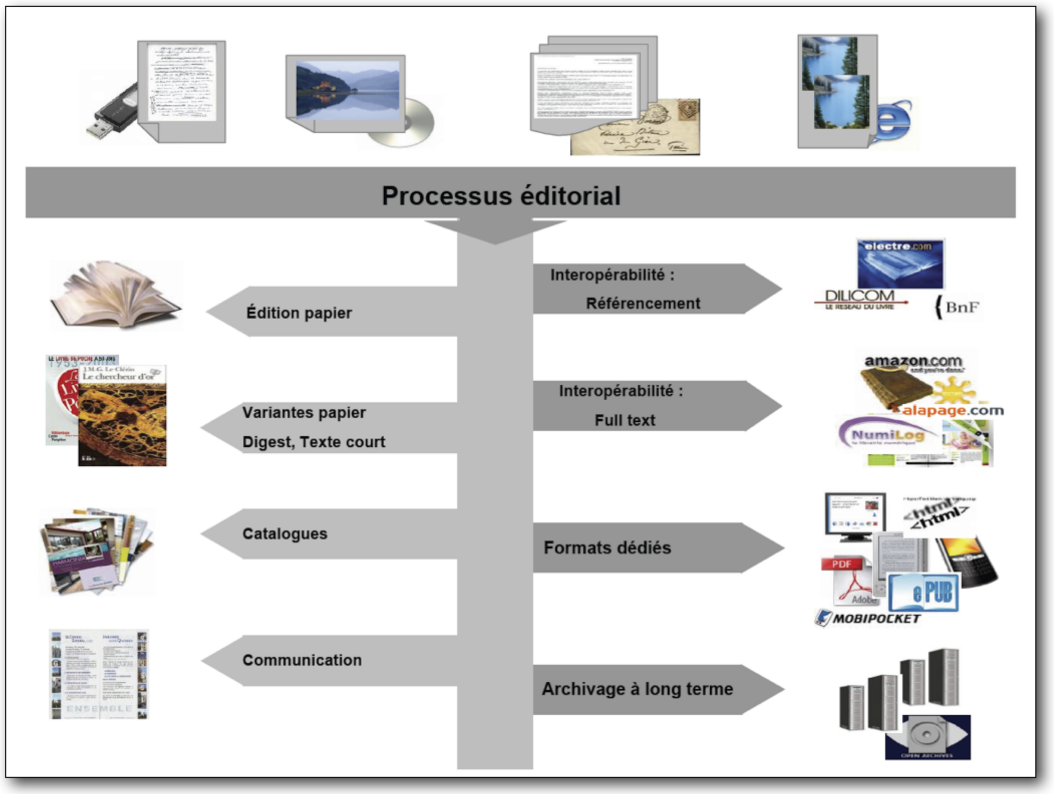

FIGURE 1 – La convergence numérique. Source : Alain PIERROT

L'éditeur doit également mettre en place des procédures pour alimenter l'ensemble des supports de communication comme des *flyers* ou des catalogues thématiques, qui ne reprendront pas l'ensemble des contenus, mais seulement quelques éléments (image de couverture, titre, auteur, résumé, etc.).

De la même manière, les outils de référencement tels qu'Electre ou Dilicom doivent faire l'objet d'alimentation de données afin de permettre la découverte par les lecteurs ou les libraires des titres produits par l'éditeur.

Par ailleurs, la multiplication des supports de lecture (liseuses, tablettes ou *smartphones*) et l'existence d'un lectorat qui souhaite lire sur ces dispositifs, contraignent l'éditeur à produire des versions électroniques pour répondre à cette demande en respectant les spécificités des appareils de lecture, tout comme en intégrant des dispositifs spécifiques d'accompagnement de la lecture. Ainsi, il est nécessaire de prévoir la production d'un certain nombre de formats dédiés (HTML, ePub, mobi, etc.) à ces nouvelles pratiques de lecture sur écran d'ordinateur, de tablettes ou de *smartphones* ou encore sur le papier électronique des liseuses.

Enfin, le dernier point, qui ne concerne pas véritablement la diffusion mais qui constitue une étape essentielle de l'activité éditoriale, est l'archivage. En effet, l'éditeur doit mettre en place des procédures qui lui assurent la possibilité de réutiliser les contenus dans les prochaines années. Plusieurs utilisations sont envisageables : produire une nouvelle édition enrichie sans être contraint de recommencer tout le travail de fabrication, produire une nouvelle forme de diffusion sans opération de rétroconversion, etc. L'enjeu de l'archivage à long terme est donc stratégique et les fichiers numériques archivés constituent aujourd'hui le trésor de l'éditeur.

Nous souhaitons affirmer que la production de ces fichiers repose en grande partie sur les principes de distinction et de séparation du fond et de la forme, ou des formes dans le contexte contemporain. Le respect de ce principe central, sur lequel nous reviendrons dans cette étude, constitue l'un des critères majeurs de mesure qualitative des fichiers produits. La qualité de ces fichiers et leur simplicité de réutilisation ou d'échange sont des caractéristiques fondamentales. Cette organisation du travail est un idéal vers lequel il est indispensable de tendre dans la mesure où la diffusion d'un même texte sur plusieurs supports est devenu un besoin et une attente d'un grand nombre de lecteurs. Pour autant, il ne s'agit pas encore d'une pratique généralisée, même si elle tend à le devenir.

1

Convergence numérique

Nous nous intéressons ici à la dimension numérique présente dans les projets d'édition de sources anciennes et médiévales en particulier. Par édition, nous entendons la fabrication d'une version mise en forme du texte directement utilisable par le lecteur. Ce type de projets concerne plusieurs acteurs qui tous travaillent avec le numérique pour assurer leurs missions. Sans entrer dans les détails de leurs pratiques numériques sur lesquels nous reviendrons plus loin [1], nous devons malgré tout préciser dès maintenant certains points sur les compétences générales mobilisées et la manière dont nous considérons les objets manipulés. Il s'agit donc ici de donner le paysage général des activités exercées dans le cadre de la production d'une édition de sources anciennes.

1.1 Acteurs

Un projet d'édition de sources anciennes fait intervenir plusieurs acteurs, du conservateur qui réalise l'inventaire au secrétaire de rédaction qui assure la préparation de copie en passant par le chercheur qui l'édite scientifiquement. Tous ces travaux sont réalisés en utilisant des outils numériques que les acteurs intègrent à leurs pratiques professionnelles. Dans le cadre d'un projet d'édition, les sources anciennes et les textes qu'elles contiennent constituent donc à la fois le matériau de travail de nombreux acteurs, mais aussi ce qui leur permet d'interagir.

Notons dès maintenant que la qualité des résultats produits est directement liée à la capacité de l'ensemble de ces acteurs à intégrer un dispositif collaboratif. En effet, le travail doit être mené en bonne intelligence car il est entendu qu'aucun d'entre eux n'est en mesure de réaliser seul l'intégralité des opérations nécessaires à la production

1. Voir p. 45 et suivantes.

d'une édition numérique, tout au moins dans le temps généralement fixé pour ce type de projet à l'heure actuelle.

1.2 Outils et compétences informatiques

Certains outils informatiques utilisés dans le domaine de la numérisation sont aujourd'hui tout à fait stables et imposent simplement un effort de formation des acteurs, comme c'est le cas des logiciels d'acquisition ou de saisie. D'autres opérations imposent un effort d'ajustement des outils existants, voire d'innovation et de production de nouvelles solutions informatiques. C'est alors que le développeur devient un acteur central. Il peut être amené à intervenir à toutes les étapes de la vie des projets, de la transcription des textes à leur diffusion en passant par leur structuration et leur annotation.

Toutes ces interventions se font en étroite collaboration avec d'autres acteurs. De ce point de vue, l'informatique est un outil incontournable dont il faut mesurer toute la portée. Bien entendu, cet outil évolué et complexe n'est pas sans influence sur la manière dont les objets sont considérés et manipulés. L'ensemble des dispositifs reposent sur la qualité des relations établies entre les développeurs et les autres acteurs. En effet, il est capital que le dialogue soit le plus clair possible de manière à limiter au maximum la portée et l'influence non maîtrisées des outils mis en place sur l'ensemble des étapes des projets.

Il s'agit, ni plus ni moins, d'éviter autant que possible l'effet "boîte noire" de telle sorte que l'influence de la conception des outils sur les objets soit explicite et choisie pour tous les acteurs.

1.3 Ressources numériques

Dans le cadre d'un projet d'édition de sources anciennes visant la mise à disposition de textes directement exploitables par le lecteur, les ressources numériques textuelles tiennent une place centrale puisque c'est à partir de celles-ci que toutes les formes de diffusion seront produites. Or la constitution de ressources numériques textuelles exploitables dans les meilleures conditions possibles repose sur la capacité des acteurs à respecter des normes et des standards. Plusieurs arguments peuvent être avancés pour justifier cette affirmation.

Tout d'abord les standards et les normes utilisés dans un domaine donné ne sont pas conçus hors contexte. La production d'un standard s'appuie sur des pratiques et

des cultures métier intégrant des représentations précises des objets manipulés dans le domaine et des interactions que les acteurs concernés entretiennent avec ces objets : en un mot sur un modèle de représentation abstrait. Certains modèles s'appuient principalement sur les objets [Baechler et Ingold, 2010] quand d'autres articulent objets et acteurs en intégrant en particulier la dimension d'événement au cœur des dispositifs comme c'est par exemple le cas du CIDOC-CRM [Crofts *et al.*, 2011]. Il est clair que la manière d'aborder les objets influence considérablement la nature du modèle. L'utilisation d'un modèle centré sur la description matérielle des manuscrits ne peut produire les mêmes résultats que l'utilisation d'un modèle centré sur l'organisation logique des textes portés par les mêmes manuscrits. Dans le premier cas, le folio va constituer une unité de base du modèle tandis que dans le second, on créera probablement un découpage plus logique en parties, chapitres, sections, etc. Si les deux approches sont parfaitement défendables, le choix est très lié à la nature du projet. Si l'objectif est de comprendre l'organisation topographique du texte sur le folio alors un modèle orienté vers la matérialité du manuscrit sera plus efficace. En revanche, si l'objectif est de diffuser le texte, c'est sur le contenu textuel que le modèle devra se focaliser.

Ensuite, pour mettre en place une véritable interopérabilité des données qui n'implique pas une maintenance constante de systèmes d'alignement plus ou moins fiables, les communautés ont besoin de s'accorder sur un vocabulaire de description commun, de préférence dans un cadre technique identifié. La norme, ou le standard, est pour cela une méthode éprouvée pour organiser ce travail d'identification des objets manipulés de manière efficace. Ajoutons que si un tel standard existe au sein d'une communauté, que cette dernière en fait usage, mais qu'il présente de réelles lacunes, des problèmes de cohérence ou encore de justesse de dénomination, il est possible, pour les membres de cette même communauté, de participer à l'évolution de l'outil déjà exploité pour en améliorer la qualité. En effet, si une base existe et présente l'énorme avantage d'être bien acceptée par une communauté, il serait vain et assez peu avisé du point de vue de l'intérêt collectif, de chercher à repartir de zéro. Le taux de pénétration d'un outil tel qu'un standard de description est bien évidemment directement lié au service rendu aux membres d'une communauté. Un outil très utilisé malgré des faiblesses manifestes doit donc plutôt faire l'objet de toutes les attentions afin d'être amélioré et rendu encore plus efficace. Ces derniers points nous concernent ici directement car l'édition de sources anciennes couvre plusieurs domaines, et la prise en compte de leurs modèles doit faire partie intégrante de notre raisonnement.

Si ces derniers points peuvent sembler relativement évidents, ils restent fondamentaux et devaient être rappelés ici. La capacité d'une communauté donnée à s'organiser pour proposer des vocabulaires de description est centrale pour lui permettre de mettre en place des ressources numériques pleinement exploitables à grande échelle au-delà des limites d'une institution ou d'un groupe d'individus.

Quand une communauté s'accorde et suit une norme ou un standard cela simplifie les échanges en mettant en place une véritable langue commune articulant un ensemble de vocabulaires de désignation associé à un corpus de règles d'organisation, tous deux mobilisés dans un cadre technique précis. Il s'agit tout simplement de s'accorder sur les modes de désignation des typologies textuelles manipulées et sur la manière de les exploiter. L'existence d'un standard et son utilisation peuvent ainsi être considérées comme des indicateurs forts de la capacité d'une communauté à s'organiser tout d'abord, et de sa capacité à trouver des accords sur la construction des objets manipulés, quelle que soit la nature, scientifique, conservatoire ou autre, des objectifs du projet à l'origine de la mobilisation des acteurs. La communauté des archivistes, avec le standard opérationnel EAD dont elle s'est dotée, constitue ici un très bon exemple. En produisant le standard la communauté construit un outil commun d'échange en même temps que des modes de désignation univoque des objets (fonds, pièces, instruments de recherche, etc.) que les archivistes manipulent.

Enfin, les normes et les standards permettent d'étendre les échanges hors d'une communauté donnée. En affirmant des usages clairs et formalisés, une communauté explicite aux autres sa démarche, l'élaboration de sa problématique et son rapport aux sources. Dans un contexte de convergence numérique, en clarifiant sa pratique, elle améliore considérablement son efficacité.

En définitive, respecter les standards et les normes dans la constitution de ressources numériques, c'est se donner le moyen de fournir un support à l'ensemble des interactions en nommant les phénomènes traités d'une manière univoque.

Rappelons aussi que ces standards et normes doivent être pensés en dehors de tout dispositif technique. Il s'agit d'organiser un vocabulaire de description et pas forcément d'enchaîner une activité à des solutions technologiques, même si, comme nous le verrons, certaines sont nettement plus adaptées à la démarche proposée ici que d'autres. Le meilleur exemple de cette indépendance est celui de la *Text Encoding Initiative* (TEI). En effet, la TEI existe depuis 1987 et sa première "instanciation" était en SGML (*Standard Generalized Markup Language*), mais en 1998, l'arrivée de XML n'a pas du tout remis en cause l'existence de ce standard. La communauté s'est adaptée à ce nouvel environnement technique en assurant l'évolution de l'instancia-

tion du vocabulaire. Pour autant le dispositif technique n'est pas à négliger car c'est lui qui fournit les modalités d'expression : de ses qualités et de sa complexité de mise en œuvre dépendent les utilisations possibles.

Il faut donc voir les normes et les standards comme des langages communs indispensables à tout échange d'un projet à l'autre, mais aussi au sein d'un même projet entre des acteurs qui ne partagent pas nécessairement la même culture ou la même approche du même objet.

Ils constituent aussi la "langue commune" au sein des communautés, parfois émergentes, et constituées autour d'objets communs par différents acteurs qui ne partagent pas toujours la même culture professionnelle. L'objectif est bien de fournir des solutions stables pour permettre la constitution de ressources hautement structurées par les spécialistes des domaines concernés. C'est une solution efficace pour mettre en place des "silos" ou "entrepôts" richement annotés et validés.

Un grand nombre de normes et de standards sont susceptibles d'entrer dans le domaine de l'édition de sources qui nous occupe ici, mais deux vont nous concerner très fortement dans cette étude, la TEI et l'*Encoded Archival Description* (EAD). Ce sont les deux seuls sur lesquels nous reviendrons plus précisément dans ce travail.

Mais avant d'en parler en détail, citons malgré tout ici quelques normes particulièrement utilisées dans le monde de l'édition, des bibliothèques ou du traitement de données qui participent à l'environnement intellectuel de notre projet.

ONIX [2] (*ONline Information eXchange*) créée en 1999 par un groupe d'éditeurs est une norme qui se propose de décrire l'ensemble des aspects de la vie d'un livre, du dépôt du manuscrit chez l'éditeur jusqu'à la mise en rayon chez le libraire. Ce vocabulaire permet donc la manipulation de l'ensemble des métadonnées descriptives, administratives, commerciales et techniques sur les produits éditoriaux de tous types, de la monographie au roman en passant par l'article de revue scientifique. Typiquement ONIX est idéal pour structurer le contenu d'un catalogue d'éditeur. ONIX fonctionne sur un principe simple de transmission de messages sur une base d'identification des produits manipulés. Il est ainsi possible de décrire un projet éditorial dès son commencement, le dépôt du manuscrit chez l'éditeur, et de mettre à jour les informations sur le projet en envoyant de nouveaux messages tout au long de la fabrication du ou des produits correspondants. Ainsi, un message ONIX embryonnaire est créé au dépôt du manuscrit chez l'éditeur. Ce message minimal est ensuite enrichi à chaque étape du traitement du dossier : de la gestion des droits

2. http://www.editeur.org

iconographiques à la rédaction des résumés, en passant par les caractéristiques de toutes les formes de diffusion et les dates de mise en vente, etc. Le flux d'information s'étoffe donc au fur et à mesure de la fabrication.

Dublin Core[3] est une norme créée en 1995 et maintenue par la *Dublin Core Metadata Initiative* volontairement extrêmement simple. Elle ne comporte que 15 éléments de description tous optionnels et tous répétables. L'objectif est ici de faciliter au maximum sa mise en place et son utilisation et pas du tout de prétendre rendre compte de la complexité des ressources aussi sommairement décrites. Il s'agit bien d'améliorer l'efficacité des moteurs de recherche. Mais au-delà de cette ambition initiale, les 15 éléments de base du *Dublin Core* ont rapidement montré leurs limites du point de vue des usages des communautés spécialisées. Celles-ci ont alors travaillé à l'enrichissement du jeu de métadonnées de base par l'ajout d'un système de raffinement pour produire le *Qualified Dublin Core* permettant d'améliorer la précision du vocabulaire initial. Il s'agit d'une opération d'enrichissement du *Dublin Core* appuyée sur l'expérience des communautés. *Dublin Core* est aussi massivement utilisé dans le cadre des archives ouvertes (OAI – *Open Archieve Initiative*), le système aujourd'hui très répandu de moissonnage et d'exposition de données, qui a pour objectif de faciliter la recherche d'information et l'accès aux ressources dans le cadre d'une thématique spécifique. Sans entrer dans les détails, l'OAI offre un système très simple d'exposition de données ainsi qu'un vocabulaire minimal et standardisé d'interrogation sous la forme d'une API REST. Ainsi un unique fournisseur de services, c'est-à-dire un moteur de recherche spécialisé, va pouvoir interroger plusieurs dépôts de données sans connaître l'organisation interne des ressources, dès l'instant où ces entrepôts sont en mesure de répondre aux interrogations de base décrites par le protocole des archives ouvertes.

Les langages du web (XHTML, HTML5, etc.) constituent autant de normes et de standards massivement utilisés. Deux orientations se sont fait face durant ces dernières années. La première approche, et la plus ancienne, celle de la *famille* XHTML place la séparation du fond et de la forme au cœur des préoccupations et consiste à mettre en avant une approche de structuration des données. Les pages écran poussées vers les ordinateurs ou les terminaux mobiles étaient ainsi produits à partir de systèmes de feuilles de styles CSS et de feuilles de transformations XSLT le cas échéant. La seconde approche, celle d'HTML5 et des services associés, renverse totalement la perspective en proposant d'organiser l'ensemble de l'information à partir

3. http://dublincore.org

des applications et des services proposés à l'utilisateur. L'organisation des opérations est réalisée à partir des besoins applicatifs. Il ne s'agit plus de mettre en place des ressources électroniques structurées, mais de proposer des services et des applications qui peuvent faire appel à des contenus. La structuration des données se trouve ainsi reléguée au second plan, derrière le service proposé. La seconde approche l'a emporté [Lecarpentier, 2011]. Il s'agit, comme dans le cas de *Dublin Core*, de simplifier la fabrication d'applications et de services web et non de proposer des solutions de constitution de ressources richement annotées ou structurées.

Les évolutions récentes des langages web affirment donc une orientation vers l'applicatif et le service. Si on peut regretter cet état de fait sur le fond, cette orientation présente pourtant une certaine cohérence : le web est un outil qu'il s'agit d'enrichir en favorisant son développement par la mise en place de normes les plus commodes pour l'implémentation d'applications et de services. En d'autres termes, la structuration de données ne concerne pas directement le web et les technologies qui lui sont liées mais agit dans le *back office* pour produire, stocker et rendre disponibles les documents qui viendront nourrir les applications. Cette partie documentaire du web laisse la question, plus large, de la structuration et de l'annotation des ressources, entière. Deux standards proposent des solutions pour répondre au mieux à ces besoins de constitution de ressources numériques finement structurées dans des domaines qui nous concernent directement dans cette étude. La TEI dans le domaine des humanités au sens large et l'EAD dans le domaine archivistique. Nous les traiterons en détail plus loin.

1.4 Interactions

Tous les acteurs des projets d'études scientifiques des textes sont bien entendu amenés à interagir les uns avec les autres. Les standards et les normes que nous avons rapidement évoqués servent en réalité de socle ou de support à ces interactions.

Il est impossible de proposer une modélisation fonctionnelle sans tenir compte des réalités des pratiques des professionnels. Chaque professionnel entretient avec les objets qu'il manipule quotidiennement des relations spécifiques. Et la nature de ces relations donne des informations sur la nature des objets eux-mêmes. De la même manière, les relations entretenues par les différents acteurs donnent de précieuses indications dont il est impératif de tenir compte dans le travail de modélisation.

1.4.1 Nature

Une étude scientifique de textes a dorénavant pour résultat la mise à disposition des textes eux-mêmes et des commentaires scientifiques sous la forme d'une simple publication ou sous celle, plus exigeante, d'une édition de textes. Pour aboutir à cette mise à disposition des résultats auprès d'un public, les différents intervenants doivent nécessairement interagir les uns avec les autres.

La figure 1.1 donne une représentation de l'organisation du travail telle qu'on peut l'observer couramment aujourd'hui. Si tous les projets ne sont pas explicitement organisés de cette façon, cette représentation donne selon nous une bonne idée de la manière dont le travail se déroule réellement dans la plupart des cas.

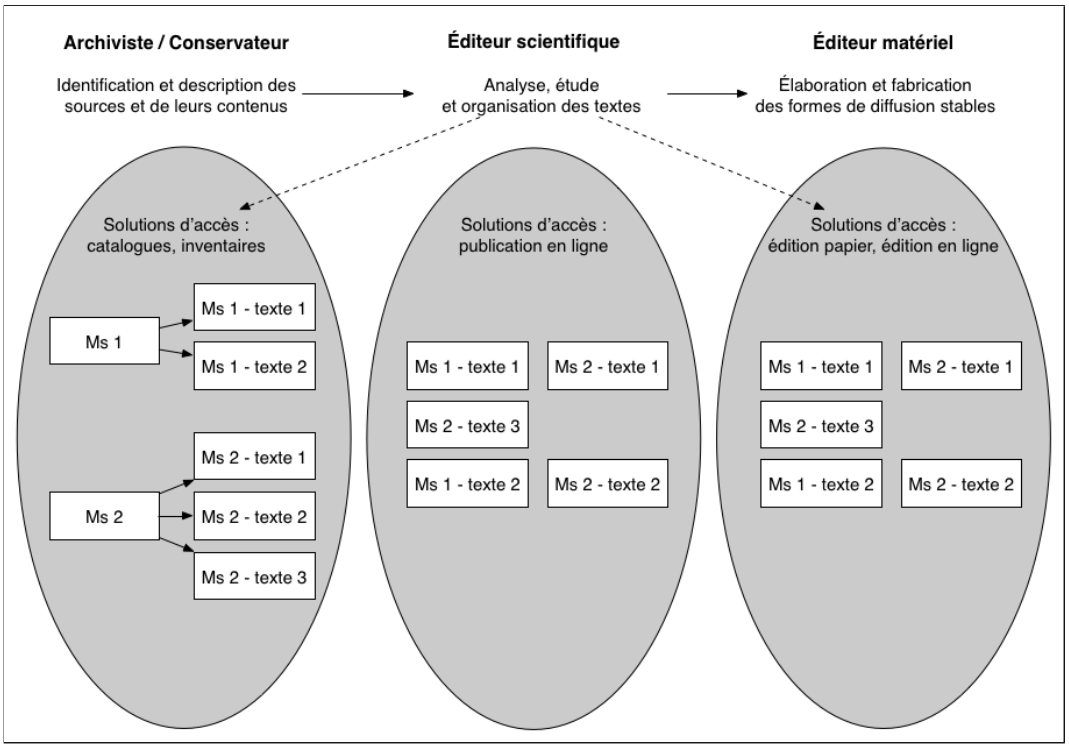

FIGURE 1.1 – De l'inventaire à l'édition.

L'archiviste assure donc ses activités de conservation, d'inventaire et de description des documents et propose des solutions électroniques de consultation des fonds. Les politiques d'inventaire, de catalogage et de diffusion peuvent éventuellement être influencées en amont par des demandes de chercheurs dans le cadre de collaborations ponctuelles ou au long cours. Dans tous les cas, un travail d'édition de sources commence forcément par un repérage, une identification et une description des documents, autrement dit, il est impératif pour tous les acteurs que les archivistes puissent travailler dans les meilleures conditions possibles. Ainsi l'éditeur scientifique a tout

intérêt à faciliter au maximum le travail de l'archiviste et à s'accorder avec lui sur les modalités de travail. De la même manière, une activité de recherche aboutissant à la mise à disposition d'études sur un fonds donné participe de sa valorisation. L'archiviste a donc également tout intérêt à faciliter le travail du chercheur et à s'accorder avec lui sur les modalités de travail.

La figure 1.1 montre comment chaque opération, de l'inventaire à l'édition multimodale, est liée à la précédente. L'archiviste indique et identifie les textes portés par les manuscrits, puis l'éditeur scientifique étudie ces textes, en propose une nouvelle organisation pour la lecture et rédige des notes de commentaires et d'explication en facilitant la compréhension. Enfin l'éditeur matériel assure les opérations de contrôle et de normalisation avant de produire les formes de diffusion.

Le rôle du développeur n'est pas indiqué sur la figure 1.1 car il n'intervient pas directement sur les données textuelles elles-mêmes, mais fournit les outils qui permettent aux autres acteurs de le faire. Son intervention est donc transversale et il collabore avec l'ensemble des acteurs. Son rôle est capital car les applications informatiques ne sont pas sans implications sur la manière dont les données sont perçues et conçues par ceux qui les manipulent. Autrement dit les outils peuvent créer des besoins et susciter de nouvelles approches chez ceux qui les utilisent. C'est pourquoi les aspects ergonomiques sont tout à fait centraux. Plus les interfaces seront commodes à manipuler et plus les acteurs pourront se concentrer sur la qualité et la richesse des structures mises en place. Le développeur doit donc proposer des solutions très proches des données qui doivent en fait guider les opérations. Les langages à balises se prêtent particulièrement bien à ce type d'approche orientée données. Toute la difficulté réside dans les solutions mises en œuvre pour accéder aux données : directement aux "sources" XML, à l'arbre XML, à une vue transformée, à une mise en forme linéaire, à une combinaison de tout ou partie de ces possibilités ? Comme nous le verrons plus loin, il existe aujourd'hui des solutions qui permettent, sans modifier l'organisation des données, de laisser une certaine marge de manœuvre aux utilisateurs dans ce domaine, tout en guidant leur travail.

Chaque acteur apporte ses compétences au service d'un même projet et sa réussite profite à chaque acteur. Bien entendu, il est indispensable de fournir un environnement technique stable pour permettre à ces interactions de se dérouler dans les meilleures conditions. Plus cet environnement est solide et plus il favorise la circulation d'informations univoques, d'où l'intérêt de privilégier des approches orientées données, maîtrisées par les utilisateurs.

1.4.2 Importance des modèles de domaine

Les communautés composées de professionnels partagent une culture métier qui regroupe aussi bien des pratiques que des points de vue sur leurs activités ainsi que sur les objets qu'ils manipulent dans le cadre de leur activité professionnelle. Ainsi, cette dimension culturelle propre à un domaine d'activité est une caractéristique capitale qu'il s'agit de formaliser dans un premier temps, puis de rendre opérationnelle dans un second temps.

Les modèles de représentation abstraits ont justement pour fonction de proposer des formalisations des objets, des activités et des acteurs d'un domaine. Ainsi les *Functional requirements for bibliographic records* [IFLA, 1999] constituent une formalisation des objets manipulés dans les bibliothèques. L'objectif étant bien entendu d'améliorer la qualité du service rendu en regard des missions de ces institutions.

Une fois un modèle élaboré, il s'agit de le rendre opérationnel. C'est le rôle que prennent les standards et les normes qui permettent d'exploiter les catégories proposées par le modèle en les intégrant aux activités d'un domaine. D'une certaine manière, il s'agit en définitive de considérer un standard comme l'expression opérationnelle d'un modèle abstrait.

Bien entendu la séparation proposée ici n'est pas aussi tranchée et il est sans doute plus exact de se représenter les choses à la manière d'un cycle empirique : la fabrication de l'expression opérationnelle du modèle abstrait nourrissant, améliorant ou modifiant ce dernier.

Le rôle des standards et des normes est central dans la qualité des interactions entre les différents acteurs. Leurs dimensions à la fois théorique, proche des cultures métier (car souvent produites par les professionnels), et technique (par leurs implémentations) font des standards et des normes un dispositif très efficace pour servir de socle à l'ensemble des échanges. Les outils fondés sur ces vocabulaires de description ne peuvent que favoriser les échanges tout en améliorant leur qualité.

La figure 1.2 propose une représentation de la circulation des données et des différents formats potentiellement manipulés. Il s'agit ici de montrer comment les standards et les normes, le plus souvent liés à des modèles abstraits, se placent au centre du système de circulation des données. Une fois les données encodées dans un format XML respectant les standards des domaines, il est possible de les transformer à moindre coût (en termes de temps) pour les adapter aux vocabulaires de référence de chaque communauté concernée. Ainsi, si le standard permet la description intégrale des ressources dématérialisées dont les formes sont des extractions rema-

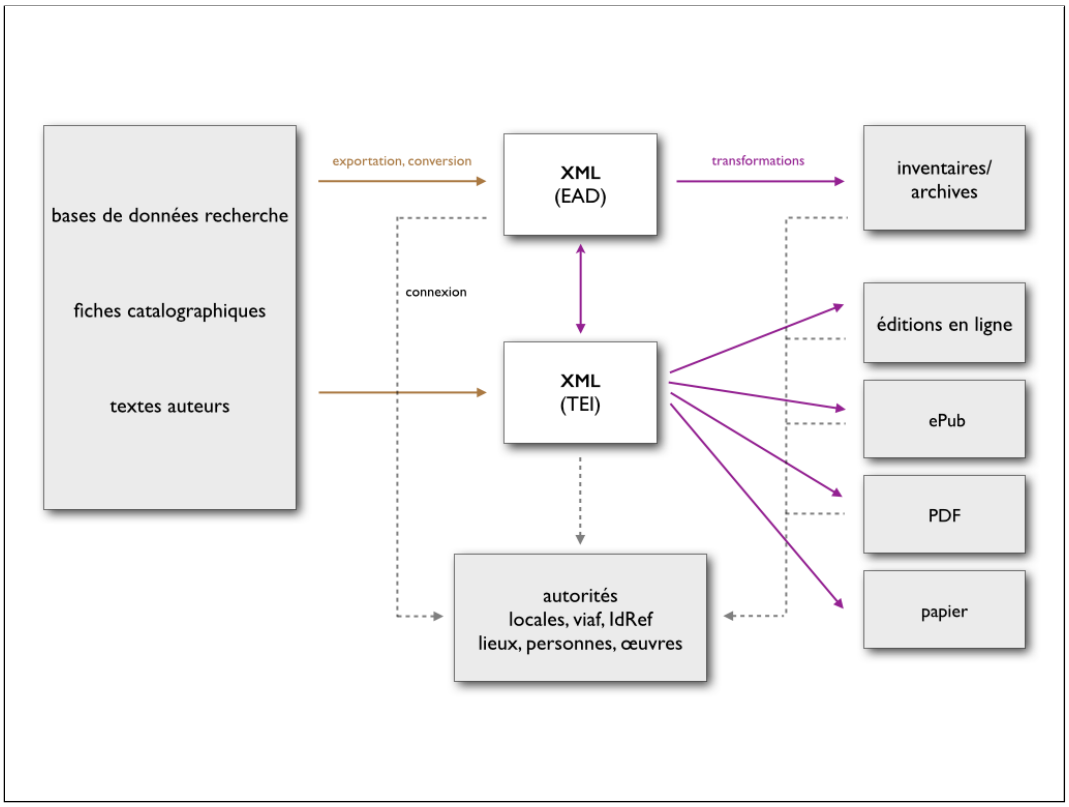

FIGURE 1.2 – Organisation et circulation technique des informations.

térialisées, des spécialistes peuvent s'accorder sur les équivalences qui permettent la construction d'outils de conversion souples, adaptables et évolutifs. Enfin, tout au long du processus, des liens avec des bases d'autorités (prosopographiques ou autres) peuvent être réalisés. De cette manière, une fois le lien établi avec une ressource centrale, il ne sera jamais perdu même au cours des différentes conversions, à condition bien entendu, que les outils de transformation en tiennent compte pendant les opérations.

Le nombre d'acteurs, les standards utilisés, les normes en vigueur et leurs interactions montrent à quel point le monde des humanités est en complète évolution. Si les bases techniques sont relativement stables et propices, par leurs caractéristiques, au développement de collaborations efficaces et pérennes, il reste tout à fait indispensable de présenter et de comprendre le contexte général de la recherche du point de vue des pratiques, mais aussi dans sa dimension institutionnelle, avec en particulier le développement de ce qu'il est aujourd'hui commun de désigner par l'appellation *humanités numériques*.

2

De la recherche en SHS aux humanités numériques

2.1 Introduction

Le contexte actuel est, comme nous venons de le voir, tout à fait favorable au développement de ce qu'il est aujourd'hui convenu d'appeler les *humanités numériques*. Cependant le développement du numérique ne se limite pas au seul domaine de la recherche et concerne également, comme nous l'avons déjà évoqué précédemment, un autre domaine qui nous concerne ici directement, celui de l'édition matérielle. Revenons ici sur la manière dont ces deux domaines, la recherche et l'édition, ou autrement dit, la recherche et la diffusion de ses résultats, intègrent le numérique dans leurs pratiques et leurs organisations.

2.2 Développement des humanités numériques

La prise en compte de l'outillage numérique des sciences humaines et sociales est maintenant claire au niveau national et international. En effet, un nombre toujours croissant de projets de recherche dans le domaine des humanités intègre une forte dimension numérique qui peut parfois aller jusqu'à la conception d'outils novateurs. Pensons par exemple au projet TXM [4], qui propose un logiciel de textométrie, c'est-à-dire un ensemble d'outils fondés sur les acquis des méthodes statistiques appliquées à l'étude textuelle ou encore au projet ProDescartes [5] qui propose un moteur de recherche exploitant les technologies du web sémantique pour l'exploration du corpus

4. http://textometrie.ens-lyon.fr/?lang=fr
5. http://www.unicaen.fr/recherche/mrsh/document_numerique/projet/anrprodescartes

cartésien. De ce point de vue, la transition numérique n'impacte pas seulement les humanités, mais remet en question les frontières d'usage de l'utilisation, de l'appropriation et de l'innovation, les projets dans le domaine des humanités pouvant se trouver à l'origine d'innovations technologiques. Le numérique est porteur et a un impact très fort et, en conséquence, le nombre de projets avec une dimension numérique est de plus en plus important. Il est donc tout à fait capital dans le cadre de cette étude de prendre en compte cet aspect.

L'essor des humanités numériques constitue une évolution marquante [6], même s'il n'est pas forcément aisé de saisir sa nature. En effet, s'agit-il d'un domaine nouveau, d'une pratique ou d'un ensemble de pratiques ou encore d'une étape dans le développement des sciences humaines et sociales ? Peut-on raisonnablement penser que les humanités sont susceptibles de se tenir à l'écart des évolutions numériques ? Autrement dit, le numérique serait-il en train de devenir un dogme auquel les humanités seraient sommées de se conformer ?

Si l'observation directe permet d'identifier relativement facilement des projets comme relevant des humanités numériques, car ils associent techniques informatiques et sciences humaines, en revanche il reste difficile de définir de façon théorique les humanités numériques. Nous proposons un élément de réponse à cette question récurrente en nous focalisant sur les enjeux de diffusion. En effet, si les humanités numériques mettent en jeu la fabrication d'outils et de ressources c'est dans le but principal de les diffuser :

> Il s'agit [...] de partage, c'est-à-dire de communication, de réflexion méthodologique collective et de mise en commun : ni de la théorie pure, ni de la pratique pure, mais un dialogue au sujet de nos représentations du savoir [Berra, 2012].

Nous avançons donc que les humanités numériques sont à considérer comme une démarche qui se propose d'outiller informatiquement la recherche en sciences humaines et sociales et d'en diffuser les résultats en prenant soin, dans un mouvement de distanciation, d'examiner comment les outils mis en place et développés sont susceptibles d'influencer le travail de recherche lui-même.

Par ailleurs, l'évolution des techniques n'est pas neutre dans l'explosion des humanités numériques. Si les projets mêlant humanités et informatique existent depuis

6. Marin Dacos et Pierre Mounier n'hésitent d'ailleurs pas à parler de révolution dans la mesure où ce mouvement risque fort, de leur point de vue, de « redéfinir l'ensemble des champs de la recherche en sciences humaines et sociales » [Dacos et Mounier, 2014].

longtemps, en particulier quand il s'agit de constitution de ressources[7], force est de constater que leur progression s'est considérablement accentuée ces dernières années. En effet, les techniques sont aujourd'hui beaucoup plus souples et plus propices à l'expérimentation. Dans ces conditions il est infiniment plus facile de tester des solutions.

L'utilisation de l'informatique dans le domaine la recherche en sciences humaines implique de numériser, c'est-à-dire d'encoder, les objets d'étude et en particulier, mais pas uniquement, les textes. Mais pour encoder de manière rationnelle ces objets il faut disposer d'un référentiel commun qui va permettre d'unifier les modes de désignation des objets manipulés. Un tel référentiel devrait permettre de désigner les chapitres, les sections, les paragraphes, etc., mais aussi les titres de ces chapitres, sections et paragraphes, sous une forme sur laquelle tout le monde pourra s'entendre. En d'autres termes, il faut disposer d'un modèle abstrait [Burnard, 2012] auquel une communauté pourra se référer de manière univoque.

Bien entendu, ce type de modèles abstraits existe en sciences humaines, même si ceux-ci ne sont pas toujours désignés de cette manière et apparaissent plutôt sous la forme d'hypothèses ou d'ensembles conceptuels.

Quand il s'agissait, il y a quelques années, de rendre opérationnels ces modèles conceptuels dans des applications informatiques, il était indispensable de faire appel à des spécialistes qui se chargeaient, une fois comprise la complexité du modèle, de construire un schéma de représentation des données [Teorey, 1990]. Le plus souvent, ce schéma servait à bâtir une base de données. Ces opérations mettaient en jeu des spécialistes de disciplines éloignées et le dialogue pouvait être particulièrement délicat. Autrement dit, les technologies étaient tellement complexes à mettre en œuvre qu'il fallait absolument faire appel à des spécialistes, le chercheur en sciences humaines et sociales se retrouvant au final assez loin de la mise en œuvre car ne disposant pas d'une connaissance assez fine des technologies mobilisées. En définitive les outils pesaient de manière considérable sur la mise en place des projets et des ressources.

L'arrivée du web marque le tournant véritable de cette évolution technologique :

> En 1994 [...] le Web est arrivé, entraînant un développement exponentiel des capacités informatiques et générant une explosion du nombre de bibliothèques numériques et de projets de numérisation en masse. À tel point que l'informatique compliquée des années 1980 devient, dans les années 1990, un outil pratique et commode [Burnard, 2012].

7. Le *Thesaurus Linguae Graecae* est par exemple né au tout début des années 70. http://www.tlg.uci.edu/about/

L'évolution technique permet en fait de réduire l'écart entre les spécialistes des technologies d'un côté, et les sciences humaines de l'autre. Accéder à un niveau de compréhension et de maîtrise des techniques suffisant pour échanger avec des spécialistes de l'informatique est devenu possible pour des non-spécialistes. On pense en particulier aux langages à balises et aux outils associés qui permettent, comme nous le verrons, d'étiqueter les textes en référence à un modèle dans des conditions aujourd'hui tout à fait accessibles.

Mais le mouvement existe également dans le sens de l'appropriation des outils par les utilisateurs. Cette *littératie numérique*, c'est-à-dire

> [l']aptitude à comprendre et à utiliser l'information écrite dans la vie courante, à la maison, au travail et dans la collectivité en vue d'atteindre des buts personnels et d'étendre ses connaissances et ses capacités [OCDE, 2000].

développée dans le cadre de la convergence numérique, participe donc d'un cercle vertueux dans lequel l'appropriation des outils par les utilisateurs répond et accompagne la modernisation des techniques.

Si ces évolutions peuvent paraître assez naturelles et relativement peu impactantes, c'est en réalité un tournant très fort qui peut, de notre point de vue, expliquer en partie l'essor des humanités numériques. En effet, la mise en place de solutions opérationnelles en référence à des modèles abstraits devenant plus simple et plus facile, l'erreur et l'hésitation deviennent de moins en moins bloquantes ou problématiques. Ainsi, il est tout à fait possible de mettre rapidement à l'épreuve une instanciation d'un modèle abstrait ou encore de procéder par étapes successives de validation. Commencer par exemple par travailler avec de grandes catégories de textes pour aller progressivement dans le détail en testant systématiquement les résultats obtenus à chaque palier.

Nous devons ici définir la notion de *laboratoire de texte*, car en définitive, l'un des grands apports des évolutions technologiques récentes aux sciences humaines et sociales, c'est la possibilité de l'expérimentation, au sens strict de la soumission à l'expérience, sous la forme de manipulations comme on en rencontre depuis très longtemps dans les sciences dures comme la chimie par exemple. Il est aujourd'hui possible de construire des outils [Quint *et al.*, 2010], [Vertan et Reimers, 2012] permettant de mettre rapidement à l'épreuve des faits des modèles de représentation abstraits. Ainsi, les chercheurs peuvent contrôler l'efficacité et la pertinence du modèle tout en le construisant et en le modifiant grâce à la souplesse des technolo-

gies mises en œuvre aujourd'hui et organisées au sein d'un véritable environnement textuel d'expérimentation.

Dans le domaine de l'édition de textes, le projet ANR ProDescartes est ici un bon exemple d'environnement capable de préserver ce qui, dans les données manipulées, doit être pérenne et d'accueillir les évolutions rendues nécessaires par l'avancée des travaux. Il s'agit de publier un corpus constitué des œuvres et de la correspondance de René Descartes avec un ensemble d'annotations scientifiques et d'outils d'aide à la lecture. Pour élaborer le modèle de représentation des textes regroupant l'ensemble des phénomènes textuels à discriminer, un groupe représentatif a été sélectionné dans le corpus pour servir de base de test. Ensuite les catégories de textes ont été choisies en fonction des objectifs scientifiques et éditoriaux, puis utilisées sur l'ensemble des tests. Chaque fois qu'un manque s'est fait jour, les catégories nécessaires ont été ajoutées et une nouvelle phase d'expérimentation a été lancée. Bien entendu, chaque phase supplémentaire profitait des avancées des précédentes et il n'était pas question de reprendre le travail d'annotation de zéro. Au terme de ce processus d'expérimentation un schéma XML TEI de référence incluant les catégories de texte pertinentes pour l'ensemble du projet a pu être mis en place, accompagné d'interfaces de transcription et de structuration ergonomiques intégrant les règles établies pendant les phases de tests. Les technologies permettent ainsi de faire évoluer les modèles manipulés en fonction des cas réels rencontrés sans avoir besoin de refondre l'ensemble du dispositif. Cette contrainte doit aussi être respectée pendant la phase de production, même si un ajustement en fin de processus n'est jamais souhaitable, car il entraîne nécessairement un travail d'adaptation de l'ensemble des outils de saisie et de lecture. Pour autant, il reste un coût d'expérimentation qu'il ne faut pas négliger et l'effort de modélisation à réaliser par les spécialistes du domaine reste une étape tout à fait centrale. Simplement, une erreur de modélisation ne provoque sans doute plus autant de retard que par le passé quand il était nécessaire de construire le modèle avant toute expérimentation de grande envergure.

Le domaine de l'étude matérielle des sources primaires fait aussi l'objet de nombreuses expérimentations. L'utilisation des technologies du web sémantique est ainsi testée pour une application des grands modèles du domaine à la recherche codicologique[8]. Ainsi Robert Kummer [Kummer, 2010] montre comment le cadre conceptuel de référence CIDOC peut être exploité pour constituer des corpus de descriptions de manuscrits hautement structurées capable de répondre à des requêtes très précises.

8. La codicologie est l'étude des manuscrits dans leur dimension matérielle.

La figure 2.1 donne sous forme graphique sa proposition de modélisation d'un manuscrit en utilisant le CIDOC-CRM. Cette représentation ne se focalise pas sur une organisation hiérarchique, mais s'attache plutôt aux relations existantes entre deux individus particuliers : le document textuel (`E31.Document`) porté (`crm:P128.carries`) par un objet manufacturé (`E22.Man-Made_Object`). Chacun de ces individus étant identifiés (`crm:P1.is_identified_by`) par des chaînes de caractères.

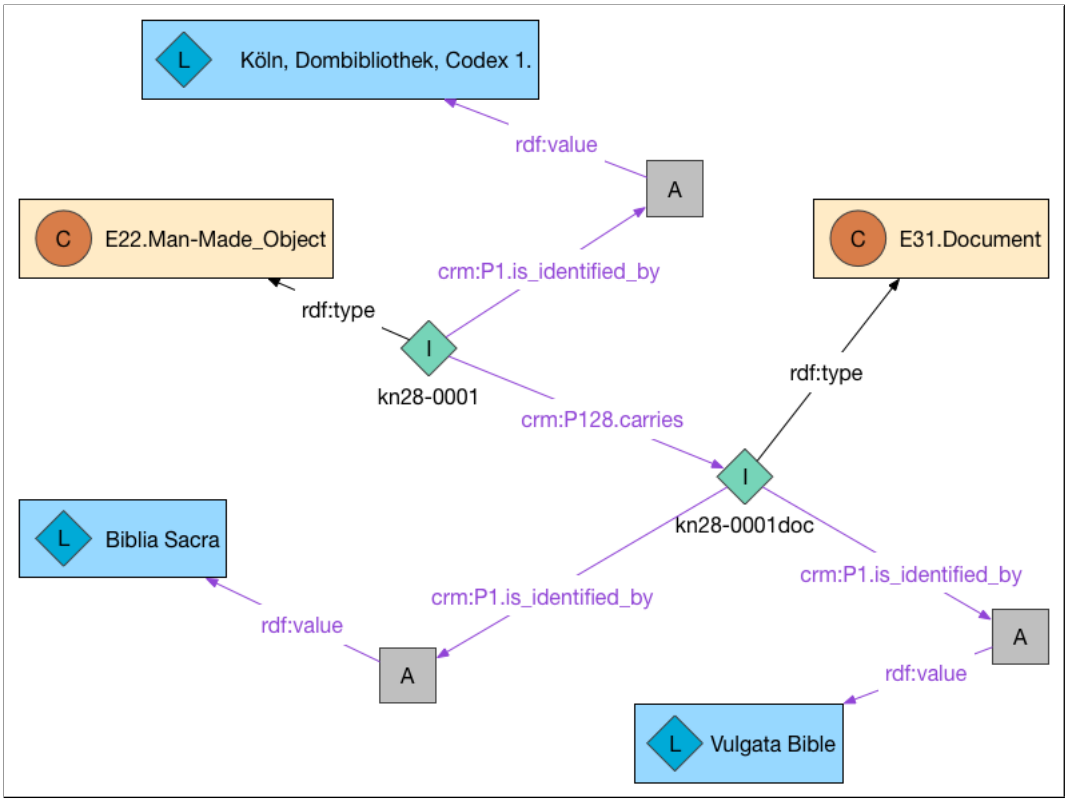

FIGURE 2.1 – Graphe des informations d'un manuscrit (d'après R. KUMMER).

Le passage d'un modèle abstrait à un outil informatique opérationnel impose une étape de numérisation des objets manipulés et c'est bien là qu'un problème se pose.

> Certains diront qu'on peut se servir d'un ordinateur pour faire du *data processing*, pour gérer des données : des chiffres, des faits, des observations, des objets, des tendances statistiques. Le texte, en revanche, est composé d'autres choses : de mots, d'une langue, de paroles, qui ont une existence tout à fait indépendante de leur représentation dans un format numérique. Les données numériques, elles, n'existent que dans leur expression informatisée. C'est précisément là que réside la tension entre les deux. Pour réussir à la lever, certains informaticiens proposent de traiter le texte comme s'il s'agissait d'une donnée [BURNARD, 2012].

Le texte présente quelques résistances à la numérisation. Il est impossible de se contenter de traiter le texte comme de la donnée brute sans en perdre une dimension capitale. Il s'agit donc de proposer des solutions pour permettre aux spécialistes des textes de contrôler le travail de numérisation en ajoutant cette couche d'information à la donnée brute, qui fait partie intégrante de l'objet texte, pour permettre par la suite des exploitations qui tirent partie de toutes les dimensions du texte.

Nous trouvons une notion similaire dans les travaux du groupe de recherche RTPDoc [Pédauque, 2003] pour qui un document numérique, considéré dans sa dimension de forme, c'est de *la structure + des données* [9]. La structure est précisément ce qu'il s'agit d'ouvrir aux chercheurs en sciences humaines dans les meilleures conditions possibles.

En d'autres termes, le chercheur en sciences humaines et sociales va pouvoir appliquer un modèle abstrait et l'expérimenter sur ses données pour rendre compte de toutes les dimensions pertinentes du texte étudié, faire une analyse critique et une évaluation des résultats obtenus et, le cas échéant, retourner au modèle pour le modifier ou l'amender. Pour autant, même s'il existe aujourd'hui des solutions techniques relativement simples à mettre œuvre, il ne faut pas non plus penser que l'historien ou le spécialiste de lettres classiques pourra tout réaliser seul ; il devra toujours collaborer avec des informaticiens pour construire les applications et bâtir les ressources numériques richement structurées. Les évolutions technologiques permettent de gagner du temps collectivement, mais ne font pas disparaître certains acteurs de la scène.

Une autre approche doit ici être mentionnée. Celle qui consiste à considérer que le numérique va permettre de découvrir de nouveaux éléments de recherche et d'observer des phénomènes invisibles sans cet équipement numérique. Le *Big Data*, que nous concevons ici comme le mouvement de numérisation en masse et à gros grains [10] d'ensembles importants de manuscrits ou d'imprimés anciens, est un élément important de ce point de vue. Pour les tenants d'une approche statistique des *Big Data*, il s'agit de considérer que la numérisation à grande échelle va permettre de réaliser des requêtes à un niveau jamais atteint et d'observer les données avec un point de vue beaucoup plus distancié apportant la possibilité de découvertes insoupçonnées. C'est la masse d'informations par elle-même et par le jeu des corrélations qui doit,

9. Nous reviendrons plus loin sur les définitions proposées par le groupe RTPDoc. Voir p. 142 et suivantes.

10. C'est-à-dire avec, dans le meilleur des cas, une couche de reconnaissance de caractères, quand elle est possible, permettant simplement une exploration rapide des textes.

en fait, apporter de nouvelles possibilités d'interprétation. Bien entendu, il n'est pas question ici de critiquer cette approche, mais simplement de préciser que ce n'est pas celle que nous choisissons dans cette étude. En effet, nous croyons que l'analyse fine d'un corpus restreint, telle qu'elle est mise en œuvre dans l'édition de sources anciennes, rend également justice aux caractères spécifiques du texte, de l'expression écrite et à la manière dont celle-ci porte l'histoire, les conceptions et les principes qui ont été mis en œuvre dans l'écriture première comme dans les choix de transmission et de recopie.

2.2.1 Pratiques et usages

L'activité autour des humanités numériques est foisonnante. On ne compte plus les manifestations ou dispositifs techniques qui leurs sont consacrés : les tables rondes, les séminaires, les listes de discussion, soit très généralistes, soit liées à un standard, à une norme, ou encore à un outil, etc. L'activité, allant de la production de ressources à la production de discours, est devenue tellement effrénée qu'il est parfois difficile de la suivre...

L'appellation *humanités numériques* désigne peut-être finalement simplement le fait d'outiller la recherche dans le domaine des sciences humaines et sociales, ce qui se déroule souvent dans le cadre de collaborations entre informaticiens et humanistes. En définitive les humanités numériques, de plus en plus souvent considérées comme une discipline à part entière, désigne en réalité la zone floue de rencontre entre l'informatique et les sciences humaines et sociales. C'est souvent dans ces zones à l'intersection entre deux disciplines que se situent la création et l'enrichissement réciproque. Pour l'humaniste, il s'agit le plus souvent d'intégrer dans sa recherche de nouvelles méthodes et de nouveaux outils. Pour l'informaticien, il peut s'agir par exemple de valider des modèles ou de constituer des corpus contrôlés. C'est enfin, pour l'un comme l'autre, la possibilité de participer à la définition pluridisciplinaire d'objets d'études. Les humanités numériques se caractérisent donc en grande partie par le caractère *pratique* qu'elles ajoutent à la recherche dans le domaine des sciences humaines et sociales.

Le développement des humanités numériques s'accompagne aussi de tentatives d'identification des acteurs qui s'en réclament d'une manière ou d'une autre. Marjorie Burghart [Burghart, 2013] propose ainsi de comprendre la communauté des humanités numériques en détournant un titre de Georges Duby [11] : *ceux qui prient,*

11. Georges Duby, *Les trois ordres ou L'imaginaire du féodalisme*, Paris, Gallimard, 1978.

ceux qui combattent, ceux qui travaillent. Les acteurs de la communauté seraient répartis dans trois catégories perméables les unes aux autres.

Ceux qui travaillent assurent à la réalisation concrète des projets, ce sont eux qui *font* les choses au sens premier du terme. Ce sont souvent des ingénieurs ou des humanistes ayant développé un intérêt fort pour l'informatique.

Ceux qui combattent sont les défenseurs des humanités numériques. Ce sont souvent des militants convaincus qui font tout leur possible pour assurer l'avenir des humanités numériques et de ceux qui les pratiquent en particulier sur le plan institutionnel. Comme le note l'auteur, l'enjeu n'est pas de savoir si nous sommes confrontés à une nouvelle discipline ou pas, mais bien de permettre à ceux qui s'y investissent d'y trouver un intérêt en terme de carrière.

Ceux qui prient développent un discours et une théorie sur les humanités numériques sans les pratiquer directement. Ils sont en quelque sorte des théoriciens qui cherchent à caractériser et à définir les humanités numériques.

Ces catégories proposées (avec humour) par Marjorie Burghart sont très opérationnelles et il est assez tentant de se les approprier en y ajoutant la notion de métier.

Comme nous l'avons dit, beaucoup de métiers différents sont concernés par ces approches, tous avec leurs cultures et leurs compétences. Les acteurs des humanités numériques ne viennent pas de nulle part et c'est justement la pluralité des cultures métier qui apporte une réelle plus-value aux projets. Les aspects numériques développés dans le cadre d'un projet viennent justement comme un trait d'union entre les métiers concernés lorsque tous les acteurs impliqués ont fait l'effort de se former à une culture numérique partagée. C'est, par exemple, grâce à cette "culture numérique" commune que l'historien *qui prie* peut échanger et travailler dans les meilleures conditions possibles avec l'éditeur matériel *qui travaille.*

2.2.2 Organisations et institutions de recherche et de pédagogie

Les humanités numériques sont aujourd'hui intégrées dans les institutions de la recherche en France. En atteste par exemple la Très Grande Infrastructure de Recherche (TGIR) Huma-Num [12] officiellement lancée en mars 2013 et qui se revendique comme "la TGIR des humanités numériques". Cette infrastructure s'organise autour d'une offre de services numériques pérennes et d'un système de concertation collective. Ainsi, elle intègre des *consortiums* labellisés qui apportent des moyens humains en

12. http://www.huma-num.fr

particulier et qui définissent des procédures et des standards partagés pour traiter des thématiques et des objets communs. Par exemple, le consortium interdisciplinaire CAHIER (Corpus d'Auteurs pour les Humanités : Informatisation, Édition, Recherche) traite des corpus fortement liés à une activité d'édition multisupport ou exclusivement électronique pour favoriser l'accès aux données. Le consortium favorise ainsi l'émergence et l'utilisation de standards et d'outils autour de sa thématique scientifique [13].

Au niveau européen, les institutions sont aussi clairement orientées dans le sens de la prise en compte des humanités numériques. DARIAH (*Digital Research Infrastructure for the Arts and Humanities*) [14], se présente comme "l'infrastructure numérique européenne en sciences humaines". L'un de ses objectifs principaux est "de valoriser les réalisations dans le domaine des humanités numériques" [TGIR-Huma-Num, 2013]. Cette infrastructure ne fait pas de choix dans les objets numériques qu'elle traite et s'intéresse donc aussi bien aux textes, qu'aux images, vidéos et sons. Là encore, la notion d'accès facilité aux données est au cœur des préoccupations de DARIAH. Il s'agit bien de faire fonctionner la recherche à l'échelle européenne selon une logique d'échange d'informations en fournissant un réseau technique.

L'articulation entre le niveau national et le niveau européen se fait via la TGIR Huma-Num.

Dans le domaine de l'enseignement également, les humanités numériques font l'objet d'une attention de plus en plus soutenue. De plus en plus de masters proposent des unités d'enseignement directement en rapport avec les techniques numériques. Pensons, par exemple, aux masters « Patrimoine écrit et édition numérique » et « Patrimoine culturel immatériel » du Centre d'Études Supérieures de la Renaissance [15] ou au master « Document spécialité Édition, mémoire des textes » de l'Université de Caen Basse-Normandie [16]. Certains diplômes sont mêmes entièrement dédiés aux humanités numériques comme le master « Technologies numériques appliquées à l'histoire » de l'École nationale des Chartes qui existe depuis 2006 et qui permet aux historiens d'acquérir les bases informatiques [17].

D'autre part, on ne compte plus les colloques, séminaires de travail et formations qui présentent une importante dimension numérique et ce dans beaucoup de

13. http://cahier.hypotheses.org

14. http://www.dariah.eu

15. http://cesr.univ-tours.fr/formations/formations-du-cesr-50697.kjsp

16. http://webetu.unicaen.fr/master-document-specialite-edition-memoire-des-textes-pro-rech-388731.kjsp ?RH=1165594563007

17. http://www.enc.sorbonne.fr/master-technologies-numeriques-appliquees-l-histoire

disciplines des SHS. Il suffit, pour s'en convaincre, de consulter la page d'annonce d'événements dédiée aux humanités numériques du calendrier des lettres et sciences humaines et sociales Calenda [18].

Toujours d'un point de vue organisationnel, mais en marge des institutions, notons aussi que la communauté des humanités numériques se dote d'associations tant au niveau national qu'international. En effet, si l'association *Alliance of Digital Humanities Organisations* existe depuis 2004, l'association *Humanistica* [19] est née en juillet 2014 à Lausanne à l'issue d'une initiative lancée à l'occasion du THATCamp [20] d'octobre 2013.

Sur un plan peut-être plus opérationnel et, en tout cas plus local, il faut aussi noter la naissance de centres ou de pôles dans certaines institutions. Ainsi, la Maison de la Recherche en Sciences Humaines de Caen (MRSH) [21] s'est dotée, dans le cadre de son axe prioritaire sur le document numérique, d'un pôle pluridisciplinaire. Ce pôle organise son activité autour de deux axes. D'une part, il apporte un soutien à la recherche en sciences humaines et sociales en proposant des outils ergonomiques de manipulations de flux de données structurées. D'autre part, il est investi dans des projets de recherche sur le numérique en tant qu'objet d'étude. Ces deux aspects de l'activité du pôle se nourrissent l'un l'autre, lui permettent d'assurer la qualité de service proposée aux chercheurs et de participer à des projets de recherche aux échelles nationale et européenne. Le soutien à la recherche permet de proposer avec toujours plus de précision les outils et les techniques numériques les plus adaptés pour servir les objectifs scientifiques. Au cours de ce travail, de nouveaux objets et de nouveaux usages numériques se font jour. L'activité de recherche prend en charge leurs définitions et leurs caractérisations pour pouvoir les intégrer à leur tour dans de nouveaux dispositifs technologiques qui deviendront par la suite les outils utilisés dans le cadre de l'activité de service. L'objectif est donc bien de mettre en place un lieu d'échange et de circulation entre l'activité d'ingénierie et la recherche scientifique. Précisons qu'ici l'ingénierie et la recherche scientifique ne renvoient pas à l'informatique et aux humanités, elles peuvent concerner les deux domaines. Ainsi, l'ingénierie concerne aussi bien l'analyste de sources anciennes de formation classique que le développeur. De la même manière, la recherche scientifique porte aussi bien sur l'étude de texte que sur la découverte automatique d'entités nommées. Nous

18. http://calenda.org/search.html ?primary=fsubject&fsubject=298

19. http://www.humanisti.ca

20. THAT pour *The Humanities and Technology*. Il s'agit d'un non-colloque car le programme n'est pas prévu à l'avance sur les humanités et l'informatique.

21. http://www.unicaen.fr/recherche/mrsh/

ne sommes donc pas ici dans une logique d'instrumentalisation d'une discipline de recherche par une autre.

Ce type d'organisation du travail correspond en réalité très bien aux humanités numériques, qui relèvent à la fois d'une activité très pratique de production et d'une approche plus théorique et réflexive. Cette dualité renvoie à la question de savoir si c'est l'outil qui crée le concept ou l'usage du chercheur qui fabrique l'outil.

Dans un récent rapport sur l'essor des humanités numériques en France commandé par l'Institut Français [Dacos et Mounier, 2014], la première recommandation des auteurs consiste justement à soutenir ce type de centres dans les établissements de recherche et d'enseignement supérieur.

2.3 Convergence numérique dans l'édition

La convergence numérique désigne le mouvement amorcé depuis bien longtemps dans le sens du tout numérique. Toutes les sphères de l'édition sont concernées par le numérique, de la rédaction à la lecture qui a été la dernière activité à intégrer cette dimension avec les ordinateurs dans un premier temps puis avec les tablettes, les liseuses et les *smartphones*.

D'un point de vue organisationnel, il faut noter que les problématiques, de la production à la consultation en passant par la conservation et le signalement, autour de l'information scientifique, dont l'édition universitaire est un important pourvoyeur, sont aujourd'hui au cœur des politiques publiques avec la création en 2009 de la Bibliothèque Scientifique Numérique (BSN) [22] qui a pour objectif général de veiller :

> [...] à ce que tout enseignant-chercheur, chercheur et étudiant dispose d'une information scientifique pertinente et d'outils les plus performants possibles.
> En accord avec les orientations de la Commission européenne, la BSN privilégie l'accès ouvert aux documents scientifiques sous différentes formes reposant sur des innovations, des négociations avec les éditeurs ou le soutien aux archives ouvertes, en tenant compte des différences entre les disciplines [BSN, 2012].

BSN est organisée en 9 segments thématiques parmi lesquels BSN 7 traite spécifiquement des questions d'édition et de publication. Ce groupe a ainsi publié une *Charte des bonnes pratiques pour l'édition numérique scientifique* [23] qui fixe les cri-

22. http://www.bibliothequescientifiquenumerique.fr
23. http://www.bibliothequescientifiquenumerique.fr/?Charte-des-bonnes-pratiques-pour-l

tères vérifiables (citabilité, interopérabilité, accessibilité, ouverture, durabilité) à respecter pour pouvoir recevoir le label bibliothèque scientifique numérique.

2.3.1 Production éditoriale

La figure 2.2 donne une représentation synthétique de l'organisation éditoriale fondée sur des formats de données structurées. La méthode consiste à produire, à partir des jeux de données d'entrée, des versions pivots respectant des modèles abstraits adaptés au domaine de l'édition matérielle. L'ensemble des formes de diffusion sera ensuite produit à partir de ces versions pivots. Notons que, dans certains cas, les versions remises par les éditeurs scientifiques sont déjà respectueuses de modèles de représentation. Dans ce cas, l'éditeur matériel doit être en mesure de réorganiser les données pour les faire correspondre aux catégories qu'il manipule habituellement dans les opérations de production des formes de diffusion. Ainsi la tendance décrite plus haut impose en quelque sorte aux éditeurs matériels d'être capable de traiter des données très finement structurées par les chercheurs.

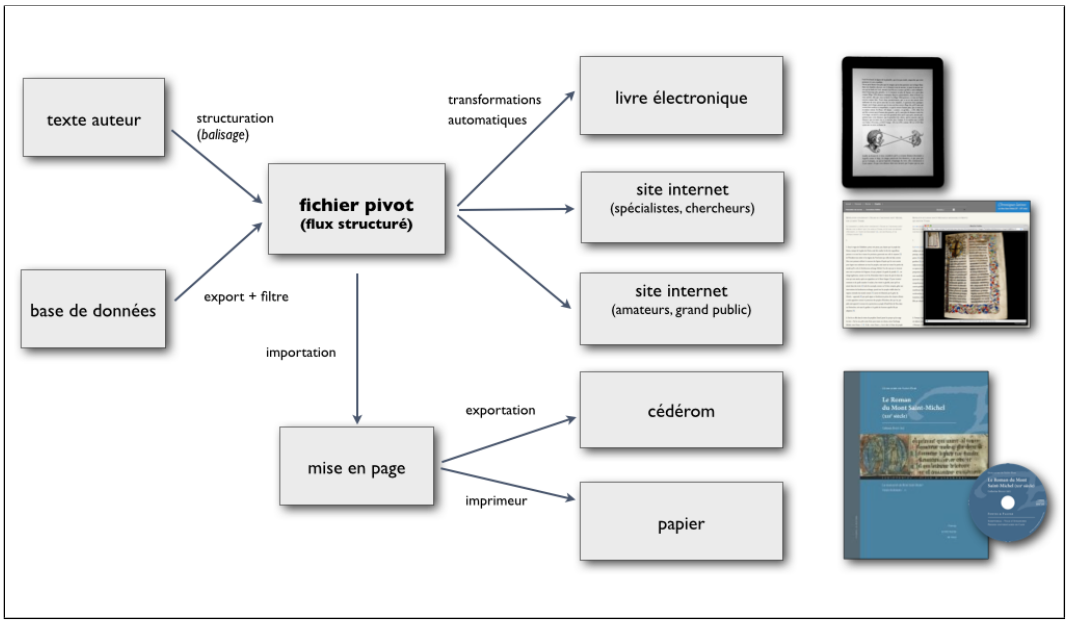

FIGURE 2.2 – Schéma général de production éditoriale.

En effet, dans la mesure où, d'une part, les humanités numériques se fixent comme objectif de constituer et de partager des ressources et que, d'autre part, ces ressources sont le plus souvent produites en exploitant des langages à balises en liens avec des modèles, les auteurs déposent de plus en plus souvent des instances structurées chez les éditeurs, là où ils remettaient auparavant des manuscrits ou des tapuscrits.

Cet effort d'adaptation ne pose pas nécessairement d'énormes problèmes et pour comprendre ce qu'il implique il faut revenir à la notion de modèle. Les modèles de représentation de texte sont bien souvent présents chez les éditeurs mais avec des méthodes d'exploitation parfois peu claires pour ceux qui les manipulent. En effet, le plus souvent, ce modèle s'instancie dans des feuilles de style de paragraphes et de caractères. Si ces feuilles de style sont correctement pensées elles reflètent bien le modèle de représentation textuel de l'éditeur en proposant des catégories de textes génériques et transférables d'un document à un autre, quelle que soit sa mise en forme, comme "titre de chapitre", "titre de section", "paragraphe", "citation", "note de bas de page", etc. Cependant, il est important de comprendre que cet outil ne présente aucune espèce de coercition et peut être utilisé d'une manière totalement libre et totalement orientée par la mise en forme avec des styles locaux difficiles à généraliser comme "Gras, corps 12, 10 points avant et 25 après". Si une certaine rigueur dans l'application de tels styles autorise à poser une correspondance avec des catégories relevant d'un modèle abstrait, l'entreprise reste risquée et relativement hasardeuse.

L'existence d'un modèle de représentation, même local, relativement peu formalisé et existant uniquement sous forme de feuilles de style, permet donc à un éditeur matériel d'aborder plus sereinement l'évolution de sa chaîne de production par l'intégration de données structurées.

De plus, le respect d'un modèle de représentation abstrait est aussi à considérer comme le révélateur d'une bonne intégration des grandes fonctions de l'édition [Schuwer, 1997]. C'est pour cette raison que les éditeurs qui ont conservé une forte culture métier ne rencontrent en général que peu de difficultés à intégrer des formats de données pivot au cœur de leurs méthodes de production. Le rapport d'étude sur l'édition numérique commandé par les éditions QUÆ [Prost, 2007] explicite clairement ce lien entre la sphère "intellectuelle" de l'édition matérielle et ses mises en œuvre techniques.

La figure 2.3 donne deux exemples de feuilles de style qui montrent bien la différence entre les deux approches qui peuvent présider à la conception de ce type d'outils. La première cherche à nommer les types de textes en faisant abstraction de leurs mises en forme locales dans un document donné alors que la seconde se contente de rassembler un certain nombre de propriétés graphiques et typographiques sous une étiquette unique pour simplifier les opérations de mise en forme. La première cherche donc à factoriser des dénominations de types de textes à une échelle plus générale, celle de la collection voire celle du fonds de l'éditeur, tandis que la seconde se limite

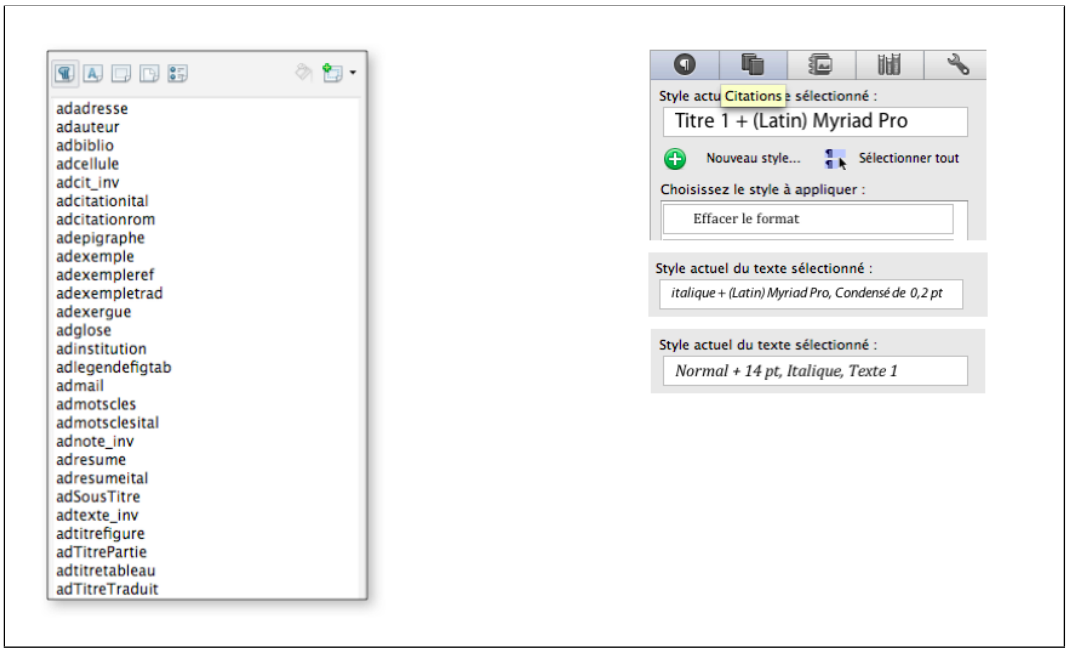

FIGURE 2.3 – Exemples de feuilles de style de logiciels de traitement de textes.

à lister les différentes formes, c'est-à-dire des caractéristiques graphiques et typographiques, à appliquer à un document spécifique. Autrement dit, dans le premier cas il s'agit d'une première ébauche de modèle tandis que la seconde prétend seulement simplifier le travail de systématisation des formes dans le document.

Si, dans un contexte de production d'une forme unique, comme c'était le cas il y a quelques années, ces deux approches sont tout à fait utiles et contribuent, chacune à leur échelle, à faciliter le travail de l'éditeur matériel, la convergence numérique bouleverse totalement la donne. La multiplication des supports de diffusion impose une nécessaire adaptation des pratiques de production éditoriale pour permettre la diffusion des mêmes contenus sous différentes formes avec la même exigence de qualité pour chacune d'elles.

Ainsi, il ne s'agit plus aujourd'hui de produire des textes pour alimenter une forme de diffusion, mais de manipuler des flux d'informations textuelles, complétés par des ressources multimédia comme des images, des vidéos ou des sons, qui sont organisés, croisés pour alimenter différents supports de lecture.

Le trésor de l'éditeur n'est plus son fonds de livres papier, mais bien un réservoir de données qui seront mobilisées en fonction des choix de politique éditoriale. L'éditeur pourra ainsi extraire des séquences de textes pour alimenter une version papier, articulée avec une version numérique qui pourra reprendre une partie du volume papier en la complétant d'images et de vidéos.

2.3.2 Diffusion

Le fonds de l'éditeur devient donc un réservoir de données numériques de plus en souvent structurées. En conséquence, les éditeurs mettent en place des stratégies de diffusion adaptées à ce nouveau paradigme. Les modèles économiques suivants sont donc à considérer comme des retombées du modèle technique d'organisation des flux de travail autour des ressources numériques construites par les différents acteurs.

Ainsi, le Centre pour l'édition électronique ouverte (CLEO) a adopté depuis 2011 le modèle économique *freemium*. Ce modèle consiste à proposer la diffusion gratuite du format HTML et payante des formats détachables PDF et ePub. Du point de vue commercial, l'idée est de faire financer la totalité de la plateforme par les utilisateurs prêts à payer pour consulter les formats détachables [24].

L'unité éditoriale de base est l'article ou le chapitre, regroupés pour constituer des numéros ou des ouvrages. Ces unités de base sont structurées en XML en suivant les recommandations de la TEI et leur organisation en numéros ou en livres est réalisée en utilisant le *Metadata Encoding and Transmission Standard* (METS) [METS Editorial Board, 2010]. C'est la rentabilité de ce processus qui rend possible l'approche économique du CLEO.

Les Presses universitaires de Caen proposent un modèle articulant version papier et version électronique pour leur collection de sources. Dans cette organisation, les versions papier sont payantes et les versions en ligne sont gratuites. Les commentaires scientifiques constituent la valeur ajoutée du papier, mais les textes des sources éditées constituent un corpus consultable et interrogeable en tant que tel gratuitement en ligne.

Cette organisation est rendue possible par la mise en place d'un fonds d'éditeur composé d'unités éditoriales atomiques qui peuvent être mobilisées et réagencées en fonction de la politique éditoriale.

Cette construction de fonds d'édition permet d'organiser la consultation en ligne des travaux scientifiques dans un cadre précis, comme dans le cas des sources du Mont Saint-Michel, mais aussi selon d'autres critères d'interrogation et à l'échelle souhaitée. Tout le fonds de l'éditeur devient une base interrogeable par thème par exemple. En réalité, tous les éléments structurés dans les textes sont susceptibles de devenir des critères d'extraction ou d'interrogation.

Sans entrer dans les détails sur lesquels nous reviendrons dans la partie IV, c'est ce type d'exploitation qui est mis en place dans l'édition en ligne de l'*Hortus Sanitatis*

24. Pour des précisions sur ce modèle commercial : http://www.openedition.org/8873

avec le *répertoire des citations*[25] qui permet de consulter le texte en fonction de la provenance des fragments à l'échelle de l'œuvre.

2.4 Bilan et postulat de travail

Le fait que la recherche s'organise de plus en plus autour de projets intégrant une dimension numérique forte contraint à penser les projets en ouverture. L'époque de la fabrication de bases de données internes au laboratoire sans ouverture sur l'extérieur autre que selon des modalités choisies par les concepteurs est révolue. Aujourd'hui, de plus de plus fréquemment, tout est pensé dans le cadre d'une ouverture maximale. Les fichiers produits sont souvent soumis à l'exercice critique de la communauté pour le bien de l'avancée de la connaissance scientifique.

Dans le cadre de cette étude nous allons donc privilégier une approche centrée sur les chercheurs et sur les modes de production des données. Autrement dit, nous postulons que, plutôt que de chercher *a posteriori* à structurer les informations en utilisant, par exemple, des solutions de fouille de données et de recherche de connaissances basées sur des systèmes probabilistes, nous obtiendrons de très bons résultats en proposant aux producteurs de données des solutions souples de structuration et d'annotation pour garantir : la qualité des annotations produites (la validation est réalisée au fur et à mesure, ce qui supprime ou diminue fortement le(s) problème(s) d'évaluation) ; les possibilités d'exploitation des données annotées ; enfin, l'utilisation de systèmes de fouille sera d'autant plus efficace que les données seront déjà riches. Il sera donc possible d'envisager l'utilisation des solutions de fouilles ou de traitement automatique soit pour faciliter le travail d'annotation des chercheurs en SHS soit dans un second temps pour enrichir les données déjà annotées par les spécialistes.

Beaucoup de compétences différentes, donc beaucoup d'acteurs différents (ceux que nous avons mentionnés qui sont directement concernés, mais aussi des chercheurs d'autres disciplines et de l'informatique en particulier) sont appelés à intervenir sur les données à un moment ou à un autre. Beaucoup travaillent avec les mêmes technologies suffisamment expressives pour représenter leurs données dans toutes leurs dimensions. Il s'agit de proposer une solution qui assure un continuum pour permettre aux projets de profiter des compétences de tous sans avoir de manipulation lourde à réaliser et en préservant la qualité du travail de chacun.

Le modèle doit donc permettre de conserver toutes les dimensions des données et proposer des solutions à chaque moment du traitement pour convoquer les méthodes

25. http://www.unicaen.fr/puc/sources/depiscibus/citations

et les outils les plus appropriés pour la constitution, la structuration, l'annotation et l'exploitation des corpus.

Deuxième partie

Métiers et pratiques

Introduction

Nous proposons de modéliser l'ensemble des activités et des objets qui sont liés à l'édition de sources anciennes, il est donc indispensable d'en faire un rapide inventaire avant de proposer un cadre conceptuel. L'ensemble de ces notions interviendra dans nos réflexions mais il ne s'agit pas pour autant obligatoirement d'objets qui seront manipulés en tant que tels.

Nous étudions donc dans cette partie les repères fondamentaux dans le cadre d'un projet d'édition de sources anciennes.

Nous donnons pour commencer quelques éléments conceptuels et méthodologiques ainsi que des définitions. Il ne s'agit pas d'une étude détaillée de chaque domaine qui pourrait chacun donner matière à travail indépendant, mais de fournir les éléments indispensables à la compréhension de notre travail.

Nous suivons ensuite le cas idéal typique du cycle de production d'une édition de sources dans lequel le travail débute par le repérage des manuscrits et la production d'un inventaire suivi par les opérations d'études scientifiques puis par la production des formes de diffusion. Nous présentons donc chacun de ces acteurs et nous examinons les aspects de leurs pratiques qui nous concernent dans le cadre de cette étude.

Rappelons enfin que nous expérimentons pour répondre à des exigences de production. La méthode s'apparente donc à un mouvement cyclique articulant expérimentation, modélisation, évaluation et critique, expérimentation, etc. Chaque production fait donc partie intégrante de l'expérimentation et apporte une brique supplémentaire à la modélisation présentée. L'évaluation du modèle se fait donc de manière presque synchrone. Pendant la production et juste après nous assurons l'analyse critique pour faire évoluer le modèle et la pratique associée si nécessaire. En conséquence, pour illustrer notre propos, nous prendrons beaucoup d'exemples tirés des modèles que nous avons mis en œuvre. Certains de ces exemples font l'objet d'une présentation détaillée dans la partie IV.

3

Repères techniques

3.1 Sources anciennes et structures de textes [26]

Dans le travail d'édition, la notion de structuration de textes prend une dimension particulièrement complexe avec les sources anciennes. Le matériau de base des spécialistes de langues anciennes est composé de manuscrits qui constituent leurs sources primaires. Les manuscrits font l'objet de plusieurs types d'études : paléographique (qui se concentre sur l'étude des écritures), codicologique (qui considère le manuscrit dans sa matérialité), historique, etc. Ces manuscrits sont les porteurs des textes étudiés par les chercheurs, ils en sont les supports.

Pour beaucoup de spécialistes, le texte ne peut être correctement entendu qu'en tenant compte du support matériel sur lequel il est inscrit [Driscoll, 2010]. Les textes sont par ailleurs très souvent lisibles dans plusieurs manuscrits. Chacun des manuscrits portant un même texte constitue un témoin de ce dernier. Les méthodes de classification des manuscrits permettent de faire apparaître les variantes proposées par chaque témoin... sachant que l'original peut avoir lui-même disparu. Les témoins manuscrits ne sont bien évidemment pas tous de la même époque et les textes qu'ils portent sont bien souvent le résultat d'un processus plus ou moins complexe de copie. Ainsi, le texte d'un manuscrit A peut avoir été copié dans un manuscrit B (avec éventuellement quelques erreurs ou autres modifications).

Ainsi les textes racontant l'histoire du Mont Saint-Michel peuvent se lire dans plusieurs manuscrits qui constituent autant de témoins pour les chercheurs. Pour la production des *Chroniques latines du Mont Saint-Michel* [Bouet et Desbordes, 2009], les auteurs se sont appuyés sur cinq témoins principaux. Ces cinq manuscrits per-

26. Les exemples que nous mobilisons ici font partie intégrante de notre étude et des expérimentations auxquelles nous avons participé.

mettent de lire les textes, parfois avec des différences dont les chercheurs vont rendre compte, comme nous le verrons un peu plus loin. La figure 3.1 propose une vue de la répartition des textes fondateurs du Mont Saint-Michel dans les manuscrits conservés au fonds ancien de la bibliothèque municipale d'Avranches. Notons qu'aucun manuscrit ne contient l'intégralité des textes de la collection. Aucun témoin ne peut donc fournir de clés de compréhension définitive de la collection de textes étudiée. Il est également important de noter que des éléments de mise en page ou de structuration des textes, utiles, ou même indispensables, à un lecteur aujourd'hui, tels que titres, ponctuations, paragraphes, etc. peuvent être absents des témoins médiévaux ou différer d'un témoin à l'autre. De plus, le contenu textuel d'une même œuvre n'est pas identique dans chacun des témoins, il peut varier dans des proportions plus ou moins grandes. Le plus souvent, l'éditeur scientifique, comme le verrons, ne se contente pas de transcrire le texte, mais il doit l'« établir ».

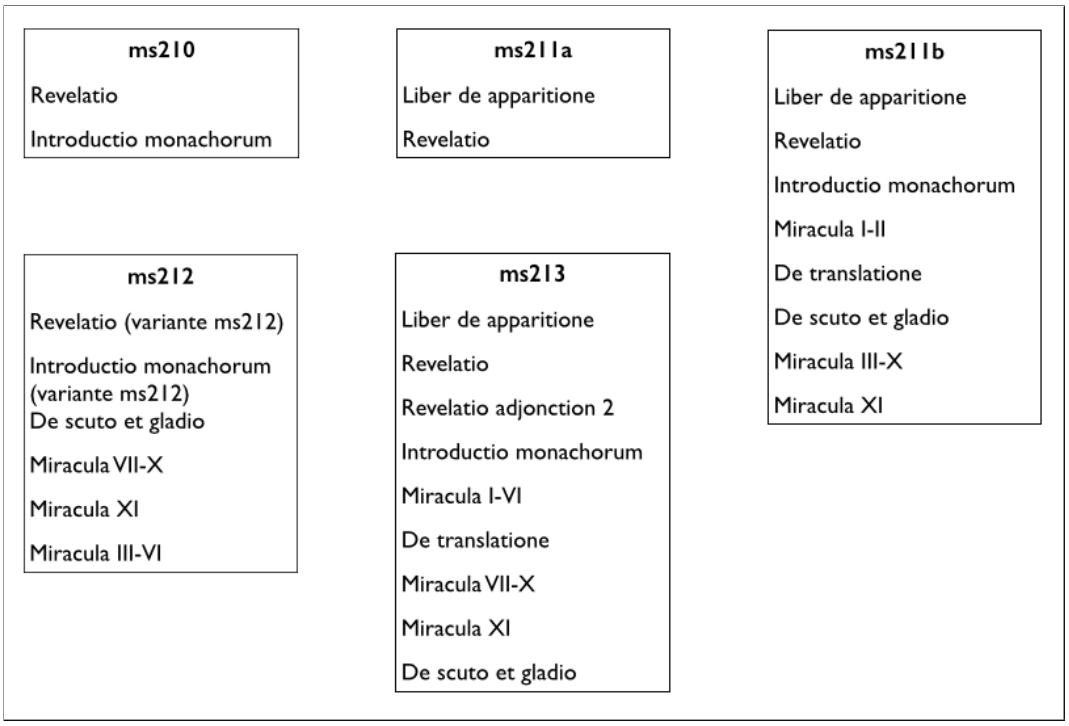

FIGURE 3.1 – Les textes fondateurs du Mont Saint-Michel dans les manuscrits d'Avranches.

Nous définissons donc les *sources primaires* comme le matériau de base utilisé par les chercheurs pour l'étude des textes anciens. Elles nous intéressent particulièrement ici car leur grande complexité en font un excellent sujet d'expérimentation du point de vue de la structuration des différents contenus à fournir au lecteur (textes établis, variantes, écarts, notes explicatives, etc.).

Mais d'autres types de complexité peuvent appraître dans l'édition de sources anciennes. L'exemple de l'*Hortus Sanitatis* [Jacquemard *et al.*, 2013] est ici particulièrement éclairant. En effet, ce traité latin d'ichtyologie de la fin du XV^{e} siècle est le résultat d'une compilation de différentes œuvres. Il s'agit en fait d'un incunable [27] dans lequel le compilateur a réalisé un montage de citations, se "contentant" d'écrire des passages de transition ou de liaison entre les citations. Nous reviendrons sur les rapports entre texte et support matériel du texte, mais nous avons ici un bon exemple de la complexité à laquelle nous aurons à faire. C'est bien le texte de l'*Hortus Sanitatis* qui intéresse ici les chercheurs mais ce texte est en réalité le résultat d'un montage de fragments de textes existants dans plusieurs témoins. Idéalement, pour pouvoir comprendre l'histoire du texte de l'*Hortus Sanitatis*, il faudrait disposer de chacune de ses sources pour pouvoir reconstituer l'ensemble du réseau de relations existant entre l'*Hortus Sanitatis* et ses sources. Ce travail, extrêmement long et difficile, reste un objectif à long terme, mais ne doit pas interdire de commencer à travailler sur le texte de l'*Hortus Sanitatis* tel qu'il est connu aujourd'hui. Les méthodes de travail mises en œuvre doivent donc permettre de construire ce réseau de relations entre le texte étudié et ses sources de manière progressive. Nous reviendrons plus loin sur l'organisation du travail et sur la modélisation des textes proposée pour atteindre cet objectif.

Ces *sources primaires* ou ces *témoins* constituent des objets centraux sur le plan logique que nous devrons placer au cœur de notre démarche. Ils sont la base même du travail des chercheurs humanistes et l'ensemble de notre modèle devra les intégrer.

3.2 Étiquetage des textes

L'étiquetage des textes n'est pas une technique nouvelle. Les documents font depuis longtemps l'objet de beaucoup d'attention en particulier pour en simplifier le traitement et la circulation. La définition d'un système d'étiquetage explicitant les structures internes à divers niveaux (physique, logique ou un subtil mélange entre les deux) est bien entendu une piste très suivie depuis des années. Il s'agit ici d'encapsuler à l'intérieur même du document un système d'étiquetage permettant typiquement, par exemple, de produire une version mise en forme et imprimable.

27. C'est-à-dire un livre imprimé au tout début de l'imprimerie occidentale, entre 1450 et les cinq décennies suivantes, quand les techniques d'impression étaient encore au berceau (*incunabula* en latin).

Le principe général est, comme nous allons le voir avec deux exemples, d'englober des séquences de caractères pour leur affecter une étiquette dont le rôle et la nature seront variables : propriété typographique, graphique, position hiérarchique, etc. ou une combinaison de ces propriétés. Autrement dit, il s'agit ici plus d'identifier des points précis que des zones de textes à proprement parler, même s'il est possible d'associer à une séquence de caractères une fonction de "début d'ensemble", comme un titre de chapitre par exemple.

Il n'est pas ici utile d'entrer dans le détail d'une typologie générale, déjà réalisée par ailleurs [Coombs *et al.*, 1987]. Examinons deux exemples d'outils typiques de la logique d'étiquetage du texte relevant de la catégorie *descriptive* car proposant des mécanismes internes de marquage du texte.

Le premier exemple est particulièrement utilisé dans le monde des sciences "dures" puisqu'il s'agit de LaTeX, qui est à la fois un langage de composition de documents et un programme capable d'interpréter ce langage pour produire des descriptions de pages. Ce langage crée en 1983 par Leslie Lamport est en fait une collection de macro-commandes TeX [28], lui aussi un langage et un système de composition de documents, mais dont les commandes sont particulièrement ardues à maîtriser. LaTeX n'est donc pas un système WYSIWYG [29], mais un langage formel qui permet de décrire le document et l'organisation du texte en donnant une série d'instructions directement encapsulées dans le fichier texte. La présence de ces instructions oriente clairement LaTeX vers une logique de structuration des textes, même si sa logique principale est celle de la mise en forme imprimée (papier ou PDF) et que le vocabulaire proposé est intégré au langage de formatage et, de ce fait, relativement délicat à étendre.

La figure 3.2 montre une séquence de texte et de commandes LaTeX [30]. On y voit très clairement que les séquences de caractères sont parfois englobées entre `\...{`, qui marque le début d'une commande, et `}`, qui en marque la fin, l'espace indiqué par les points de suspension étant occupé par le nom de la commande. Le travail s'effectue donc bien à l'échelle de la séquence de caractères. Ainsi, la commande `\section` permet de marquer le début d'une nouvelle section et d'en donner le titre entre accolades, mais elle ne permet pas d'englober l'ensemble des éléments (textes,

28. Crée en 1977 par Donald Knuth pour pallier le manque d'outils de qualité pour la gestion informatique de la typographie.

29. *What You See Is What You Get.*

30. La coloration syntaxique n'est pas l'œuvre directe de LaTeX, mais du package AUCTeX, https://www.gnu.org/software/auctex/.

```
\DontFrameThisInToc
\chapter[De la recherche en SHS aux humanités numériques]{De la recherche en SHS\\ aux humanités numériques}

\section{Introduction}

\noindent Le contexte actuel est, comme nous venons de le voir, tout à fait favorable au développement de ce qu'il est aujourd'hu
i convenu d'appeler les \emph{humanités numériques}. Cependant le développement du numérique ne se limite pas au seul domaine de
la recherche et concerne également, comme nous l'avons déjà évoqué précédemment, un autre domaine qui nous concerne ici directeme
nt, celui de l'édition matériel. Revenons ici sur la manière dont ces deux domaines, la recherche et l'édition, ou autrement dit
sur la recherche et la diffusion de ses résultats, intègrent le numérique dans leur pratiques leurs organisations.

\section{Développement des humanités numériques}

\noindent La prise en compte de l'outillage numérique des sciences humaines et sociales est maintenant clair au niveau national e
t international. En effet, un nombre toujours croissant de projets de recherche dans le domaine des humanités intègre une forte d
imension numérique qui peut parfois aller jusqu'à la conception d'outils novateurs\footnote{On pense, par exemple, aux projets TX
M, http\string://textometrie.ens-lyon.fr/\string?lang=fr ou ProDescartes, http\string://www.unicaen.fr/recherche/mrsh/document\st
ring_numerique/projet/anrprodescartes, qui proposent pour le premier une plateforme de textométrie et pour le second un moteur de
 recherche sémantique.}. De ce point de vue, la transition numérique n'impacte seulement les humanités, mais remet en question le
s frontières d'usage de l'utilisation, de l'appropriation et de l'innovation. Les projets dans le domaine des humanités pouvant s
e trouver à l'origine d'innovation technologique. Le numérique est porteur et a un impact très fort et en conséquence le nombre n
ombre de projet avec une dimension numérique est de plus en plus important. Il est donc tout à fait capital dans le cadre de cett
e étude de prendre en compte cet aspect.

L'essort des humanités numériques constitue une évolution marquante\footnote{Marin \textsc{Dacos} et Pierre \textsc{Mounier} n'hé
sitent d'ailleurs pas à parler de révolution dans la mesure où ce mouvement risque fort, de leur point de vue, de <<~redéfinir l'
ensemble des champs de la recherche en sciences humaines et sociales~>>~\cite{Dacos:HumanitesNumeriques:2014}.}, même s'il n'est
pas forcément aisé de saisir sa nature. En effet, s'agit-il d'un domaine nouveau ou d'un étape dans le développement des sciences
 humaines et sociales? Peut-on raisonnablement penser que les humanités sont susceptibles de se tenir à l'écart des évolutions nu
mériques ? Autrement dit, le numérique serait-il en train de devenir un dogme auquel les humanités seraient sommées de se conform
er?

Si l'observation directe permet d'identifier relativement simplement des projets relevant des humanités numériques, car associant
 techniques informatiques et sciences humaines, la question lancinante de la définition reste cependant au cœur des humanités num
ériques~: s'agit-il d'une discipline à part entière~? D'un moment dans l'évolution des sciences humaines et sociales~? D'une prat
ique ou d'un ensemble de pratiques~? Nous proposons ici d'apporter des éléments de réponse à cette question récurrente en nous fo
calisant sur les enjeux de diffusion. En effet, si les humanités numériques mettent en jeu la fabrication d'outils et de ressourc
es c'est dans le but principal de diffuser ces outils et ces ressources~:

\myCitation{Il s'agit [...] de partage, c'est-à-dire de communication, de réflexion méthodologique collective et de mise en commu
n : ni de la théorie pure, ni de la pratique pure, mais un dialogue au sujet de nos représentations du savoir~\cite{berra2012fair
e}.}

Nous avançons donc que les humanités numériques sont à considérer comme un mouvement qui se propose d'outiller informatiquement l
a recherche en sciences humaines et sociales et d'en diffuser les résultats en prenant soin, dans un mouvement de distanciation,
d'examiner comment les outils mis en place et développés sont susceptibles d'influencer le travail de recherche lui-même.

Par ailleurs, l'évolution des techniques n'est pas neutre dans l'explosion des humanités numériques. Si les projets mêlant humani
tés et informatique existent depuis longtemps, en particulier quand il s'agit de constitution de ressources\footnote{Le \emph{The
saurus Linguae Graecae} est par exemple née au tout début des années 70. http\string://www.tlg.uci.edu/about/}, force est de cons
tater que leur progression s'est considérablement accentuée ces dernières années. En effet, les techniques sont aujourd'hui beauc
oup plus souples et plus propices à l'expérimentation. Dans ces conditions il est infiniment plus facile de tester des solutions.

L'utilisation de l'informatique dans le domaine la recherche en sciences humaines implique de numériser, c'est-à-dire d'encoder,
les objets d'étude et en particulier, mais pas uniquement, les textes. Mais pour encoder de manière rationelle ces objets il faut
 disposer d'un référentiel commun qui va permettre d'unifier les modes de désignation des objets manipulés. Un tel référentiel de
vrait permettre de désigner les chapitres, les sections, les paragraphes, etc., mais aussi les titres de ces chapitres, sections
et paragraphes, sous une forme sur laquelle tout le monde pourra s'entendre. En d'autres termes, il faut disposer d'un modèle abs
trait~\cite{burnard2012literary} auquel une communauté pourra se référer de manière univoque.
```

FIGURE 3.2 – Exemple d'une séquence de code LaTeX.

images, tableaux, etc.) qui la composent. Il s'agit là d'une différence majeure avec les langages appuyés sur un DOM comme HTML.

Notons malgré tout que LaTeX offre la possibilité de créer des ensembles en marquant le début et la fin et en englobant l'intégralité des éléments constitutifs de la partie choisie. Cependant, ce principe n'est pas systématique et ne concerne que le document global, on indique explicitement le début et la fin du document ou des portions réduites comme les figures ou listes par exemple. Toutes les parties internes du document (parties, chapitres, sections, sous-sections, etc.) sont simplement indiquées par leurs titres, donc par leurs commencements. C'est donc la présence d'une commande de création de partie, quel que soit son niveau hiérarchique, qui termine implicitement la partie précédente si elle est d'un niveau équivalent ou supérieur ; dans le cas contraire, la commande marque l'entrée dans un niveau inférieur. Les

commandes `\chapter` et `\section` de la figure 3.2 donnent une illustration de ce dernier cas de figure.

Le second exemple que nous souhaitons présenter est le format RTF (*Rich Text Format*), développé par Microsoft dans les années 80. Ce format est exemplaire d'une évolution majeure dans le domaine de la manipulation de textes et dans le type d'interactions qu'il implique entre la machine et l'utilisateur.

La figure 3.3 montre une séquence de code RTF. Nous y voyons une syntaxe très largement inspirée du langage TEX qui présente donc également une importante proximité avec LATEX. Nous retrouvons en effet les accolades qui permettent dans le langage RTF de marquer des groupes imbricables les uns dans les autres. Les \ sont utilisés pour marquer les codes de contrôle. Ainsi, la séquence `\b` indique le début de séquence qui devra être formatée en gras (*bold* en anglais). Il est donc possible pour un humain de comprendre et de lire une séquence de code RTF sans dispositif d'interprétation spécifique. Cependant, reconnaissons que ce travail reste assez délicat et que le code de ce langage présente malgré tout quelques résistances à la lecture. Cette caractéristique n'est cependant pas à considérer comme une faiblesse intrinsèque du langage RTF puisqu'il a été pensé pour être manipulé *via* une interface de saisie qui interprète le code au cours du travail d'écriture. Ainsi l'auteur n'est pas sensé avoir un accès direct au code pendant la rédaction, mais il doit passer par une interface WYSIWYG.

Cette souplesse apparente proposée par ce type d'interface, qui a eu le mérite de faciliter l'accès aux outils de saisie par sa simplicité d'utilisation, présente cependant un inconvénient majeur car elle détourne en partie l'auteur de son travail de production d'un texte indépendamment de sa, ou de ses formes.

Ainsi, les interfaces WYSIWYG ne proposent pas seulement des solutions ergonomiques. Elles bouleversent totalement la manière dont les textes sont produits. Le travail de production du fond, c'est-à-dire d'écriture du texte, est ainsi continuellement perturbé par l'interface qui présente une version mise en forme détournant le rédacteur de la manière dont les éléments constitutifs du texte doivent interagir les uns avec les autres d'un point de vue logique.

Ce type d'outils propose malgré tout des solutions pour discriminer les différents éléments textuels mais avec quelques faiblesses car une même étiquette peut convenir à plusieurs éléments. Une étiquette *italique* pourra par exemple aussi bien convenir à une séquence en emphase qu'au titre d'une référence bibliographique. De plus, l'utilisation des feuilles de style proposées par les systèmes WYSIWYG, dont le succès est aujourd'hui incontestable, est le plus souvent totalement facultatif :

```
{\rtf1\ansi\deff3\adeflang1025
{\fonttbl{\f0\froman\fprq2\fcharset0 Times New Roman;}{\f1\froman\fprq2\fcharset2 Symbol;}{\f2\fswiss\fprq2
\fcharset0 Arial;}{\f3\froman\fprq2\fcharset128 Times New Roman;}{\f4\fswiss\fprq2\fcharset128 Arial;}{\f5\
fnil\fprq0\fcharset128 SFBX2488;}{\f6\fnil\fprq0\fcharset128 SFBX1440;}{\f7\fnil\fprq0\fcharset128 SFRM1095
;}{\f8\fnil\fprq0\fcharset128 SFTI1095;}{\f9\fnil\fprq0\fcharset128 SFRM0800;}{\f10\fnil\fprq0\fcharset128
SFRM0900;}{\f11\fnil\fprq0\fcharset128 SFRM1000;}{\f12\fnil\fprq0\fcharset128 SFCC1000;}{\f13\fnil\fprq0\fc
harset128 SFCC0900;}{\f14\fnil\fprq0\fcharset128 SFTI0900;}{\f15\fnil\fprq0\fcharset128 SFCC1095;}{\f16\fni
l\fprq2\fcharset128 Arial Unicode MS;}}
{\colortbl;\red0\green0\blue0;\red128\green128\blue128;}
{\stylesheet{\s0\snext0\nowidctlpar{\*\hyphen2\hyphlead2\hyphtrail2\hyphmax0}\cf0\kerning1\hich\af16\langfe
2052\dbch\af16\afs24\lang1081\loch\f3\fs24\lang1036 Standard;}
{\*\cs15\snext15 Caract?res de note de bas de page;}
{\*\cs16\snext16{\*\updnprop5801}\up10 Appel de note;}
{\*\cs17\snext17{\*\updnprop5801}\up10 Appel de note de fin;}
{\*\cs18\snext18 Caract?res de note de fin;}
{\s19\sbasedon0\snext20\sb240\sa120\keepn\hich\af16\dbch\af16\afs28\loch\f4\fs28 Titre;}
{\s20\sbasedon0\snext20\sb0\sa120 Corps de texte;}
{\s21\sbasedon20\snext21\sb0\sa120 Liste;}
{\s22\sbasedon0\snext22\sb120\sa120\noline\i\afs24\ai\fs24 L?gende;}
{\s23\sbasedon0\snext23\noline Index;}
{\s24\sbasedon0\snext24\li283\ri0\lin283\rin0\fi-283\noline\afs20\fs20 Note de bas de page;}
}{\info{\author Pierre-Yves Buard}{\creatim\yr2014\mo12\dy4\hr11\min34}{\revtim\yr0\mo0\dy0\hr0\min0}{\prin
tim\yr0\mo0\dy0\hr0\min0}{\comment OpenOffice}{\vern4110}}\deftab709

{\*\pgdsctbl
{\pgdsc0\pgdscuse195\pgwsxn11906\pghsxn16838\marglsxn1134\margrsxn1134\margtsxn1134\margbsxn1134\pgdscnxt0
Standard;}}
\formshade\paperh16838\paperw11906\margl1134\margr1134\margt1134\margb1134\sectd\sbknone\sectunlocked1\pgnd
ec\pgwsxn11906\pghsxn16838\marglsxn1134\margrsxn1134\margtsxn1134\margbsxn1134\ftnbj\ftnstart1\ftnrstcont\f
tnnar\aenddoc\aftnrstcont\aftnstart1\aftnnrlc
\pgndec\pard\plain \s20\sb0\sa120{\b\afs40\ab\rtlch \ltrch\loch\fs40
De la recherche en SHS aux humanite\uc2 \u769\'cc\'81s nume\u769\'cc\'81riques\uc1 }
\par \pard\plain \s20\sb0\sa120{\rtlch \ltrch\loch\fs28\loch\f6
2.1 Introduction }
\par \pard\plain \s20\sb0\sa120{\rtlch \ltrch\loch\fs22\loch\f7
Le contexte actuel est, comme nous venons de le voir, tout a\uc2 \u768\'cc\'80 fait favorable au de\u769\'c
c\'81veloppement de ce qu\uc3 \u8217\'e2\'80\'99il est aujourd\u8217\'e2\'80\'99hui convenu d\u8217\'e2\'80
\'99appeler les \uc1 }{\rtlch \ltrch\loch\fs22\loch\f8
humanite\uc2 \u769\'cc\'81s nume\u769\'cc\'81riques\uc1 }{\rtlch \ltrch\loch\fs22\loch\f7
. Cependant le de\uc2 \u769\'cc\'81veloppement du nume\u769\'cc\'81rique ne se limite pas au seul domaine d
e la recherche et concerne e\u769\'cc\'81galement, comme nous l\uc3 \u8217\'e2\'80\'99avons de\uc2 \u769\'c
c\'81ja\u768\'cc\'80 e\u769\'cc\'81voque\u769\'cc\'81 pre\u769\'cc\'81ce\u769\'cc\'81demment, un autre doma
ine qui nous concerne ici directement, celui de l\uc3 \u8217\'e2\'80\'99e\uc2 \u769\'cc\'81dition mate\u769
\'cc\'81riel. Revenons ici sur la manie\u768\'cc\'80re dont ces deux domaines, la recherche et l\uc3 \u8217
\'e2\'80\'99e\uc2 \u769\'cc\'81dition, ou autrement dit sur la recherche et la diffusion de ses re\u769\'cc
\'81sultats, inte\u768\'cc\'80grent le nume\u769\'cc\'81rique dans leur pratiques leurs organisations. \uc1
 }
```


FIGURE 3.3 – Exemple d'une séquence de code RTF.

> *Unfortunately, the style-sheet metaphor orients authors toward the presentation instead of toward the role of entities in the document. Thus, the block style might seem appropriate for any of a number of entity types, and nothing motivates the author to make distinctions that may be important later. Furthermore, style sheets tend to be an optional feature instead of a standard interface* [COOMBS *et al.*, 1987].

Nous verrons plus bas comment nous proposons d'exploiter ces interfaces pour "contraindre" les producteurs de textes [31] à utiliser les feuilles de style pour améliorer la qualité technique des textes manipulés en ajoutant explicitement des éléments de structuration dans les fichiers informatiques.

Les langages *descriptifs* comme LaTeX, lisibles directement par l'humain sans dispositif d'interprétation dynamique du code, présentent l'avantage énorme d'accompagner les auteurs dans leur travail de rédaction ou d'annotation des textes :

31. Il s'agit, comme nous le montrerons, de faire en sorte que le temps passé à l'étiquetage, c'est-à-dire à la discrimination des phénomènes textuels les uns par rapport aux autres, soit le plus rentable possible pour les producteurs de textes.

> *One of the more subtle advantages of descriptive markup is it supports authors in focusing on the structure and content of documents. Both presentational and procedural markup tend to focus authors' attention on physical presentation* [Coombs *et al.*, 1987].

Il s'agit donc en définitive, dans les langages *descriptifs*, de rendre explicite la structure des textes manipulés en appliquant finalement le principe de séparation du fond et de la forme. Les langages de type RTF sont très orientés vers la mise en forme et perturbent les opérations de production et de structuration explicite du fond, quand les interfaces de manipulation ne rendent pas cette dernière totalement secondaire voire facultative.

Ces différentes approches sont aussi liées à des modèles de représentation dont les fonctions générales sont marquées. RTF considère le texte avant tout comme une succession de mots à mettre en forme alors que des langages comme LaTeX proposent de considérer le texte comme une série d'informations à articuler explicitement à travers la catégorisation des éléments et des relations qu'ils entretiennent.

Dans le contexte actuel qui tend à la multiplication des supports, mettre l'accent sur la forme du texte au moment de sa rédaction pose des problèmes évidents. Comment imaginer la meilleure façon de traduire formellement un texte sur un support totalement différent de celui sur lequel il a été pensé et écrit ? En effet, les solutions WYSIWYG, en simulant la page imprimée, cadrent fortement l'ensemble de la démarche et masquent un certain nombre de problèmes potentiels comme les renvois internes par exemple. Une référence telle que "voir p. 12" n'aura plus aucun sens hors du contexte de la page, comme, par exemple, lorsque le lecteur choisit de consulter le texte sous la forme d'une page web ou d'un livre numérique au format ePub.

La multiplication des supports de diffusion oriente donc clairement la préférence vers les langages *descriptifs* qui se focalisent sur la mise en place de structures textuelles explicites pouvant faire l'objet de traitements postérieurs pour la production de formes de diffusion.

Mais ces langages *descriptifs* imposent d'utiliser leur propre syntaxe de description des contenus pour étiqueter les éléments constitutifs des textes. L'utilisateur peut, dans le meilleurs des cas, ajouter une sur-couche aux catégories gérées par les langages sous la forme de macros-commandes mais il ne peut en aucun cas mettre en place son propre modèle de référence.

3.3 SGML et les langages à balises

Le SGML[32] (*Standard Generalized Markup Language*) introduit en 1986, pousse la logique des langages *descriptifs* un peu plus loin [Bryan, 1992], en offrant justement aux utilisateurs et, plus précisément, aux communautés d'utilisateurs, la possibilité de construire leurs propres modèles formels de données en séparant la définition de ce modèle de l'instanciation des données, d'une part, et la production des formes de diffusion, d'autre part.

En conséquence, SGML ne propose pas en propre de systèmes de vocabulaires, c'est-à-dire des ensembles de balises descriptives, pour tel ou tel type de document, c'est pour cette raison que l'on qualifie SGML de méta-langage : cette tâche est réservée aux applications SGML qui vont assurer la production de grammaires de référence pour un domaine donné. Les informations sur les types d'éléments constitutifs des textes et les systèmes d'organisation hiérarchiques de ces éléments entre eux sont ainsi renseignés dans des fichiers distincts ne contenant aucune donnée, ce sont les *Document Type Definition*[33] (DTD). Ces modèles formels peuvent ainsi circuler au sein des communautés et être amendés indépendemment des données.

SGML introduit également un système de contrôle et de validation des données qui va permettre d'éviter un certain nombre d'erreurs dans le travail de structuration des données.

Les principes introduits par SGML ont incontestablement marqué une évolution fondamentale dans la production et la gestion des documents numériques qui a produit un héritage et un ensemble de pratiques toujours en vigueur aujourd'hui. Sans entrer dans les détails, nous allons ici nous intéresser à deux points saillants : une application SGML et une évolution de ce méta-langage.

3.3.1 *Hypertext Markup Language*

Le HTML est incontestablement l'application SGML la plus connue. En constante évolution depuis sa première apparition à la fin des années 80, ce langage s'est imposé comme l'une des bases du *World Wide Web*. Le HTML est présenté dès 1991 comme une application SGML dont il respecte les principes en proposant un vocabulaire de base pour décrire les documents. Mais c'est en 1998 avec l'annonce de HTML 4.0 [Raggett *et al.*, 1999] que l'effort de normalisation entrepris pendant de longues années va payer. Les spécifications du DOM (*Document Object Mo-*

32. ISO standard 8879.

33. Ou Définition Type de Document.

del) [Le Hors *et al.*, 2004], dont la première publication date de 1998, apportent des solutions normalisées pour accéder, lire et modifier le contenu des documents balisés. Il est alors possible de mettre en place et de manipuler des structures complexes en respectant (dans la majorité des cas) les principes de séparation du fond et de la forme, déjà recommandées dans la norme SGML, dans un cadre de travail cohérent, normalisé et documenté. Le modèle de représentation des informations contenues dans les documents et d'accès à ces informations permet en effet de fabriquer des sites en assurant une reconnaissance maximale par les navigateurs ayant correctement implémentés les normes.

Mais HTML 4.0 constitue toujours une application SGML qui conserve encore quelques éléments de mise en forme comme `<i>`, `<b>` ou `<big>` et `<small>` par exemple et qui ne respecte donc pas totalement les principes de séparation fond/forme recommandés.

Le passage à XHTML [Pemberton, 2002], dont les premières spécifications datent de 2000, permet d'achever le mouvement de séparation du fond et de la forme tout en assurant le passage à un nouveau format de données, lui aussi dérivé de SGML : le XML (*eXtensible Markup Language*) [Bray *et al.*, 2008], sur lequel nous reviendrons plus en détail (voir 3.3.2). Cette évolution marque une orientation de quelques années plaçant la qualité du document au centre des préoccupations, avec le respect des principes de rigueur de structuration de la recommandation XML.

HTML 5 [Berjon *et al.*, 2014] marque en revanche un virage fort avec la prise en compte de la dimension applicative. C'est en effet, le service proposé à l'utilisateur qui s'impose face à la qualité de structuration des données. Il ne s'agit donc plus de construire et de structurer des documents dont les formes seront produites éventuellement par des transformations, réalisées à l'aide des langages de la famille XSL [Berglund, 2006] et en particulier de XSLT [Kay, 2007], et des feuilles de styles CSS, mais de produire des applications intégrant directement des dispositifs d'interactions entre les utilisateurs et les contenus (menus, barre d'outils, animations, etc.). Même s'il est toujours possible d'utiliser une syntaxe XML avec la déclinaison XHTML 5, le fait de ne pas la mettre en avant est révélateur du choix de mettre l'accent sur la commodité du travail de développement et non d'encourager la production de documents finement annotés et structurés.

Cependant, si le web s'est orienté vers l'applicatif avec l'apparition de la norme HTML 5 [Lecarpentier, 2011], l'approche centrée sur la structuration des données reste pertinente en particulier dans le cas de l'édition de sources qui nous occupe dans

cette étude surtout dans la mesure où le web n'est pas le seul vecteur d'utilisation envisagé.

3.3.2 *eXtensible Markup Language*

Descendant de SGML dont il reprend les principes généraux en les simplifiant pour faciliter les travaux d'implémentation d'une part et l'interopérabilité des données d'autre part, le XML, datant de 1998, est très lié au web et à la famille des langages HTML qui lui est associée [Bray *et al.*, 2008].

Les efforts de simplification portaient en particulier sur les questions de permissivité de syntaxe. En SGML, il est possible d'ouvrir des éléments sans les fermer alors que XML impose de fermer tout élément ouvert. Il s'agit donc d'une forme de radicalisation de la syntaxe qui facilite le développement des applications en s'appuyant en réalité sur la clarification des procédures d'étiquetage des textes : plus les données sont claires, plus il est facile et rapide de les interpréter. En effet, préciser le début et la fin de chaque phénomène lève toute ambiguïté, ce qui est certes utile pour le développement applicatif mais c'est aussi capital pour toute personne qui aurait besoin de comprendre la nature des données manipulées. L'échange de données est considérablement facilité lorsque le formalisme de structuration contraint à la clarté. Si cet argument conserve encore aujourd'hui toute sa force pour simplifier la circulation d'information entre humains, il perd de son importance pour ce qui concerne le développement. En effet, un aspect important du XML du point de vue du développement est de limiter les temps de calcul en "purifiant" les données. Mais avec la montée en puissance des ordinateurs cette dimension est de moins en moins vraie : l'application de règles d'interprétation des éléments ouverts mais pas fermés ne demande plus autant de temps qu'avant, quand les machines étaient moins performantes. La syntaxe HTML 5 de base, autorisant les raccourcis, est une parfaite illustration de ce phénomène. Nous verrons plus bas que la dimension technique ne doit pas être la seule à prendre en compte lorsqu'il s'agit de choisir une syntaxe et que la compréhension [34] par l'humain doit être un critère de choix important.

L'application de XML dépasse aujourd'hui complètement les usages strictement liés au cadre du web justement grâce à ses qualités de simplicité et d'indépendance technique provenant de son contexte d'invention très fortement connecté au réseau. Beaucoup de fichiers de configuration de logiciels ou de systèmes d'exploitation sont

34. Cette compréhension ne doit pas pour autant forcément passer par un accès direct au code et il s'agit aussi de proposer des interfaces ergonomiques.

maintenant codés en XML. Apple, par exemple, utilise ainsi massivement le XML pour les fichiers de configuration internes de son système d'exploitation Mac OS.

Mais le XML est aussi un format de données très utilisé dans le domaine de la production et de la gestion de textes. Ainsi, le logiciel grand public OpenOffice utilise depuis longtemps le XML nativement pour l'enregistrement de ses fichiers[35]. La norme *Open Document Format for Office Applications* [Durusau et Brauer, 2011] utilisée par OpenOffice est depuis 2006 certifiée par l'Organisation internationale de normalisation (ISO). Rapidement, Microsoft, contraint par cette certification qui a mis fin à l'hégémonie de son standard de fait, le format *.doc*, a suivi, en modifiant le format de fichier de son logiciel de traitement de texte Word pour adopter aussi le XML et la norme *Office Open XML* [ECMA, 2008] pour la structuration des données textuelles. Citons aussi pour terminer le format IDML (*InDesign Markup Language*) utilisé par la firme Adobe pour décrire les documents de son logiciel de publication assisté par ordinateur (PAO) InDesign. Il ne s'agit pas ici du format utilisé par défaut par Adobe InDesign, mais du format d'échange recommandé par la firme pour faciliter la circulation des données entre deux versions différentes du logiciel. Dans le même domaine de la PAO, notons que presque tous les logiciels[36] proposent aujourd'hui des solutions plus ou moins efficaces pour importer du XML.

Le XML a acquis aujourd'hui une telle importance, comme nous venons de le voir, qu'il n'est pas sans influence sur la manière dont le texte est aujourd'hui perçu et étudié dans différents domaines de recherche. Avec la prépondérance du web, auquel il est encore très lié, il occupe souvent une place centrale dans les réflexions et les travaux de recherche. Il en est ainsi, par exemple, dans les efforts de définition de la notion de document dans le contexte du réseau mondialisé [Pédauque, 2003].

Le XML s'est donc imposé comme un moyen efficace de manipuler de l'information structurée dans de nombreux domaines. Les modèles de représentation abstraits sont aussi très nombreux et émanent souvent de communautés ayant pris en compte l'importance de XML. Ainsi, la logique de représentation en arbre d'XML[37] n'est pas sans influencer les modèles abstraits, dont on pourrait pourtant penser qu'ils sont conçus sans tenir directement compte des contingences "techniques". Mais on

35. Avant que ce format devienne une norme puisque les premières versions d'OpenOffice développées par Sun Microsystems utilisaient déjà le XML.

36. C'est le cas par exemple de Quark Xpress, Adobe InDesign ou encore, et depuis longtemps, de FrameMaker qui était un logiciel pionnier en matière d'intégration des langages à balises puisqu'il existait dès les années 1990 une version FrameMaker+SGML.

37. Cette notion d'arborescence n'est pas présente seulement ici, elle est par exemple la base organisationnelle de la plupart des systèmes de fichiers informatiques.

peut estimer que si ce modèle d'arbre s'impose, même au niveau de la conception de modèle abstrait, c'est aussi en raison de la grande souplesse qu'il propose. En effet, cette caractéristique lui permet souvent d'être appliqué à un grand nombre de données différentes, qu'il s'agisse de décrire l'information ou de la structurer.

3.4 DOM et HTML 5

C'est avec le *Document Object Model*, dont la première version (*DOM Level 1*) date de la fin 1998, que le document balisé est modélisé comme un arbre en tant que tel. Ainsi l'ensemble des *nœuds*, c'est-à-dire les éléments d'un document comme les paragraphes, les liens, les séquences de textes, etc., constitutifs d'un document est organisé dans un arbre DOM. Cette modélisation constitue un tournant important qui implique des modes de parcours et d'accès spécifiques à l'information textuelle contenue dans les documents. Le DOM propose une API dédiée qui participe d'une conception spécifique du document avec des moyens d'extraire des nœuds, c'est-à-dire des fragments de textes arborescents, d'en insérer de nouveaux, bref de manipuler les contenus textuels.

Notre travail s'inscrit pleinement dans le modèle de document proposé par le DOM : les textes que nous manipulons seront considérés et traités comme des arbres. C'est une approche fondamentale dans notre étude, en particulier du point de vue de l'édition des documents. En effet, la manipulation d'un arbre est totalement différente de la manipulation d'une séquence de caractères sérialisant un arbre.

Le langage *XML Path Language*, ou XPath [Clark et DeRose, 1999], dont la syntaxe n'est pas XML, fixe les règles pour accéder aux éléments constitutifs des arbres DOM. XPath permet donc de manipuler des fragments de texte sur la base de la détermination de chemins à parcourir dans les arbres DOM, depuis la racine jusqu'au nœud voulu.

Le *Document Object Model* et le langage *XML Path Language* fournissent donc un modèle de document spécifique dans lequel les documents textuels sont des arbres, avec des règles d'organisation et de parcours qui seront au cœur de nos réflexions et de nos propositions.

Par ailleurs, une autre composante importante dans la modélisation des textes est introduite par le HTML 5 [Berjon *et al.*, 2014] avec la notion de *flow content*. Il s'agit d'un type de contenu qui rassemble la plupart des éléments rencontrés dans le corps d'un document. Le texte est donc ici considéré dans sa dimension fluide : il a un volume déterminé et s'étend sur le support de lecture écran, en s'adaptant à l'espace

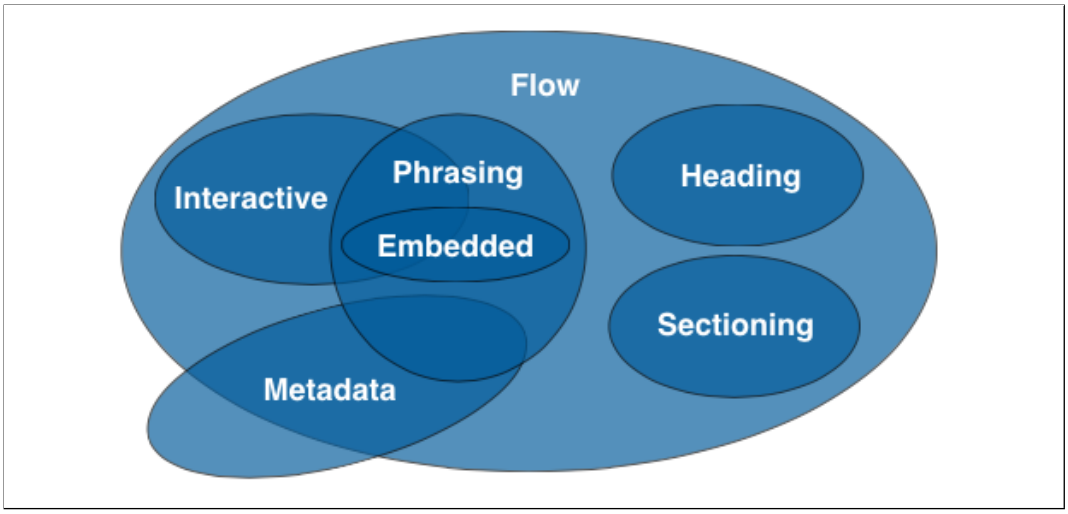

FIGURE 3.4 – Les modèles de contenus en HTML 5.

dont il dispose en fonction du dispositif sur lequel il est consulté : largeur de fenêtre, résolution de l'écran, etc. Si nous retrouvons ici le caractère résolument applicatif de HTML 5, il n'en reste pas moins que cette conception du texte comme fluide à contraindre sur des supports de lecture prend en compte une dimension importante du texte.

La figure 3.4 présente l'articulation des différents types de contenus proposés par le HTML 5. Le *flow content* occupe bien entendu une place importante car il englobe l'ensemble du texte constitutif du document. Il n'est plus question de considérer l'organisation hiérarchique interne des documents, que l'on retrouve dans les types *heading* et *sectioning*, mais d'envisager le texte du document comme une entité fluide unique regroupant et articulant différents types d'éléments participant tous à sa composition.

Avec la notion de *flow content* de HTML 5 et le DOM, nous sommes confrontés à deux points centraux sur lesquels nous reviendrons dans cette étude : la nature fluide du texte et son organisation hiérarchique et arborescente.

Dans les trois chapitres suivants, nous examinerons comment les communautés des différents domaines concernés par notre étude ont construit des modèles de représentation décrivant les objets qu'ils manipulent ainsi que l'activité de leurs domaines. Nous montrerons que beaucoup de ces communautés ont choisi XML pour proposer des solutions de mise en œuvre de ces modèles.

4

Archivistique et conservation

4.1 Introduction

Les archives et les institutions de conservation ont pour mission de conserver les documents qui leur sont confiés. Le cadre de leur activité est donc clairement identifié et impacte directement la manière dont les objets manipulés sont considérés et perçus.

La conservation s'accompagne également d'un rôle important de signalement des pièces. Il s'agit donc pour les archivistes et les conservateurs de mettre en place des systèmes d'organisation documentaire permettant de retrouver le plus efficacement possible une pièce ou un groupe de pièces parmi les ensembles conservés.

La pièce conservée, qui peut être de toute nature, est donc l'unité de travail de base. Le niveau de description de cette unité de base est directement lié aux missions de l'institution [38]. La description des pièces rassemble donc les informations nécessaires à la découverte des documents dans les fonds d'archives.

L'archiviste se préoccupe de décrire précisément les réalités documentaires dont il a la responsabilité mais il ne réalise pas ce travail pour le contenu des documents, même si, bien entendu, il utilise ce contenu pour ordonner les ensembles. En définitive, le conservateur ou l'archiviste réalise un travail d'organisation des pièces décrites dans des ensembles eux aussi identifiés précisément et méthodiquement.

Les institutions de conservation ont aussi une mission de valorisation des fonds conservés. Cette valorisation prend aujourd'hui souvent la forme d'exposition de données sur le web. Dans cette optique, la numérisation se limite dans la plupart des cas à un mode image (avec, parfois et quand c'est techniquement possible, une

38. Ainsi, le niveau de description archivistique sera sans doute insuffisant du point de vue d'un codicologue, même si l'objet décrit est le même.

couche de reconnaissance optique des caractères (OCR[39])) des pièces. Ces images numériques sont ensuite placées dans des entrepôts auxquels l'utilisateur peut accéder via des inventaires virtuels reprenant l'organisation établie par les conservateurs et les archivistes.

4.2 Objets manipulés

D'une manière générale, les objets manipulés par les institutions de conservation sont des documents au sens très général du terme, c'est-à-dire dans la mesure où ils transmettent une information. Il est donc aisé d'imaginer que les responsables de fonds d'archives peuvent être confrontés à diverses réalités matérielles. Pour se faire une idée de cette diversité, on peut ici revenir à l'exemple célèbre de l'antilope qui une fois capturée, étudiée et cataloguée devient un document :

> L'antilope cataloguée est un document initial et les autres documents sont des documents seconds ou dérivés [Briet, 1951].

Dans le cadre de la présente étude, nous nous concentrerons malgré tout sur des documents matériels précis : les manuscrits et imprimés anciens qui concernent donc les archivistes dans leur dimension matérielle. Nous nous intéresserons également aux ensembles composés de ces pièces élémentaires : les fonds.

La figure 4.1 donne une représentation de l'organisation d'un fonds jusqu'aux pièces qui le composent. Les niveaux hiérarchiques ne sont pas tous ici représentés et il est possible d'en ajouter. Un fonds particulièrement complexe et riche pourra exiger d'ajouter autant de niveaux que nécessaire pour aboutir à un résultat satisfaisant pour faciliter sa compréhension et la découverte des pièces qui le composent. Ce schéma permet très bien de voir que le modèle archivistique propose des relations de composition complexes entre le fonds et la pièce avec un certain nombre d'objets intermédiaires.

Nous ne pourrons pas ici entrer dans tous les détails car l'archivistique est un domaine complexe. Pour notre étude, nous décrirons deux objets centraux : la pièce, et pour ce qui nous concerne, les témoins (manuscrits et imprimés anciens) et le fonds, c'est-à-dire un ensemble de pièces rassemblées en un tout cohérent.

39. *Optical character recognition.*

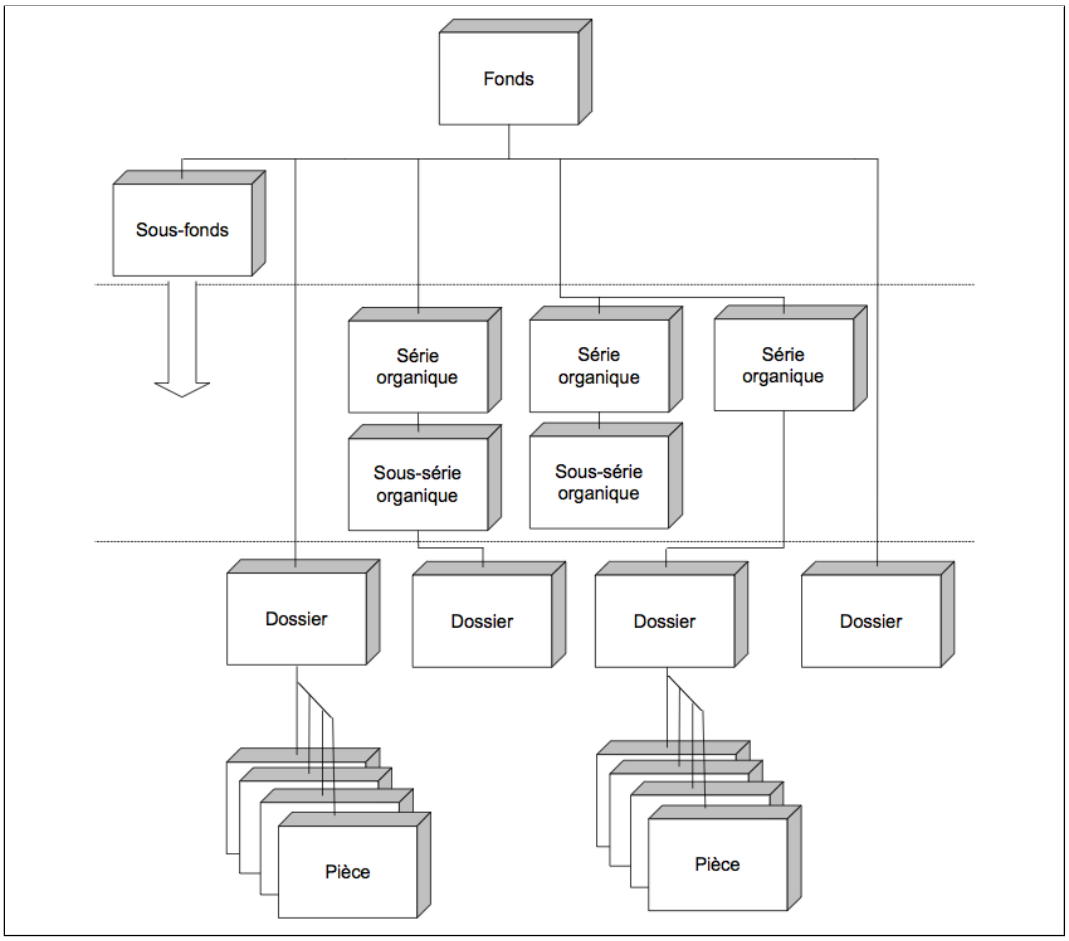

FIGURE 4.1 – Schéma des niveaux de classement d'un fonds [ICA, 2000].

4.2.1 Manuscrits et imprimés anciens (support matériel)

Cette notion est probablement la plus évidente à concevoir puisqu'il s'agit de désigner une réalité : celle du support matériel du texte, de l'objet que l'on a sous les yeux ou entre les mains, c'est-à-dire le livre, le manuscrit ou l'écran.

Il s'agit en fait du porteur du texte, ou plus exactement d'une version du texte dans le cas des sources anciennes. Cette version se caractérise par ses propres spécificités : format, matières utilisées, reliures, organisation du texte dans la page, erreurs de copie, détérioration, etc.

Les images numériques produites en masse de nos jours présentent des vues de ces supports matériels, manuscrits, exemplaires imprimés ou incunables. Ainsi, chaque image numérique d'un texte présente en réalité une vue d'une des versions du texte, celle de tel ou tel manuscrit par exemple, ou encore de tel exemplaire annoté par telle ou telle personne. Ainsi, même si le texte porté par un support imprimé ne diffère pas d'un autre exemplaire imprimé conjointement, la présence d'une annotation

unique consultable uniquement sur un support précis peut justifier une acquisition numérique. C'est en fait à ce texte unique que l'on porte attention, en tant que tel, mais aussi dans ses rapports avec le texte imprimé. C'est par exemple le cas du *first folio* des œuvres de Shakespeare, le 233e exemplaire sur les 800 imprimés à l'origine, récemment découvert à Saint-Omer. Cet exemplaire semble présenter des annotations utilisées pour monter des pièces de théâtre [40]. Enfin, même s'ils sont un peu en dehors de notre périmètre, les versions imprimées modernes entrent bien entendu totalement dans cette catégorie. En effet, elles constituent des éléments matériels supportant du texte produits selon des modalités différentes des manuscrits ou des incunables.

La notion d'exemplaire peut alors avoir une importance prépondérante, pour être en mesure de distinguer deux instanciations matérielles. Ce dernier point peut être capital, par exemple si un exemplaire d'un fonds contient des notes de lecture en marge du texte imprimé.

Nous désignerons ici par *support matériel*, *artefact* ou *item* ces réalités matérielles, manuscrits, incunables ou exemplaires.

4.2.2 Fonds d'archives

Pour les institutions de conservation, un fonds est un ensemble documentaire constitué par un producteur, c'est-à-dire par une personne physique ou morale. La cohérence du fonds est donc donnée par le producteur, c'est pourquoi l'archivistique s'interdit en principe de modifier l'organisation des documents. En principe, car des difficultés peuvent apparaître et contraindre à déplacer les documents d'un fonds, par exemple lorsque deux fonds se sont trouvés mélangés accidentellement. Cependant le principe de respect du fonds, capital en archivistique, reste une règle générale théorique que les spécialistes cherchent au maximum à respecter. Le document n'est jamais, dans un contexte archivistique, considéré isolément,

> il se situe au sein d'un processus fonctionnel, dont il constitue lui-même un élément [...]. Il a toujours un caractère utilitaire, qui ne peut apparaître clairement que s'il a gardé sa place dans l'ensemble des autres documents qui l'accompagnent [Duchein, 1977].

En forçant le trait, nous pouvons considérer que, pour l'archiviste, le document est en réalité un fragment d'un fonds qui ne peut être exploité ou interprété indépendamment de son contexte archivistique.

40. Voir : https://lejournal.cnrs.fr/articles/first-folio-de-shakespeare-le-regard-dun-expert

Du point de vue de la répartition et de l'organisation des informations, un fonds est décrit de manière hiérarchique en procédant du général au particulier. Ainsi, à chaque niveau de la figure 4.1 sera donc associée une description comprenant toutes les informations le concernant spécifiquement. Le travail consiste à définir autant de niveaux que nécessaire et à donner une description de chacun de ces niveaux jusqu'à la pièce elle-même.

4.3 Modèle du domaine

Pour normaliser et faciliter le travail, l'archivistique s'est doté d'un modèle de référence qui explicite les objets du domaine, leur organisation ainsi qu'un ensemble de règles générales pour orienter et cadrer les pratiques.

Nous avons déjà évoqué la norme ISAD(G) (*General International Standard Archival Description*) [ICA, 2000] qui présente ce modèle archivistique international qui peut être adapté au niveau national. Mais il s'agit toujours d'un modèle abstrait qui ne fournit pas directement d'outils pour la réalisation du travail. Cependant, ce modèle abstrait fournit toutes les catégories nécessaires pour la construction d'un standard de structuration.

L'EAD (*Encoded Archival Description*) [41] a été développée pour structurer de manière hiérarchique des inventaires électroniques, quel que soit leur degré de complexité et dans le respect du modèle ISAD(G).

Ce standard est donc tout naturellement à considérer comme la langue naturelle des conservateurs et des archivistes dans la mesure justement où il respecte les usages du domaine et en particulier cette norme internationale de description archivistique ISAD(G) [42]. Même si la TEI permet aussi de décrire des inventaires de manière hiérarchique, comme on peut le voir avec le projet Handrit [43] sur la collection arnamagnéenne, force est de constater qu'une très grande partie des inventaires électroniques sont aujourd'hui encodés en EAD. Cet état de fait peut sans doute aussi s'expliquer par l'existence d'outils EAD qui facilitent considérablement les opérations techniques inhérentes à ce type de projets.

41. http://www.loc.gov/ead/
42. http://www.ica.org/ ?lid=10207
43. http://handrit.is

Historique et caractéristiques techniques

La première version de l'EAD est sortie en 1993 sous la forme d'une DTD utilisant le langage SGML dans le cadre d'un projet de recherche de l'Université de Berkeley. Le groupe de recherche est devenu international. Il est composé de bibliothèques universitaires publiques ou privées, de la Bibliothèque du Congrès, des archives nationales américaines, du secteur commercial privé. La Société Américaine des Archivistes [44] et la Bibliothèque du Congrès [45] ont rapidement apporté leur soutien à l'entreprise et en assurent encore aujourd'hui la maintenance et la diffusion. La version actuellement utilisée est celle de 2002 qui est diffusée sous forme de DTD, de schéma XML ou de schéma Relax NG.

C'est la relative simplicité du standard EAD, composé de 143 éléments en tout, qui fait sa force car le nombre de solutions de structuration est mathématiquement limité. En effet, comme nous le verrons, la relativement faible étendue des possibilités favorise le développement de solutions techniques tant pour la saisie et la structuration des données que pour leur diffusion. Ainsi il existe des outils qui permettent de mettre en œuvre l'EAD dans des conditions relativement simples sans investissements de développement trop lourd.

Organisation et outils

Malgré le nombre modéré d'éléments proposés par le standard EAD, l'ouverture de ce format laisse malgré tout une large place à l'interprétation qui peut elle-même conduire à d'importants écarts de pratiques. Or, comme nous l'avons déjà vu, l'un des intérêts majeur de l'utilisation des standards est justement de simplifier l'échange et la circulation des données en les structurant de la même manière. C'est pour éviter cet écueil que la communauté a produit des guides de bonnes pratiques pour harmoniser les usages. C'est le cas dans plusieurs pays et notamment en France où le guide des bonnes pratiques a été confié à un groupe de travail composé d'experts du domaine à l'initiative du Ministère de l'Enseignement supérieur et de la Recherche, et du Ministère de la Culture [46].

Il existe aujourd'hui de nombreux outils d'exploitation pour la production et l'exposition de données en EAD, qui peuvent être commerciaux comme les solu-

44. http://www2.archivists.org
45. http://www.loc.gov/index.html
46. http://bonnespratiques-ead.net

tions Mnesys[47], ou *open source* comme Pleade[48], ica-atom[49], ou encore *archivists toolkit*[50]. Ces outils présentent parfois des caractéristiques ergonomiques très marquées. Ainsi, *archivists toolkit*, très utilisé dans le monde anglo-saxon, est souvent peu apprécié par les archivistes et conservateurs français.

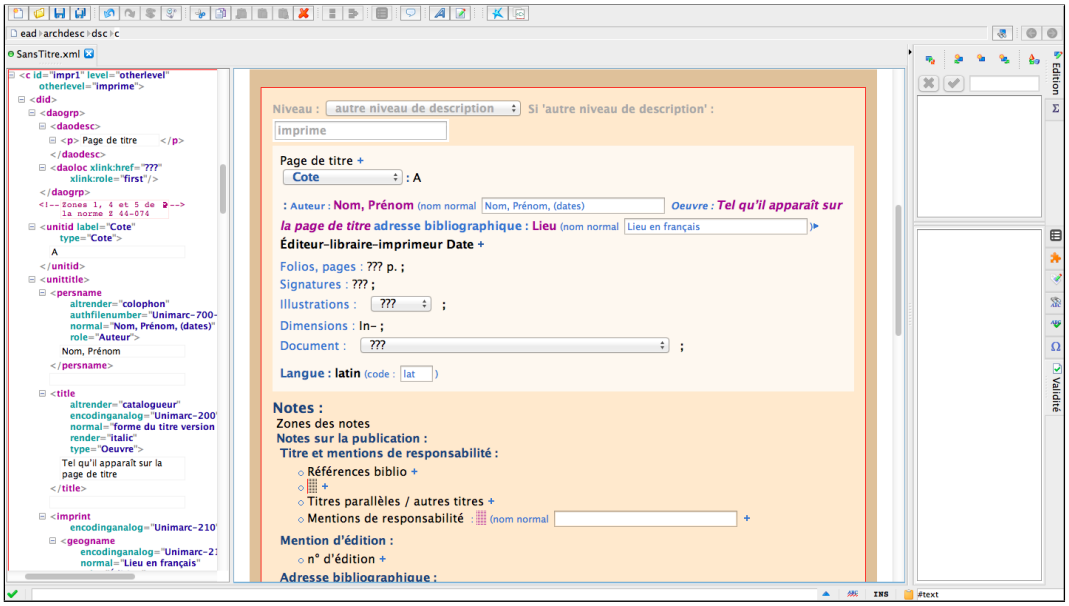

FIGURE 4.2 – Interface de structuration EAD respectant le *Guide des bonnes pratiques* sous XMLmind XML Editor.

En suivant les indications des guides des bonnes pratiques, il est relativement aisé de configurer certains logiciels d'édition XML pour construire des interfaces de saisie qui vont faciliter le travail des archivistes et des conservateurs en prenant en charge presque tous les aspects techniques de la structuration des données. La figure 4.2 montre l'interface de saisie développée dans le logiciel XMLmind XML Editor[51] dans le cadre du projet de bibliothèque virtuelle du Mont Saint-Michel. L'interface proposée permet aux opérateurs qui le souhaitent de rédiger leurs notices descriptives en recourant à des masques de saisie adaptés au plan de classement préalablement défini ainsi qu'à la granularité et à la hiérarchisation des éléments de structuration retenus. Avec ce type d'outil, l'ensemble des aspects de structuration de données, c'est-à-dire l'ensemble des choix de vocabulaire dans le respect du guide des bonnes pratiques, est géré au moment du développement de l'interface. Le travail de saisie se

47. http://www.mnesys.fr
48. http://www.pleade.com
49. https ://www.ica-atom.org
50. http://archiviststoolkit.org
51. http://www.xmlmind.com

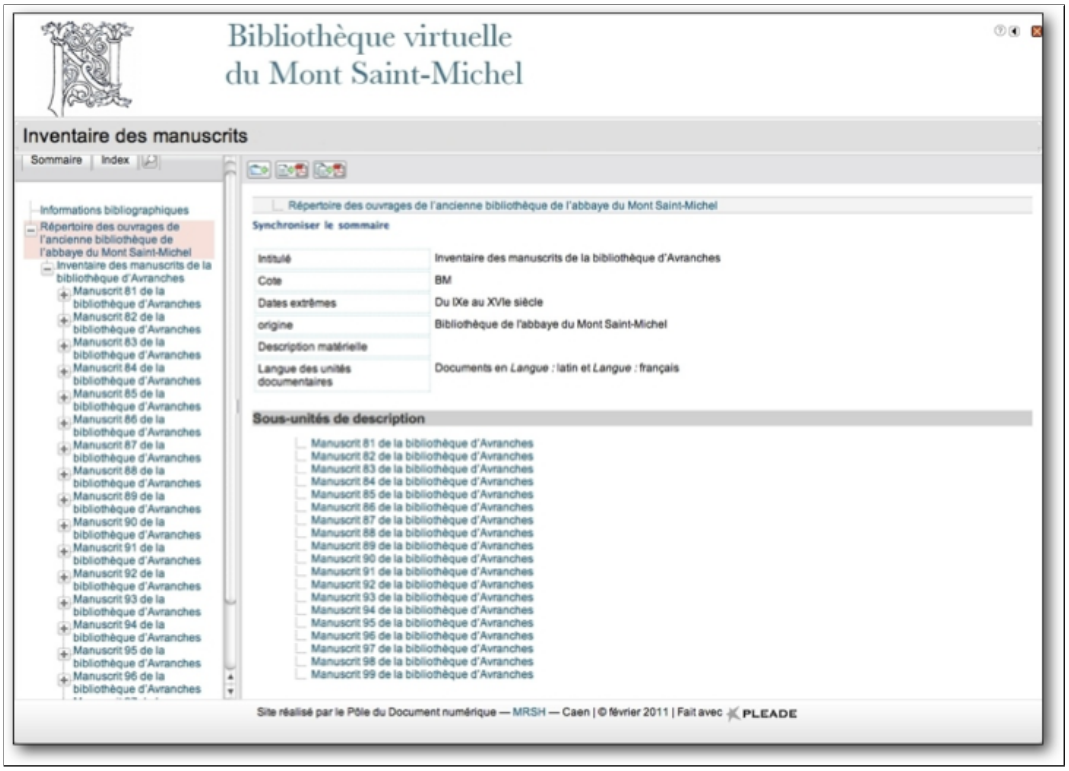

FIGURE 4.3 – La bibliothèque virtuelle du Mont Saint-Michel.

fait simplement en utilisant le formulaire de la zone centrale et c'est le logiciel qui se charge de l'écriture de l'ensemble du code XML correspondant. Si jamais, au cours du travail de saisie, un manque se fait jour, il suffit de modifier l'interface pour que le formulaire produise le code approprié. Il s'agit bien de simplifier et d'accélérer le travail, mais pas de construire des outils opaques pour les professionnels du domaine.

C'est pour toutes ces raisons que l'EAD est aujourd'hui massivement utilisé par les archives et les institutions de conservation pour mettre en ligne leurs inventaires électroniques.

La figure 4.3 donne un exemple d'inventaire en ligne produit avec le logiciel *open source* Pleade. On y voit clairement le respect de l'organisation recommandée par l'ISAD(G).

4.4 Conclusion

Le monde de la conservation et des archives est donc une communauté organisée qui dispose à la fois d'un modèle abstrait, l'ISAD(G) et d'un langage de description

formel, l'EAD. Ces deux éléments vont dans le sens de la collaboration et de la diffusion des informations manipulées, au cœur du métier comme nous l'avons vu.

Dans le cadre d'un projet d'édition de sources, les chercheurs s'appuient sur l'organisation des informations réalisée par les professionnels de la conservation pour débuter leur travail et organiser les opérations. La chapitre suivant est consacré à l'ensemble des opérations effectuées pendant le travail d'édition des sources anciennes proprement dit.

5

Édition de sources

Idéalement les chercheurs disposent d'un inventaire numérique respectant le standard de description EAD pour identifier les témoins porteurs des textes à étudier et à éditer. Il est donc possible de les rassembler pour constituer un corpus et mettre en œuvre les solutions nécessaires à leur édition. Nous présentons ici l'ensemble des opérations ainsi que les objets manipulés pendant leur exécution.

5.1 Introduction

La notion d'édition de sources, dont l'édition critique est sans doute l'une des formes les plus complexes, ne se définit pas simplement. Nous allons ici nous contenter d'une définition relativement générale sans doute insuffisante pour les spécialistes, mais satisfaisante pour nos besoins.

Nous parlerons donc d'édition critique pour désigner le résultat d'un travail scientifique et éditorial visant à mettre à la disposition du lecteur une version immédiatement utilisable, c'est-à-dire lisible et compréhensible, de la documentation collectée. Elle doit donc contenir une explicitation très claire de son propre mode de construction. Une édition critique est le résultat d'opérations de sélection dans les variantes proposées par les différents témoins. Bien entendu, elle doit s'accompagner de toutes les explications et justifications nécessaires pour montrer la cohérence de l'ensemble des choix réalisés. Il ne s'agit donc pas seulement de livrer un matériau brut (une "simple" transcription des textes en quelque sorte), mais bien de procéder à un véritable travail d'établissement d'un texte en apportant systématiquement une explication précise du cheminement qui a permis au chercheur de privilégier tel témoin contre tel autre pour chaque segment proposant plusieurs versions.

De ce point de vue, nous pouvons donc considérer qu'une édition critique a vertu à devenir l'édition de référence pour les textes qu'elle prétend traiter. En conséquence, la question de la citabilité de ces éditions est primordiale. Une édition de référence doit être stable et permettre aux lecteurs de la citer de manière pérenne quel que soit son mode de diffusion (électronique ou papier).

Le chercheur, éditeur du texte établi, est ainsi le garant de la qualité du texte proposé dans l'édition critique, résultat de multiple mesure de variantes entre les différentes versions portées par les témoins. Nous y reviendrons plus loin dans ce mémoire, mais notons dès maintenant que le chercheur, avec le développement des techniques numériques, participe aussi à la définition des fonctionnalités des éditions en ligne ou autres éditions numériques. Le lecteur doit-il pouvoir accéder au texte de chacun des témoins ? Quelles doivent être les modalités d'accès aux textes (chapitrage, moteur de recherche, etc.) ? Quel rapport entre le texte et les images numérisées des témoins ? Autant de questions auxquelles le chercheur doit apporter des éléments de réponse pour pouvoir, en étroite relation avec l'éditeur matériel, construire l'édition la plus adaptée possible aux objectifs initialement fixés.

La forme papier conserve toujours une grande importance pour les éditions critiques. En effet, la production d'un volume papier reste un élément central à plus d'un titre. Tout d'abord de grandes collections parfaitement identifiées par les communautés depuis de nombreuses années facilitent le travail de repérage et d'information. Elles sont devenues de véritables outils pour les chercheurs dont il est inutile d'essayer de se passer dans la mesure où elles remplissent parfaitement leur fonction. D'autre part, le support papier présente certaines qualités fondamentales dans le cas des éditions critiques. La principale est sa stabilité. En effet, comme nous l'avons déjà noté plus haut, la stabilité est capitale pour permettre au discours scientifique de se construire. En proposant une version papier, les éditeurs sont certains de permettre facilement la citation. Si le numérique permet également d'apporter une certaine stabilité, c'est souvent au prix d'une limitation dans les choix des fonctionnalités proposées au lecteur.

Nous donnons ici quelques exemples d'articulations possibles entre les différents supports pour la production d'éditions critiques. Nous reviendrons sur certains d'entre eux en détails, notamment sur les opérations d'expérimentation et de validation dans la partie IV de notre étude.

Le *Roman du Mont Saint-Michel* [Bougy, 2009] est une édition critique multimodale publiée sur trois supports différents : papier, cédérom et web. Le papier est la forme de référence, particulièrement adaptée à la citation, qui contient l'intégra-

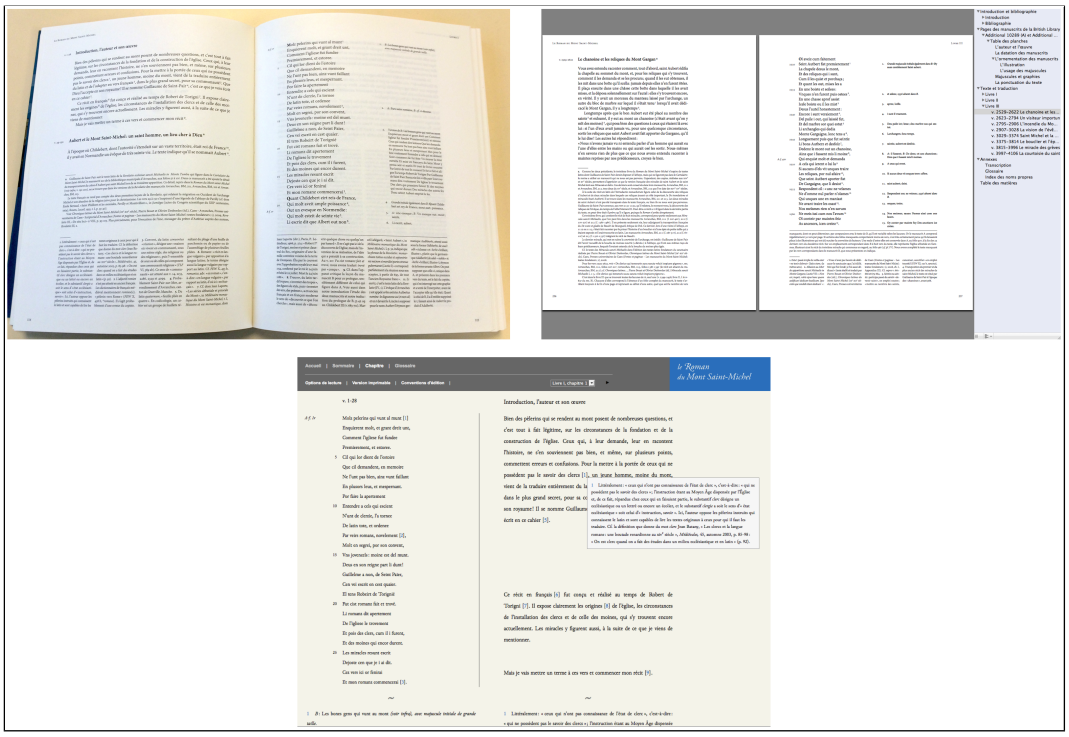

FIGURE 5.1 – Les trois formes de diffusion du *Roman du Mont Saint-Michel*.

lité des textes, complétée par une introduction et des commentaires scientifiques. La version cédérom, commercialisée avec le livre, reprend l'ensemble de ces contenus en ajoutant un système de liens hypertextes pour faciliter la circulation dans le texte. La fonctionnalité du glossaire, en particulier, s'en est trouvée considérablement accrue. La version web ne propose que les textes des sources en ancien français versifié et leur traduction en prose en français moderne et les outils associés (glossaire interactif, notes de variantes et notes scientifiques). Dans toutes les versions, le texte original et la traduction sont alignés comme le montre la figure 5.1. Dans les formes papier et PDF (cédérom), côte-à-côte dans la partie supérieure de la figure, la traduction en prose est sur la page de gauche et la transcription sur la page de droite. La version web propose une organisation similaire avec la transcription à gauche comme le montre la partie basse de la figure.

L'édition de *Hortus Sanitatis*, traité latin d'ichtyologie de la fin du XV^{e} siècle, propose le texte original latin en regard de sa traduction en français contemporain sur deux supports, le papier et le web, avec un mode de répartition des types de textes (commentaires scientifiques et sources) très similaire.

Cependant, cette source se caractérise par la très faible quantité de passages originaux. Cette compilation de sources a donc fait l'objet d'un important travail

d'identification de chaque citation primaire et secondaire à l'origine chaque segment de texte. L'édition doit donc rendre compte de ce travail d'analyse pour permettre au lecteur de comprendre comment le compilateur de l'*Hortus Sanitatis* a construit son œuvre et comment (et pourquoi), dans certains cas, des erreurs se sont glissées dans les textes. Encore une fois, l'édition doit rendre le texte immédiatement utilisable par le lecteur quel que soit le support de diffusion que ce dernier a choisi pour accéder au texte. Ainsi, dans la version papier, les notes de marge de gauche du texte latin placées en face de chaque paragraphe, donnent les informations sur les origines de chacun des paragraphes et de chacun des segments qui les composent. Dans l'édition en ligne, ces informations sur l'histoire des fragments de texte sont données sous la forme de fenêtres *pop-ups* dont les points d'appel sont placés au début de chacun des paragraphes ou segments concernés.

Le traitement de l'illustration est différent dans le volume papier et dans l'édition numérique. Le volume papier reproduit en tête de chaque chapitre, dans une taille réduite, la vignette illustrative qui figurait à l'origine dans l'*Hortus Sanitatis* (très précisément ce sont les images de l'exemplaire conservé à la bibliothèque municipale de Valognes qui sont reproduites). La version en ligne permet de mieux rendre justice au rôle prépondérant joué par l'illustration dans la conception de l'*Hortus Sanitatis*. Les images de deux exemplaires numérisés sont accessibles au lecteur en cliquant sur les liens insérés directement dans le texte, aux endroits précis de chaque rupture de page. Ainsi, le lecteur connait très précisément le positionnement de chaque début et fin de page de chaque exemplaire numérisé et peut, très simplement, consulter la version en mode image de cette page originale.

Enfin, l'édition en ligne de l'*Hortus Sanitatis* propose au lecteur d'accéder aux textes d'une manière totalement différente, en instanciant de nouveaux documents (des pages web) à la volée en fonction de l'origine des fragments de citations réunis par l'auteur. Autrement dit, l'édition permet au lecteur de rassembler sur une même page l'ensemble des citations provenant de telle ou telle source et présents dans l'*Hortus Sanitatis*. Il suffit au lecteur de sélectionner dans le répertoire des citations une autorité pour composer un nouveau document rassemblant tous les fragments provenant de cette source et utilisés par le compilateur. La figure 5.2 propose une série de captures écran des différents modes de lecture proposés par les deux versions de l'édition de l'*Hortus Sanitatis*.

Une édition critique doit donc permettre au lecteur d'exploiter directement les textes traités par tous les moyens disponibles sur chacun des supports de lecture exploités. Ainsi, une édition papier peut fournir un excellent moyen de stabiliser un

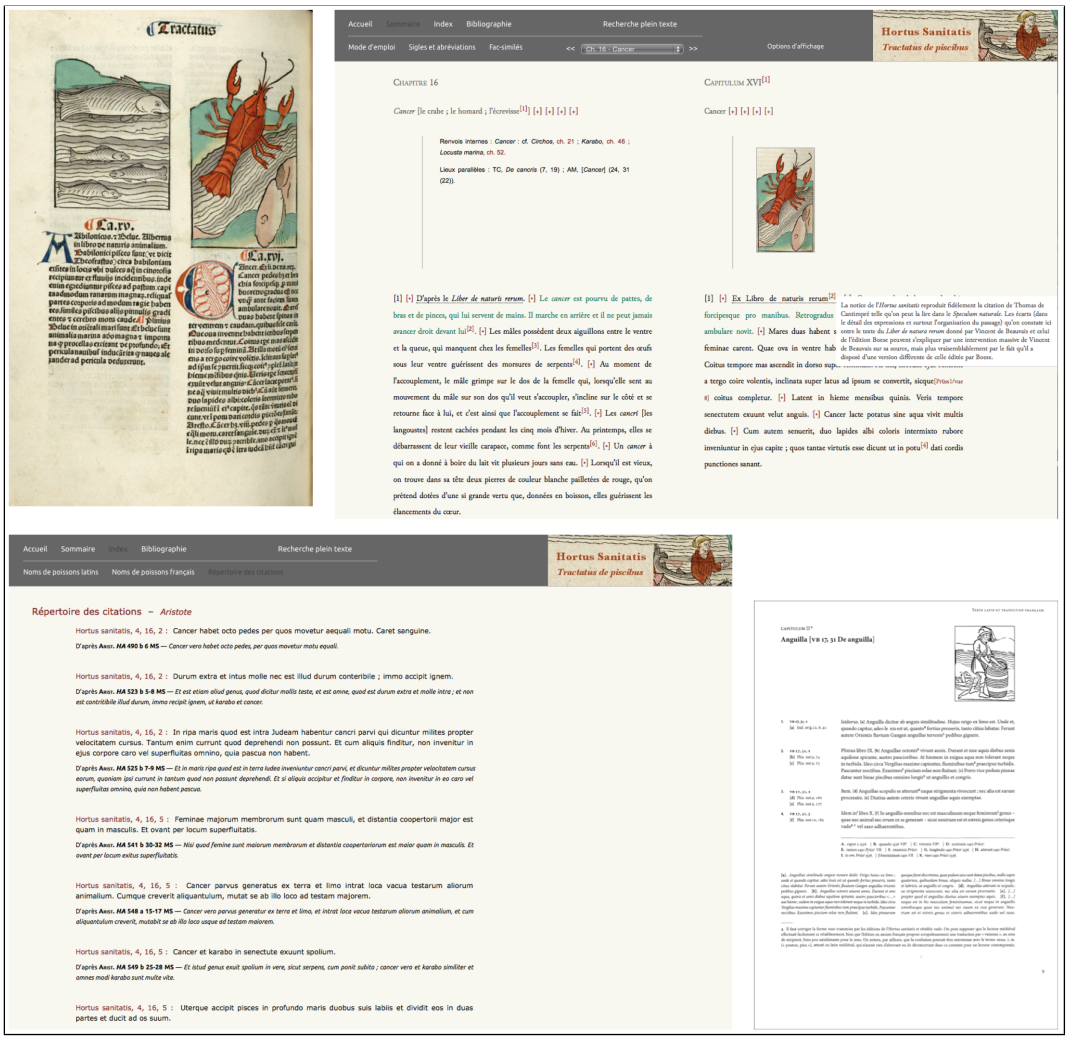

FIGURE 5.2 – Les modes de lecture de l'édition de l'*Hortus Sanitatis*.

système de citation, une édition en ligne peut proposer un accès aux images numérisées des témoins collationnés ainsi que des systèmes d'alignement ou de restitution des textes de l'édition, etc. La richesse des fonctionnalités de chacun des supports est étroitement liée à la qualité et à la finesse de l'annotation, ou de la caractérisation, des phénomènes textuels observés. C'est en effet la qualité des vocabulaires mis en œuvre pour décrire les éléments constitutifs des textes qui permet d'enrichir les outils d'exploitation et d'exploration intégrés dans les éditions. La discrimination des éléments constitutifs des textes les uns par rapport aux autres est un élément central qui impacte directement la qualité des éditions produites. Une large part de cette discrimination repose sur l'analyse préalable des textes réalisée par les spécialistes (historiens, philologues, etc.). Il est donc tout à fait indispensable d'associer le plus fortement possible ces experts aux opérations de structuration et d'annotation des

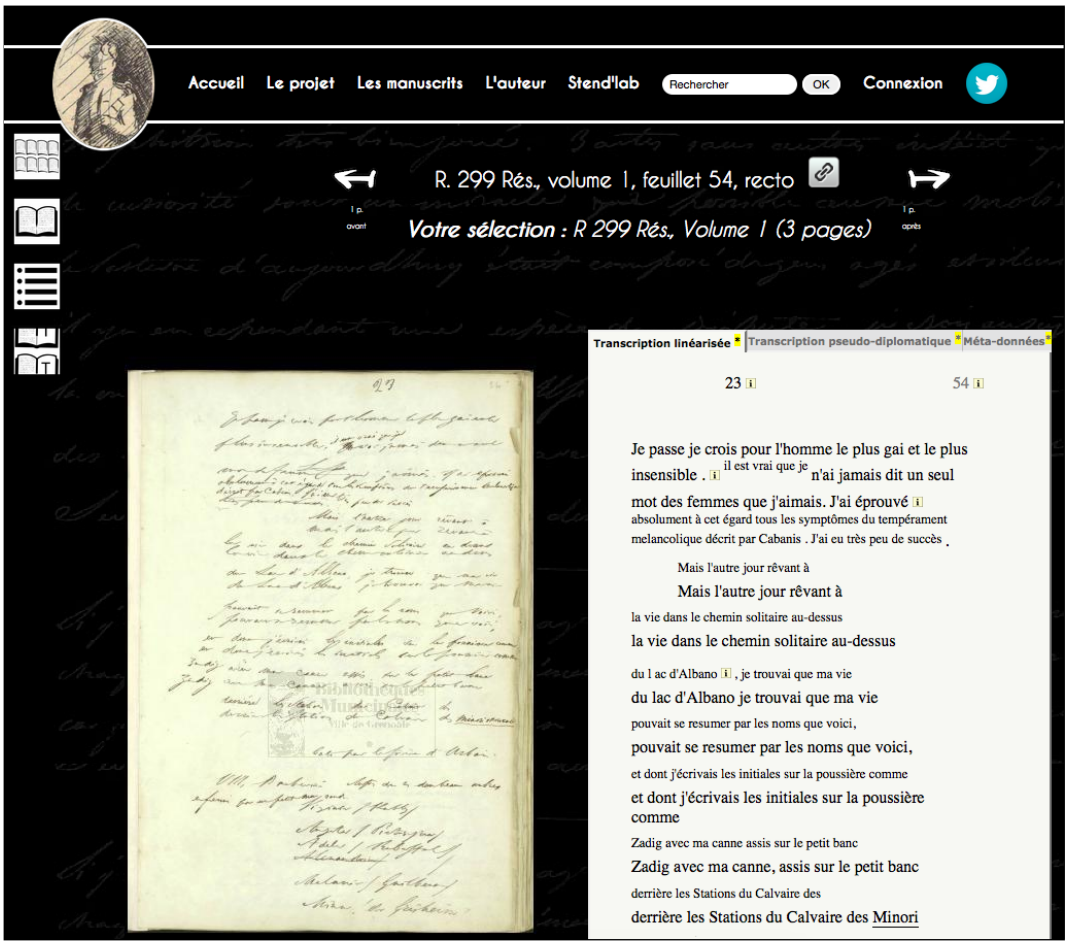

FIGURE 5.3 – L'interface de lecture des *Manuscrits de Stendhal*.

textes. Comme nous le verrons un peu plus loin, l'ergonomie des outils joue un rôle fondamental de ce point de vue.

L'exemple des *Manuscrits de Stendhal*[52] [LEBARBÉ *et al.*, 2008] montre une approche différente. En effet, l'approche des chercheurs dans le cadre de ce projet est fondée sur les unités matérielles : les pages manuscrites de l'auteur. L'ensemble de la modélisation est construite autour de ces objets matériels qui sont aussi la base de l'organisation matérielle du fonds puisque l'ensemble est agencé en fonction du format des pages et des cahiers [LEBARBÉ et MEYNARD, 2009]. Tout le travail de transcription et d'annotation a donc été guidé par l'organisation matérielle des objets étudiés. Cependant, des corpus littéraires ont été reconstitués en ajoutant des informations au cœur des transcriptions pour permettre le regroupement et l'affichage d'ensembles de pages liées thématiquement, par exemple, les unes au autres. Les transcriptions ainsi produites ont permis d'alimenter l'édition en ligne et de fabriquer un premier

52. http://stendhal.msh-alpes.fr/manuscrits/index.php

volume papier [Meynard *et al.*, 2013]. La figure 5.3 donne l'interface de lecture de la version électronique avec l'image numérique de la page manuscrite en regard de la transcription linéarisée (une version pseudo-diplomatique respectant les lignes et les abréviations est aussi disponible).

Dans ces perspectives d'exploitation multimodale et multisupport, tous les supports de lecture, des écrans au papier, sont autant de possibilités offertes au lecteur. Le papier ne concurrence en rien les dispositifs électroniques de lecture. Il s'agit au contraire de penser l'articulation de ces outils pour améliorer la qualité d'ensemble de l'édition et permettre au lecteur d'utiliser au mieux les textes édités en fonction de ses besoins, qui peuvent changer selon ses propres objectifs de recherche. Un même lecteur peut utiliser la version papier d'une édition pour une lecture immersive et la version en ligne pour une extraction d'information sur une thématique précise.

5.2 Objets manipulés

Les chercheurs manipulent des textes qui se lisent dans un ou plusieurs témoins. Ainsi un même texte peut se trouver à plusieurs endroits matériels. Ces textes peuvent aussi constituer des parties d'ensembles regroupant plusieurs des unités partageant une même thématique ou un même sujet. Les textes peuvent aussi participer de la même œuvre, en tant que production intellectuelle.

5.2.1 Œuvre

Dans le domaine de l'édition de sources anciennes, il convient de manipuler avec précaution la notion d'*œuvre* et de vérifier le sens que lui donnent les porteurs de chaque projet. Cependent, nous pouvons ici considérer l'*œuvre* comme l'entité abstraite produite par une personne indépendemment de toute forme d'instanciation : texte écrit, discours, etc.

Cette conception de la notion d'œuvre, ou cette catégorie conceptuelle, est aujourd'hui très répandue et nous pouvons trouver beaucoup d'exemples de définitions totalement en phase avec notre conception. Nous retiendrons ici par exemple la définition émanant du monde des bibliothèques et plus particulièrement du groupe de travail IFLA sur les *Functional requirements for bibliographic records* :

> *A work is an abstract entity ; there is no single material object one can point to as the work. We recognize the work through individual realizations or expressions of the work, but the work itself exists only in the commonality of content between and among the various expressions of the work. When we speak of Homer's Iliad as a work, our point of reference is not a particular recitation or text of the work, but the intellectual creation that lies behind all the various expressions of the work* [IFLA, 1999].

Ainsi, il est impossible d'accéder à l'œuvre en tant que telle. On accède à l'*œuvre* par l'une des *manifestations* d'une de ses *expressions*. On accède à l'œuvre d'un poète en lisant un exemplaire d'un recueil ou à celle d'un compositeur en écoutant une interprétation d'un orchestre. *L'œuvre* est donc réifiée d'une manière ou d'une autre et nul ne peut y accéder sans cette réification. Elle peut prendre plusieurs formes, s'intancier de plusieurs manières. Pour reprendre l'exemple d'un recueil de poésie, d'aucuns pourraient préférer accéder à l'œuvre *via* une version audio lue à haute voix par un acteur plutôt que de lire un recueil.

Les spécialistes des sources anciennes, utilisent également une notion tout à fait similaire, voire identique, pour désigner la création intellectuelle indépendamment de ses formes d'expressions (texte, discours oral, etc.). Ainsi M.-J. Driscoll propose la définition suivante [Driscoll, 2010] :

> *The "work", being an abstraction, is perhaps hardest to pin down. By "Hamlet, the work" I mean simply the sum of all the Hamlets that have ever been, printed, staged, filmed or otherwise manifested would say it was* [Driscoll, 2010].

Prenons ici l'exemple de l'auteur moderne qui travaille sur son œuvre : l'examen des manuscrits qui résultent de ce travail (les premières manifestions des œuvres en quelque sorte) montre à quel point l'instanciation d'une œuvre peut être complexe. Il suffit par exemple de consulter les pages d'un manuscrit de Samuel Beckett pour s'en convainvre [Hulle, 2008].

Si la notion d'œuvre est relativement intuitive, elle n'en est pas moins complexe à manipuler. En effet, quels sont les critères qui permettent d'attribuer une œuvre à un auteur ? Dans beaucoup de cas, la réponse à cette question est évidente, mais à partir de quelle quantité de modifications assiste-t-on à la création d'une nouvelle œuvre plutôt qu'à une recopie ? Cette question, qui échappe aux notions contemporaines d'œuvres et d'auteurs, garde toute sa pertinence pour les sources anciennes.

Nous appelons donc *œuvre* une abstraction regroupant l'ensemble des réalités qu'un texte a pu prendre au cours de son histoire sans qu'aucune d'elles ne suffise à

désigner l'œuvre dans son entier. Une conception heuristique qui s'oppose à la notion d'original peu adaptée aux types de travaux qui nous intéressent ici.

5.2.2 Texte

Nous désignerons par *texte* l'expression d'une œuvre indépendamment de tout support. Ainsi un texte pourra aussi bien être porté par un manuscrit, un incunable ou un livre imprimé.

Le texte présente une série de mots ordonnés selon un ordre précis. Ces mots constitutifs du texte peuvent être analysés de deux manières différentes et complémentaires. D'une part, il est possible de se concentrer sur le sens du texte, sur ce que l'auteur veut dire, sur ce que W. W. Greg propose de nommer la substance du texte (*subtantives*) [Greg, 1966]. D'autre part, le texte peut être considéré de façon beaucoup plus pratique comme une série de mots sur un support, quelle que soit sa nature, c'est-à-dire dans sa dimension purement matérielle. Cette matérialité du texte, *les mots sur la page* [Driscoll, 2010], appelée *accidentials* par W. W. Greg [Greg, 1966], recouvre donc toutes les formes tangibles (ou intangibles quand le texte est lu sur un écran) que peut prendre un texte au cours d'une instanciation sur un support donné. Tous ces *accidentials* qui peuvent survenir au moment de la production d'un texte sur un support : modification de l'ordre des mots, fautes d'orthographes, ponctuation problématique, abréviations... peuvent être considérés comme relevant de la pure matérialité du texte car, si ces *accidentials* peuvent rendre plus délicate la compréhension de la pensée de l'auteur, l'accession au sens, ils ne changent pas pour autant la nature profonde de l'œuvre exprimée par le texte.

Dans le cas des sources anciennes, on peut donc considérer que l'ensemble des dégradations subies par un manuscrit ancien tel que les dégâts dues à l'humidité, l'effacement des encres, ou les morceaux de parchemins déchirés, etc. sont autant d'*accidentials* : le texte, dans son essence, le texte idéal en quelque sorte, n'en est pas pour autant changé.

La difficulté peut commencer à apparaître lorsqu'une intervention humaine est la cause de la variation. Ainsi, l'ensemble des variations ou des erreurs introduites par le copiste d'un texte, au moment de la copie d'un manuscrit vers un autre, doivent-elles être considérées comme des événements de pure forme ou des changements de sens ? On voit bien ici qu'il est tout à fait impossible de fixer des principes généraux et que l'analyse devra être menée au cas par cas, variante par variante, pour décider si le sens du texte est modifié ou s'il subit simplement quelques *accidentials*. Dans tous les cas,

il sera probablement indispensable de considérer, en définitive, deux textes : le *texte réel*, porté par le support matériel, et le *texte idéal* [Driscoll, 2010], ce que l'auteur veut dire. Tout le problème réside dans le processus de (re)constitution du *texte idéal* à partir *du*, ou plus exactement, dans le cas des sources anciennes, *des textes réels*. Le résultat de ce travail sera toujours à manipuler avec une grande précaution dans la mesure où il sera systématiquement fondé sur des interprétations. Une large partie du travail d'édition de sources consiste précisément à donner explicitement la méthode choisie pour établir le texte dans le but de permettre au lecteur de saisir, à chaque étape, les choix opérés par l'éditeur et les critères qui ont présidé à ces choix. Cette méthode de travail dans l'établissement d'un texte idéal est encore beaucoup appliquée aujourd'hui dans le cadre des éditions critiques.

La nouvelle philologie propose, pour aller vite, de considérer les textes dans leur matérialité, c'est-à-dire d'étudier les textes réels que nous avons déjà évoqués, pour les décrire très précisément. Dans ce cadre, chaque version du texte est placée sur un même plan. Ici, il n'est plus vraiment question d'établir un texte idéal mais plutôt de témoigner de l'ensemble des différences existant entre toutes les versions d'un même texte. La connaissance de l'œuvre ne pouvant alors s'acquérir qu'en lisant l'ensemble des textes éventuellement dans une version superposant toutes les versions : factorisant les dénominateurs communs et donnant partout où elles existent les différences de telle ou telle version. Dans cette approche, les supports matériels, sur lesquels nous reviendrons plus loin, ont une importance tout à fait considérable, puisqu'on reconnaît le texte et son support comme totalement indissociables. Le présupposé suivant est ici appliqué à la lettre : on ne peut comprendre un texte indépendamment de son support.

Dans le cadre de ce mémoire, nous appellerons *texte* l'ensemble des littéraux constitutifs du texte indépendamment de son instanciation sur un support donné.

5.2.3 Support matériel

Nous avons déjà évoqué plus haut[53] la notion de support. Les caractéristiques objectives n'ont donc pas besoin d'être redonnées ici.

Cependant, il est important de noter que le rapport entretenu par l'archiviste avec les supports matériels n'est pas de même nature que celui que le chercheur entretient avec lui dans le travail d'édition de sources. Ainsi, si l'archiviste se préoccupe du contenu des documents pour en rationaliser le classement et la conservation et

53. Voir. p. 63.

pour en faciliter le repérage et l'identification, le chercheur, dans le contexte d'une édition de sources, s'intéresse au texte porté par les manuscrits. Ainsi les caractéristiques physiques des manuscrits l'intéressent d'abord de ce point de vue. Quels sont les textes qui environnent le texte étudié ? Le manuscrit est-il décoré sur les folios sur lesquels le texte étudié est reproduit ? Autrement dit, les informations sur la matérialité du support de texte seront le plus souvent très fortement orientées par la nature du texte étudié et l'ensemble des informations disponibles sur un manuscrit sera majoritairement interprété à l'aune de ce même texte. Le support a donc une importance capitale en tant que contexte d'occurrence du texte, si l'on peut dire, et c'est sur les relations entre le texte et son support que le chercheur va, dans le cadre d'une édition de sources, concentrer son travail. En définitive, et comme nous l'avons déjà évoqué, le support matériel est le plus souvent, dans le cadre d'une édition de sources, considéré comme un témoin et un espace de dépôt du texte avec sa propre organisation et son propre processus de production.

5.3 Importance du support papier

Une édition de sources est un travail long et difficile qui compte beaucoup dans la carrière d'un chercheur. Il est donc tout à fait capital pour un chercheur qui se lance dans une telle entreprise d'être certain de pouvoir en tirer tous les bénéfices. Or, à l'heure actuelle, même dans le contexte de forte poussée du numérique et des nouveaux dispositifs de lecture électronique, les livres imprimés conservent une place centrale de ce point de vue. À tort ou à raison, le livre papier est toujours un critère pour évaluer la qualité d'une édition de sources ; comme si le support électronique était réservé aux travaux de moindre qualité ou de moindre importance.

Ainsi, pendant le travail de production il faut tenir compte de l'impératif de production d'une version papier en respectant les règles en vigueur dans les communautés qui vont en évaluer la qualité. Autrement dit, la version papier n'est pas seulement produite à cause des qualités objectives de ce support (pas de consommation d'énergie une fois produite, confort d'utilisation, etc.) mais aussi pour des raisons d'évaluation des travaux au sein des communautés.

5.4 Convergence numérique : impact sur les chercheurs

L'arrivée des interfaces WYSIWYG a provoqué un bouleversement dans les méthodes de travail des chercheurs. Le fait qu'il soit possible de produire une matérialité en-

tretenant une relative proximité avec le résultat final, le livre, dès les premières étapes de conception du texte n'a pas été sans conséquence. Les chercheurs se sont, comme nous l'avons déjà évoqué, trouvés quelque peu perturbés par la présence prématurée d'une forme donnée au texte en rédaction. La synchronicité entre la production du fond et la production d'une forme détourne en effet l'attention de l'auteur [Coombs *et al.*, 1987].

Mais dans le mouvement de la convergence numérique, la production informatisée des textes n'est qu'une étape. Depuis plusieurs années, le numérique concerne aussi la sphère de la lecture et un même contenu peut-être consulté et lu sur plusieurs supports. Si le chercheur construit et pense son texte avec en point de mire un seul de ces supports, le livre le plus souvent comme l'y incite la majorité des outils utilisés en sciences humaines et sociales, les autres formes de diffusion risquent d'en pâtir considérablement. Par exemple, la construction d'un index basé sur des références à des numéros de pages pose d'incontestables problèmes pour le rendre opérant sur une page web ou un livre électronique. S'il est toujours possible de laisser à un autre le soin de prendre en charge ces opérations, un éditeur matériel *a priori*, la richesse des exploitations s'en trouvera largement limitée et le temps de production global risque fort d'être allongé.

La figure 5.4 donne un bon exemple de l'effet pervers des outils WYSIWYG qui détournent l'auteur de l'établissement du fond indépendamment de toute forme de diffusion. Il s'agit du "tapuscrit" remis par l'auteur du *Roman du Mont Saint-Michel*. Cette édition propose une transcription du texte original en ancien français versifié et une traduction en prose en français contemporain. Dans les formes de diffusion ces deux versions seront en vis-à-vis. Dès les étapes de rédaction l'auteur a composé son document en intercalant des séquences en ancien français et des séquences en français moderne. Bien entendu le volume occupé par ces deux langues est totalement différent et rapidement il devient impossible de conserver un rythme d'apparition cohérent assurant la correspondance entre un passage en ancien français et sa traduction en français moderne. Les notes sont toutes traitées indifféremment, même si elles sont de nature profondément variables : notes de variantes ou commentaires scientifiques. Le logiciel de traitement de texte utilisé ici est tout simplement incapable de réaliser correctement le type d'opérations véritablement nécessaires. Il est difficile de lui reprocher cet état de fait : cet outil n'est tout simplement pas adapté à la tâche. Tous les problèmes ne sont pas des clous, même si l'outil à disposition est un marteau...

Mais la figure 5.4 révèle aussi tout simplement l'impasse dans laquelle se trouve l'auteur : quelle autre solution mettre en œuvre pour réaliser la traduction d'un texte

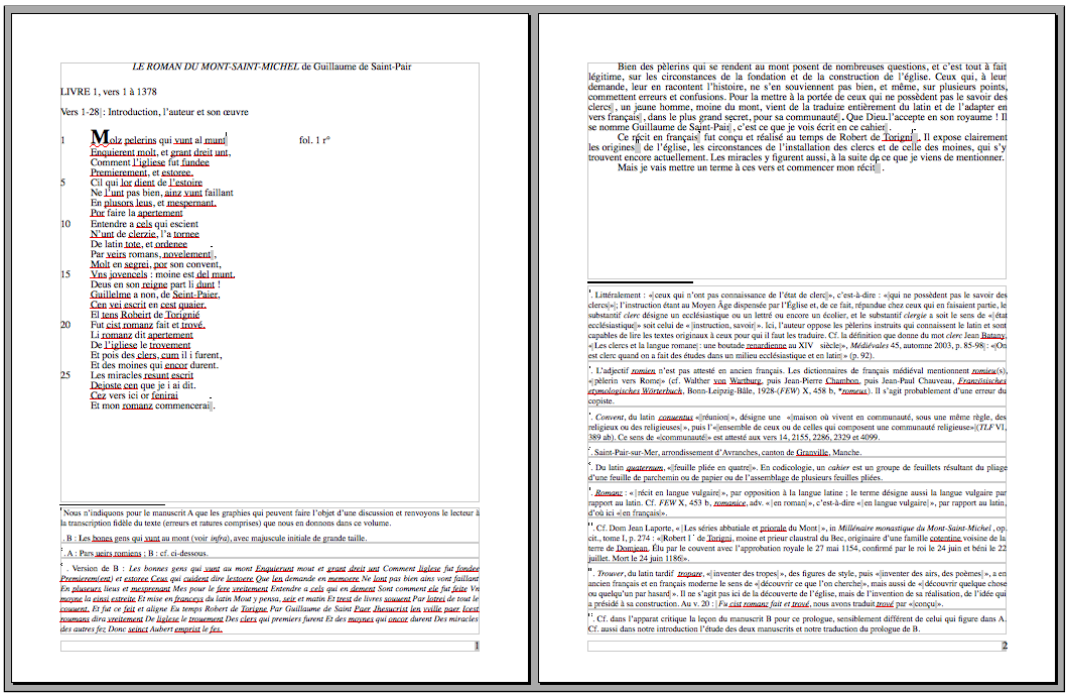

FIGURE 5.4 – Extrait du tapuscrit d'auteur du *Roman du Mont Saint-Michel*.

dans les meilleures conditions sans être contraint de se former à l'utilisation d'un outil trop complexe ? Nous revenons ici à la nécessité de proposer aux chercheurs des solutions pour établir le fond sans être perturbé par la forme que nous avons traitée plus haut [54]. L'outil idéal de manipulation de texte

> n'exige pas que l'utilisateur joue à la fois le rôle de l'écrivain, de l'éditeur, du maquettiste et du typographe. Il laisse l'auteur exprimer ses intentions et permet à chacun de jouer son rôle au cours de la vie du document dans la chaîne éditoriale [QUINT, 1987].

La convergence numérique bouleverse donc les habitudes, et les auteurs doivent en quelque sorte apprendre, ou réapprendre [55], à travailler et à préparer leurs textes indépendamment d'une forme de diffusion spécifique.

54. Voir p. 49.

55. En effet, avant l'informatique, l'écriture manuscrite n'entretenait au final que très peu de rapport avec les formes de diffusion.

5.5 Modèles de représentation de textes

5.5.1 Projets

Il existe de nombreux efforts de modélisation des textes dans le domaine de l'édition. Étant donnée l'ampleur de la tâche, beaucoup de projets limitent les approches à un type spécifique de sources ou à un angle scientifique précis. Nous donnons ici quelques exemples illustrant différentes approches.

Avec une approche très fortement liée à la matérialité du texte, le projet *Armarius* [Doumat *et al.*, 2008] propose un modèle organisé autour de l'image de la source sur laquelle des zones peuvent être définies pour être ensuite associées à des annotations. Le texte est donc ici considéré comme une information associée à une image de manuscrit.

Le projet *Shared Canvas*[56] [Sanderson *et al.*, 2011] propose une approche similaire avec un objectif d'interopérabilité des données très forte. Cependant, les *canevas* manipulés ne sont pas directement liés aux images et agissent comme des réceptacles à l'ensemble des annotations qui peuvent être des coordonnées d'images par exemple, mais aussi des commentaires scientifiques, des transcriptions, etc.

Le projet CLELIA (Corpus littéraire et linguistique assisté par des outils d'informatique avancée) s'intéresse plus particulièrement aux manuscrits modernes et propose une approche articulant le texte avec son support matériel. Le prototype s'appuie sur les manuscrits de Stendhal, que nous avons évoqués plus haut[57], et propose des solutions de balisage et d'annotation ainsi que des systèmes d'organisation des pages et des textes portés par ces pages. Il s'agit donc d'une véritable plateforme de traitement et d'exposition de données quelque soit leur nature : textes ou images.

Mentionnons pour terminer l'exemple du projet *Fragmentary Texts*[58] qui prend pour objet d'étude les fragments de textes trouvés sous forme de discours rapportés ou de citations dans d'autres sources [Romanello *et al.*, 2009b]. La caractéristique principale de ces fragments textuels est donc l'absence de support matériel donnant des informations sur leur contexte "originel" en quelque sorte. Autrement dit, *Fragmentary Texts* s'intéresse aux textes pour lesquels nous ne disposons plus de témoins directs. Bien entendu, la spécificité de l'objet a poussé les chercheurs à proposer un modèle centré sur le fragment de texte et son histoire. Ils donnent ainsi des solutions d'exploitation du modèle pour collecter et identifier les

56. http://www.shared-canvas.org
57. Voir p. 76.
58. http://www.fragmentarytexts.org

fragments [Romanello *et al.*, 2009a], [Berti *et al.*, 2009] dans les sources encore consultables aujourd'hui et s'appuient sur le protocole *Canonical Text Services*[59] pour fournir des identifiants uniques aux fragments [Smith, 2009].

5.5.2 *Text Encoding Initiative*

Parmi les multiples efforts de modélisation des textes dans le domaine des humanités au sens large, la *Text Encoding Initiative* (TEI)[60] exige un examen particulièrement attentif. En effet, ce projet s'est fixé des objectifs tellement ambitieux qu'il est indispensable d'y accorder toute l'attention nécessaire.

La TEI est un ensemble de recommandations pour la structuration de l'intégralité des textes dans le domaine des sciences humaines. Il s'agit de proposer des vocabulaires de description et de dénomination ainsi que des systèmes d'articulation de tous les phénomènes textuels observés dans le monde des humanités. Lou Burnard propose d'étendre encore le périmètre de la TEI à l'ensemble des documents, tous supports confondus :

> *The TEI emphasizes what is common to every kind of document, whether physically represented in digital form on disk or memory card, in printed form as book or newspaper, in written form as manuscript or codex, or in inscribed form on stone or wax tablet. This continuity facilitates the migration of text from older manifestations such as print or manuscript to newer ones such as disk or display* [Burnard, 2014].

Historique et périmètre

Lancée en 1987 par l'*Association for Computers and the Humanities*, l'*Association for Computational Linguistics* et l'*Association for Literary and Linguistic Computing*, la TEI se fixe donc pour objectif de fournir à la communauté des listes de mots pour caractériser les types de textes rencontrés dans les sciences humaines en décrivant précisément la nature des phénomènes textuels mais aussi les relations possibles avec l'ensemble des autres. Ainsi la TEI propose un système cohérent composé de collections de mots entretenant des relations claires entre eux pour la structuration de tous les types de textes rencontrés dans les sciences humaines, de la pièce de théâtre au dictionnaire en passant par l'article de littérature et la transcription de manuscrit médiéval.

59. http://wiki.digitalclassicist.org/Canonical_Text_Services.
60. http://www.tei-c.org.

L'objectif initial est donc bien d'apporter une réponse théorique à la question de la discrimination univoque des phénomènes textuels observés et manipulés dans les sciences humaines. Le projet est bien de mettre en place un vocabulaire commun, c'est-à-dire un modèle en réalité, avant de proposer une solution technique.

Pour autant et très naturellement, ces ensembles de vocabulaires, pour être efficaces, doivent être implémentés sous une forme exploitable par les communautés concernées. Ainsi, en 1987, les premières implémentations de la TEI ont été produites en SGML puis adaptée en XML après la sortie de la norme en 1998. Le standard TEI a donc déjà fait la preuve de sa capacité à suivre les évolutions techniques, ce qui pour tous ceux qui l'utilisent ne peut être qu'extrêmement rassurant : investir du temps dans l'appropriation des recommandations de la TEI est un gage de stabilité malgré l'évolution des outils et des techniques.

Le vocabulaire général est organisé en groupes thématiques ou fonctionnels, les *tagsets* ou modules. Il existe ainsi un module de base par défaut contenant ce qui est *a priori* le plus commun dans les textes rencontrés dans les humanités (divisions, paragraphes, titres, notes, etc.). Vient ensuite s'ajouter à ce module par défaut un certain nombre d'extensions thématiques comme les vocabulaires pour les dictionnaires, l'apparat critique, la transcription de textes oraux, la description de manuscrits, etc. Cette organisation facilite considérablement la compréhension de l'ensemble en permettant d'y accéder par blocs, mais elle simplifie aussi la production des outils associés.

Organisation

Le consortium qui gère aujourd'hui les vocabulaires est issu des communautés d'utilisateurs de la TEI. Ainsi, ce sont les acteurs eux-mêmes qui président aux décisions d'évolutions de la grammaire de référence. Cette organisation sociale doit être perçue comme un garant de la pérennité des recommandations. En effet, si un acteur estime que la direction prise par des évolutions n'est pas la bonne, il n'a pas à sortir du standard pour en fonder un autre, solution extrême certes, mais qui peut parfois être la seule. Il peut au contraire s'investir dans les structures de décision et faire valoir son point de vue pour faire évoluer le vocabulaire dans une direction qui lui semble la meilleure. C'est incontestablement l'une des seules solutions pour construire et maintenir une langue commune en évitant la prolifération de projets concurrents et à terme de standards qui n'en portent plus que le nom. La TEI est

donc un ensemble de recommandations pour les sciences humaines organisé et géré par les communautés de sciences humaines elles-mêmes.

D'un point de vue plus technique, en suivant des principes de modularité, la TEI propose aux utilisateurs de sélectionner des mots ou des groupes de mots nécessaires à la structuration de textes dans un cadre précis. L'idée part du constat relativement simple qu'aucun texte n'exige la mobilisation de tous les vocabulaires gérés par les recommandations. Donc, pour un ensemble de textes clos et identifiés à minima, il est possible de choisir les vocabulaires nécessaires et de produire des outils de description techniques (DTDs, schémas, etc.), et la documentation qui les accompagne, qui vont permettre la mise en œuvre concrète de la structuration. La sélection se fait d'abord par module puis, à l'intérieur de chaque module, mot par mot si c'est nécessaire. Les outils proposés par le consortium rendent ces opérations très simples et l'enrichissement d'un vocabulaire lacunaire se fait très rapidement. Ainsi, si les opérations de structuration révèlent un manque dans le vocabulaire pour annoter correctement un phénomène textuel, il suffit simplement d'ajouter le ou les termes nécessaires et de produire un nouveau schéma enrichi des dernières modifications. La figure 5.5 présente une vue de la page d'accueil de Roma, l'outil de production de grammaire de référence et de documentation proposé par le consortium TEI, avec la liste des modules existants.

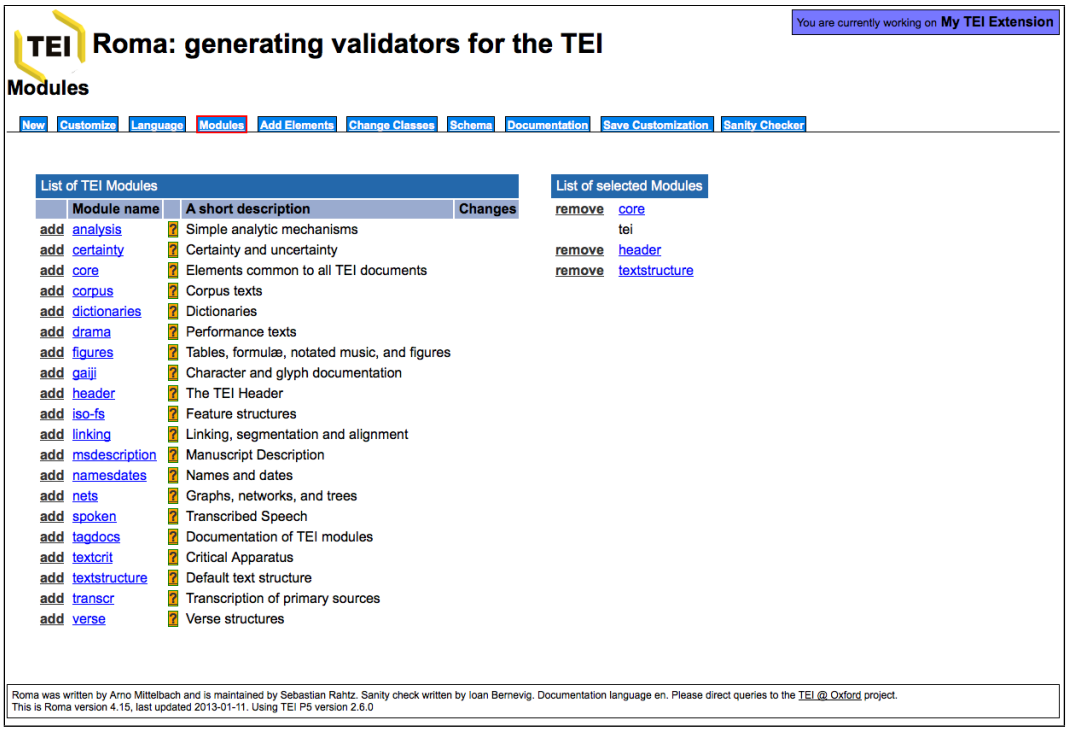

FIGURE 5.5 – L'outil Roma du consortium TEI.

La TEI est un standard aujourd'hui extrêmement utilisé dans le domaine des humanités dès qu'il s'agit de constituer des ressources numériques. Les communautés l'utilisent maintenant très massivement et s'écarter de la TEI implique de se couper d'une grande partie de l'activité, ce qui serait contre-productif dans la mesure où la dimension d'interopérabilité est de plus en plus souvent considérée plutôt comme un impératif catégorique que comme une évolution souhaitable.

Par ailleurs, il est assez difficile d'analyser les raisons pour lesquelles les porteurs de projet comprenant un volet de structuration de données pourraient décider de se priver de 35 ans de recherche et d'analyse en choisissant de s'écarter des recommandations de la TEI.

5.6 Environnements techniques

Le consortium TEI propose des solutions pour constituer des modèles de représentation adaptés à des cas précis ainsi que des outils pour les rendre opérants. Une fois ces modèles opérationnels, quelles sont les solutions techniques pour la manipulation et la structuration des textes valides contre les modèles produits ?

Nous proposons ici de distinguer deux grands types d'approche, qui ne sont pas nécessairement exclusives : le traitement de texte utilisé à des niveaux d'expertise variables et la manipulation "directe" des instances XML. Le choix de l'approche est très lié à la complexité du modèle de texte établi.

5.6.1 XML sans le savoir

Lorsque le niveau d'encodage souhaité est simple, c'est-à-dire lorsque les catégories textuelles abstraites sont relativement peu nombreuses, il est tout à fait possible de mettre en place très rapidement des solutions simplifiant la production de flux de textes structurés.

Il est en effet possible d'utiliser un logiciel de traitement de texte pour produire un encodage éditorial du texte en utilisant les fonctions de styles de paragraphes et de styles de caractères des logiciels modernes connus de tous aujourd'hui. La première étape consiste donc, en référence à un modèle de représentation du texte, à développer des feuilles de style raisonnées, pour les paragraphes et les séquences de caractères, afin de discriminer les différents types de textes rencontrés. Il faut suffisamment d'étiquettes pour rendre compte des titres, des paragraphes de textes standards, des citations, des notes, des tableaux, des figures, c'est-à-dire des pointeurs vers des ressources iconographiques externes accompagnés de titres et éventuellement de

légendes, etc. Ces styles peuvent prendre en charge le sens de lecture (gauche/droite ou droite/gauche) en s'appuyant sur le contexte de séquences étiquetées. Le sens de lecture général est fixé à l'échelle du document traité, les styles dédiés se contentent de préciser si la séquence concernée doit voir son sens de lecture inversé par rapport à celui du document. Cette méthode permet de traiter un maximum de cas de figure : un document fonctionnant en droite/gauche contenant des séquences gauche/droite et réciproquement. Ainsi, une séquence de texte composée de droite à gauche dans un texte composé de gauche à droite sera gérée par la simple application d'un style.

La transformation de ces documents stylés se fait en utilisant les logiciels de traitement de texte modernes, en s'appuyant notamment sur le fait que leurs formats de fichiers internes sont aujourd'hui souvent en XML comme nous l'avons vu[61]. La norme *open document* d'OpenOffice.org par exemple offre une stabilité suffisante pour permettre le développement de feuilles de transformation pour passer d'un encodage *open document* à un encodage TEI. Une fois les documents passés en XML et respectant le formalisme XML du logiciel de traitement de texte utilisé, il reste à le transformer pour le faire correspondre au modèle de représentation de texte choisi en utilisant par exemple une feuille de transformation XSLT dédiée. La base de la production de l'instance réside dans la correspondance entre les styles et les balises XML auxquelles on souhaite parvenir au terme du processus de conversion. Le développement de la feuille de style est donc une étape capitale qui doit être réalisée avec le plus grand soin en respectant le modèle de représentation des textes.

L'exécution de la transformation peut être réalisée directement depuis l'interface du logiciel s'il le permet, comme c'est le cas de la suite OpenOffice.org qui propose de configurer autant de "filtres" d'exportation que l'utilisateur le souhaite. C'est la solution la moins coûteuse en terme d'infrastructure logicielle et matérielle puisque toutes les opérations sont exécutées localement sur une même machine.

Une autre solution est de mettre en place un serveur pour se charger de la transformation. C'est la solution retenue par le Centre pour l'Édition Électronique Ouverte (CLEO) avec le système OpenText [Dacos, 2010]. Il s'agit en réalité d'un webservice qui permet à l'utilisateur d'envoyer un fichier stylé avec les feuilles de style du CLEO. Une fois le fichier transmis, le serveur exécute l'ensemble des transformations. Le cœur du dispositif se compose en particulier d'une instance d'OpenOffice.org exécutée en mode serveur qui répond aux demandes de transformation.

61. Voir p. 57 et suivantes.

Cette seconde solution, plus coûteuse en terme d'architecture, offre cependant l'avantage de donner accès à plus de techniques de production et de manipulation des instances que la première qui repose exclusivement sur XSLT.

5.6.2 XML natif et les outils dédiés

Pour la mise en place d'un balisage complexe, donc d'un modèle de représentation plus riche, les styles sont totalement insuffisants et ne permettent pas un niveau de profondeur de description satisfaisant. Il est dans ce cas indispensable d'intervenir directement sur le texte encodé en XML. Travailler sur un texte balisé en XML implique, pour être réalisé dans les meilleures conditions, l'utilisation d'un outil dédié : un éditeur XML. Il en existe aujourd'hui plusieurs, souvent payant et proposant plusieurs logiques de manipulation. Les publics qu'ils visent sont parfois très différents et il convient de choisir son outil avec soin. Certains éditeurs proposent un accès direct au code XML, d'autres prennent en charge l'ensemble des aspects de syntaxe et imposent l'usage de fonctions dédiées pour la mise en place de la structure. En définitive, sur la question de l'encodage XML natif des textes, nous pouvons considérer que deux logiques ergonomiques, avec deux présupposés totalement différents, existent.

La première, la plus répandue, postule qu'il faut voir le code et les balises pour vraiment comprendre les opérations réalisées par l'utilisateur, qu'il s'agisse d'un chercheur, d'un archiviste ou de tout autre acteur. Autrement dit, il faut que ce soit l'utilisateur qui s'approprie les règles de syntaxe, éventuellement avec l'assistance d'un outil qui va guider ces opérations, signaler les erreurs de structuration ou de syntaxe, etc. La seule concession de forme acceptée par cette méthode d'accès direct au code, est l'indentation. Il s'agit en effet bien d'une mise en forme destinée aux humains, qui ne présente aucun intérêt technique pour les machines. Mais, nous verrons que cette mise en forme, qui ne devrait avoir aucun impact, peut pourtant présenter des problèmes d'exploitation lourds pour la diffusion sur support papier.

La seconde méthode accorde en fait moins d'importance à la technique et privilégie l'accès au modèle de représentation en plaçant les modalités informatiques (règles d'écriture et de syntaxe) au second plan. Il s'agit donc de proposer des interfaces qui permettent à l'utilisateur de savoir qu'il manipule un arbre de données, de voir la nature des éléments qui composent cet arbre, de pouvoir ajouter, modifier ou supprimer ces éléments et leurs contenus mais sans accéder directement au code XML correspondant. Ce sont les logiciels qui prennent en charge tous ces as-

pects. Cette approche implique une connaissance des contraintes de syntaxe, pour comprendre le comportement général des outils, mais pas une maîtrise de ces principes. C'est surtout la connaissance et la maîtrise du modèle de représentation des textes traités qui va permettre à l'utilisateur d'utiliser l'outil de manière raisonnée. Cette seconde approche favorise aussi le contrôle de la qualité des structures XML produites : le logiciel, pour fonctionner correctement, doit manipuler des arbres bien formés et n'autorise donc pas le passage par un arbre mal formé. Mais cette nécessité peut aussi devenir une contrainte : le passage par une structure mal formée pouvant permettre de réaliser certaines opérations plus rapidement.

Ces deux approches ne se limitent donc pas à de simples choix ergonomiques, mais renvoient aussi à des conceptions de l'accès aux données, et aux modèles de représentation abstraits qui leurs sont associés, ainsi qu'à la place accordée à la dimension technique dans les opérations d'édition et de saisie, et à la responsabilité laissée aux utilisateurs sur ce point.

La figure 5.6 donne deux exemples d'éditeurs XML. Le premier (partie haute de la figure), oXygen [62], s'oriente plutôt vers la manipulation du code. Même s'il est possible d'associer une feuille de style CSS à l'instance manipulée, il s'agit surtout de proposer une forme de base et non des solutions d'interactions complètes. Le second éditeur (partie basse de la figure), XMLmind XML Editor [63], interdit purement et simplement l'accès au code : l'utilisateur ne saisit aucune balise, ni aucun nom ou valeur d'attribut, mais utilise des fonctions dédiées du logiciel pour manipuler l'arbre XML. Pour l'utiliser dans les meilleures conditions, il est donc indispensable de créer des environnements de travail adaptés aux modèles de texte.

Le choix d'un éditeur XML ne doit pas être négligé. En effet, c'est un outil particulièrement utile qui peut rapidement devenir un point de blocage si la sélection ne correspond pas aux usages. Les outils plutôt dédiés aux développeurs ne sont, par exemple, pas forcément les plus adaptés aux chercheurs qui souhaitent avant tout travailler sur les textes.

L'utilisation d'un éditeur XML peut aussi intervenir dans une organisation du travail dans laquelle la structure XML de base est mise en place grâce à un logiciel de traitement de texte, comme nous l'avons vu pour l'encodage éditorial. L'éditeur XML n'est alors utilisé que pour l'enrichissement de la structure produite à l'export d'OpenOffice.org. C'est une solution qui peut considérablement simplifier la produc-

62. http://www.oxygenxml.com. Notons que le mode "author", proposant de manipuler du XML sans éditer le code directement, se développe avec les dernières versions.

63. http://www.xmlmind.com/xmleditor/

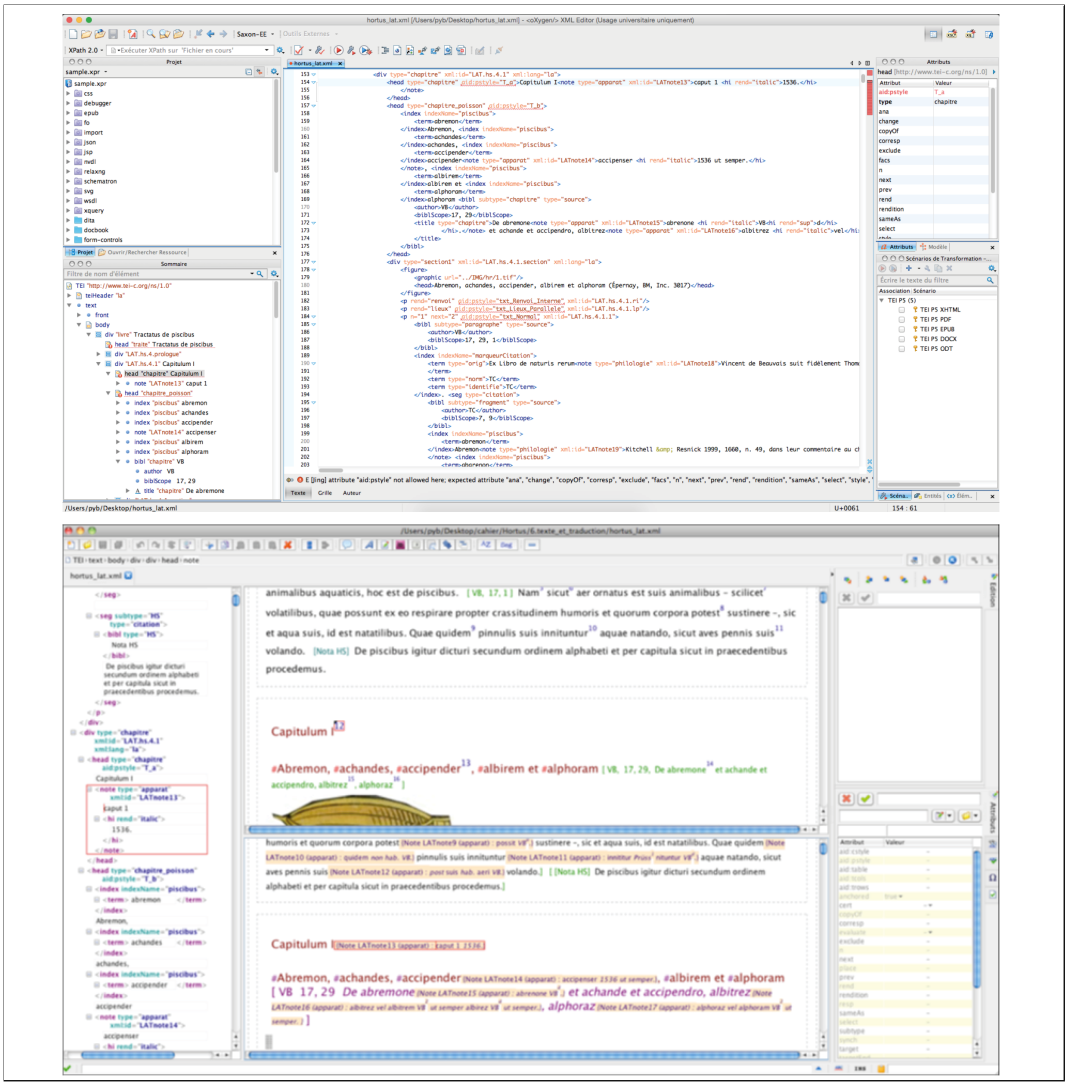

FIGURE 5.6 – Édition XML : aide à la saisie de code ou interface dédiée.

tion des flux structurés. Cependant, le développement ou l'adaptation des feuilles de transformation peut prendre beaucoup de temps, il convient donc de mesurer précisément le temps de développement et de le confronter au temps d'annotation manuelle directe afin de choisir la méthode la plus efficace.

Les modes de production sont naturellement très liés au niveau d'encodage : plus l'encodage souhaité est riche, plus les outils devront être perfectionnés. Le plus souvent un chercheur ne pourra pas faire l'impasse sur le balisage de certains éléments et sera contraint d'utiliser un éditeur XML pour annoter aussi richement qu'il le souhaite les textes sur lesquels il travaille. Cependant, nous verrons qu'il est possible de trouver des solutions logicielles pour rendre ce travail le moins long et difficile possible.

5.7 Conclusion

Les objectifs de recherche sont variables selon les projets et leurs problématiques, il existe donc des modèles et des pratiques différentes en fonction des approches. Malgré tout il existe un système de description qui tend à s'imposer dans le domaine : la TEI et plusieurs outils pour la manipuler. Les chercheurs disposent donc souvent pour leurs sources de fichiers XML TEI contenant les traductions, les transcriptions et toutes les annotations nécessaires à leur compréhension (mesures d'écarts, notes de commentaires scientifiques, etc.).

Cependant, l'importance du support papier pour l'évaluation des chercheurs ne semble pas faiblir. Or ce sont les éditeurs matériels qui assurent la production de cette forme, mais aussi sa diffusion, le plus souvent à travers des collections parfaitement identifiées par les communautés scientifiques. Mais les éditeurs matériels sont-ils en mesure de produire des livres à partir des fichiers XML TEI encodés par les chercheurs ? Leurs chaînes éditoriales ont-elles fait l'objet d'évolutions leur permettant de gérer la complexité descriptive de la recherche dans un processus production de livres complexes ?

6

Édition matérielle

6.1 Introduction

Beaucoup de chercheurs travaillent aujourd'hui en intégrant des solutions de structuration de données telles que la TEI dans la manipulation de leurs sources. Le résultat des travaux de recherche prend donc de plus en plus souvent la forme de fichiers XML richement encodés contenant la source elle-même ainsi que des renseignements sur la manière dont cette source a été étudiée. De plus ce fichier contient également l'ensemble du discours scientifique produit par les chercheurs.

Ces matériaux, c'est-à-dire les fichiers XML TEI de recherche, constituent donc parfois la matière textuelle que les éditeurs matériels doivent exploiter pour produire des formes de diffusion adaptées et en particulier la forme papier, dont les communautés de chercheurs se servent encore massivement pour l'évaluation des travaux.

Les activités d'édition matérielle sont généralement regroupées en trois grandes fonctions.

L'éditeur matériel a d'abord en charge la sélection des textes. Dans le monde de l'édition institutionnelle et scientifique, cela prend le plus souvent la forme de la mise en place d'un système d'évaluation et d'expertise dont la fiabilité participera de la crédibilité et de la réputation de la maison d'édition. Dans le monde académique ces experts sont en général des chercheurs reconnus dans leur domaine, le système prend alors une forme d'évaluation par les pairs.

La seconde grande fonction de l'éditeur matériel est la fabrication. Il s'agit ici de mettre en place l'ensemble des opérations nécessaires à la production des formes de diffusion et d'assurer, ou de faire assurer par des prestataires externes, leurs réalisations. Autrement dit, il s'agit de mettre en place une chaîne de production éditoriale allant de la relecture du texte et de la préparation, à la production des

différents supports de lecture. Quels que soient les acteurs réalisant les opérations, l'éditeur est celui qui assure le suivi de l'ensemble du processus et de sa cohérence. Cette fonction fera l'objet d'une attention particulière dans notre étude, en particulier du point de vue de l'évolution des techniques de production.

L'éditeur assume enfin les activités de promotion des produits éditoriaux. Cette activité centrale consiste à organiser la diffusion et la distribution des livres et des autres formes de lecture. Nous verrons que le numérique a un impact important sur cette sphère d'activité et exige de la part des éditeurs et de leurs partenaires une certaine capacité d'adaptation. En effet, la multiplication des supports de lecture n'est pas sans influencer la manière dont la promotion est assurée. De plus, les éditeurs doivent aussi trouver des modèles économiques satisfaisants pour articuler toutes les manières de lire les textes édités.

Dans ce chapitre, nous étudions les modes de production utilisés dans le monde de l'édition matérielle à travers les objets manipulés et les objets produits. Nous examinons également l'impact des techniques de structuration sur ce domaine professionnel ainsi que les normes existantes. Nous étudions enfin l'orientation générale vers le modèle de production du *single source publishing*, son influence sur l'activité de production et, dans une moindre mesure, de promotion.

6.1.1 Définition

L'activité d'édition matérielle recouvre un large spectre d'activités qu'il est nécessaire de préciser. Nous ne cherchons pas à désigner une réalité matérielle mais bien une activité humaine qui aboutit à la production de supports de lecture imprimés ou numériques.

L'édition matérielle comprend l'ensemble des opérations de normalisation et de mise en forme des textes. Ainsi la qualité de la typographie, la correction de copie, la complétude et la normalisation des références bibliographiques, la composition des pages d'une édition papier ou encore la bonne organisation graphique du texte et des figures relèvent de l'activité d'édition matérielle.

Dans le mouvement général de développement du numérique, c'est sans doute l'activité qui a le plus souffert. Avant que les ordinateurs, et les logiciels de traitement de texte en particulier, n'aient envahi aussi bien les sphères privées que les sphères publiques, tous ceux qui, à divers titres, participaient à un projet d'édition, avaient une assez bonne compréhension de la part qui revenait précisément à l'éditeur matériel. Mais les outils informatiques ont introduit une grande confusion qui

consiste à faire croire que produire un livre (sur support papier en particulier) revient en fait simplement à pouvoir justifier un texte dans un bloc à gauche *et* à droite. Tant que le papier était en fait le seul support de lecture mis à la disposition du lecteur, la confusion était assez difficile à lever.

Mais, bien entendu, la convergence numérique a poursuivi son développement en s'étendant aux autres *activités humaines autour du texte*, toute interaction avec un texte peut se faire via la médiatisation d'un dispositif technique numérique : l'écriture, comme nous l'avons déjà dit, mais aussi la production des formes de diffusion, et enfin la lecture depuis que nous disposons de liseuses, de *smartphones*, de tablettes, etc.

L'activité d'édition matérielle est toujours la même, le métier n'a, en définitive, pas du tout changé sur le fond. C'est au moment de l'édition matérielle d'un texte que se posent souvent les questions des supports de diffusion, en relation avec les publics visés. Mais si une même œuvre doit être diffusée sur plusieurs supports comment organiser le travail ? Doit-on mettre en place un flux de production par support ? Il est évident que c'est tout à fait impossible et totalement déraisonnable, ne serait-ce que pour des raisons économiques évidentes. Il convient donc de trouver des solutions au niveau de l'organisation du travail éditorial qui permettent de réaliser l'ensemble des opérations sur une *version pivot* du texte à partir de laquelle toutes les formes de diffusion pourront être produites, soit automatiquement, pour les versions consultées en flux, soit semi-automatiquement, pour les versions contraintes dans un espace fixé tel que la page d'un livre imprimé. Ce modèle d'organisation, couramment nommé *single source publishing*[64], est encore souvent considéré par les éditeurs comme un idéal vers lequel il faut tendre plutôt qu'une réelle pratique. Comme nous le verrons plus loin, il s'agit en fait d'une réalité incontournable pour réaliser le travail d'édition matérielle correctement, particulièrement dans le cas des sources anciennes qui nous occupe ici.

6.1.2 Édition et publication

Revenons un moment sur la notion d'édition au sens large pour préciser la nature de cette activité dans ses différences avec la publication.

La publication consiste à mettre à disposition du public un contenu éventuellement mis en forme. L'internet a considérablement favorisé le développement de la publication : on ne compte plus les réservoirs proposant un nombre de plus en

64. Nous donnons p. 113 une représentation de ce modèle d'organisation extraite du *Chicago manual of styles* [UCP, 2010].

plus important de textes à consulter. L'effort de publication s'accompagne souvent, comme nous le disions, d'un travail de mise en forme. Cependant, le texte n'est souvent ni relu, ni préparé, ni corrigé dans le cadre des opérations de publication. Il s'agit simplement de diffuser un contenu tel qu'il a été déposé par un auteur. Ces contenus ne sont pas nécessairement réorganisés en un tout cohérent éditorialement tel qu'un numéro de revue ou une monographie. Ainsi, la mise en ligne d'un ensemble de textes sans contrôle de structure, normalisation typographique et bibliographique ou prise en charge de la stabilité des adresses pour la référenciation, est une activité de publication. La publication peut ainsi se caractériser comme une opération de mise en forme des textes sans réelle préoccupation d'identification et de normalisation des catégories éditoriales à manipuler.

L'activité d'édition peut se définir en miroir par rapport à la publication : l'effort est justement porté sur les activités dont la publication fait l'impasse. Ainsi, l'édition se préoccupe de constituer des unités cohérentes, normalisées et stables afin de faciliter le travail de lecture et de permettre les références aux contenus traités, c'est-à-dire en réalité, de permettre à la *disputatio* universitaire de se développer dans les meilleures conditions possibles. Il est en effet tout à fait évident qu'un système de référencement stable est indispensable pour que les chercheurs puissent citer les textes, les confirmer ou les contredire, en un mot : les discuter.

6.2 Objets manipulés

L'édition matérielle impose la manipulation de beaucoup d'objets de nature très différente : livre papier, chapitre, articles, numéros de revue, bref des produits diffusés. Ces produits sont, la plus grande partie du temps, composés majoritairement de textes. Nous nous intéresserons donc à ces deux objets : le texte, remis par les auteurs du point de vue des éditeurs, et les supports de diffusion produits au terme des opérations d'édition.

Les éditeurs manipulent aussi de nombreuses métadonnées commerciales, c'est-à-dire des données commerciales sur les données éditoriales. C'est un enjeu capital pour les éditeurs qui fait l'objet de beaucoup d'attention et qu'ils doivent maintenant renseigner très tôt dans le processus de production, avant même que le travail d'édition ne débute, c'est-à-dire dès l'initiation du projet. Nous pensons ici en particulier à la norme ONIX que nous présenterons plus loin. Cependant, aussi centrales soient-elles, elles ne seront abordées ici que superficiellement, par le biais des modèles

de représentation qui les traitent, dans la mesure où elles ne sont pas au cœur du sujet qui nous occupe dans le cadre de cette étude.

6.2.1 Texte

Le texte en tant que tel présente des spécificités et des "résistances" quand il s'agit de le numériser [Burnard, 2012] sur lesquelles nous nous sommes déjà penchés plus haut [65]. Rappelons simplement que le texte articule, sans s'y limiter, des données et des informations supplémentaires qui viennent le qualifier et l'expliciter.

Le texte manipulé par l'éditeur ne présente pas de différence objective majeure avec celui manipulé par le chercheur. Cependant, le rapport au texte est d'une tout autre nature. En effet, l'éditeur matériel manipule les textes pour s'assurer de leurs cohérences logique et structurelle, réaliser les opérations de normalisation et produire les formes de diffusion. Contrairement au chercheur, il ne travaille donc pas directement à l'établissement du fond ; l'éditeur s'assure seulement que le texte à éditer est intelligible et il fabrique les formes les plus adaptées pour le diffuser. D'une certaine manière son rapport au texte est donc plus distancié, ce qui lui permet d'ailleurs d'en être le premier lecteur et d'être en mesure d'assurer ses tâches dans les meilleures conditions possibles.

L'éditeur repère dans le texte remis par l'auteur les catégories qui doivent être discriminées afin de faire l'objet de traitements formels particuliers. Si le texte est déjà porteur de catégories scientifiques, il doit sélectionner celles qui correspondent à de futures formes spécifiques. Son travail sur le texte, d'un point de vue technique, consiste donc à opérer des distinctions entre les différents éléments constitutifs des textes pour être en mesure de leur affecter des formes sur les différents supports qui seront utilisés pour acheminer le texte jusqu'au lecteur. Comme nous l'avons vu, il y a toujours un minimum de distinction opéré par les auteurs entre les éléments de texte, parfois signalés par de simples différences formelles (graisse, taille des caractères, espacements, etc.), mais qui renvoient toujours à des catégories. L'éditeur prend en charge la mise en cohérence et la systématisation de ces distinctions dans le cadre d'une collection et d'une politique éditoriale.

6.2.2 Support de diffusion

Tout texte diffusé l'est nécessairement sous une forme particulière et sur un support spécifique, qu'il soit imprimé ou numérique. Un texte va ainsi circuler soit sur les

65. Voir p. 79.

pages d'un livre, sur le web, dans un fichier PDF ou ePub, etc. Chacun de ces dispositifs techniques constitue un support de diffusion à part entière. Chaque support va imposer un certains nombre de contraintes et de qualités intrinsèques qu'il s'agira pour les différents acteurs d'exploiter au mieux pour atteindre leurs objectifs.

Nous désignons ici par *support de diffusion* tous les dispositifs techniques de lecture et de consultation qu'ils soient matériels, comme les livres ou les brochures, ou immatériels, comme les sites web ou les fichiers informatiques.

Web

Il semble ici important d'opérer une distinction dans les types de produits de consultation sur le web. En effet, le web est un espace très ouvert qui offre de nombreuses possibilités de diffusion. Nous proposons ici de différencier deux types de solutions de lecture en ligne : l'édition électronique et l'application web. Ces deux catégories ne sont ni opposées ni exclusives, il s'agit simplement de distinguer deux grandes familles pour éclairer l'ensemble des solutions existantes. Le premier type, celui de l'édition électronique, doit permettre de lire en continu des textes et de les citer de manière pérenne et le second, celui de l'application web, est quant à lui, pourvu de nombreux outils d'accès aux textes, de fouilles de données, d'outils de lecture dynamique, etc. Autant de fonctionnalités assez peu compatibles avec un système de référenciation simple et efficace à même de servir de base à la *disputatio* universitaire. Bien entendu, ces deux niveaux ne s'excluent pas nécessairement l'un l'autre d'un point de vue théorique. Cependant, dans la pratique, il est assez peu aisé de proposer des produits de diffusion richement outillés et faciles à citer. En effet, il peut être assez délicat de citer le texte d'une édition critique, par exemple, en précisant l'ensemble des options de lecture à activer pour permettre à un autre lecteur de lire le texte dans l'état dans lequel on souhaite le citer. En revanche, il est tout à fait envisageable de mettre en place plusieurs utilisations web d'un même flux de données. On peut par exemple envisager d'exploiter les données pour produire une édition électronique citable mais aussi, dans le même temps, d'exploiter ces mêmes données pour alimenter un laboratoire de textes pourvu de nombreux outils de fouille, d'extraction et de réorganisation mis à la disposition du lecteur. Tous les arbitrages entre ces fonctionnalités doivent être réalisés pour chaque projet par les responsables.

L'édition électronique est une forme de diffusion adaptée à la lecture immersive et linéaire d'un texte. Une édition électronique propose un accès aux textes relativement simple avec un nombre limité d'outils. Cependant, elle peut se voir enrichie d'index,

de tables des matières, de systèmes d'accès aux notes dynamiques, etc. C'est-à-dire d'un ensemble de petites solutions dont l'objectif est toujours de faciliter l'activité de lecture et l'accès aux textes dans ce but.

Une édition électronique doit permettre le référencement. Il peut être réalisé *a minima* sur la base des URLs ou en utilisant des identifiants pérennes du type DOI [66]. Dans les deux cas, on voit bien ici pourquoi des systèmes trop dynamiques peuvent être coûteux et complexes à mettre en place lorsque l'on souhaite permettre aux lecteurs de citer les textes et les fragments qui les composent.

Le référencement implique également, même si c'est une évidence, que le texte soit stabilisé au moment de sa mise en ligne. En effet, si le texte évolue après le début de sa diffusion, les citations existantes pourront se trouver sans objet, là encore, il s'agit d'une condition essentielle pour permettre une bonne circulation des discours sur le texte.

Une édition électronique est relativement simple à mettre en place à partir d'un flux XML peu complexe, c'est-à-dire d'un balisage éditorial [67] correspondant seulement aux catégories nécessaires pour les mises en forme. Bien entendu, il est tout à fait possible de fabriquer ce type de produit à partir d'un balisage scientifique en le ramenant à un niveau éditorial. En fixant un niveau de balisage connu et contrôlé (et en ramenant les formes plus complexes à ce vocabulaire), on peut donc prévoir des systèmes génériques pour ce niveau de complexité. Il existe d'ailleurs déjà plusieurs solutions : modules dédiés pour Drupal comme *TEICHI*[68] ou l'utilisation du module *Feed*[69] adapté à la TEI, Wordpress avec le module en cours de développement *WordPress-TEI-XML*[70], ou encore l'application *TEI Boilerplate*[71] qui se caractérise par sa simplicité d'utilisation et de mise en œuvre.

L'application web est un système de diffusion qui profite pleinement des évolutions technologiques ; en cela, nous pouvons considérer qu'elle constitue une évolution de l'édition électronique. En effet, en ajoutant de nombreux outils aux éditions électroniques, les applications web proposent toutes sortes de possibilités aux lecteurs, qui deviennent alors davantage des utilisateurs. On trouvera ainsi des solutions pour reconstituer différents états d'un texte en fonction des étapes de corrections et

66. Un *Digital Object Identifier* est un identifiant unique attribué à un objet numérique. Gérés par la fondation à but non-lucratif *International DOI Foundation* les DOI fournissent une solution pour assurer l'accès aux ressources numériques. Voir : http://www.doi.org.

67. Nous traitons en détails des différents niveaux de balisage p. 146.

68. http://www.teichi.org

69. https://drupal.org/project/feeds_xpathparser

70. https://github.com/davekelly/WordPress-TEI-XML

71. http://dcl.slis.indiana.edu/teibp/

d'amendements identifiées par les chercheurs. Une application web pourra aussi permettre de constituer des parcours de lecture en fonction de critères choisis par l'utilisateur. Les possibilités sont en définitive très ouvertes et les responsables de projet pourront mettre en place les plus adaptées aux textes diffusés et au projet scientifique. Ces applications web sont de véritables outils de recherche qui nécessitent cependant beaucoup de travail. En effet, plus les fonctionnalités sont nombreuses et spécifiques à un projet, plus le temps de développement (ou d'adaptation) d'outils sera long. De plus, un balisage très fin est souvent indispensable pour mettre en place ces fonctionnalités de manière satisfaisante.

Livre électronique et autres formats détachables

La notion de livre électronique désigne aujourd'hui le plus souvent les fichiers lisibles sur les plateformes de lecture comme les liseuses, les tablettes ou les *smartphones*. S'il ne faut pas oublier le format PDF qui reste très utilisé pour le téléchargement de textes et la lecture à l'écran, ce format de fichier ne permet pas l'adaptation au support sur lequel il est consulté et sort en partie du strict cadre de la notion de livre électronique. Les formats ePub (le plus couramment utilisé aujourd'hui), mobi ou azw pour des liseuses kindle d'Amazon renvoient plus directement à la notion de livre numérique.

La production de ce type de fichiers est assez simple et ne pose plus aujourd'hui de problème majeur. Il s'agit le plus souvent de fichiers d'archives (au format zip ou autre) avec quelques règles internes d'organisation des fichiers. De simples transformations XSL des flux de textes encodés en XML permettent sans aucune difficulté d'obtenir des livres électroniques qui respectent les normes. Il faut cependant noter que ces formats sont encore aujourd'hui très dépendants des dispositifs de lecture disponibles sur le marché. Si la norme est très claire sur un certain nombre de points, ce sont encore les industriels qui choisissent de respecter, ou pas, l'ensemble des règles formulées. La prudence commande donc de tester sur un maximum d'outils de lecture les fichiers produits afin de s'assurer de la qualité d'affichage. De plus, les mises en page complexes restent relativement risquées à l'heure actuelle pour les mêmes raisons de respect variable de la norme.

La plus grande difficulté pour contrôler la qualité éditoriale des livres numériques réside dans le caractère recomposable des fichiers diffusés. En effet, le lecteur reste maître d'un grand nombre de paramètres lors de l'affichage du texte sur son outil de lecture : choix de la police de caractères, taille du texte, etc. D'autres paramètres

sont directement liés au support de lecture, certains industriels imposent la taille des marges, la couleur des liens hypertextes, etc. Il faut donc tenir compte de ces contraintes lors de la production des livres numériques et faire en sorte d'orienter l'affichage plus que de le contrôler. C'est là l'enjeu majeur pour les éditeurs : apprendre à maîtriser des environnements de lecture en donnant des orientations formelles et non plus en fixant des règles de présentation et de consultation.

Enfin, le caractère recomposable des livres numériques provoque également la disparition de la notion de page au sens classique du terme, ce qui impacte aussi directement les habitudes de référencement. En effet, dans la mesure où c'est le lecteur qui fixe la taille des caractères affichés en fonction de ses propres critères, le nombre de pages d'un livre est variable d'une personne à une autre, il n'est donc pas envisageable de citer un ePub en faisant référence à une page précise. Restent donc seulement des solutions de référencement absolu (c'est-à-dire lié au découpage structurel du texte et non au découpage formel d'un livre) basées sur des numérotations de paragraphes qui apportent une réponse pragmatique, mais qui exigent de prendre l'habitude de citer tous les textes scientifiques comme les textes sacrés[72]... On peut aussi prévoir des solutions adoptant un codage des pages telles qu'elles figurent dans l'édition imprimée.

Papier

Le support papier reste un support important car il présente des qualités techniques tout à fait singulières qu'il ne faut pas écarter, comme le souligne Geoffrey Nunberg, cité par Frédéric Barbier :

> Si le livre avait été inventé après l'ordinateur il aurait constitué une avancée majeure. Ses qualités sont remarquables : légèreté, disponibilité, faible coût, fonctionnement sans consommation d'énergie, qualité d'affichage. De plus, le livre constitue une interface particulièrement bien adaptée à l'homme. Le cerveau de ce dernier possède en effet une excellente mémoire spatiale, qui lui permet de localiser approximativement une information ou une page après lecture [Barbier, 2000].

Ajoutons que sa conservation et sa diffusion sont parfaitement maîtrisées, et qu'il propose un système de référencement extrêmement stable et totalement intégré par les communautés de recherche, etc.

72. Les éditeurs et les lecteurs de textes classiques manipulent couramment des systèmes de référencement de ce type.

Cependant, contraindre un texte dans une planche papier n'est pas une tâche aisée et impose la maîtrise d'un grand nombre de paramètres. Outre le poids réel de la tradition du monde de l'édition, pour produire un volume papier agréable à lire, c'est-à-dire, dans le monde scientifique, un outil efficace, il est impératif de connaître et savoir appliquer les règles de composition typographique pour obtenir un gris efficace, ou encore d'être familier des règles spécifiques de tel ou tel type d'édition scientifique (édition critique de sources anciennes, édition génétique, etc.). Toutes ces règles ou traditions doivent être respectées pour que les volumes produits soient bien acceptés par le public visé et il faut bien reconnaître que les techniques numériques ne changent pas grand chose sur le fond de ce point de vue. Si les techniques ont évoluées, et continuent d'évoluer, le fond du métier n'a en revanche pas changé (et pas seulement pour la production papier).

6.3 Modèles et normes du domaine

6.3.1 Métadonnées

Les métadonnées peuvent être présentées comme des données qui visent à définir ou à caractériser d'autres données pour les référencer et les manipuler. Ainsi la *National Information Standards Organization* propose la définition suivante :

> Les métadonnées sont des informations structurées qui décrivent, expliquent, localisent ou encore facilitent la découverte, l'utilisation ou la gestion d'une ressource d'information [NISO, 2004].

Les métadonnées ont donc pour objectif de décrire les ressources numériques pour en assurer la meilleure exploitation possible dans différents domaines d'applications, de la recherche au commerce en passant par la valorisation. Nous avons déjà rapidement évoqué [73] des standards et des normes sur lesquels il convient maintenant de revenir pour préciser l'utilisation spécifique qui en est faite dans le domaine de l'édition.

La norme Dublin Core, ISO 15836 et NISO Z39-85, est utilisée pour le signalement et l'amélioration des résultats des moteurs de recherche. Comme nous l'avons déjà signalé, elle sert de base au protocole des archives ouvertes.

Pour un éditeur, c'est le moyen de proposer un accès à l'ensemble de son fonds avec un moteur de recherche unique soit pour obtenir des informations sur des res-

73. Voir p. 14.

sources (livres papiers, numéros de revues, etc.), soit pour y accéder directement lorsqu'elles sont diffusées en ligne. Dans le cas d'une diffusion sur plusieurs supports, il est possible que les unités ne soient pas les mêmes. Par exemple, chaque article d'un numéro de revue papier peut faire l'objet d'une diffusion en ligne autonome. Sur le papier l'unité est le numéro composé d'articles alors qu'en ligne chaque article est une unité indépendante. L'utilisation de la norme Dublin Core, et en particulier de sa capacité à décrire des relations entre des ressources, va permettre de conserver le lien entre les unités que sont le numéro de revue papier et les articles diffusés en ligne. Ainsi, un utilisateur sera directement informé que l'article qu'il est en train de télécharger a été diffusé dans le contexte d'un numéro de revue.

Ce type de résultats pourrait être obtenu avec beaucoup d'autres solutions, mais le grand avantage de Dublin Core est de simplifier considérablement le changement d'échelle. Ainsi construire un portail de recherche, en s'appuyant sur Dublin Core et le protocole OAI, pour un groupement d'éditeurs ayant tous intégré le Dublin Core dans leurs systèmes, est extrêmement simple sur le plan technique. Il existe d'ailleurs des solutions logicielles complètes et performantes dédiées à ce type d'opération comme *Open Harvester Systems*[74].

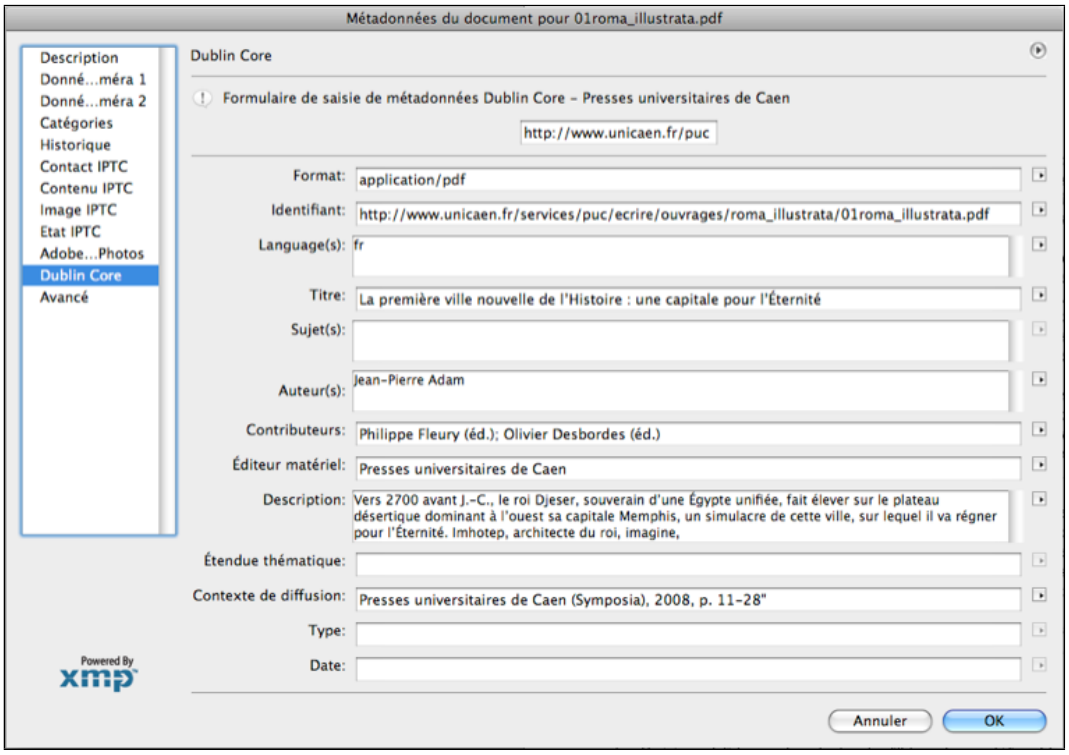

FIGURE 6.1 – Saisie de métadonnées Dublin Core dans un fichier PDF.

74. http://pkp.sfu.ca/ohs/

Enfin, l'effort d'intégration de métadonnées Dublin Core est relativement peu coûteux puisqu'il existe de nombreuses solutions pour les encapsuler dans les ressources elles-mêmes. La figure 6.1 donne un exemple d'un formulaire de saisie de métadonnées Dublin Core directement dans un fichier PDF. Bien entendu, il est aussi tout à fait possible d'injecter ces données automatiquement, le formulaire permettant simplement de les contrôler ou de les modifier en cas d'erreur.

Standard de description de la vie du produit éditorial du dépôt du manuscrit jusqu'à la diffusion, ONIX (*ONline Information eXchange*) [EDItEUR, 2014], émane du monde industriel et commercial (*Book Industry Communication* et *Book Industry Study Group*) et d'un groupe international EDItEUR [75] qui assure la coordination de développement de standards pour le commerce électronique du livre. ONIX est soutenu par le Cercle de la Librairie en France qui participe à la traduction du guide français. L'objectif d'ONIX est de permettre la circulation de messages de synchronisation des données éditoriales chez tous les partenaires.

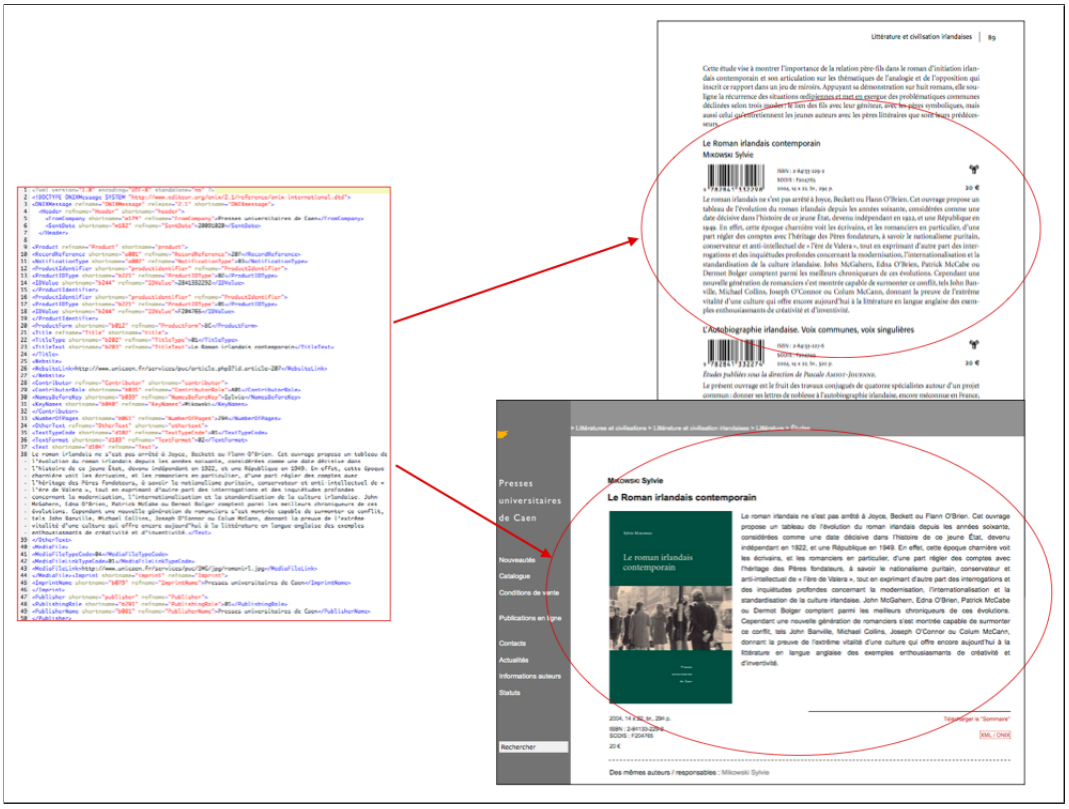

FIGURE 6.2 – Exemples d'exploitations de données ONIX.

75. http://www.editeur.org

La figure 6.2 donne un exemple de code XML ONIX et d'exploitations possibles. Il s'agit d'un message ONIX décrivant un ouvrage de recherche que l'éditeur, ici les Presses universitaires de Caen, exploite dans deux contextes différents. Dans un premier temps ce message ONIX, avec tous les autres concernant l'ensemble du fond, est intégré dans un flux de données unique pour permettre la production du catalogue papier, c'est la partie supérieure de la figure 6.2. Dans un second temps, ce même message est interprété pour compléter les informations sur le catalogue en ligne. Cependant, toutes les informations du message ONIX ne sont pas traitées systématiquement dans les deux contextes. Ainsi, dans la version papier du catalogue, l'image de couverture, c'est-à-dire le lien vers la ressource numérique dans l'instance XML ONIX, n'est pas exploitée, en particulier pour des raisons de coût d'impression. En revanche, dans la version en ligne ce lien est exploité pour intégrer l'image dans la page de présentation de l'ouvrage. La même logique est à l'œuvre pour l'intégration du code-barre, cette fois dans la version papier et non dans la version en ligne.

6.3.2 Contenus

Dans le domaine de la structuration des textes pour l'édition il existe de nombreuses initiatives avec des portées et des prétentions variables : de la solution technique locale interne à une maison d'édition au standard international adapté aux besoins de l'édition matérielle. Nous ne dressons pas ici un catalogue exhaustif de ces solutions, si tant est que ce soit possible, mais nous nous concentrons sur deux solutions majeures du domaine : DocBook et TEI.

Si ces deux exemples proposent tous les deux une dimension technique, il y a bien des modèles de représentation des données et des aires d'applications en quelque sorte. Dans les deux cas, il s'agit de décrire la structure logique des documents manipulés. Cependant la portée n'est pas tout à fait la même. DocBook, comme nous allons le voir, se concentre sur la documentation technique dans un contexte d'édition matérielle, ce qui n'est bien évidemment pas sans influence sur les vocabulaires qu'il propose pour la structuration des documents. La TEI considère l'ensemble des textes dans le domaine des sciences humaines et ne présuppose pas d'utilisation particulière, même si elle est particulièrement riche pour la description des sources.

DocBook

DocBook [Walsh, 2010] est un projet né en 1991 à l'initiative de *HaL Computer Systems* et *O'Reilly & Associates*. Il propose des vocabulaires précis plutôt orientés vers la documentation technique et a clairement pour objectif de décrire des livres.

Nous trouvons donc, en plus des éléments d'étiquetage de base des articles, chapitres, paragraphes, etc. des vocabulaires spécialisés dédiés à la documentation technique avec des distinctions très précises comme `<programlisting>` ou `<screen>`. Ces deux éléments sont destinés à l'intégration de texte en mode verbatim mais le premier, comme son nom l'indique, est destiné au marquage de code informatique et le second doit être utilisé pour du texte lu sur un écran. L'exemple de `<screenshot>` est aussi significatif puisqu'il est dédié à l'insertion dans le texte d'un lien vers une capture d'écran et pas une autre ressource iconographique. Ces distinctions très fines illustrent bien l'orientation de DocBook vers les livres avec une forte dimension technique.

DocBook propose aussi des solutions pour manipuler des notions informatiques comme les commandes (`<cmdsynopsis>`), les fonctions (`funcsynopsis`) et les classes (`classsynopsis`) des programmes avec, pour chacun de ces éléments des modèles de contenus dédiés parfaitement adaptés.

Le projet est donc très technique, avec des visées opérationnelles très fortes dès son initiation, de telle manière qu'il s'agit finalement avant tout de traiter un problème posé par une catégorie de documents bien particulière. Cependant l'extension de DocBook est possible [76], mais c'est une opération complexe et délicate sur le plan de l'analyse qui peut rapidement conduire à construire un modèle dédié en dehors de DocBook. Si, d'un point de vue strictement fonctionnel rien ne l'interdit, sortir du standard peut en revanche poser de sérieuses difficultés pour suivre les évolutions du standard et profiter des dernières avancées.

Un des points forts de DocBook est son excellente intégration dans les outils du domaine ; XMLmind XML Editor, par exemple, propose par défaut des environnements de travail pour DocBook, et OpenOffice est capable de manipuler des documents DocBook. La figure 6.3 donne un exemple de fichier XML respectant le schéma DocBook. De plus il existe de nombreuses solutions d'exploitation de ce vocabulaire comme les feuilles de transformation dédiées [77]. Toutes ces solutions de saisie et d'exploitation font de DocBook un outil d'étiquetage de documents très

76. DocBook utilise par exemple le vocabulaire CALS pour la structuration des tableaux.

77. http://docbook.sourceforge.net

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE article
  PUBLIC "-//OASIS//DTD DocBook XML V4.1.2//EN" "http://www.oasis-open.org/docbook/xml/4.1.2/doc
bookx.dtd">
<article lang="">
   <sect1>
      <title>Lorem ipsum dolor sit amet</title>
      <para>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec a diam lectus. Sed si
t amet ipsum mauris. Maecenas congue ligula ac quam viverra nec consectetur ante hendrerit. Done
c et mollis dolor. Praesent et diam eget libero egestas mattis sit amet vitae augue. Nam tincidu
nt congue enim, ut porta lorem lacinia consectetur. Donec ut libero sed arcu vehicula ultricies
a non tortor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean ut gravida lorem.
Ut turpis felis, pulvinar a semper sed, adipiscing id dolor. Pellentesque auctor nisi id magna c
onsequat sagittis. Curabitur dapibus enim sit amet elit pharetra tincidunt feugiat nisl imperdie
t. Ut convallis libero in urna ultrices accumsan. Donec sed odio eros. Donec viverra mi quis qua
m pulvinar at malesuada arcu rhoncus. Cum sociis natoque penatibus et magnis dis parturient mont
es, nascetur ridiculus mus. In rutrum accumsan ultricies. Mauris vitae nisi at sem facilisis sem
per ac in est.</para>
      <para>&amp;#8232;Vivamus fermentum semper porta. Nunc diam velit, adipiscing ut tristique vita
e, sagittis vel odio. Maecenas convallis ullamcorper ultricies. Curabitur ornare, ligula semper
consectetur sagittis, nisi diam iaculis velit, id fringilla sem nunc vel mi. Nam dictum, odio ne
c pretium volutpat, arcu ante placerat erat, non tristique elit urna et turpis. Quisque mi metus
, ornare sit amet fermentum et, tincidunt et orci. Fusce eget orci a orci congue vestibulum. Ut
dolor diam, elementum et vestibulum eu, porttitor vel elit. Curabitur venenatis pulvinar tellus
gravida ornare. Sed et erat faucibus nunc euismod ultricies ut id justo. Nullam cursus suscipit
nisi, et ultrices justo sodales nec. Fusce venenatis facilisis lectus ac semper. Aliquam at mass
a ipsum. Quisque bibendum purus convallis nulla ultrices ultricies. Nullam aliquam, mi eu aliqua
m tincidunt, purus velit laoreet tortor, viverra pretium nisi quam vitae mi. Fusce vel volutpat
elit. Nam sagittis nisi dui.</para>
      <para>&amp;#8232;Suspendisse lectus leo, consectetur in tempor sit amet, placerat quis neque.
Etiam luctus porttitor lorem, sed suscipit est rutrum non. Curabitur lobortis nisl a enim congue
 semper. Aenean commodo ultrices imperdiet. Vestibulum ut justo vel sapien venenatis tincidunt.
Phasellus eget dolor sit amet ipsum dapibus condimentum vitae quis lectus. Aliquam ut massa in t
urpis dapibus convallis. Praesent elit lacus, vestibulum at malesuada et, ornare et est. Ut augu
e nunc, sodales ut euismod non, adipiscing vitae orci. Mauris ut placerat justo. Mauris in ultri
cies enim. Quisque nec est eleifend nulla ultrices egestas quis ut quam. Donec sollicitudin lect
us a mauris pulvinar id aliquam urna cursus. Cras quis ligula sem, vel elementum mi. Phasellus n
on ullamcorper urna.</para>
      <para>&amp;#8232;Class aptent taciti sociosqu ad litora torquent per conubia nostra, per incep
tos himenaeos. In euismod ultrices facilisis. Vestibulum porta sapien adipiscing augue congue id
 pretium lectus molestie. Proin quis dictum nisl. Morbi id quam sapien, sed vestibulum sem. Duis
 elementum rutrum mauris sed convallis. Proin vestibulum magna mi. Aenean tristique hendrerit ma
gna, ac facilisis nulla hendrerit ut. Sed non tortor sodales quam auctor elementum. Donec hendre
rit nunc eget elit pharetra pulvinar. Suspendisse id tempus tortor. Aenean luctus, elit commodo
laoreet commodo, justo nisi consequat massa, sed vulputate quam urna quis eros. Donec vel.</para
>
   </sect1>
</article>
```

FIGURE 6.3 – Exemple basique de fichier DocBook exporté depuis OpenOffice.

puissant. Cependant, il reste très lié à son contexte de production, l'édition matérielle de documents techniques à laquelle il est parfaitement adapté, et présente souvent quelque résistance pour être mis en œuvre dans d'autres domaines comme les sciences humaines par exemple et l'édition de sources anciennes en particulier.

TEI

Contrairement à DocBook, la TEI, déjà présentée plus haut[78], n'est pas dédiée explicitement à l'activité d'édition matérielle mais sa richesse lui permet de répondre aux besoins des éditeurs, même si sa complexité peut effrayer au premier abord [PROST, 2011].

78. Voir p. 85.

En effet, les recommandations de la TEI sont de plus en plus utilisées pour structurer les textes par les éditeurs publics qui les placent de plus en plus souvent au cœur de leurs chaînes de traitement, soutenus dans ce sens par la Bibliothèque Scientifique Numérique. Ainsi, les formations à l'édition structurée basée sur la TEI se multiplient depuis quelques années, dans le cadre de l'AEDRES et de BSN, mais aussi dans celui du réseau MÉDICI (Métiers de l'ÉDition sCIentifique publique), qui rassemble les professionnels de l'édition publique et qui organise aussi de plus en plus des opérations de formation aux techniques de structuration basées sur la TEI. Rappelons que le CLEO a également fait du XML TEI son format interne, même si, comme nous le verrons plus loin, les éditeurs qui utilisent la plateforme proposée par ce centre ne sont pas dans l'obligation de le manipuler directement.

```
<div type="chapitre">
  <div type="section" xml:id="AFR.1" xml:lang="afr">
    <lg xml:id="AFR.1.1">
      <l n="1" aid:pstyle="txt_Original_Vers" xml:id="vers1"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/"><pb ed="A"
      n="1r" />Molz pelerins qui vunt al munt<note type="marginal"
      xml:id="AFRftn1"> <emph aid:cstyle="typo_Italique">B </emph>: Les
      bones gens qui vunt au mont <emph aid:cstyle="typo_Italique">(voir
      infra), avec majuscule initiale de grande taille.</emph>
      </note></l>

      <l n="2" aid:pstyle="txt_Original_Vers" xml:id="vers2"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Enquierent
      molt, et grant dreit unt,</l>

      <l n="3" aid:pstyle="txt_Original_Vers" xml:id="vers3"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Comment
      l'igliese fut fundee</l>

      <l n="4" aid:pstyle="txt_Original_Vers" xml:id="vers4"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Premierement,
      et estoree.</l>

      <l n="5" aid:pstyle="txt_Original_Vers" xml:id="vers5"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Cil qui lor
      dient de l'estoire</l>

      <l n="6" aid:pstyle="txt_Original_Vers" xml:id="vers6"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Que cil
      demandent, en memoire</l>

      <l n="7" aid:pstyle="txt_Original_Vers" xml:id="vers7"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Ne l'unt pas
      bien, ainz vunt faillant</l>

      <l n="8" aid:pstyle="txt_Original_Vers" xml:id="vers8"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">En plusors
      leus, et mespernant.</l>
```

FIGURE 6.4 – Exemple de code TEI pour l'édition matérielle.

Par ailleurs, l'utilisation toujours plus importante de la TEI dans le monde de la recherche scientifique [79] est un levier puissant pour que ces recommandations soient également adoptées pour l'édition matérielle des textes. En effet, l'adoption des mêmes vocabulaires par l'ensemble des acteurs ne peut que faciliter l'ensemble des opérations, nous y reviendrons.

Enfin, le consortium propose de multiples outils de conversion [80] pour faciliter la manipulation de données encodées en TEI. Si ces solutions ne prétendent pas traiter tous les cas de figure, ce que la richesse du vocabulaire interdit presque totalement, elles constituent une excellente base de travail déjà très avancée.

La figure 6.4 donne un exemple de code XML TEI produit par les Presses universitaires de Caen pour la transcription en ancien français versifié du *Roman du Mont Saint-Michel*.

6.4 Convergence numérique : impact sur les éditeurs

La généralisation du numérique à toutes les étapes de l'édition des textes, de la saisie à la lecture, n'est pas sans provoquer de changement dans le monde de l'édition matérielle. Cependant, s'il ne faut pas négliger cet impact, il convient également de ne pas l'exagérer. Comme toutes les évolutions, cette généralisation impose un certain nombre d'adaptations aux éditeurs, mais c'est aussi une occasion d'affirmer, ou de réaffirmer, la spécificité du métier et des compétences qui lui sont associées. En revanche l'activité de promotion des éditeurs matériels est plus directement impactée : il s'agit d'apprendre à faire connaître et à vendre des objets nouveaux.

Si l'on considère que l'activité d'édition matérielle consiste à attribuer une forme physique à un texte pour un support donné, la seule véritable évolution introduite par la convergence numérique est de multiplier les formes de diffusion.

Autrement dit, et d'un point de vue plus théorique, l'édition matérielle ne consisterait plus à produire *une structure physique* adaptée à la *structure logique* d'un document [Quint, 1987], mais *des structures physiques* toujours en étroite relation avec une *structure logique* en fonction des spécificités des supports de diffusion choisis. On retrouve bien entendu ici le modèle de production du *single source publishing* déjà évoqué et le principe de séparation du fond et des formes. Nous reviendrons sur ce point et ses implications en détail dans la partie III.

79. Voir p. 71 et suivantes.

80. http://sourceforge.net/projects/tei/files/?source=navbar

Le travail d'édition consiste donc en quelque sorte à superposer à la structure logique produite par l'auteur plusieurs structures physiques adaptées aux contraintes des supports de diffusion.

Le papier impose ainsi de contraindre le texte dans une page, interdit d'intégrer des contenus dynamiques, oblige à limiter la quantité de textes traités pour que le volume papier reste manipulable, etc. L'éditeur doit donc fixer et appliquer les règles qui vont permettre d'assurer cette mise en correspondance de la structure logique et de la structure physique du livre papier. Nous ne pouvons pas reprendre ici toutes les règles de composition de la tradition typographique française [Imprimerie Nationale, 2002], mais citons par exemple : tous les titres de chapitres doivent correspondre à la création d'une nouvelle page de droite, ou encore les lignes des paragraphes doivent toujours être conservées en groupe de deux en haut et en bas des pages, etc. Certaines de ces règles pourront faire, et font d'ailleurs déjà, l'objet d'un traitement informatique plus ou moins facilement, mais certaines d'entre elles peuvent aussi entrer en conflit, conflit qui devra être arbitré par une intervention humaine pour obtenir un résultat satisfaisant. C'est tout le travail de composition typographique qui, comme le disait Maximilien Vox [81],

> est un métier ancien et très simple, aussi simple que de jouer du violon, mais guère plus.

La diffusion en ligne, ou le web en tant que support, n'est pas moins contraignante. En effet, l'éditorialisation d'un flux de texte diffusé en ligne nécessite un travail important pour se prêter dans de bonnes conditions à une lecture immersive. Ainsi, une page longue impose de fournir des solutions de navigation et de repérage dans le texte sans que ces dispositifs ne pèsent trop lourdement sur la lecture. Il faut donc établir des règles pour préserver une longueur de ligne acceptable et un gris typographique harmonieux tout en l'outillant de liens et de systèmes de comptage par exemple. La principale difficulté réside dans la liberté laissée au lecteur qui a toujours le dernier mot sur les arbitrages. Ainsi les éditeurs doivent apprendre à composer avec les contraintes du web.

Revenons pour terminer sur la tension entre structure logique et structure physique avec l'exemple des références bibliographiques qui changent de forme en fonction de leur ordre dans le texte et de la zone d'apparition sur les pages. Autrement dit, le texte du document peut changer en fonction de la structure physique. Le sys-

81. Maximilien Vox était un spécialiste de typographie. Il est en particulier connu pour sa classification des polices de caractères.

tème de description de la structure logique doit donc tenir compte de ces aspects pour permettre la production de structures physiques correctes et efficaces.

6.5 État des pratiques et modes de production

La variété des méthodes de production dans le monde de l'édition matérielle est difficile à mesurer et en dresser un inventaire exhaustif pourrait faire l'objet d'une étude à part entière. Nous postulons donc ici que l'impératif de diffusion sur plusieurs supports est aujourd'hui accepté par les éditeurs qui cherchent des solutions pour adapter leurs méthodes de production à ce nouveau paradigme. Les solutions consistant à mettre à jour un cahier des charges de sous-traitance imposant la livraison de n formes de diffusion ne nous concernent pas ici. Il s'agit bien des modes de production techniques.

L'acceptation du mode de production s'articulant autour d'un fichier source unique à partir duquel sont produites, automatiquement ou semi-automatiquement, toutes les formes de diffusion est maintenant presque unanime [Prost, 2007]. C'est le modèle vers lequel tous les éditeurs tendent aujourd'hui. Le nombre de formations sur ces questions proposées aux professionnels du domaine en atteste [82].

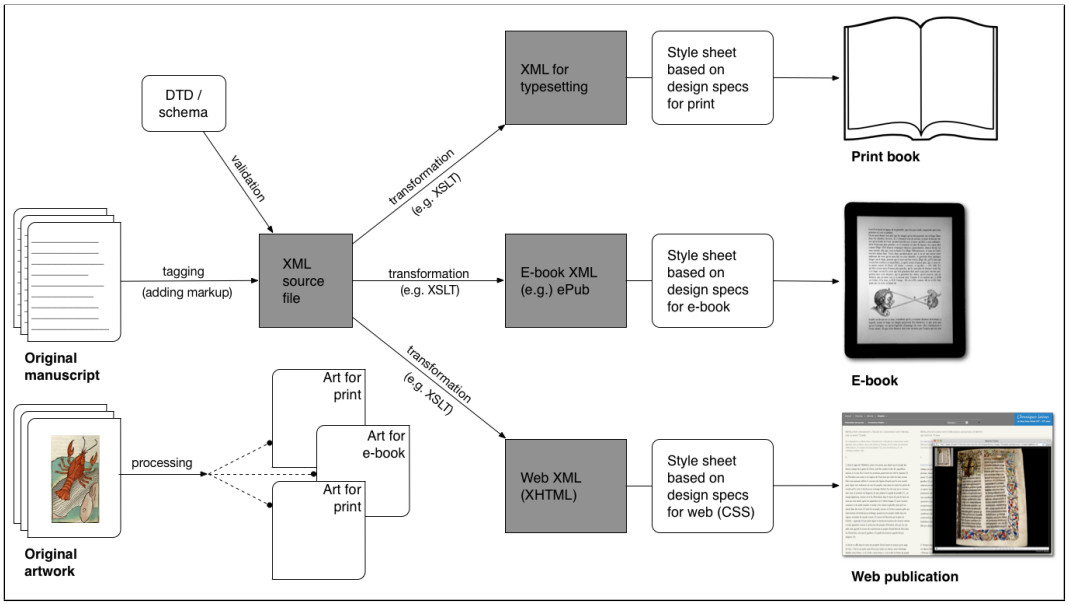

Figure 6.5 – L'organisation éditoriale (d'après le modèle du *single source publishing* vu par le *Chicago manual of style* [UCP, 2010]).

82. Voir par exemple la rubrique "Concevoir, structurer et éditer les contenus" du catalogue des formations de l'Asfored (Association nationale pour la formation et le perfectionnement professionnels dans les métiers de l'édition), http://www.asfored.org.

La figure 6.5 donne une représentation de ce modèle de production extraite du *Chicago manual of style* [UCP, 2010]. Cette figure illustre très bien la méthode idéale d'organisation du travail en centrant le point de vue sur les objets manipulés par les éditeurs matériels. Au centre, se place un fichier XML pivot produit à partir du document, analogique ou numérique, remis par l'auteur. Une fois le fichier pivot produit, toutes les formes de diffusion sont générées au moyen d'outils de transformation ou d'importation dans des logiciels dédiés à la production d'une forme particulière, comme typiquement un logiciel de PAO.

Il s'agit bien entendu d'une vision générale qui doit être adaptée par chaque éditeur en fonction des contextes de production locaux : forces informatiques disponibles, nature de la production, niveau de spécialisation, etc. Un éditeur spécialisé dans le domaine informatique aura sans doute moins de difficulté à développer ses propres outils pour intégrer DocBook par exemple, qu'un autre, spécialisé en poésie contemporaine. En effet, ses personnels ne disposeront pas, *a priori*, des mêmes compétences et expériences.

Dans tous les cas, la notion centrale à la base de la méthode de travail décrite par la figure 6.5 est le *tagging* du texte remis par l'auteur. L'ensemble des opérations repose donc sur la discrimination des éléments constitutifs du texte les uns par rapport aux autres. Nous allons donc examiner deux types de solutions avec des niveaux d'implication technique différents.

6.5.1 Étiquetage des textes

Il s'agit ici de s'appuyer exclusivement sur la forte intégration de la pratique du stylage chez les éditeurs et de conserver à l'extérieur du service d'édition tous les systèmes de conversion et de transformation. Ainsi, l'éditeur peut se concentrer sur la discrimination des éléments constitutifs du texte avec les outils dont il a une parfaite maîtrise sans se préoccuper des implications techniques des outils informatiques de conversion.

La figure 6.6 donne une représentation du travail éditorial s'appuyant sur un service extérieur de conversion. L'éditeur commence par un étiquetage systématique du document remis par l'auteur et réalise les opérations habituelles (relecture, normalisation bibliographique, contrôle de la structure, etc.) avec les outils traditionnels du secrétariat de rédaction : un logiciel de traitement de texte et, le plus souvent, un outil de contrôle et de correction ortho-typographique.

Une fois le texte stabilisé sur le fond, l'éditeur envoie le fichier à un service de conversion externe qui se charge de transformer le fichier en une version pivot archivable et pérenne, c'est-à-dire un fichier XML que l'éditeur pourra stocker. C'est à partir de cette version pivot que les formes de diffusion numériques seront produites à partir du même service de conversion. Le format PDF peut être exploité pour une diffusion numérique, il présente alors des différences par rapport aux fichiers utilisés pour l'impression : absence de marques de coupes, compression des images pour réduire leur poids, etc.

La version papier, quant à elle, est produite à partir du document stylé de manière traditionnelle en important le fichier dans un logiciel de PAO. Si l'éditeur souhaite contrôler les fichiers électroniques reproduisant une mise en page, comme typiquement des fichiers PDF, il peut les produire à l'issue du travail de PAO.

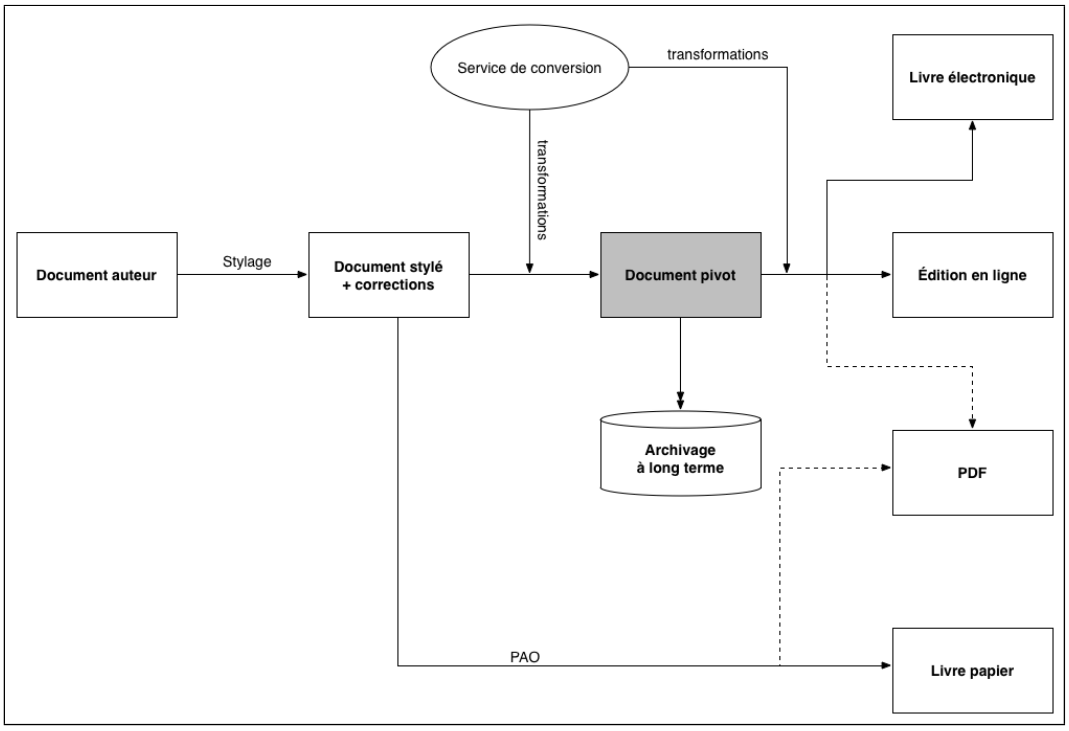

FIGURE 6.6 – Organisation éditoriale basée sur l'utilisation de styles.

Du point de vue de la production, l'inconvénient majeur de cette méthode réside dans le risque de divergence entre les versions diffusées sur supports numériques (ici livre et édition en ligne) et la version papier car toutes ne sont pas produites à partir du même fichier. L'éditeur doit donc être extrêmement rigoureux tout au long du travail de production. Idéalement, toute intervention sur le document stylé devrait

être proscrite après la production des formes électroniques de diffusion pour éviter au maximum les risques de divergence.

Cette organisation induit aussi une véritable dépendance de l'éditeur vis-à-vis du service de conversion externe utilisé quant aux vocabulaires mobilisés. Autrement dit, le contrôle de l'éditeur sur la structure logique des textes reste relativement limité. De la même manière, les structures physiques des formes électroniques risquent d'être fortement liées à ce même service. En définitive ce type de solution doit nécessairement rester généraliste pour répondre au plus grand nombre de besoins et si un éditeur se trouve confronté à des cas spécifiques il devra soit trouver une solution pour les gérer en interne, soit faire en sorte de les ramener à des cas gérés par le service externe. L'absence de contrôle de l'outil de conversion impose donc une certaine souplesse du point de vue des exigences éditoriales.

Il existe plusieurs services de conversion externes avec des offres plus ou moins grandes allant de la stricte conversion de document jusqu'à la diffusion dans des bouquets spécialisés. Citons par exemple OxGarage [83] qui s'appuie très fortement sur la TEI et OpenText que nous avons déjà évoqué. Notons enfin que revues.org [84] propose un système pour les revues scientifiques très proche de ce que nous venons de présenter ici.

6.5.2 Modèle XML intégré

Le modèle XML intégré se distingue du modèle d'étiquetage des textes par le fait que le fichier pivot constitue un véritable fichier de travail autour duquel l'ensemble des opérations s'organise.

La figure 6.7 donne une représentation du modèle XML intégré mis en place dans un premier temps aux Presses universitaires de Caen puis chez d'autres éditeurs institutionnels dans le cadre de l'AEDRES.

Quand l'auteur remet un fichier de traitement de texte saisi au kilomètre, la première opération, comme dans le modèle d'étiquetage des textes, consiste à styler le document pour introduire une première discrimination des éléments constitutifs du texte les uns par rapport aux autres. Mais cet étiquetage n'est en réalité que la première étape du traitement dans le modèle XML intégré. En effet, une fois le document étiqueté, OpenOffice est utilisé comme plateforme de conversion pour faire la correspondance entre les styles et les balises. Il s'agit en fait d'utiliser les logiciels

83. https://github.com/sebastianrahtz/oxgarage. Une version d'évaluation est accessible en ligne : http://oxgarage.oucs.ox.ac.uk:8080/ege-webclient/.

84. http://www.revues.org

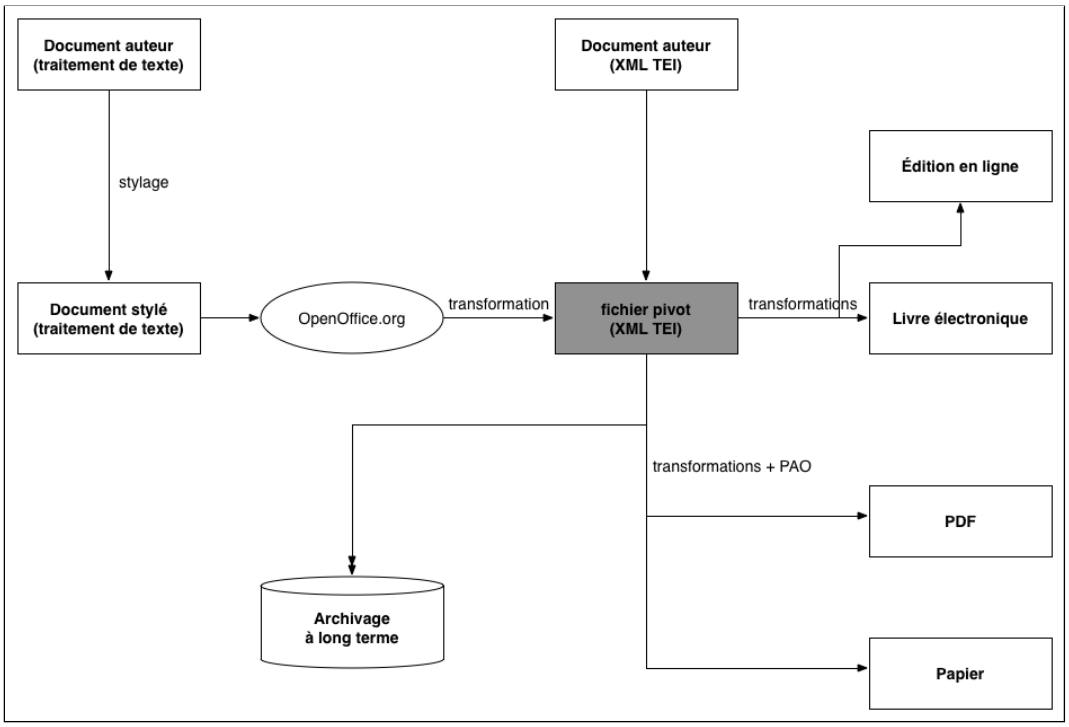

FIGURE 6.7 – La chaîne éditoriale de l'AEDRES.

de traitement de texte pour simplifier la mise en place d'une structure XML de base [85].

Le fichier pivot doit contenir toutes les corrections et proposer un état stable du texte. Ces corrections peuvent être effectuées avant la transformation sous un logiciel de traitement de texte ou directement sur le fichier pivot. Dans ce cas, il est indispensable de mettre en place des solutions logicielles qui permettent aux secrétaires de rédaction d'interagir avec la structure logique des documents sans subir les contingences inhérentes au XML. La figure 6.8 propose un exemple d'une telle interface, ici du logiciel XMLmind XML Editor, articulant plusieurs vues synchronisées d'un même document XML qui permettent d'intervenir sur le contenu textuel ainsi que sur la structure logique du document.

L'intégration du XML au cœur du dispositif autorise sa manipulation, ce qui implique aussi que les secrétaires de rédaction soient formés à cette tâche : ce dispositif permet d'enrichir les structures manipulées si les styles s'avèrent insuffisants pour décrire les textes assez précisément. En effet, d'une manière générale, les styles ne permettent de distinguer que deux niveaux hiérarchiques, sauf dans le cas des titres, en affectant et en interprétant leurs niveaux : niveau 1 pour les titres

85. Voir p. 88.

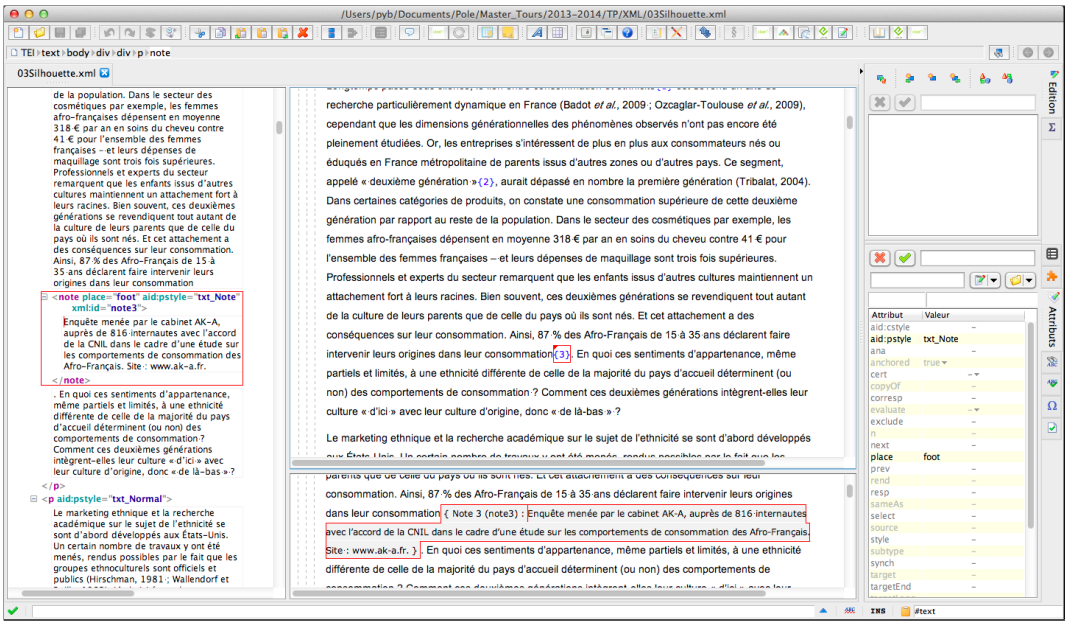

FIGURE 6.8 – Interface de travail éditorial (XXE).

de chapitre ou d'article qui deviendront par exemple des `<div type="chapitre">` ou `<div type="article">`; niveau 2 pour les sections `<div type="section">`; niveau 3 pour les sous-sections `<div type="sous-section">`, etc.

Mais ce modèle permet aussi d'intégrer des documents XML de recherche directement dans la chaîne de traitement éditorial, sans avoir de transformation ou de conversion supplémentaire à introduire. Sans entrer dans les détails sur lesquels nous reviendrons plus loin [86], l'opération consiste pour l'essentiel à assurer l'attribution de formes déterminées par l'éditeur matériel aux catégories de textes discriminées par les chercheurs. Techniquement, cela revient en fait, pour le secrétaire de rédaction, à renseigner, pendant la relecture, les rôles qui seront affectés à certaines catégories de textes au cours des étapes suivantes dans la chaîne de traitement. Ces rôles correspondent à des formes spécifiques dans chaque support de diffusion produit.

Une fois le fichier pivot finalisé, c'est-à-dire lorsqu'il contient l'ensemble des corrections et que le rôle éditorial de chaque élément constitutif du texte est défini et la structure logique stabilisée, le travail de production des formes de diffusion peut débuter.

Les formes de diffusion numériques, telles que l'édition en ligne ou le livre numérique sont produites à partir du fichier pivot sur la base de l'application de transformation. Sur le principe, il s'agit de passer du système de vocabulaire utilisé pour le

86. Voir p. 146 et suivantes.

fichier pivot à un autre, typiquement un langage de la famille HTML utilisé pour le web ou pour les livres au format ePub. Ce dernier va aussi nécessiter d'organiser les fichiers dans une archive d'une manière spécifique [Pritchett et Gylling, 2011]. Cette organisation des fichiers se fait également sur la base de l'analyse et de la transformation du fichier pivot au cours desquelles toutes les ressources externes liées au texte, comme les images par exemple, sont copiées dans l'archive et les liens vers ces mêmes ressources adaptés à la nouvelle organisation des fichiers. Une fois le fichier ePub produit, il peut encore nécessiter certains calages et ajustements formels. Pour ces opérations purement liées à la forme de diffusion, des logiciels d'édition d'ePub peuvent être utilisés. Citons par exemple l'application en ligne Polifile[87] développée avec le framework Sydonie (SYstème de gestion de DOcuments Numériques pour l'Internet et l'Édition) [Lecarpentier, 2011] ou le logiciel Sigil[88].

La production des formes papier peut être réalisée en utilisant deux types d'outils tout à fait différents : d'une part, les langages comme XSL-FO ou LaTeX et, d'autre part, les metteurs en page comme Adobe InDesign, FrameMaker ou Quark Xpress.

Si les premiers, et LaTeX en particulier, ont fait la preuve de leur efficacité dans le domaine de la production de documents papier de qualité, leur usage et leur maîtrise restent assez peu répandus chez les éditeurs matériels, en particulier dans le domaine des SHS qui nous occupe ici. Nous allons donc nous concentrer sur l'utilisation d'un metteur en page pour la production des formes papier et d'Adobe InDesign en particulier qui est le plus répandu et le plus utilisé.

InDesign dispose d'un espace de nom dédié à la mise en forme des textes balisés en XML qui propose des attributs destinés à automatiser l'application de styles de paragraphes, de tableaux, de séquences de caractères et d'objets graphiques, à des éléments XML. Une fois ces attributs renseignés, soit automatiquement sur la base de l'analyse de la structure soit manuellement lors de la relecture, InDesign est capable d'affecter un style, et donc toutes les propriétés graphiques et typographiques qui lui sont associées, aux éléments constitutifs du fichier pivot initial.

Cependant, InDesign présente aussi quelques contraintes qu'il s'agit de prendre en compte pour que l'importation se déroule dans de bonnes conditions.

Ainsi, si le logiciel dispose d'un espace de nom dédié à l'association d'un élément à un style de mise en forme, il s'avère en revanche incapable de faire la différence entre les *block level elements* et les *inline elements*. Il est donc nécessaire de formater le fichier en regroupant les *block level elements* sur des lignes séparées. La partie

87. http://www.polifile.fr

88. https://code.google.com/p/sigil/

haute de la figure 6.9 donne l'exemple de code XML indenté à gauche et formaté pour l'importation dans InDesign.

Dans la même optique, le traitement des espaces doit aussi faire l'objet d'une attention particulière. La partie basse de la figure 6.9 montre le comportement d'InDesign lors de l'importation d'un fichier XML indenté. En fait, Adobe InDesign ne prend pas de décision sur le statut des espaces : pour lui, il n'y a que des espaces typographiques, il les conserve donc toutes.

Il faut donc importer un texte qui ne contient effectivement que des espaces typographiques et des retours à la ligne correspondants à des fins de *block level elements*, typiquement, à des fins de paragraphes.

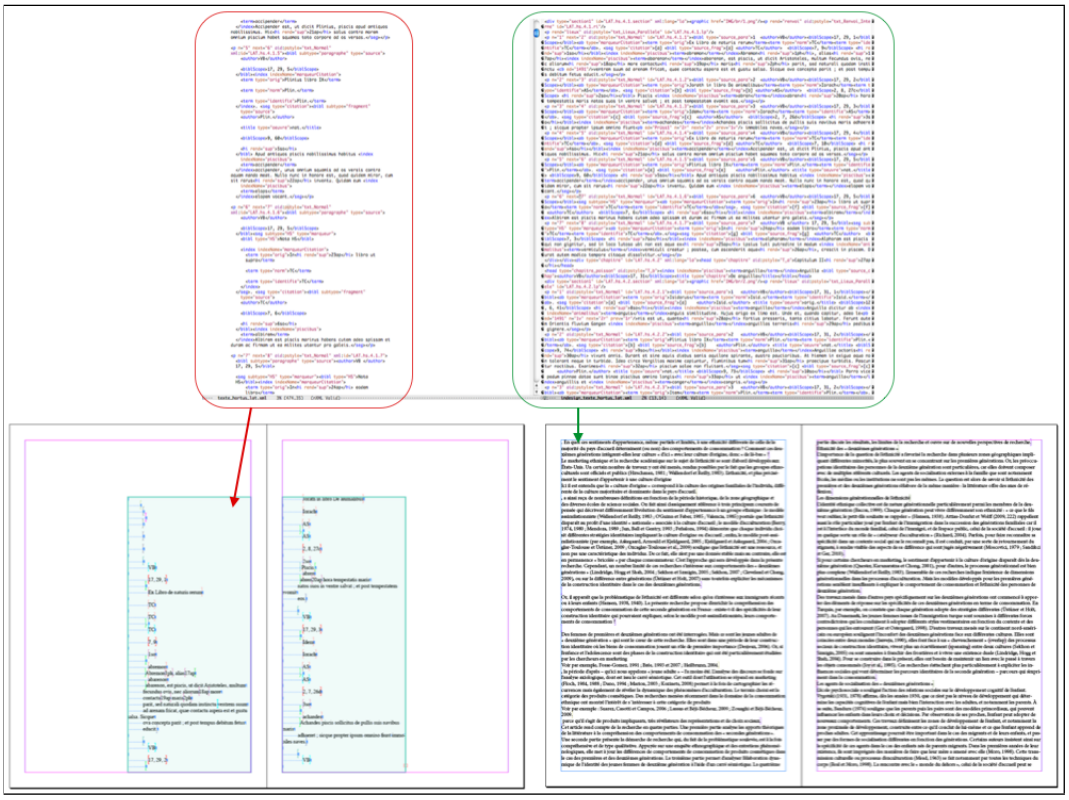

FIGURE 6.9 – Contraintes d'importation de fichiers XML dans Adobe InDesign.

Si ces limitations pourraient être moins importantes — après tout, l'indentation et les espaces qu'elle ajoute ne perturbent pas autant d'autres solutions de production de pages[89] — il est aussi possible de penser qu'elles sont en partie liées au modèle `complextype` de XML qui autorise les contenus mixte structure/donnée qui rend difficile de faire la différence entre des caractères de données et des caractères

89. Même si, pour produire un document LaTeX par exemple, il faudra aussi tenir compte des *block level elements* pour reconstituer des paragraphes cohérents.

de présentation. Nous revenons ici aux questions de l'accès aux données et de la complexité des relations entre la *structure logique* et la *structure physique* des documents manipulés qui, comme nous le voyons, peuvent impacter fortement les étapes de production des formes de diffusion.

6.6 Édition électronique, multisupport et multimodale

Donner une définition précise de l'édition électronique est presque impossible tant cette expression est polysémique. Tout d'abord elle désigne aussi bien des produits diffusés qu'une activité ou un ensemble d'activités, ainsi Pierre Mounier et Marin Dacos notent qu'il n'est

> [...] guère possible de donner une définition [...] précise de ce qu'est l'édition électronique. Car, de définitions, il y a autant que d'acteurs, pour ainsi dire, tant le domaine est divers et peu uni. Qu'y a-t-il de commun entre la numérisation de dizaines de milliers d'ouvrages de bibliothèques par Google ou Gallica, [...] la diffusion sur Internet de milliers de revues scientifiques [...] ou encore le développement d'une plateforme de blogs scientifiques [Dacos et Mounier, 2010].

Nous ne nous lançons donc pas ici dans une tentative de définition de l'activité tant l'entreprise est, aujourd'hui encore, plus ou moins hasardeuse. En revanche, nous proposons de tracer des frontières entre différentes formes de diffusion produites en nous concentrant sur leurs fonctionnalités respectives.

Usuellement, l'expression "*une* édition électronique" désigne tout produit diffusé sur un support numérique. Il s'agit donc d'une notion très générale et peu précise qui regroupe tout autant les livres destinés aux liseuses ou tablettes (ePub) que les textes diffusés sur internet, comme les revues accessibles sur le portail *revues.org* par exemple. Cette désignation n'apprend donc en réalité que peu de choses sur l'objet en question. Cependant, la fréquence de son utilisation est un bon indicateur de l'impact de la convergence numérique dans le domaine et sur l'ampleur des inquiétudes qu'elle provoque.

Une édition multisupport propose la consultation d'un même texte, ou des parties d'un même texte, sur plusieurs supports différents comme le papier et l'électronique par exemple. Certaines éditions multisupports proposent en effet des variations de contenus en fonction des supports. Ainsi, dans la collection de sources *Fontes & Paginæ*, les Presses universitaires de Caen proposent une ventilation des contenus en

fonction des supports. La version papier contient l'ensemble des textes, les transcriptions, les traductions et les introductions scientifiques ainsi que des reproductions des témoins. La version en ligne ne propose que l'ensemble des transcriptions et des traductions et, quand les droits le permettent, l'intégralité des images numériques des manuscrits ou imprimés anciens.

Il s'agit donc de tirer profit des qualités intrinsèques des différents supports dans le cadre d'un projet éditorial donné. Ajouter l'intégralité des images dans un volume papier est techniquement possible mais économiquement difficile et commercialement risqué ; pourtant les avantages de stabilité et de simplicité de consultation du papier présentent des avantages incontestables[90]. La version web permet, elle, de rendre opérationnel, sous forme de liens hypertextes, l'ensemble des renvois internes et externes pour améliorer la consultation du texte. L'objectif est donc bien d'articuler les supports de diffusion les uns avec les autres plutôt que les opposer.

Une édition multimodale présente beaucoup de proximité avec une édition multisupport dont elle généralise en fait les principes. Il s'agit en effet le plus souvent d'une édition multisupport enrichie de modalités d'accès aux textes variées et qui ne se contente donc pas de proposer des solutions de lecture immersive. Dans cette optique le support peut être considéré comme une simple modalité : le texte se lit soit sur le papier, soit en ligne, soit sur un livre numérique par exemple. Une édition multimodale de manuscrit peut donc être exclusivement web et proposer de lire les textes de chaque témoin ou le texte idéal édité à partir des multiples variantes. Les modalités peuvent varier en fonction des cas d'utilisation du texte, de la lecture immersive à l'étude de texte.

Autrement dit, un même texte peut faire l'objet d'autant de modalités d'accès et de lecture que voulus en fonction de sa nature et du projet éditorial. Ainsi, le web, en tant que support de diffusion, permet de construire autant de modalités d'accès aux textes et de lecture que nécessaires pour satisfaire les usages (lecture immersive, étude de textes, recherche, etc.). Mais une telle situation ne va pas sans poser certains problèmes de modélisation. En fait, pour permettre la production d'éditions multimodales dans des conditions satisfaisantes il faut trouver une modélisation adaptée.

90. Voir p. 103.

Troisième partie

Modélisation

Introduction

Comme nous l'avons vu, un certain nombre d'éléments sont déjà en place dans les différents domaines pour manipuler les textes dans des conditions satisfaisantes.

Ainsi, les bases techniques sont maintenant données, avec XML et toutes les technologies qui l'accompagnent, même s'il y a une réelle tendance générale de fond à placer les applications et les volumes de données au centre des débats et des réalisations comme nous pouvons le voir avec le développement de HTML5 et du *big data*. Il n'en reste pas moins que toute une partie de l'activité, en particulier celle que nous avons décrite plus haut, a besoin de centrer son attention sur la qualité des contenus manipulés et sur la manière dont ces contenus sont décrits et indéxés.

De plus, les vocabulaires de description métiers sont aussi bien identifiés. Il est donc maintenant nécessaire de trouver un mode d'organisation approprié ainsi que les objets les plus adaptés à manipuler. Comme nous le verrons, sur ce point, le réseau fait vaciller les bases sur lesquelles nous avions tous pris l'habitude de nous appuyer.

Ainsi, la notion centrale du document, document qui est aujourd'hui pourtant tellement variable qu'on est en droit de s'interroger sur ce qu'il désigne exactement. En effet, si tout est toujours un document, il devient assez difficile de savoir de quoi il est question exactement...

En définitive, dans le domaine de l'édition en particulier, mais peut-être aussi plus généralement, nous nous trompons tout simplement d'objet. Avec l'omniprésence du réseau, c'est sans doute sur le texte en tant que tel qu'il faut travailler et sur lequel se concentrer, mais plus vraiment sur les documents qui les portent.

Pour être en mesure de repenser le document, lui redonner sa juste place, nous proposons de repenser les concepts qui président à sa production et les systèmes qui les mettent en œuvre. Si l'on arrive bien, en bout de chaîne, à la création d'un document, ou d'un *néo-document*, pour reprendre le vocabulaire de Jean-Michel Salaün [Salaün, 2012], hyper-connecté à un grand nombre de ressources (banques

d'images, de sons, de textes, etc.), il s'agit de comprendre en détail les processus de création à l'œuvre aujourd'hui afin, soyons clairs, de permettre aux acteurs des domaines concernés d'en prendre la maîtrise. Car effectivement, qui mieux qu'un archiviste ou un bibliothécaire peut décrire ses collections ? Ou qui mieux qu'un historien peut éditer et expliciter les sources dont il est spécialiste en explicitant les typologies textuelles qui les constituent ?

Nous proposons dans les prochains chapitres des solutions d'organisation des flux de données et des flux de travail en replaçant la qualité des données au centre de nos réflexions. Si cette notion de flux dépasse très largement la question du texte — on pense en particulier aux flux radiophoniques par exemple — elle s'y applique néanmoins parfaitement. Il s'agit tout autant de proposer des grilles de lecture d'une situation déjà en partie établie que des solutions pour donner aux acteurs le contrôle et la compréhension des techniques mises en œuvre.

L'enjeu est de contrôler la finesse d'annotation des textes dans tous les sens du terme : aussi bien sur le versant technique, c'est-à-dire quelles étiquettes et quelles méthodes pour désigner les typologies textuelles rencontrées, qu'intellectuel, autrement dit, comment tenir un discours historique, philologique, etc. sur ces ensembles de données richement annotés ?

Comme nous l'avons vu, le domaine de l'édition de sources primaires concerne et mobilise plusieurs acteurs qui interviennent à divers degré d'un projet. Les objets manipulés se recoupent et changent parfois en fonction du point de vue et des objectifs de chacun de ces acteurs.

Chacune des disciplines concernées dispose d'un modèle plus ou moins formel qui encadre et organise son activité. L'archivistique dispose ainsi d'un cadre de référence parfaitement établi avec ses objets et ses solutions de manipulation. Les communautés de recherche et d'édition mettent en place des solutions d'organisation de leur travaux et de manipulations de leurs objets, avec des solutions partagées de plus en plus robustes.

La convergence numérique impacte l'ensemble des activités concernées par l'édition de sources primaires à divers niveaux et pousse à organiser la multiplicité des pratiques et des modèles en un tout cohérent.

Notre modèle doit donc articuler l'ensemble des modèles et des pratiques existants. Autrement dit, il s'agit pour nous d'identifier les zones de recouvrement pour être en mesure de les organiser entre elles. Certains objets sont les mêmes, mais chaque acteur les considère dans des dimensions différentes qu'il s'agit de systéma-

tiser au sein d'un même continuum d'informations, de l'inventaire à l'édition. Cela revient alors en partie, sur ce point précis, à prévoir des ponts entre les formats et les modèles de l'archive et ceux de la recherche et donc, de l'EAD vers la TEI et réciproquement par exemple.

Nous proposons donc ici des solutions d'articulation des modélisations existantes dans chacun des domaines concernés. Les archivistes travaillent sur des documents matériels, les chercheurs (nous nous situons plus précisément dans le contexte des éditions de sources) sur des textes portés en tout ou en partie par plusieurs de ces documents et les éditeurs matériels produisent de nouveaux documents à partir des textes. Il ne s'agit pas de remplacer les modèles existants dans chacun des domaines concernés par une sorte de méta-modèle, mais bien de proposer un nouveau modèle qui articule les modèles des domaines pour permettre de faire circuler les informations dans les meilleures conditions possibles.

Du point de vue méthodologique, précisons que la modélisation proposée dans cette étude a été réalisée de manière incrémentale depuis une quinzaine d'années. Toutes les expérimentations réalisées, dont seulement certaines sont présentées plus loin [91], ont permis de régler un certain nombre de problèmes, et parfois d'en dévoiler de nouveaux. Chaque expérimentation s'accompagne d'objectifs spécifiques et de solutions pour les atteindre. Toutes les solutions trouvées et développées permettent d'améliorer le modèle sous-jacent et de le rendre explicite. Ainsi chaque expérimentation permet d'enrichir le modèle.

91. Voir p. 173 et suivantes.

7

Du modèle de document au modèle de flux

Nous proposons dans ce chapitre d'interroger la notion de document pour la replacer dans un système plus vaste et la définir explicitement. Avec l'omniprésence des réseaux sur lesquels circule un volume toujours plus important de données, cette notion doit en effet faire l'objet d'un examen afin d'en tracer les limites. Il s'agit ici, d'une certaine manière, de caractériser les flux d'informations circulant sur les réseaux en regard des documents produits à l'aide de logiciels de traitement de texte par exemple.

7.1 Limites de la notion de document

Dans le contexte de l'édition de sources primaires, qui nous occupe ici, nous sommes confrontés à un réseau d'acteurs qui entretiennent des relations de travail dans le cadre de projets d'étude, de valorisation ou de conservation, bref, qui collaborent, autour des documents qu'ils manipulent.

Mais ces mêmes documents entretiennent également des relations entre eux. Pensons par exemple aux manuscrits du Mont Saint-Michel dont les textes ont été recopiés d'un manuscrit à l'autre.

Nous nous trouvons ici confrontés à une situation très similaire à celle décrite par Jean-Michel Salaün :

> Les documents ne sont plus maintenant interprétés pour ce qu'ils disent explicitement, mais comme des traces à mettre en relation avec d'autres qui témoignent d'évolutions sans rapport nécessaire avec leur contenu. Ils sont pris dans un système signifiant plus vaste [Salaün, 2012].

À la suite de Jean-Michel Salaün nous proposons donc de concentrer l'effort sur le *système signifiant plus vaste*, c'est-à-dire sur le réseau de textes dans lequel circule l'information dont le fil conducteur est le *texte* lui-même que l'on retrouve par exemple d'un manuscrit à l'autre.

Il s'agit donc pour nous de proposer des solutions pour permettre à un réseau d'acteurs d'éditer un réseau de textes et de métadonnées décrivant ces textes et leurs supports.

Dans cette perspective, le document agit plus comme un élément bloquant qui enferme le texte dans un état donné à un instant précis dans l'histoire du texte. En effet, s'il existe des solutions de travail collaboratif de type traitement de texte, comme *Google Docs*, nous avons vu qu'elles n'offrent pas de solutions de marquage assez précises pour rendre compte de la complexité des types de textes rencontrés.

En revanche, le travail en local et personnel, c'est-à-dire non-collaboratif, sur des documents richement structurés est possible avec des éditeurs XML intégrant une gestion des modèles métiers comme la TEI et l'EAD. Nous avons vu qu'il existait plusieurs solutions pour manipuler des documents XML avec des approches très différentes : de la présentation de l'arbre XML à l'accès direct à la sérialisation du code.

Ces différentes approches ne se limitent pourtant pas à de simples écarts ergonomiques et il nous faut ici entrer dans les détails qui ne seront pas sans impact pour notre étude.

Vincent Quint propose de considérer le document comme deux structures qui interagissent continuellement l'une avec l'autre : la *structure logique* et la *structure physique* [Quint, 1987]. Nous retrouvons ici le principe déjà évoqué de distinction du fond et de la forme, mais en intégrant le fait qu'on ne peut jamais accéder au fond sans forme. Autrement dit, il s'agit, ni plus ni moins, d'intégrer le fait que l'accès aux données est *toujours* médiatisé par une forme apportée à ces mêmes données. Dès lors, la conception de cette médiatisation n'est plus seulement une question d'ergonomie, mais doit donner une compréhension la plus juste possible du modèle qui préside à l'organisation des données au sein du document. De ce point de vue l'accès à la sérialisation du code XML est à considérer comme une forme, qui n'est, en outre, probablement pas la plus efficace pour donner une représentation de l'organisation logique du document en arbre. Nous ne prétendons pas ici que la manipulation du code XML est une mauvaise chose en général, mais simplement qu'il ne s'agit pas de la solution la plus adaptée pour manipuler le texte du point de

vue des éditeurs scientifiques et matériels. L'intervention directe sur le code reste, en revanche, souvent une solution terriblement efficace pour l'ingénieur par exemple.

En suivant la distinction de Vincent Quint, nous proposons ici de considérer l'*arbre* XML, et non sa sérialisation, comme la *structure logique* du texte et de lui ajouter une (ou des) *structure(s) physique(s)* permettant l'accès aux données.

Il faut ici faire la distinction entre plusieurs types de structures physiques : les *structures physiques synchrones* et les *structures physiques asynchrones*. Les premières peuvent se calculer en temps réel et proposent des solutions d'interaction et de modification tandis que les secondes se calculent à un instant donné et ne proposent pas de solution d'intervention sur le contenu du document.

Les structures physiques synchrones sont totalement dynamiques et sont calculées en temps réel. Ainsi, chaque modification de la structure logique du document ou de son contenu est prise en compte de manière à proposer une vue à jour à l'utilisateur. C'est au moment du calcul d'une structure physique synchrone que les liens vers les ressources externes comme les images par exemple pourront être intégrées pour simplifier la compréhension et la manipulation du document.

Les structures physiques asynchrones sont, quant à elles, calculées à la demande et ne permettent pas d'intervention sur le document. Elles peuvent rester dynamiques, mais ne sont pas calculées en temps réel. C'est la logique de fonctionnement des formateurs de textes. Un fichier PDF issu de la compilation d'un code LaTeX pour lire le résultat *pendant* la rédaction est un exemple de ce que nous proposons d'appeler une structure physique asynchrone.

La figure 7.1 donne une représentation de l'organisation d'un document tel que nous proposons de le manipuler. Les structures physiques ne sont pas toutes représentées ici, il s'agit d'en donner quelques exemples. En fonction de ses besoins un utilisateur peut convoquer telle ou telle structure physique pour obtenir une vue articulant une structure logique et une ou plusieurs structures physiques.

Cette distinction entre structure physique synchrone et asynchrone se retrouve toujours, même dans le cas d'un choix porté sur un éditeur de code. En effet, répétons le, l'indentation doit être considérée comme une structuration physique : elle n'apporte rien sur le plan logique et a pour seul objectif de permettre la compréhension de la structure logique par l'utilisateur humain. Pour les machines, l'indentation ne sert strictement à rien [92]. D'ailleurs la plupart des logiciels, comme OpenOffice.org, qui manipulent du XML le sérialisent sur une ligne unique...

92. Nous avons d'ailleurs vu comment l'indentation peut poser problème, notamment pour la production de certaines formes. Voir p. 120.

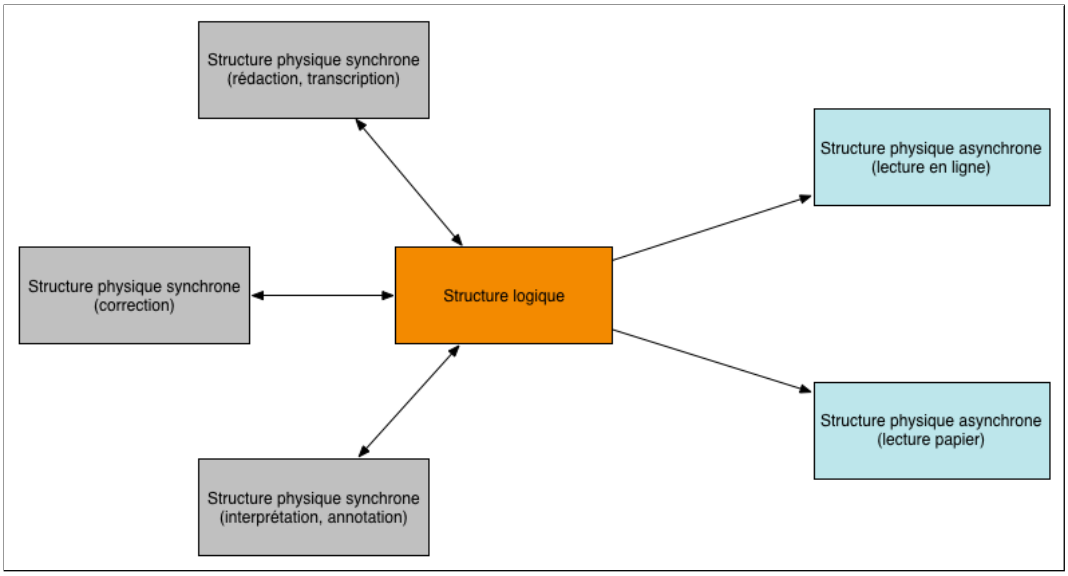

FIGURE 7.1 – Structure logique et structures physiques de document.

Ainsi, le travail sur des documents richement structurés est possible et il existe des solutions logicielles dont les logiques de fonctionnement sont très proches de celle que nous proposons. Nous avons déjà évoqué XMLmind XML Editor, mais il en existe d'autres comme Xmetal [93] par exemple.

Mais comment adapter ces solutions à des flux de données, à des réseaux de textes ? Nous posons l'hypothèse que c'est l'objet texte, tel que nous l'avons défini plus haut [94] qui offre la base la plus sûre pour bâtir notre modèle en servant en quelque sorte de fil conducteur.

7.2 Flux de texte et fragments de texte

Pour dépasser les limites imposées par la notion de document dans le cadre de l'édition numérique de sources anciennes, nous proposons ici de nous inspirer des modèles documentaires sous-jacents des systèmes d'agrégation d'informations, les flux *Really Simple Syndication* (RSS) [95] et de l'outil de réseau social Twitter [96]. Il s'agit d'une approche métaphorique pour mieux illustrer notre notion de flux de texte.

La production des informations contenues dans un flux RSS est générée à partir des unités documentaires qui composent un site web, typiquement un article d'un

93. http://xmetal.com.
94. Voir p. 79 et p. 99.
95. http://www.rssboard.org/
96. https://twitter.com

journal électronique ou d'un blog par exemple. L'organisation des unités en catégories peut ou non être reprise pour proposer plusieurs flux d'informations différents. Autrement dit, le flux RSS est constitué à partir de l'ensemble des informations contenues dans les bases. Ces informations font l'objet d'un formatage et d'une exposition de données dans un nouveau format, celui des flux RSS. Chaque unité devient une partie d'un ou de plusieurs flux d'informations éventuellement typés en fonction des catégories auxquelles il appartient dans le contexte du site dont il est issu.

L'utilisateur intéressé par un flux RSS peut s'y *s'abonner* en utilisant l'application de son choix. Le flux d'informations structuré, composé d'unités de base, sera intégré dans un ensemble au cœur de l'application, constitué par tous les flux sélectionnés par l'utilisateur, et visualisé d'une manière spécifique en fonction de l'application utilisée par le lecteur. La figure 7.2 donne deux exemples de visualisations, donc deux mises en forme différentes, d'un même flux d'informations en utilisant deux applications différentes : l'application Flipboard au premier plan et le navigateur web Firefox au second. Notons aussi que toutes les applications ne traitent pas le contenu des unités de la même manière. Certaines proposent un affichage synthétique des titres pour permettre la sélection des unités que l'utilisateur souhaite consulter, d'autres comme Flipboard tronquent le contenu pour occuper un volume donné dans son interface.

C'est donc l'application finale qui recompose l'ensemble des informations récupérées dans un tout cohérent correspondant au choix de l'utilisateur pour aboutir à un *document*, un exemplaire de journal électronique personnalisé, dynamique et produit à partir d'un ensemble de flux récupérés auprès de sources variées. Ou, pour le formuler avec les catégories proposées, l'application récupère la *structure logique* exposée par l'émetteur (le site source), en filtrant éventuellement au passage certaines informations, et produit une *structure physique asynchrone* proposée à l'utilisateur.

Ces applications sont très utilisées dans le cadre de veille informative sur un sujet précis. En effet, il est très simple d'organiser des flux par thème et ainsi de centraliser les informations sur une thématique particulière sans avoir à parcourir plusieurs sites différents.

Un flux RSS se compose donc de deux objets principaux : un canal (`channel`), qui est en réalité le flux en lui-même, et qui se compose d'items (`item`), chacun de ces derniers correspondant, *a priori*, à une unité de base originale dont le flux est extrait. Autrement dit, quelle que soit la nature des informations dont provient le flux elles doivent être ramenées à ces deux niveaux de hiérarchisation. Bien entendu, il est possible d'encapsuler dans un `item` des données structurées respectant l'organisation

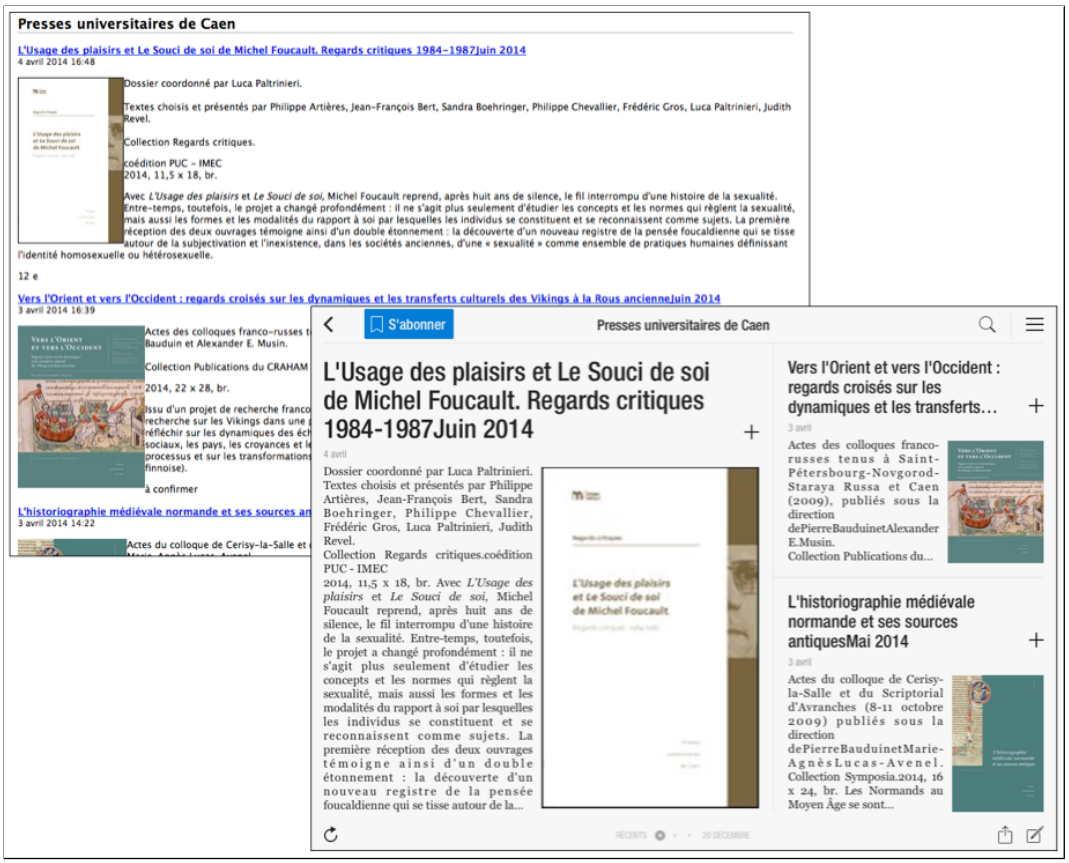

FIGURE 7.2 – Deux visualisations d'un même flux RSS.

interne des unités en utilisant `content:encoded` ou `description` en fonction de la version de la syntaxe RSS.

La figure 7.3 donne en exemple extrait de code RSS de base issu du flux correspondant aux visualisations de la figure 7.2. C'est en interprétant ce code RSS que Flipboard et Firefox produisent les deux visualisations présentées plus haut. Nous reviendrons plus loin sur les méthodes d'adaptation des notions retenues pour les documents aux flux de textes.

Bien entendu, les flux RSS se limitent à la consultation et à la lecture des contenus. Examinons maintenant une solution qui propose aussi l'ajout d'information à un flux d'informations : twitter.

Si les flux RSS se caractérisent par leur multiplication en fonction des sources, chaque source produisant ses propres flux pour permettre leur lecture, Twitter propose un flux unique, une sorte de grande conversation centralisée à laquelle chacun, c'est-à-dire chaque utilisateur enregistré, peut participer. Mais il est impossible de tout lire, ce flux unique ne se prête pas à une lecture immersive bien entendu et

```
<?xml version="1.0" encoding="utf-8"?>
<rss version="2.0"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:content="http://purl.org/rss/1.0/modules/content/">

<channel>
        <title>Presses universitaires de Caen</title>
        <link>http://www.unicaen.fr/puc/</link>
        <description></description>
        <language>fr</language>
        <generator>SPIP - www.spip.net</generator>



 <item>
                <title>Centres de pouvoir et organisation de l'espaceJuin 2014</title>
                <link>http://w3.unicaen.fr/services/puc/spip.php?article942</link>
                <guid isPermaLink="true">http://w3.unicaen.fr/services/puc/spip.php?article942</guid>
                <dc:date>2014-04-29T09:02:06Z</dc:date>
                <dc:format>text/html</dc:format>
                <dc:language>fr</dc:language>


 <content:encoded>&lt;img class=&quot;spip_logos&quot; alt=&quot;&quot; align=&quot;left&quot; src=&quot;http://w3.unicaen.fr/service
s/puc/IMG/arton942.jpg?1398763462&quot; width=&quot;139&quot; height=&quot;209&quot; /&gt;
                &lt;div class='rss_chapo'&gt;Actes du X&lt;sup&gt;e&lt;/sup&gt; colloque international sur l'histoire et l'arch&#233;
ologie de l'Afrique du Nord pr&#233;historique, antique et m&#233;di&#233;vale (Caen, 25-28 mai 2009) r&#233;unis par &lt;contri
buteur role=&quot;directeur&quot;&gt;&lt;prenom&gt;Claude&lt;/prenom&gt; &lt;nom&gt;Briand-Ponsard&lt;/nom&gt;&lt;/contributeur&gt;.&
lt;br/&gt;&lt;br/&gt;Collection Symposia&lt;/div&gt;
                &lt;div class='rss_chapo'&gt;&lt;p&gt;2014, 16 x 24, br., 650 p., ill.&lt;/p&gt;&lt;/div&gt;
                &lt;div class='rss_texte'&gt;&lt;p&gt;&lt;span class='spip_document_1775 spip_documents spip_documents_left' style='f
loat:left; width:207px;'&gt;
&lt;/span&gt;Sur le th&#232;me f&#233;d&#233;rateur de l'organisation des territoires, ce dixi&#232;me colloque sur l'histoire et
 l'arch&#233;ologie de l'Afrique du Nord antique et m&#233;di&#233;vale s'inscrit tant dans la continuit&#233;é des grandes r&#233
;éunions scientifiques organis&#233;es par la Soci&#233;t&#233;é d'&#233;tudes pour le Maghreb pr&#233;historique antique et m&#2
33;édi&#233;val, ancien CTHS-Afrique, que dans le renouvellement. Avec des analyses stimulantes et l'expos&#233;é de d&#233;couvert
es r&#233;centes, historiens et arch&#233;ologues ont orient&#233;é leurs travaux selon quatre angles d'approche : les centres de p
ouvoir et la hi&#233;rarchie des r&#233;seaux urbains ; l'organisation de son territoire par la cit&#233;é ; le r&#233;seau routie
r, le bornage et la circulation de productions ; la religion comme &#233;l&#233;ment structurant.&lt;/p&gt;&lt;/div&gt;
                &lt;div class='rss_ps'&gt;&lt;p&gt;42 e&lt;/p&gt;&lt;/div&gt;
                </content:encoded>

        </item>
 <item>
                <title>n&#176;° 45 :  L'&#233;ducation en exercice(s)Juin 2014</title>
                <link>http://w3.unicaen.fr/services/puc/spip.php?article941</link>
                <guid isPermaLink="true">http://w3.unicaen.fr/services/puc/spip.php?article941</guid>
                <dc:date>2014-04-29T08:15:22Z</dc:date>
                <dc:format>text/html</dc:format>
                <dc:language>fr</dc:language>


 <content:encoded>&lt;img class=&quot;spip_logos&quot; alt=&quot;&quot; align=&quot;left&quot; src=&quot;http://w3.unicaen.fr/service
s/puc/IMG/arton941.jpg?1398759294&quot; width=&quot;137&quot; height=&quot;210&quot; /&gt;
                &lt;div class='rss_chapo'&gt; &lt;/div&gt;
```


FIGURE 7.3 – Exemple de code d'un flux RSS.

l'accès se fait par sujet, les fameux *hashtags*, ou par auteur. Chacun peut suivre les interventions d'une personne spécifique ou au contraire toutes les interventions sur un sujet précis. Enfin, la lecture se fait également sur la base d'une fenêtre temporelle donnée : la consultation à un instant t ne donnera pas accès au même contenu qu'à instant $t + 1$.

Ici, on ne trouve donc plus de document : nous sommes vraiment totalement dans un modèle de flux qui agrège des fragments. Chaque utilisateur émet ses fragments en précisant directement dans le message les mots-clés auxquels il se rattache. Les autres utilisateurs pourront ainsi accéder soit aux flux de l'ensemble des fragments rattachés au même mot-clé, soit au flux d'un utilisateur sans distinction de sujet.

Cependant, si twitter ne se base plus sur une quelconque notion de document, il reste possible de produire des documents à partir d'un flux de fragments, par exemple à des fins d'archivage. Par ailleurs, il est aussi possible de considérer que le simple

fait de consulter un ensemble de fragments ordonnés de manière chronologique sur un sujet particulier et regroupés dans une interface de lecture constitue déjà une instanciation de document.

La structure logique est donc ici celle d'un flux unique composé de fragment eux-mêmes composés de texte, de mots-clés, d'images et de liens. Les structures physiques sont générées par les applications au moment de la consultation : l'ensemble des fragments correspondants aux critères d'interrogation, c'est-à-dire au sujet ou à l'auteur demandé, sont regroupés dans l'interface de lecture. Bien entendu, il est possible de proposer de consulter les flux des sujets qui suscitent le plus d'interventions, ou de suivre les interventions des utilisateurs les plus populaires, etc. Même si cette souplesse s'appuie en grande partie sur la simplicité des briques d'information (un seul type de contenu, absence de hiérarchie, etc.), cette dimension dynamique est aussi apportée par la présence d'un flux d'information malléable et réorganisable en fonction des choix des utilisateurs.

Les flux d'informations correspondent à la nature profondément fluide du texte et les deux exemples d'exploitation de flux que nous venons de décrire peuvent servir dans l'effort d'adaptation de cette logique aux données qui nous concernent ici. Pensons par exemple au *codex* qui n'a pas toujours été présent et au *rotulus* et au *volumen*, qui l'ont précédé, qui sont en définitive plus proches de cette logique de flux.

D'autre part, le réseau vient en quelque sorte bouleverser nos méthodes de représentation et de travail en rendant, potentiellement au moins, opérationnelle la dimension connectée des textes manipulés par le biais des documents. Le système de références bibliographiques doit ainsi être considéré comme une forme de lien hypertexte de plus en plus souvent actif dans les textes scientifiques. Ainsi, un chercheur peut suivre les références d'un article au fil de sa lecture. Pourtant, la logique de flux reste peu utilisée lors des phases de conception des textes : nous restons enfermés dans une logique de production des textes dans des documents.

Nous proposons maintenant de nous appuyer sur les types de solutions de syndication et de réorganisation évoqués plus haut pour ordonner des flux d'informations éditoriaux avec des unités atomiques de données d'importance variable. Dans ce contexte, le document doit être perçu comme un instant dans la vie du texte qui devra être replacé dans un flux plus vaste. En réalité, nous ne manipulons plus les documents que dans le cadre de la constitution d'un flux ou de fragments d'informations. Ainsi, le document est soit un instant dans la vie de ce flux, par exemple pour

lui donner naissance, soit un moment de création de sens, résultat de l'extraction d'une séquence d'informations contenues dans un flux plus vaste. Bref le document ne doit plus être considéré autrement que comme une version bornée d'un flux de connaissances qu'il devra à terme réintégrer.

Il s'agit pour nous de proposer un nouveau cadre logique pour l'organisation du travail : celui du flux de texte qui nous semble beaucoup plus opérationnel que celui de document. Nous posons donc l'hypothèse que le travail d'édition de sources s'effectue sur des fragments de texte intégrés à des flux, qui sont manipulés, interrogés et organisés en fonction des besoins : transmission des connaissances, étude d'un phénomène, histoire des textes, enseignement, etc. Nous proposons donc de lever cette tension entre flux d'informations et documents, résultats d'extraction ou fragments de flux. Il s'agit aussi de l'intégrer le plus tôt possible dans les pratiques pour bénéficier de ses avantages.

7.2.1 Flux de texte

Qu'est-ce qu'un flux de texte dans le cadre de l'édition de sources ?

Les artefacts originaux, typiquement des manuscrits et des imprimés anciens, doivent être considérés comme des porteurs de textes, qui sont eux-mêmes constitutifs d'un flux cohérent, autrement dit, ce sont des documents qui instancient un flux de texte. Dans cette optique, les bas de pages doivent être considérés comme des événements matériels liés au support "choisi" pour le texte. De la même manière, une collection d'informations prosopographique doit être considérée comme un flux d'informations sur des personnes. Il s'agit donc de fournir au chercheur un modèle qui lui permette de constituer ou de reconstituer des flux de textes à partir des artefacts qui les portent. Nous verrons plus loin en détail[97] l'exemple des manuscrits du Mont Saint-Michel utilisés pour *Les Chroniques latines du Mont Saint-Michel* [Bouet et Desbordes, 2009][98].

C'est le texte qui sera utilisé comme objet central car c'est lui qui circule le plus entre les acteurs, c'est lui qui va permettre d'organiser une continuité entre l'ensemble des autres objets. De plus, le texte porté par les témoins concerne aussi bien l'archiviste que le chercheur ou l'éditeur matériel.

97. Voir p. 175.

98. http://www.unicaen.fr/services/puc/sources/chroniqueslatines/, pour consulter la version en ligne de cette édition.

Le modèle FRBR du monde des bibliothèques est ici utile pour étayer notre propos. Ce modèle conceptuel propose quatre niveaux de description dans une notice catalographique :

— caractéristiques d'exemplaire [item / *item*] ;
— caractéristiques de la publication [manifestation / *manifestation*] ;
— caractéristiques du contenu intellectuel [expression / *expression*] ;
— caractéristiques de la création abstraite à laquelle est attaché le contenu [œuvre / *work*].

Il s'agit donc de mettre en place un système de flux de textes qui vont s'articuler autour d'un flux "central" complet correspondant au texte "établi" de l'œuvre, ou encore, pour reprendre une formulation FRBR, de l'expression de l'œuvre, si tant est qu'il y en ait une, établie par un chercheur ou un groupe de chercheurs. Les autres flux, des manifestations dans le modèle FRBR, pourront être créés par des opérateurs, par exemple pour proposer de nouvelles fonctionnalités, d'autres parcours de lecture, la restitution du texte d'un artefact particulier, un nouvel établissement du texte ou encore pour des exploitations spécifiques comme la production d'une forme de lecture [99] à partir du flux principal. Il faut donc ici distinguer deux types de flux différents. Les premiers sont les flux contenant l'information dans toute sa complexité, indépendamment de toute forme et de toute exploitation, correspondant véritablement à l'état des connaissances sur le texte souhaité par le chercheur et incluant donc par exemple l'ensemble des variantes des témoins retenus pour l'établissement du texte, toutes les notes quels que soient leurs types (commentaires scientifiques, philologiques, etc.), liens vers les ressources iconographiques, etc. Les seconds sont des flux de production qui sont extraits des premiers. Leur rôle est de permettre une exploitation donnée dans un contexte donné, par exemple la diffusion du texte sous une forme papier.

Dans ce contexte, un flux de texte peut se voir attribuer un certain nombre de caractérisations qui peuvent prendre la forme de métadonnées destinées à faciliter sa découverte ou son exploitation mais aussi les principes qui ont présidé à sa constitution. Ces ensembles de métadonnées pourront être associés à chacun de ces flux en fonction de leur nature (œuvre, texte, artefact, édition) : titres, auteurs, éditeurs scientifiques, cotes de manuscrits, ISBN, date d'impression, ville d'édition, etc.

99. Nous reviendrons sur cette question quand nous examinerons la contrainte d'un flux de texte dans une planche de PAO pour la production d'un livre papier. Voir p. 205 et 224.

7.2.2 Fragments de textes / unités logiques

Les flux se composent de fragments, c'est-à-dire d'unités atomiques, dont ils proposent des organisations logiques correspondant à des analyses scientifiques, à des artefacts, etc.

Les fragments peuvent être de volume très variable comme nous le verrons lors de l'examen détaillé des expérimentations développées plus loin. Ces unités atomiques doivent être considérées comme des briques d'informations qui s'articulent les unes avec les autres pour composer des flux cohérents en fonction des objectifs scientifiques et éditoriaux. Il peut par exemple s'agir de chapitres définis par la source, c'est-à-dire l'artefact ou par l'éditeur scientifique. Ces fragments peuvent aussi être des personnes dans un flux d'informations prosopographique, etc. Mais il peut aussi s'agir de fragments plus petits : segment de texte pour marquer une séquence de discours rapporté, empans de texte traitant d'une question spécifique, etc. Ainsi, des fragments peuvent eux-mêmes se composer d'unités de moindre importance. L'exemple le plus évident est celui d'un chapitre qui se compose de sections contenant elles-mêmes des paragraphes. Dans ce contexte, une unité atomique désigne un fragment d'importance variable en fonction du niveau de structuration et des objectifs éditoriaux.

Ces fragments pourront donc correspondre à des unités logiques (chapitres ou sections par exemple) ou à des unités de traitement. En effet, dans la mesure où toute séquence de texte qui fait l'objet d'un marquage peut être exploitée, manipulée et extraite de son flux, toutes les séquences ainsi discriminées doivent être considérées comme des unités atomiques.

En réalité, la définition des fragments et de leur taille est avant tout dépendante de l'utilisation envisagée. Autrement dit, chaque séquence de texte marquée peut faire l'objet d'une extraction pour être réintégrée dans un nouveau flux d'informations correspondant à une nouvelle logique ou à une exploitation précise. Ainsi, pour prendre un exemple trivial, pour proposer un nouveau parcours de lecture d'un texte composé de chapitres, il est indispensable d'extraire chacun des fragments correspondant aux chapitres [100]. Un nouveau flux ajoutant ou supprimant des fragments peut aussi être nécessaire pour permettre ou pour faciliter l'indexation des données dans un outil d'analyse statistique ou de fouille par exemple. Cependant, il est important de noter que les fragments peuvent aussi avoir une dimension logique et significative et pas seulement technique quand ils correspondent à des chapitres ou des sections

100. Nous reviendrons sur cet exemple avec le cas des *Chroniques latines du Mont Saint-Michel*. Voir p. 175 et suivantes.

par exemple. D'un point de vue plus théorique, nous définissons les fragments comme des portions de texte déterminées. Les fragments sont totalement dépendants du sens que l'on souhaite donner au texte. Ainsi les questions qui doivent présider à la détermination du niveau de fragmentation vont être de l'ordre suivant : quel découpage est le plus efficace pour donner le texte à lire ? Quel parcours de lecture est le plus pertinent dans un réseau de textes ? Quels types de requêtes doivent être possibles sur le corpus et quel découpage impliquent-ils (grain paragraphe, phrase, mot...) ?

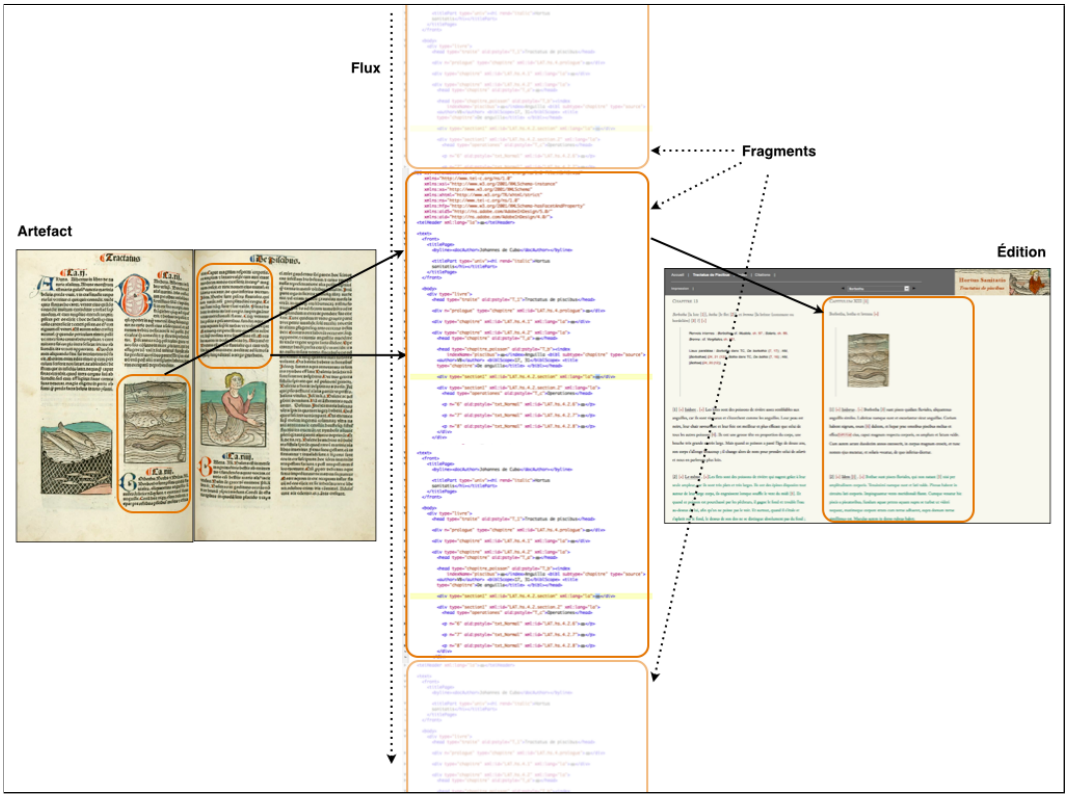

FIGURE 7.4 – De l'artefact à l'édition en passant par le flux de texte.

La figure 7.4 donne un exemple de représentation des rapports entre flux de texte, fragments de texte et exploitation éditoriale en ligne. Ici, le flux est constitué à partir de la transcription d'un texte porté par un artefact. Pendant cette phase du travail plusieurs types de fragments sont définis et marqués. Le niveau mis en exergue est ici celui du chapitre, qui se compose d'un titre, d'une image et d'une séquence de texte plus ou moins longue. L'ensemble des fragments de ce type est agrégé pour composer un flux correspondant à la totalité du texte porté par l'artefact d'origine. Le résultat des opérations est donc un flux de texte dans lequel tous les chapitres sont précisément marqués de manière systématique et contenant également des liens vers les versions numériques des illustrations présentes dans l'artefact. On notera au

passage que les ruptures de page de l'artefact de base peuvent être indiquées dans ce flux, mais simplement sous la forme d'un point signalant le changement. Chacun des fragments correspondant à un chapitre fait ensuite l'objet d'une extraction pour être intégré à l'édition en ligne finale dans laquelle l'affichage se fait par chapitres, qui constituent les unités de lecture proposées au lecteur. En termes FRBR, le flux peut, en réalité, être considéré comme l'*Expression* et l'édition en ligne comme une *Manifestation*[101], l'artefact étant lui, puisqu'il s'agit d'un imprimé, un *Item*[102].

Sans entrer dans des notions d'implémentation trop prématurément, notons que dans l'état actuel des techniques les fragments sont le plus souvent des sous-arbres XML. C'est de notre point de vue la solution la plus satisfaisante sur les plans technologique et intellectuel. Chaque niveau des arborescences correspond potentiellement, ou virtuellement, à un niveau fragmentaire et est donc manipulable en tant que tel.

Chaque fragment peut se voir associé une série de métadonnées pour le caractériser comme la source dont un passage est issu, les informations bibliographiques concernant une citation, un titre de chapitre, etc.

Enfin pour que chaque fragment soit discriminé par rapport aux autres il se voit attribuer un identifiant unique calculé sur la base de sa situation dans la structure logique du flux de texte borné dont il est issu au début du travail. Le mode de calcul de ces identifiants de fragments exploite la structure en arbre et s'inspire du système de références absolues utilisé dans le domaine de l'édition de textes classiques dans lequel chaque portion de texte, du chapitre à la ligne si nécessaire, est numérotée. Ainsi une référence comme `TC 6, 3` renvoie au livre 6, chapitre 3 de l'œuvre de Thomas de Cantimpré. Le calcul des identifiants reprend cette logique de numérotation du texte en exploitant sa structure arborescente.

Un préfixe de langue est utilisé en début de fragment pour lever toute ambiguïté sur ce point. Ensuite une série de compteurs est ajoutée dont la longueur est définie en fonction du niveau de profondeur à identifier.

Ainsi, par exemple, l'identifiant `FR.4.1.4` désigne le quatrième segment du premier paragraphe du quatrième chapitre du flux français de l'*Hortus Sanitatis*, en suivant le modèle `langue.chapitre.paragraphe.segment`. Les notions de chapitre, de paragraphe et de segment sont associées, dans l'arbre XML, aux éléments correspondants dans le système d'attribution.

101. nous considérons ici que les pages web de cette édition seront identiques quel que soit le navigateur utilisé pour les consulter.

102. Nous retenons ici la terminologie anglophone dans la mesure où la traduction proposée est *Document*, ce qui pourrait, à ce stade de notre étude, introduire quelque confusion.

7.2.3 Documents

Bien entendu, nous ne prétendons pas ici que la notion de document est à proscrire. Simplement, il nous semble que cette dénomination, très massivement utilisée pour désigner une grande variété d'objets informatiques, doit être manipulée avec une grande précaution.

Par ailleurs, définir la place du document au sein de notre modèle va aussi nous permettre d'en préciser le périmètre.

Ainsi, dans le modèle de flux et de fragments que nous proposons, quelle place devons-nous donner au document ? Dans le contexte de cette étude, nous définissons le document comme la combinaison d'une *structure logique* et d'une ou de plusieurs *structure(s) physique(s) synchrone(s)* ou *ashyncrone(s)*. Cette combinaison est une instanciation temporaire qui articule du texte et des ressources externes, comme des images, des sons ou des vidéos. La temporalité est bien évidemment un paramètre central qu'il ne faut en aucun cas écarter. Le document est donc en réalité uniquement un instant dans la vie du texte. Ainsi, par exemple, un document est, dans notre modèle, un fragment en cours d'édition ou en cours de lecture. Dans tous les cas, c'est, comme le dit Jean-Michel Salaün, « le résultat de requêtes et de calculs » sur les bases de structures logiques mises en place par des spécialistes.

On retrouve ici une des propositions avancées par Pédauque, qui définit le document comme un ensemble de données structurées associé à une mise en forme (*document = données structurées + mise en forme*[103]) et qui précise :

> Un dernier pas serait ainsi en train de se franchir : un document n'aurait de forme à proprement parler qu'à deux moments : celui de sa conception par son auteur qui devra le visualiser ou l'entendre, pour s'assurer qu'il correspond à ses choix [...] et celui de sa re-construction par un lecteur [Pédauque, 2003].

Nous proposons bien ici de sauter ce pas. Mais alors, si l'on considère que le document est effectivement la combinaison de données structurées et d'une mise en forme, s'il n'y a plus de mise en forme, s'agit-il encore d'un document ? Autrement dit, avons-nous toujours bien à faire à des documents entre la *création* et la *re-construction* pour la lecture ? Car, en réalité, une fois ôtée la mise en forme, il ne reste bien que les données structurées de la définition de Pédauque et nous obtenons alors une équation du type *données structurées = document – mise en forme*. Dans

103. Pédauque propose aussi cette définition longue : « Un document numérique est un ensemble de données organisées selon une structure stable associée à des règles de mise en forme permettant une lisibilité partagée entre son concepteur et ses lecteurs. ».

notre modèle, entre ces deux moments, *création* et *re-construction*, il ne s'agit plus vraiment de document, mais bien de données fragmentaires structurées et intégrées dans un flux.

Toujours du point de vue chronologique, dans le cycle de vie des données, les documents sont présents dès les premières étapes de traitement, ce sont les sources anciennes, et à la fin des opérations, quand le lecteur choisit son interface de lecture pour accéder au texte sous la forme de page web par exemple. Enfin, les documents apparaissent aussi de manière temporaire à chaque intervention sur un fragment, après son extraction du flux.

Autrement dit, le travail débute par l'étude de documents, les témoins, et vise justement à séparer le texte de sa mise en forme sur son support, c'est-à-dire des événements matériels tels que rupture de pages, de lignes de colonnes, dégradations, etc., pour lui adjoindre ensuite, lors de l'édition, de la lecture en ligne ou de la production d'une version papier, de nouvelles caractéristiques matérielles. Il s'agit de faciliter le travail des éditeurs de sources anciennes, dont l'objectif est précisément de simplifier la compréhension du texte en le rendant directement utilisable par les lecteurs. En définitive, du point de vue technique, l'édition de sources anciennes consiste à extraire le texte de son, ou de ses support(s) pour le transmettre dans des conditions plus adaptées à son utilisation. Le travail consiste donc à manipuler le texte en distinguant sa structure logique et sa forme initiale.

Ainsi, il s'agit finalement de permettre le travail sur le texte sans les contraintes de forme de tel ou tel support. D'une certaine manière, et en poussant le raisonnement à l'extrême, le modèle de flux et de fragments que nous proposons cherche précisément à sortir de la logique de document, conçu comme *document = données structurées + mise en forme*, pour être en mesure d'instancier correctement les bons documents aux bons moments sans que les formes initiales, celles des sources anciennes, pèsent continuellement et entravent le travail de l'éditeur. Il s'agit bien de profiter du mouvement de *dématérialisation* propre à la numérisation pour en tirer un profit maximum du point de vue de l'efficacité. En dissociant *structure logique* et *structures physiques*, nous proposons de laisser aux utilisateurs, du chercheur qui établit le texte au lecteur en passant par l'éditeur matériel, le choix de la forme la plus appropriée à son utilisation. En définitive, il s'agit de poser que, selon les utilisations et les moments (découverte, établissement du texte, étude du contexte, etc.), toutes les formes ne se valent pas pour lire, voir et savoir. Il est par exemple moins aisé de lire un texte ancien sur son support original que de lire sa transcription

rigoureuse accompagnée des notes qui décrivent les principes qui ont présidé à son établissement et de liens vers les images numériques des sources primaires.

La distinction entre *structure logique* et *structures physiques* est aussi complexe à établir : la contrainte des textes sur des pages de codex est ancienne, parfaitement fonctionnelle et totalement intégrée dans nos esprits. Ces deux types de structures sont tellement imbriqués dans nos usages que Jean-Michel Salaün parle même de la « structure logique en *pages*[104], en chapitres » [Salaün, 2012] de l'intérieur des livres. Pour autant l'impératif de séparation de ces deux structures, imposé en particulier par la multiplication des supports de diffusion, ne doit pas être interdite par la complexité de leur articulation. Raisonner en terme de fragments logiques liés les uns aux autres dans des flux permet justement de penser et de manipuler la structure logique indépendamment de toute forme. Dans ce modèle, il faut considérer les pages, les cahiers et les dégradations, bref l'ensemble des aspects matériels, comme des événements, des ruptures, qui relèvent d'une forme spécifique et non de l'organisation logique du texte, en parties, chapitres, sections, etc., et qui devra elle, trouver une expression dans toutes les formes de diffusion associées au texte.

C'est parce que les fragments n'ont pas de forme stable associée, qu'ils échappent en partie aux dimensions décrites par Jean-Michel Salaün, en revanche les fragments peuvent se voir attribuer une forme et réamorcer le cycle, devenant potentiellement à ce moment, de nouveaux témoins. En définitive, ils contiennent *en puissance* des formes et sont prêts à devenir des documents avec toutes les dimensions proposées par Jean-Michel Salaün ; autrement dit les fragments sont à considérer comme des documents *virtuels*[105]. C'est un point central et c'est tout l'intérêt du modèle de flux et de fragments. Débarrassés de toute contrainte de forme, les flux sont adaptés à la conservation et à l'archivage à long terme. Du point de vue de l'économie éditoriale et dans un contexte d'évolution rapide des supports de lecture, c'est incontestablement plus un point fort qu'une faiblesse.

Prenons l'exemple d'un chercheur qui travaille sur une séquence de texte. Il extrait uniquement la partie du flux qui l'intéresse et intervient sur le fragment résultant. D'une certaine manière, il faut considérer que pendant cet espace de temps, le fragment en cours d'édition, c'est-à-dire un ensemble de type *données + structure logique*, qui s'est vu attribuer une *structure physique synchrone*, est un document, donc, *données + structure logique + structure physique synchrone*. Une fois les modifications

104. C'est nous qui soulignons.

105. C'est-à-dire que ces documents n'ont pas d'existences actuelles.

terminées, le fragment est réintégré dans le flux dont il provient sans la *structure physique synchrone* qui lui été associée uniquement pour le temps de l'édition.

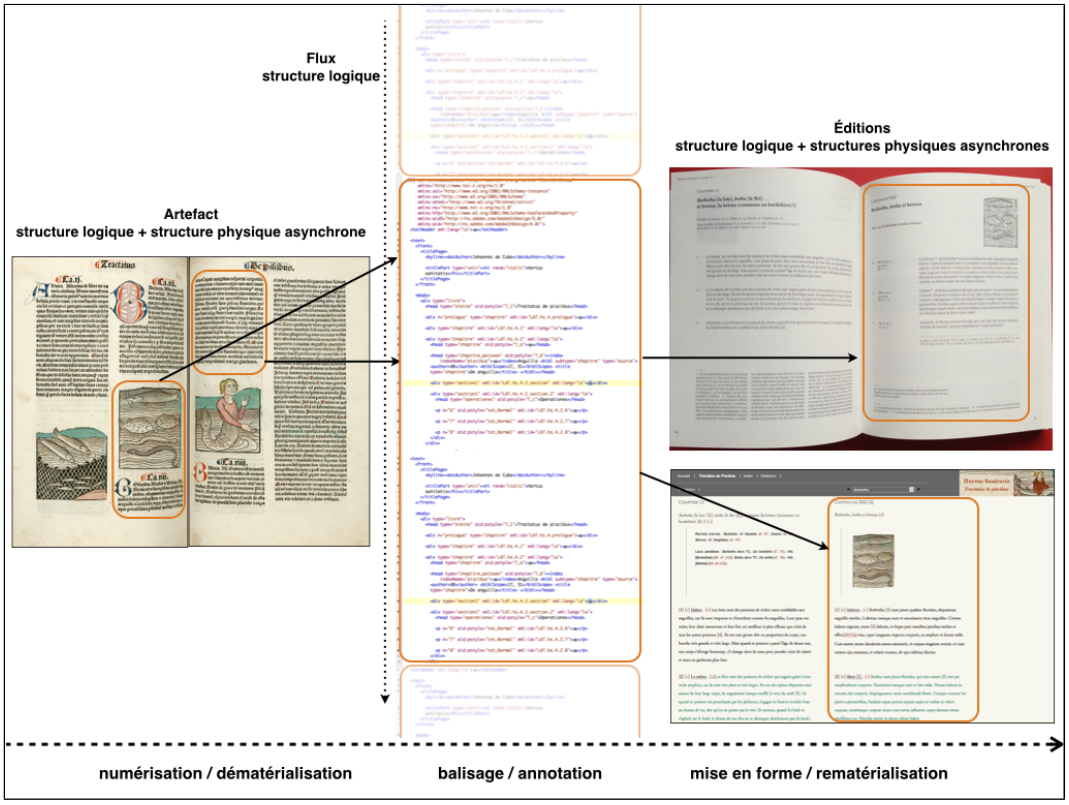

FIGURE 7.5 – Articulation de structures logiques et de structures physiques.

La figure 7.5 reprend les éléments de la figure 7.4 en ajoutant les différents types de structures et les principales étapes d'un projet d'édition de sources anciennes de la dématérialisation du document original jusqu'à la rematérialisation du texte sur de nouveaux supports de diffusion. Nous voyons donc comment le document original se compose d'une structure logique et d'une structure physique asynchrone. Cette dernière est écartée au cours du travail de dématérialisation et la structure logique est organisée en flux sur lequel le travail d'annotation scientifique est réalisé par les chercheurs. Enfin, de nouvelles structures physiques asynchrones sont associées à la structure logique enrichie pendant le travail d'édition matérielle pour produire de nouveaux documents dédiés à la diffusion. Rien n'interdit alors, si le code XML constitutif du flux est diffusé en tant que tel, à des chercheurs (les éditeurs de l'édition ou d'autres) d'amorcer un nouveau cycle d'annotation, éventuellement avec d'autres objectifs scientifiques et éditoriaux.

7.3 Niveaux de balisages

Comme nous l'avons déjà évoqué, la meilleure solution technique pour manipuler des données à n niveaux de structure et sans connaissance *a priori* de la complexité des données et des textes à traiter reste le XML. Les flux et les fragments de texte sont donc le plus souvent encodés en utilisant ces technologies.

La question du niveau d'annotation ou d'encodage, que l'on désigne aussi par la granularité, est centrale dans tout le travail que nous menons ici. Nous proposons de distinguer deux niveaux d'annotation répondant en fait à deux objectifs distincts, même s'ils peuvent parfois se recouvrir comme nous le verrons.

7.3.1 Encodage éditorial

Ce niveau d'encodage correspond au marquage des informations indispensables pour produire des formes intelligibles pour le lecteur. Il s'agit ici de mettre en place un système de balisage permettant à l'éditeur matériel de proposer des formes correspondant aux catégories de textes manipulées par les éditeurs scientifiques. Autrement dit, tous les phénomènes textuels indispensables à la compréhension du propos de l'auteur doivent faire l'objet d'un étiquetage spécifique.

Le balisage éditorial se caractérise par sa simplicité. En effet, pour assurer une mise en forme minimale du texte, l'éditeur matériel a besoin de discriminer les paragraphes de textes, les citations, les séquences en italique ou en gras, les différents niveaux de titres, différents systèmes de notes, etc. Bref un ensemble d'éléments relativement restreint et bien connu dans le monde scientifique. De plus, cet ensemble peut rester assez descriptif. En effet, pour l'éditeur matériel, en tout les cas pour la production de formes élémentaires, l'important est d'encoder le fait que telle séquence de caractères doit être traitée en italique et pas nécessairement la raison pour laquelle elle est en italique. Précisons dès maintenant qu'il ne s'agit en aucun cas de dire que l'éditeur matériel ne doit pas se préoccuper du sens du texte, bien au contraire. Il s'agit ici uniquement du processus de production technique des formes de diffusion et non des tâches de préparation et de correction des textes qui incombent à l'éditeur matériel. Sur le fond, le travail de préparation de copie n'est pas directement impacté par les changements de techniques de production.

Avec un ensemble de 30 éléments environ il est possible de décrire les textes simples les plus courants dans le domaine des sciences humaines et sociales. Comme nous le voyons, le nombre de catégories de textes reste relativement limité et tout à fait manipulable.

L'encodage éditorial va permettre de produire des formes de diffusion simples, c'est-à-dire dépourvues d'outils avancés. Les formes web par exemple ne disposeront que d'instruments de navigation classiques (tables des matières, circulation appels de notes/notes, etc.). Les outils plus fins permettant, par exemple, la restitution du texte d'un témoin spécifique dans le cadre d'une édition de sources seront impossibles à mettre en œuvre.

En définitive, l'encodage éditorial permet simplement de caractériser les éléments qui devront faire l'objet de mises en forme particulières sur tel ou tel support de diffusion.

Ce niveau de balisage, s'il présente une composante formelle forte, ne peut pour autant pas se réduire à l'expression en XML d'une mise en forme. En effet, il est plus juste de le considérer comme une annotation pour des formes de diffusion. Il respecte ainsi le principe de séparation du fond et des formes. Cependant, les éléments balisés sont ceux qui devront être traités formellement de manière spécifique sur un support ou sur un autre.

```
<div type="chapitre">
  <div type="section" xml:id="AFR.1" xml:lang="afr">
    <lg xml:id="AFR.1.1">
      <l n="1" aid:pstyle="txt_Original_Vers" xml:id="vers1"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/"><pb ed="A"
      n="1r" />Molz pelerins qui vunt al munt<note type="marginal"
      xml:id="AFRftn1"> <emph aid:cstyle="typo_Italique">B</emph>: Les
      bones gens qui vunt au mont <emph aid:cstyle="typo_Italique">(voir
      infra), avec majuscule initiale de grande taille.</emph>
      </note></l>

      <l n="2" aid:pstyle="txt_Original_Vers" xml:id="vers2"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Enquierent
      molt, et grant dreit unt,</l>

      <l n="3" aid:pstyle="txt_Original_Vers" xml:id="vers3"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Comment
      l'igliese fut fundee</l>

      <l n="4" aid:pstyle="txt_Original_Vers" xml:id="vers4"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Premierement,
      et estoree.</l>

      <l n="5" aid:pstyle="txt_Original_Vers" xml:id="vers5"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Cil qui lor
      dient de l'estoire</l>

      <l n="6" aid:pstyle="txt_Original_Vers" xml:id="vers6"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Que cil
      demandent, en memoire</l>

      <l n="7" aid:pstyle="txt_Original_Vers" xml:id="vers7"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">Ne l'unt pas
      bien, ainz vunt faillant</l>

      <l n="8" aid:pstyle="txt_Original_Vers" xml:id="vers8"
      xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">En plusors
      leus, et mespernant.</l>
```

FIGURE 7.6 – Exemple d'encodage éditorial.

La figure 7.6 donne un exemple de balisage éditorial. En plus de la hiérarchie de base définie par les éléments `div`, imbriqués sur autant de niveaux que nécessaire, la structure se compose de vers, c'est l'élément `l`, et de groupe de vers, c'est l'élément `lg`, ainsi que de notes destinées à être placées en marge et contenant les variantes des vers concernés. Il s'agit donc d'un encodage très simple dont le vocabulaire ne correspond qu'en partie à la complexité du texte manipulé. On ne trouve en effet pas de caractérisation des rapports entretenus entre le contenu textuel de l'élément `l` et celui de l'élément `note` par exemple. Une étude poussée sur ce point est donc impossible avec un encodage de ce type. En revanche, cette structure permet sans aucun problème de mettre en place des traitements éditoriaux pour la production de formes de diffusion, nous y reviendrons un peu plus loin.

7.3.2 Encodage scientifique

Il s'agit ici, non plus de rendre compte du plus petit dénominateur commun à l'ensemble des formes à produire, mais bien d'annoter avec la précision exigée par la rigueur scientifique l'ensemble des phénomènes textuels en exploitant toute la richesse correspondante dans les recommandations de la TEI. C'est l'encodage mis en place par les communautés de recherche dans le cadre de la plupart des projets menés à l'heure actuelle. Ce type d'encodage est très fortement lié aux types de textes traités et les particularités des objets étudiés ainsi que les objectifs scientifiques contraignent souvent les chercheurs à mettre en place des solutions spécifiques modulées en fonction de chaque projet.

Toutes les solutions de diffusion sont bien entendu envisageables avec un encodage scientifique. Il est ainsi possible de produire aussi bien des éditions simples favorisant une lecture immersive que des applications en ligne richement outillées permettant d'accéder aux textes par un ensemble d'opérations exploitant toute la complexité textuelle rencontrée par les chercheurs.

La figure 7.7 propose une séquence de code correspondant à une granularité scientifique. Il apparaît au premier regard que les catégories de texte manipulées sont beaucoup plus nombreuses que dans le balisage éditorial ; il suffit pour s'en rendre compte d'observer le rapport de volume occupé par le texte (en noir) d'un côté, et par le code XML de l'autre, en particulier au début de la séquence. Nous avons ici à faire à un paragraphe de texte (élément `p`) dont le contenu est lourdement structuré. Outre l'annotation scientifique catégorisée par l'utilisation de l'élément `note` avec des valeurs d'attribut `type` variant en fonction de la nature du commentaire

```
<p n="1" next="2" aid:pstyle="txt_Normal"
xml:id="LAT.hs.4.1.1"><bibl subtype="paragraphe" type="source">
<author>VB</author> <biblScope>17, 29, 1</biblScope> </bibl><index
    indexName="marqueurCitation">
    <term type="orig">Ex Libro de naturis rerum</term>

    <term type="norm">TC</term>

    <term type="identifie">TC</term>
  </index>. <seg type="citation"><bibl subtype="fragment"
type="source"> <author>TC</author> <biblScope>7, 9</biblScope>
<note type="sources" xml:id="LATnote18">Vincent de Beauvais suit
fidèlement Thomas de Cantimpré.</note> </bibl><index
    indexName="piscibus">
    <term>abremon</term>
  </index>Abremon<note type="philologie"
xml:id="LATnote19">Kitchell &amp;amp; Resnick 1999, 1660, n. 49, dans
leur commentaire au chapitre d'Albert le Grand (AM 24, 9 (10))
consacré au poisson <hi rend="italic">abremon</hi>, suggèrent
d'expliquer le nom <hi rend="italic">abremon</hi> par une
corruption de l'expression <hi rend="italic">ab arena</hi>, «_qui
vient du sable_». Un étymon latin, même séduisant, est délicat à
justifier pour un terme qui, de toute évidence, s'inscrit, avant
le latin, dans une tradition gréco-orientale.</note>, alias<note
type="apparat" xml:id="LATnote20">alas <hi
rend="italic">1536.</hi></note> <index indexName="piscibus">
    <term>abarenon</term>
  </index>abarenon, est piscis, ut dicit Aristoteles, multum
fecundus ovis, nec aliorum<note type="apparat"
xml:id="LATnote21"><hi rend="italic">post</hi> aliorum <hi
rend="italic">hab.</hi> ea <hi rend="italic">1491 VB</hi>.</note>
more contactu<note type="apparat" xml:id="LATnote22">contractu <hi
rend="italic">1491</hi>.</note> maris<note type="philologie"
xml:id="LATnote23">Le traducteur médiéval de l'<hi
rend="italic">Hortus sanitatis</hi> a compris <hi
rend="italic">maris</hi> comme le génitif de <hi
rend="italic">mas</hi>, «_le mâle_», en voyant peut-être dans ce
passage une allusion à la fécondation externe des œufs de la
```

FIGURE 7.7 – Exemple d'encodage scientifique.

des chercheurs, l'objectif est ici de donner et d'étudier l'histoire du texte. Pour cela, chaque partie du texte se voit enrichie d'un élément `bibl` qui contient la référence de la source dont provient la zone de texte concernée. La portée de cette référence dépend du contexte d'occurrence de l'élément `bibl`. Ainsi, l'élément `bibl` contenu dans l'élément `p` donne l'origine du texte de l'ensemble du paragraphe tandis que ceux qui sont contenus dans des éléments `seg` renseignent sur la provenance du texte marqué par le segment. On constate très vite qu'une contradiction semble présente dans le code de l'exemple puisque le paragraphe est attribué à Vincent de Beauvais (VB 17, 29, 1) alors que le segment interne visible est lui attribué à Thomas de Cantimpré (TC 7, 9). Il n'y a là en réalité aucune contradiction. L'explication réside dans la nature du texte traité qui reprend majoritairement un texte de Vincent de Beauvais lui-même emprunté à d'autres textes, et à ceux de Thomas de Cantimpré en particulier. Les deux informations bibliographiques donnant les références des paragraphes et des segments correspondent en définitive à deux niveaux d'identification

des sources : le paragraphe provient de Vincent de Beauvais et les segments qui le constituent sont copiés sur Thomas de Cantimpré. Nous reviendrons précisément sur ce point plus bas [106].

7.3.3 Articulation des niveaux d'encodage

Comme nous l'avons vu, l'encodage scientifique est presque systématiquement lié à un type de texte ainsi qu'à une série d'objectifs scientifiques donnés. Il est donc tout à fait déraisonnable de penser pouvoir développer des solutions génériques de traitement formel de telles données, tant la diversité des cas est élevée.

En revanche, l'encodage éditorial constitue une base suffisamment simple pour permettre le développement d'outils génériques au moins concernant les besoins les plus couramment demandés et rencontrés, quel que soit le support de diffusion concerné.

Il est en effet tout à fait possible de produire des formes de diffusion tout à fait satisfaisantes à partir d'un encodage relativement simple. Autrement dit, il est possible d'identifier des éléments de structure logique de textes réguliers dont la forme sera, elle aussi, régulière.

Dès lors, l'articulation de ces deux niveaux d'encodage peut se faire de deux façons qui ne s'excluent pas l'une l'autre.

La première se focalise sur la production des flux et consiste à produire une version initiale du flux encodé au niveau éditorial en utilisant des solutions de stylage comme celles que nous avons présentées plus haut [107]. Il s'agit d'une première étape permettant d'obtenir une structuration à gros grain sur laquelle les chercheurs pourront travailler et ajouter des informations pour produire, *in fine*, une nouvelle version du flux manipulé qui sera encodée au niveau scientifique. Dans cette première solution d'articulation des niveaux d'encodage, le niveau d'encodage éditorial est donc un point de passage permettant de gagner du temps dans le processus de production d'un flux encodé scientifiquement.

La seconde articulation inverse les deux étapes et traite de l'exploitation éditoriale d'un flux encodé scientifiquement. Il s'agit alors pour l'éditeur matériel d'identifier, dans le système d'annotation scientifique, les éléments qui doivent faire l'objet d'un traitement formel.

106. Voir p. 219 et suivantes.
107. Voir p. 88.

Ainsi, nous proposons ici d'exploiter la proximité existant entre les catégories textuelles manipulées par les chercheurs et les formes éditoriales qu'il faut leur apporter. En effet, si les chercheurs manipulent une très grande quantité de types de textes, l'éditeur se doit de ramener ces catégories à des groupes de formes intelligibles pour le lecteur. Cette approche se focalise donc sur la production de formes éditorialisées et pas forcément sur l'exploitation de l'ensemble des phénomènes textuels annotés par les chercheurs. Des telles exploitations imposent le plus souvent des développements tout aussi spécifiques que les textes que l'on souhaite exploiter. Nous verrons cependant que les deux démarches peuvent s'articuler.

Nous proposons donc ici de produire l'encodage éditorial à partir de l'encodage scientifique, qui rappelons le, est le plus riche, au moyen le plus souvent d'une feuille de transformation XSL écrite spécifiquement dans le cadre d'un projet de recherche donné. Une fois cet encodage éditorial obtenu, il est alors possible d'exploiter l'ensemble des outils développés pour l'exploiter dans le cadre de la production de toutes les formes de diffusion.

Dans cette première approche, il s'agit donc de produire dans un premier temps une structure de base, avec un encodage éditorial qui va servir de socle à l'encodage scientifique. En d'autres termes, l'encodage éditorial est en quelque sorte un premier pas dans la mise en place d'un encodage scientifique.

Dans les deux cas, l'encodage éditorial comme l'encodage scientifique, il s'agit de pratiques de structuration en profondeur des sources concernées, même si ce niveau de profondeur est variable d'un encodage à l'autre. Mais il faut également considérer une autre dimension du travail d'encodage qui vise plus à repérer les sources plutôt qu'à les structurer au sens strict du terme. Il s'agit en définitive de l'encodage correspondant à l'inventaire. Si l'on considère que l'encodage éditorial et l'encodage scientifique relèvent de l'annotation verticale des données, il faut alors traiter ce travail d'inventaire comme un encodage horizontal.

L'encodage horizontal est capital car c'est lui qui permet d'organiser des ensembles de données identifiés sans entrer dans une grande finesse de description structurelle. Il sert en réalité de maillage initial de corpus en proposant une sorte de topographie des textes peuplant un corpus. Prenons l'image d'un repère orthonormé pour se représenter cette articulation. L'encodage horizontal permet de positionner les éléments du corpus, les textes, sur un repère à deux dimensions. Ajouter un balisage éditorial ou scientifique aux textes qui le nécessitent en fonction des objectifs de recherche, revient à ajouter une troisième dimension.

Deux projets du CRAHAM illustrent bien l'articulation entre balisage horizontal et balisage vertical. Il s'agit de Scripta, le site caennais de recherche informatique et de publication des textes anciens, dirigé par Pierre Bauduin et d'E-Cartæ dirigé par Gregory Combalbert.

Le premier, Scripta, se fixe comme objectif de fournir à la communauté un riche choix d'actes médiévaux normands, notamment du X^{e} au XIII^{e} siècle. Scripta propose donc le plus souvent des textes peu structurés accompagnés d'un ensemble de métadonnées permettant de réaliser des recherches sur l'ensemble du territoire normand pour la période concernée. C'est une précieuse source d'information pour les chercheurs qui rassemble une grande quantité de données sur plus de 6000 actes, mais qui ne propose pas d'encodage fin des textes (la tâche serait d'ailleurs immense étant donné le nombre d'actes présents dans la base).

Le second projet, E-Cartæ, se focalise à la fois sur une zone géographique précise et sur un type de document particulier puisqu'il traite des chartes des évêques d'Évreux. Il s'agit donc d'un sous-ensemble du corpus traité par Scripta. L'objectif du projet est d'étudier l'ensemble du corpus d'Évreux et d'en fournir une édition. L'ensemble des données est très finement annoté avec en particulier un balisage systématique du discours diplomatique et des variantes.

Les deux projets entretiennent d'étroites relations et sont complémentaires dans leurs objectifs. Toutes les chartes d'Évreux de la base Scripta pointent vers l'édition dans E-Cartæ. En réalité, Scripta se positionne comme un inventaire virtuel avec un grand nombre de documents indexés et permettant d'identifier des ensembles qui pourront faire l'objet d'études spécifiques. E-Cartæ constitue le parfait exemple de l'une de ces études.

Scripta et E-Cartæ proposent donc des textes encodés et indexés à différents niveaux en fonction de leurs objectifs propres. Cependant, certains textes entrent dans les champs des deux projets ce qui permet de lier les textes des deux projets les uns aux autres. Ces deux exemples permettent de voir émerger les bases d'un réseau de ressources textuelles interconnectées.

8

Réseau de textes et réseau d'acteurs

8.1 Notion de réseau de textes

Nous traitons donc des flux de texte initiaux portés par les témoins, ces textes sont manipulés par les différents acteurs, les témoins sont précisément décrits et indexés puis les textes sont réorganisés pour être présentés au lecteur en toute transparence, ce qui lui permet de choisir son mode d'accès et éventuellement d'examiner et de critiquer les choix réalisés par les éditeurs scientifiques lorsque l'accès à la structure logique est proposé.

L'ensemble des flux de textes constitutifs d'un même corpus est stocké dans des réservoirs (usuellement des bases de données XML natives comme eXist [108] ou BaseX [109]). Nous définissons comme un *réseau de textes* ces réservoirs contenant des flux encodés entretenant des relations entre eux. Ces relations peuvent être de différentes natures : liens directs ou références, citations, reprises et autres discours rapportés, thématique commune, etc. Il s'agit donc de rassembler des textes entretenant des relations particulières pour les traiter avec une méthode uniforme favorisant une exploitation globale.

Un réseau de textes articule donc plusieurs flux entre eux. Il s'agit de véritables ressources textuelles proposant un système complexe dont les composants, les flux et les fragments qui les composent sont interconnectés. Nous reviendrons sur cette notion par l'exemple lorsque nous étudierons le laboratoire *Ichtya* [110].

108. http://exist-db.org
109. http://basex.org
110. Voir p. 230

8.2 Le réseau de textes comme base de collaboration

8.2.1 Vers un système distribué et collaboratif

Il s'agit en fait de mettre en place à partir des flux structurés en réseau ce que Jean-Michel Salaün appelle

> un vaste système bureautique, un organiseur ou un traitement de texte intelligent où le document n'est créé qu'en bout de chaîne, comme le résultat des requêtes et calculs [Salaün, 2012]

quand il parle du web. Ici la perspective est à la fois plus large et plus restreinte : il ne s'agit pas du web dans son entier et du point de vue de la finesse de description on dépasse un système bureautique aussi sophistiqué soit-il, pour ajouter, en amont des requêtes et des calculs, un maximum d'intelligence et de connaissances directement dans les contenus sous la forme d'une caractérisation précise des types de textes rencontrés.

Ce réseau de textes richement annoté, qui prend par exemple la forme d'un ensemble de fichiers XML, devient le cœur d'un serveur de fragments capable de répondre à des requêtes plus générales que celles portées implicitement par sa structure. On pense par exemple à des recherches plein texte dont le résultat pourrait être une série de fragments dont la taille serait choisie par l'utilisateur au moment de la fabrication de la requête. Une telle requête pourrait articuler un mot accompagné d'un contexte d'occurrence ainsi qu'un niveau d'organisation des réponses à un grain particulier comme le paragraphe, la section, le chapitre, ou tout autre niveau de structuration existant dans les flux textuels interrogés.

La figure 8.1 propose une représentation d'ensemble de la circulation et de l'exploitation des informations depuis la saisie des fichiers initiaux jusqu'à la mise à disposition des fragments. Le travail peut commencer avec un éditeur XML et des documents XML stockés de manière traditionnelle sur des ordinateurs personnels. Les fichiers produits peuvent être ensuite déposés sur un serveur. À partir de ce moment, les textes contenus dans les documents peuvent faire l'objet de requêtes indépendemment des limites des documents originaux et devenir des fragments inclus dans des flux. En définitive, dans le contexte de notre étude, il faut concevoir les documents comme des flux bornés à cause de contingences matérielles. Techniquement il s'agit de s'appuyer sur l'arborescence des fichiers XML manipulés. Une fois un texte déposé dans une base XML, la notion de document n'a plus la même

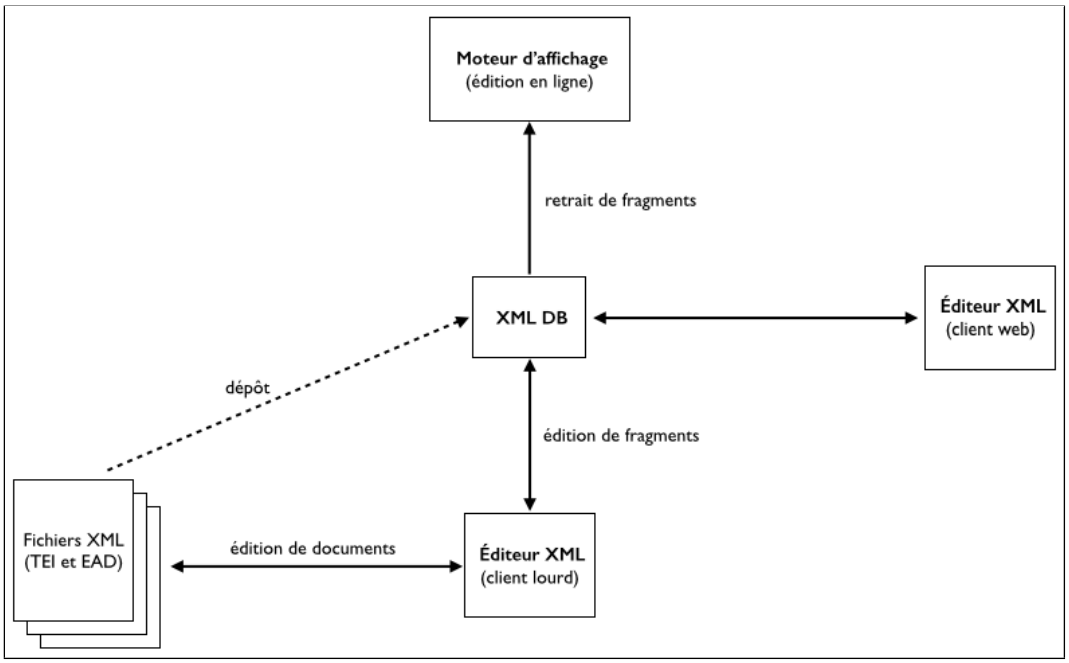

FIGURE 8.1 – Représentation du fonctionnement du serveur de fragments.

importance : elle n'est plus qu'un niveau de requête comme un autre et en particulier comme n'importe quel élément de l'arborescence XML mise en place.

Bien entendu des droits sont associés aux fragments en lecture ou lecture/écriture selon les besoins et les applications utilisées. En passant par un moteur d'affichage, il sera uniquement possible de lire les textes, éventuellement en configurant l'affichage ou en sélectionnant des types de texte souhaités selon la complexité de l'application utilisée. En passant par un éditeur ou un système d'annotation en ligne, il sera possible de modifier les flux : soit en modifiant profondément les textes et leurs structures, soit en ajoutant des notes de commentaire par exemple.

Le modèle de flux et de fragments ainsi que le projet dans le cadre duquel il est mis en place agissent respectivement comme un objectif partagé et comme un guide pour l'ensemble des interventions. Une fois le flux mis en place au sein d'un serveur de fragments, chaque acteur peut intervenir sur tel ou tel fragment qui le compose. L'articulation des métiers et des tâches est mise en cohérence par le flux dans le cadre du projet.

La convergence numérique touche tous les acteurs qui partagent aujourd'hui un même cadre technique : celui des langages à balises et d'XML en particulier. Ils disposent, nous l'avons vu, de véritables langages, proposant des vocabulaires et des règles d'organisation associées, s'appuyant sur ces technologies de structuration.

Il s'agit donc d'exploiter cette situation pour construire, dans le cadre de projets communs, des organisations du travail autour des flux d'informations à manipuler.

Du point de vue des vocabulaires, la circulation de l'information est facilitée par le partage de la base technique commune. Ainsi, un projet d'édition portant sur des manuscrits intégrés dans un fonds inventorié en EAD pourrait se dérouler de la manière suivante :

$$\text{EAD} \xrightarrow{\text{XSL}} \text{TEI sc} \xrightarrow{\text{XSL}} \text{TEI ed} \xrightarrow{\text{XSL}} \text{formats de diffusion}$$

Nous retrouvons ici, à travers les vocabulaires, tous les métiers évoqués plus haut, de l'archiviste à l'éditeur matériel en passant par le chercheur. L'inventaire en EAD est converti, au moyen de transformations en TEI, avec un niveau de structuration basique qui est enrichi par les chercheurs. Une fois le travail d'analyse scientifique terminé, les flux subissent une réduction d'encodage pour parvenir à un niveau de description conforme aux impératifs éditoriaux, afin de permettre la production de formes de diffusion adaptées aux objectifs du projet.

Cette continuité technique dans l'articulation des vocabulaires apporte la stabilité technologique nécessaire aux cyberinfratructures souhaitées par Marin Dacos et Pierre Mounier [Dacos et Mounier, 2014]. En effet, dans le cadre de projets de coopérations, les différents acteurs vont pouvoir collaborer dès les premières étapes. Cette collaboration permet d'intégrer au cœur des projets une réflexion sur les outils et les interactions exploitant pleinement les compétences de l'ensemble des acteurs. Il s'agit donc de participer au développement des humanités numériques au-delà de la dimension statistique.

Ainsi, l'archiviste, ou le conservateur, est un acteur essentiel dans tout projet de valorisation, d'étude ou de promotion de textes anciens. Il est pourtant soumis à des formes d'injonctions contradictoires qui peuvent être assez difficiles à articuler. En effet, il doit à la fois assurer la conservation, c'est-à-dire la protection physique des documents dont il a la garde, et proposer des solutions de valorisation par exemple sous la forme d'expositions.

L'utilisation des techniques numériques, de l'acquisition à la diffusion en passant par l'étude, peut apporter des solutions extrêmement satisfaisantes pour répondre à ces exigences en apparence opposées. En effet, la numérisation des sources anciennes est une solution pour assurer la conservation des documents tout en facilitant leur consultation en proposant un accès illimité à des reproductions numériques. Bien entendu, pour certaines disciplines comme la codicologie, qui imposent un accès

direct à la matérialité des sources, la numérisation n'apporte rien. En revanche pour l'ensemble des opérations de mise à disposition des textes portés par les supports anciens, la numérisation, en mode image et en mode texte, offre des solutions tout à fait satisfaisantes pour l'archiviste et le conservateur.

C'est pour ces raisons qu'il existe aujourd'hui une énorme quantité de projets de valorisation patrimoniale qui commencent par une numérisation des fonds. Ces projets sont très régulièrement initiés et portés par des institutions de conservation et des archives.

Pour autant ces projets ne sont pas toujours menés par les archives seules. En effet, il existe souvent des interactions avec d'autres acteurs. Par exemple, les programmes de numérisation d'une institution de conservation peuvent être en partie orientés en fonction de problématiques de recherche. Dans ce cas les projets sont le cadre de collaboration entre chercheurs et archivistes pour le choix des outils et des méthodes de travail. Il faut alors trouver des solutions pour harmoniser les pratiques, les points de vue sur les objets et les activités dans un environnement technique et théorique cohérent. Le modèle de flux et de fragments que nous proposons se fixe justement cet objectif.

L'activité d'édition de sources doit mobiliser plusieurs acteurs pour être réalisée dans de bonnes conditions. L'archiviste ou le conservateur identifie et signale les documents et les textes, le chercheur travaille sur ces textes, les replace dans leurs contextes, produit une transcription et l'éditeur matériel se charge des opérations de contrôle, de normalisation et de la production des formes de diffusion.

Mais, l'éditeur scientifique est incontestablement un acteur central et la plupart du temps initiateur des projets d'édition de sources. Il fixe l'ensemble des objectifs scientifiques. Spécialiste des textes étudiés et manipulés, il est le référent tout au long des étapes de production des outils d'étude et des formes de diffusion. Il est bien entendu le producteur du discours scientifique et garant de sa qualité.

Le chercheur détermine donc la vision d'ensemble des projets éditoriaux. Sa perception et sa connaissance des objets manipulés sont centrales car c'est lui qui devra trancher lors des choix stratégiques. Il fixera ainsi les règles de transcription, le niveau de structuration, etc. en fonction de ses objectifs scientifiques.

Bien entendu, la qualité des résultats dépendra aussi de sa capacité à entendre et à comprendre les contraintes et impératifs de chacun des autres acteurs. Il devra par exemple être en mesure d'intégrer les obligations des archivistes du point de vue de la valorisation, les contraintes de collection de l'éditeur matériel ou les impératifs et limitations techniques du développeur informatique.

L'éditeur matériel est en général identifié, à tort ou à raison, comme la cause de tous les retards car il intervient souvent à la fin des opérations éditoriales. Cette intervention tardive est pourtant souvent problématique car, en tant que garant de la cohérence entre le fond et les formes, l'éditeur matériel est à même d'apporter des solutions techniques qui permettraient d'organiser les opérations de structuration de données en évitant au maximum les redondances et les répétitions de manipulations. En effet, maîtrisant les spécificités des formes de diffusion, il est l'acteur le mieux placé pour organiser au mieux les manipulations (structuration, normalisation et correction) de données afin de simplifier la production de l'ensemble de ces formes.

Le rôle de l'éditeur matériel ne se limite pas seulement à assurer une simple mise en forme des textes diffusés. En effet, l'éditeur matériel assure également un contrôle de la structure générale des textes, il réalise les choix typographiques les plus appropriés en fonction de la nature des textes et des collections dans lesquels ils s'inscrivent. De plus, l'éditeur matériel contrôle et complète les listes de références bibliographiques, il assure un contrôle syntaxique et orthographique des textes.

Enfin, il est le garant de la pérennité des formes diffusées et doit donc mettre en place et maintenir des dispositifs techniques stables capables de donner au lecteur un accès aux textes sous des formes identiques et citables à long terme. Autrement dit, les formes mises en place par l'éditeur matériel ne doivent pas changer dans le temps. Si pour le livre papier, ça ne pose évidemment aucun problème, pour les formes de diffusion numérique et pour les versions web en particulier, c'est un peu plus compliqué. Il faut bien entendu que les adresses de consultation demeurent inchangées, pour que les références à ces pages restent valables, mais aussi que les dispositifs d'affichage, c'est-à-dire les programmes informatiques qui produisent les pages à partir des textes, soient constamment entretenus de manière à produire toujours les mêmes résultats indépendamment des évolutions des langages informatiques et des outils de navigation.

Il est donc tout à fait capital de ne pas négliger les opérations de maintenance exigées par les éditions électroniques, et plus les outils embarqués dans ces éditions sont sophistiqués, plus le coût de maintenance est lourd.

8.2.2 Humanités numériques et *rich data*

Comme nous l'avons déjà évoqué, l'une des méthodes courantes dans le domaine de la constitution de ressources en sciences humaines et sociales consiste à mettre en place des entrepôts de données en masse, avec peu ou pas de structuration des infor-

mations, associés à l'utilisation des techniques de fouilles de données pour extraire de l'information. La tendance est aussi à la convocation de techniques statistiques issues du traitement automatique de la langue.

Si cette approche, dans le droit fil du *big data*, est parfaitement justifiable et obtient de bons résultats sur des volumes de données importants et dans des langues maîtrisées, il ne s'agit pas de la seule approche possible, d'une part, et la question des données dans d'autres langues, parfois plus difficiles à manipuler, reste entière.

Il ne s'agit donc pas de remettre en cause une approche qui a fait ses preuves et qui continue à les faire, mais seulement de proposer une autre voie. Cette dernière consiste à s'appuyer sur le travail des spécialistes des textes anciens et des historiens, qui utilisent aujourd'hui massivement des techniques numériques pour travailler leurs données. Il s'agit donc de proposer des évolutions des méthodes utilisées par ces spécialistes pour favoriser la constitution de corpus de données interopérables et disposant d'un haut niveau de structuration et d'annotation.

Dans ce contexte, nous proposons de concevoir les humanités numériques comme une ouverture des sciences humaines et sociales aux techniques d'expérimentation au sens des disciplines réputées plus "dures" scientifiquement [111] et pas seulement comme l'utilisation d'outils informatiques pour la fouille de données en très grand nombre. Une fois encore, il ne s'agit pas de dire que ce n'est pas pertinent ou efficace, mais tout simplement d'avancer que c'est une approche trop limitative et qu'il faut enrichir ce point de vue sur les humanités numériques en partant des humanités justement.

Il s'agit d'apporter aux chercheurs en SHS les solutions pour expérimenter et mettre en place des systèmes d'annotations contrôlés par eux (les spécialistes tout de même) pour tendre vers la mise en place de ce que l'on pourrait qualifier de *fine* ou de *rich data*. En définitive, il s'agit de compléter l'approche orientée *big data* totalement appuyée sur la fouille de données et le traitement automatique de la langue, qui est aujourd'hui majoritaire, par une approche centrée avant tout sur l'expertise des spécialistes des humanités concernées. Le travail se fait en général pour l'édition des textes, mais les typologies identifiées au cours de ce travail dédié à l'édition peuvent parfaitement être exploitées dans un autre contexte [112].

111. La possibilité de répéter les expériences est même, d'une certaine manière, plus avancée, dans la mesure où la matière première de l'expérimentation, le flux XML, est diffusable en même temps que le protocole de test.

112. Voir par exemple l'exploitation du balisage de l'*Hortus Sanitatis* dans le cadre du projet ANR Sourcencyme, p. 232.

Par ailleurs, une autre caractéristique des solutions appuyées sur les pratiques des chercheurs que nous proposons concerne la construction des objets de recherche. La structuration et l'annotation mobilisant des langages *descriptifs* pour les sources textuelles primaires et leurs relations entre elles, permettent l'interrogation des corpus ou des réseaux de textes ainsi constitués avec un minimum de *présupposés*.

Il ne s'agit plus de construire des dispositifs avant le travail d'organisation des sources comme lors de la définition d'un schéma de base de données relationnelle avec un SGBD traditionnel qui impose une très bonne connaissance des sources mais aussi d'avoir des idées précises sur les exploitations qui pourront, ou devront, en être faites. Si la très bonne connaissance des sources reste centrale dans notre modèle, les possibilités d'exploitation ne doivent pas forcément être connues *a priori* : il n'est en effet pas nécessaire de définir à l'avance la manière dont le corpus devra être interrogé. Il faut simplement prévoir un niveau d'annotation permettant la caractérisation des traits étudiés dans le cadre d'un projet, sans fermer la porte aux autres possibilités. Si d'autres aspects des sources apparaissent et soulèvent un intérêt scientifique pendant le travail d'annotation, les descripteurs appropriés peuvent toujours être ajoutés.

Dans le pire des cas, c'est-à-dire lorsqu'un aspect n'est pas traité dans un cadre donné, il suffit que les responsables du projet mettent à disposition les flux encodés pour permettre à la communauté d'ajouter le balisage nécessaire et de mettre en place les solutions d'exploitation appropriées. Non seulement le modèle que nous proposons permet l'erreur et l'expérimentation, mais il permet également de construire un matériau potentiellement plus "neutre" et extensible. En définitive, les méthodes permettent d'objectiver les processus de recherche et fournissent des solutions pour permettre la critique de la construction des objets de recherche. Ainsi la communauté dans son ensemble peut examiner la manière dont les objets ont été construits dans le cadre d'un projet mené par certains de ses membres, et la critiquer pour améliorer les résultats obtenus ainsi que les conclusions proposées.

Pour élargir la portée potentielle de notre modèle, permettons-nous ici un écart par des sources d'une autre nature que celles que nous avons traitées jusqu'à présent : les transcriptions d'entretiens dans le cadre d'enquête qualitative. Dans le domaine des sciences sociales les enquêtes peuvent prendre la forme de longs entretiens qui sont le plus souvent transcrits pour être analysés à l'issue de la phase de recueil de données sur le terrain. Ces transcriptions d'entretien sont facilement structurables à un niveau de balisage éditorial pour obtenir des flux d'information interrogeables de manière groupée et pas seulement entretien par entretien. De plus une annotation, même à gros grain, permet de circonscrire les recherches à des zones du discours

pour limiter le bruit dans les résultats. Là encore, si un trait particulier intéresse les chercheurs, il suffit d'ajouter le niveau d'annotation correspondant dans chaque transcription d'entretien pour permettre l'exploitation et l'interrogation de cet aspect des discours à l'échelle du corpus complet.

Enfin, toutes ces étapes de structuration de matériau primaire de la recherche peuvent être exploitées jusqu'à la synthèse scientifique. Ainsi, tout le travail réalisé pendant la phase d'élaboration du projet de recherche peut être intégré dans les opérations de publication et de diffusion des résultats scientifiques. La convocation des propos recueillis pendant les entretiens dans une publication scientifique des extraits d'entretien, sous la forme de discours rapportés, revient, dans notre modèle, à référencer des fragments dans le flux d'un article scientifique. D'une certaine manière ces transcriptions peuvent donc être traitées comme des sources primaires.

Les ressources constituées, sous la forme de flux et de fragments, dans le cadre de projets d'édition de sources anciennes peuvent donc être exploitées dans des perspectives de recherche plus vastes. L'ensemble des textes balisés et annotés par les chercheurs dans l'optique de les diffuser accompagnés de riches commentaires scientifiques peuvent devenir de véritables bases de connaissances interrogeables. Toutes les catégories de texte deviennent alors des critères d'interrogation potentiels. Nous proposons dans le chapitre suivant d'étudier des perspectives d'exploitation de ces flux richement encodés par les spécialistes.

9

Perspectives

Nous proposons, pour les projets d'éditions de sources anciennes, un modèle de flux et de fragments, mais les ressources constituées dans ce cadre peuvent servir de base à la constitution d'outils de recherche en s'appuyant sur les technologies du web sémantique. Toujours dans une logique de *fine data*, nous étudions ici des solutions pour tirer bénéfice du travail d'annotation mis en place par les chercheurs pendant leur travail d'études des sources.

9.1 De l'arbre XML au graphe RDF

Le passage d'un arbre XML à un graphe RDF semble être une évolution technologique naturelle dans la mesure où l'arbre XML est un cas particulier de graphe RDF. En effet tandis qu'un arbre XML permet, dans sa plus simple expression, de décrire des relations d'inclusion entre des nœuds, un graphe RDF ajoute la possibilité de qualifier n'importe quel type de relation entre deux sommets sous la forme d'un triplet *(sujet, prédicat, objet)*. Ainsi, l'expression XML suivante :

$$< A >< B/ >< /A >$$

pourrait être exprimée dans un graphe RDF de type :

$$A \xrightarrow{\text{contient}} B$$

Les deux expressions étant en réalité, parfaitement équivalentes. Les graphes RDF permettent d'ajouter une couche de qualification des relations existantes entre les types de données manipulés. C'est ce gain d'expressivité qui motive le passage d'un arbre XML à un graphe RDF.

Il s'agit donc pour nous d'exploiter la richesse de l'annotation XML TEI mise en place par les spécialistes pour produire un graphe RDF exprimant explicitement l'ensemble des relations existant entre les types de textes discriminés. En fonction de la nature des éléments TEI mobilisés par les chercheurs, il faut définir les relations correspondantes à établir dans le graphe RDF.

Ainsi, traiter une structure XML TEI comme celle-ci,

$$< bibl >< title/ >< author/ >< publisher/ >< /bibl >$$

qui définit un certain nombre d'éléments et un type de relation unique entre l'élément contenant et les éléments contenus, en exploitant les noms des nœuds permet de mettre en place un graphe explicitant la relation spécifique entretenue par `bibl` et `publisher` par exemple. La relation d'inclusion entre ces éléments XML, puisqu'elle intervient entre ces deux éléments précis, permet d'obtenir un graphe de la forme :

$$publisher \xrightarrow{\text{est éditeur de}} bibl$$

et, en poussant la logique un peu plus loin, en intégrant l'élément `title` au processus :

$$publisher \xrightarrow{\text{est éditeur de}} bibl \xrightarrow{\text{a pour titre}} title$$

Enfin, le même raisonnement peut être tenu en partant, non plus du `publisher`, mais du `title` par exemple, pour obtenir la relation inverse :

$$title \xrightarrow{\text{est le titre de}} bibl \xrightarrow{\text{est édité par}} publisher$$

Les graphes RDF permettent la mise en place de réseaux de relations beaucoup plus complexes et plus fins en définissant des ensembles de triplets *(sujet, prédicat, objet)*.

D'un point de vue applicatif, nous observons donc un réel gain d'expressivité, et donc de finesse d'interrogation possible notamment en utilisant les techniques de requête de l'univers RDF comme SPARQL. Mais les graphes RDF apportent également des modes de visualisation et d'accès aux données différents. Les représentations graphiques apportent aux chercheurs un point de vue "décalé" sur les données qu'ils manipulent qui peut leur permettre d'opérer une mise à distance, ou une objectivation, elle-même propice à la formulation de nouvelles hypothèses de travail ou d'interprétation.

9.2 Vers la sémantisation des textes

Les structures XML mises en place dans le cadre du modèle de flux et de fragments portent en germe tous les objets d'études du chercheur, mais aussi toutes les relations

possibles entre ces objets. L'objectif est d'exprimer explicitement ces relations sous la forme de triplets.

Il existe de grands modèles applicables aux sources anciennes que nous traitons ici. Nous avons déjà évoqué FRBRoo, qui émane du monde des bibliothèques, ou CIDOC-CRM, provenant du monde de la conservation patrimoniale. Il serait possible d'établir des relations entre la TEI et ces modèles. Cependant, nous favorisons ici une approche centrée sur les données annotées par les chercheurs. Ainsi, plutôt que de partir des grands modèles du domaine (comme FRBRoo ou CIDOC-CRM) nous proposons de constituer des micro-ontologies, c'est-à-dire des ensembles réduits de concepts et de termes associés aux sources traitées. Ces micro-ontologies sont ensuite étendues et peuplées par les informations structurées par les chercheurs eux-mêmes dans les flux encodés en XML TEI. En définitive, plutôt que de penser ou d'intégrer *a priori* une ontologie ultracomplexe, il s'agit de s'appuyer sur l'expertise des spécialistes du domaine pour produire des ontologies avec un minimum de présupposés.

```
1     <owl:Class rdf:about="&pddn;Personne">
2     </owl:Class>
3     <owl:Class rdf:about="&pddn;Fonction">
4     </owl:Class>
5     <owl:Class rdf:about="&pddn;Oeuvre">
6     </owl:Class>
7     <owl:Class rdf:about="&pddn;Texte">
8     </owl:Class>
9     <owl:Class rdf:about="&pddn;Item">
10    </owl:Class>
11    <owl:Class rdf:about="&pddn;Lieu">
12    </owl:Class>
13    <owl:Class rdf:about="&pddn;Pays">
14    </owl:Class>
15    <owl:Class rdf:about="&pddn;Region">
16    </owl:Class>
17  [...]
18    <owl:ObjectProperty rdf:about="&pddn;occupe">
19      <rdfs:domain rdf:resource="&pddn;Personne"/>
20      <rdfs:range rdf:resource="&pddn;Fonction"/>
21      <owl:inverseOf rdf:resource="&pddn;occupeePar"/>
22    </owl:ObjectProperty>
23    <owl:ObjectProperty rdf:about="&pddn;occupeePar">
24      <rdfs:domain rdf:resource="&pddn;Fonction"/>
25      <rdfs:range rdf:resource="&pddn;Personne"/>
26      <owl:inverseOf rdf:resource="&pddn;occupe"/>
27    </owl:ObjectProperty>
```

FIGURE 9.1 – Extrait de l'ontologie de base utilisée dans les inventaires anciens.

La figure 9.1 donne un exemple d'une micro-ontologie utilisée pour produire un graphe à partir de la source XML TEI d'un inventaire ancien. Ce type de source est constitué d'une liste de livres. Les présupposés consistent à poser que ces sources mettent en jeu les titres de ces livres, des personnes, auteurs de ces livres, des lieux pour leurs publications, les régions et les pays de ces lieux,etc. Toutes ces catégories, de la ligne 1 à 16, et les relations qu'elles entretiennent, de la ligne 18 à 27, constituent les *classes* et les *propriétés* de base de notre micro-ontologie.

Une fois cette micro-ontologie définie, elle est étendue et peuplée sur la base de l'exploitation de l'ensemble des informations données par les spécialistes (archivistes/conservateurs ou éditeurs scientifiques) dans le flux XML. L'exploitation utilise aussi bien l'arborescence et la sémantique des éléments mobilisés que la richesse des informations donnée dans le système d'attribution des éléments XML.

L'objectif est donc bien de mobiliser les vocabulaires de mise en relation renseignés par les chercheurs pendant le travail de structuration et d'annotation. Ainsi nous obtenons, par étapes successives, une ontologie plus complexe, construite à partir des cas rencontrés dans les textes des sources. La figure 9.2 donne une visualisation sous l'éditeur d'ontologie *Protégé*[113] du résultat obtenu au terme des opérations.

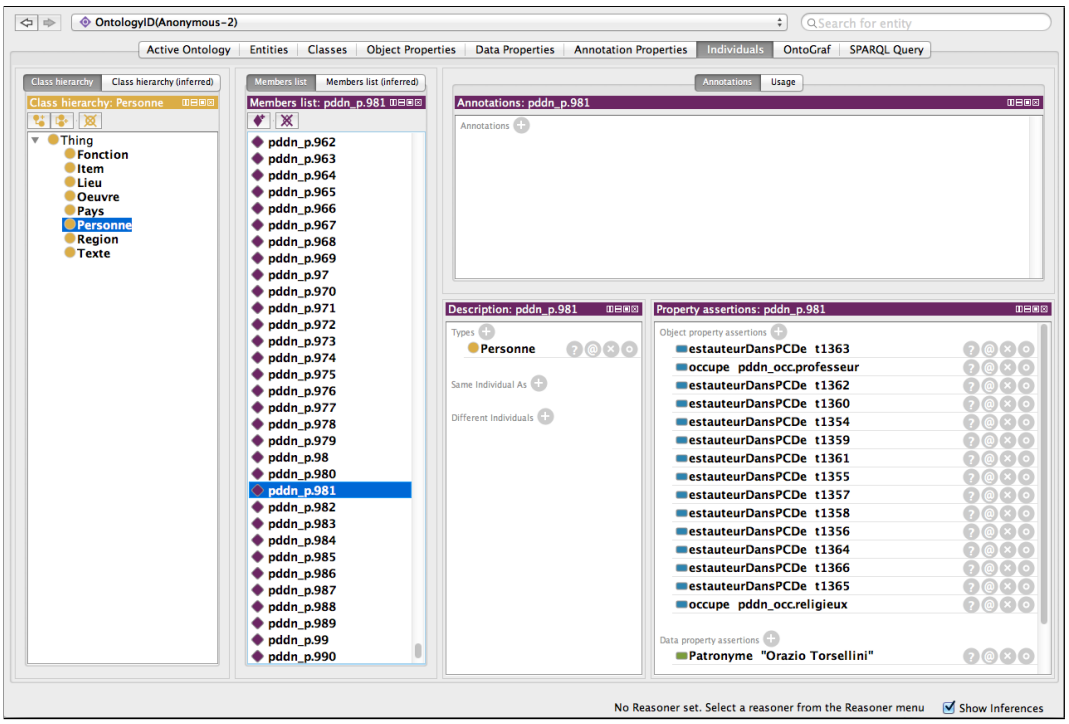

FIGURE 9.2 – Visualisation de l'ontologie peuplée sous *Protégé*.

113. http://protege.stanford.edu

9.3 Des flux de textes aux bases de connaissances

Nous proposons ici d'examiner en détails la procédure d'exploitation d'une arborescence XML TEI pour produire un graphe RDF à partir de l'inventaire de la bibliothèque du Mont Saint-Michel par Pinot-Cocherie tel qu'il figure dans un manuscrit conservé à la bibliothèque municipale d'Avranches (manuscrit Avranches BM 246). Cette source propose une liste de livres donnant l'état de la bibliothèque au moment des confiscations révolutionnaires.

Il s'agit de parvenir à la création d'un graphe RDF riche de l'ensemble des informations disponibles dans de multiples flux de données ordonnées, avec la transcription structurée et annotée de l'inventaire de Pinot-Cocherie, et non-ordonnées, avec les index de personnes et de lieux. Nous ne revenons pas sur la base des opérations qui consiste à étendre et à peupler une micro-ontologie élémentaire en exploitant les données et les annotations contenues dans le corpus produit par les spécialistes.

```
1  <item n="3" rend="p" xml:id="Avranches_BM_246.26.3">
2    <label n="théologie">Theo<choice><am>l.</am><ex>logie</ex></
       choice></label>. <idno>3</idno> : <title ref="wpddn/
       indexOeuvres.xml#pddn_w.40" type="oeuvre" xml:lang="la"><
       choice><orig>a</orig><reg>A</reg></choice>cta sanctorum or
       <choice><am>d.</am><ex>dinis</ex></choice><choice><am>s.</
       am><ex>sancti</ex></choice> Benedicti in sæculorum classes
        distributa</title> au<choice><am>t.</am><ex>thoribus</ex>
       </choice>
3    <name ref="wpddn/indexPersonnes.xml#pddn_p.74" role="auteur"
       type="personne">d'<choice><orig>a</orig><reg>A</reg></
       choice>chery</name> et
4    <name ref="wpddn/indexPersonnes.xml#pddn_p.263" role="auteur"
        type="personne"><choice><am>j.</am><ex>Johanne</ex></
       choice> <choice><orig>m</orig><reg>M</reg></choice>abillon
       </name>.
5    <name ref="wpddn/indexLieux.xml#pddn_l.16" role="publication"
        type="lieu"><choice><orig>p</orig><reg>P</reg></choice>
       aris</name>.
6    <date type="publication">1668</date>.
7    <measure quantity="1" unit="volume">1 vo<choice><am>l.</am><
       ex>lume</ex></choice></measure>
8    <choice><orig>  </orig><reg>.  </reg></choice>
9    <measure n="in-folio" type="format">in f<choice><orig>  </
       orig><reg>olio</reg></choice></measure>.<note resp="MB"
       type="identification"><ref target="http://localhost:8080/
       msm/ead.html?id=FR_UCBN_MSM_impr_av&amp;c=
       FR_UCBN_MSM_impr_av_B303">Avranches BM B 303</ref> ; ex-
       libris montois.</note>
10 </item>
```

Figure 9.3 – Extrait de l'encodage XML TEI de l'inventaire de Pinot-Cocherie.

La figure 9.3 donne un extrait du code XML TEI mis en place au cours du travail de structuration et d'annotation de la source. Les noms de personnes y sont clairement structurés, lignes 3 et 4, et une référence à une ressource prosopographique externe est présente dans l'attribut `ref`. Sans entrer dans les détails, notons simplement que cette ressource externe factorise un certain nombre de précisions sur la personne concernée : histoire, postes occupés au cours de sa vie, lieux d'exercice, filiation, etc. Bien entendu, elle prend, elle aussi, la forme d'un flux d'informations encodé en XML TEI auquel pourront se référer toutes les sources concernées. Le même type de ressource est également mis en place pour les lieux et la structuration est similaire à celle utilisée pour les personnes, comme le montre la ligne 5.

La transcription de l'inventaire est utilisée comme base de traitement et les références aux ressources externes sont résolues au moment où elles sont rencontrées. Chaque `item` de l'inventaire (ligne 1, figure 9.3) correspond à un individu de type `Item` (ligne 9, figure 9.1) défini dans la micro-ontologie et doit donc être créé dans le graphe en sortie. Les éléments `name` caractérisés par `type=personne` (lignes 3 et 4, figure 9.3) correspondent à des individus de la classe `Personne` (ligne 1, figure 9.1).

```
1     <xsl:for-each-group select="//descendant::tei:name[@type='
          personne'][@role]" group-by="@role">
2       <xsl:variable name="currentRole">
3         <xsl:text>est</xsl:text><xsl:value-of select="@role"/><
              xsl:text>DansPCDe</xsl:text>
4       </xsl:variable>
5       <xsl:variable name="currentRoleInv">
6         <xsl:text>aPour</xsl:text><xsl:value-of select="@role"/
              ><xsl:text>DansPC</xsl:text>
7       </xsl:variable>
8       <owl:ObjectProperty rdf:about="&pddn;{$currentRole}">
9         <rdfs:domain rdf:resource="&pddn;Personne"/>
10        <rdfs:range rdf:resource="&pddn;Texte"/>
11        <owl:inverseOf rdf:resource="&pddn;{$currentRoleInv}"/>
12      </owl:ObjectProperty>
13      <owl:ObjectProperty rdf:about="&pddn;{$currentRoleInv}">
14        <rdfs:domain rdf:resource="&pddn;Texte"/>
15        <rdfs:range rdf:resource="&pddn;Personne"/>
16        <owl:inverseOf rdf:resource="&pddn;{$currentRole}"/>
17      </owl:ObjectProperty>
18    </xsl:for-each-group>
```

FIGURE 9.4 – Création des rôles dans une ontologie à partir des informations encodées en XML TEI.

Si la création des individus ne nécessite pas d'étendre la micro-ontologie, il n'en va pas de même pour les propriétés. L'objectif est justement de s'appuyer sur les cas réellement rencontrés dans la source pour étendre l'ontologie.

La figure 9.4 présente le traitement XSL pour la création des propriétés correspondant aux rôles attribués aux personnes par l'auteur de l'inventaire. Pour chaque valeur de `role` rencontré pour les éléments `name type='personne'` (ligne 1) deux propriétés sont créées. La première définit la relation entre une `Personne` et un `Texte` (ligne 8) la seconde entre un `Texte` et une `Personne` (ligne 13) c'est-à-dire la relation inverse. Notons qu'il s'agit bien de la propriété donnée par l'auteur de la source comme le montrent les lignes 3 et 6 avec les nœuds `xml:text` ajoutant les valeurs `DansPCDe` et `DansPC`[114] à la valeur de l'attribut `role` trouvé dans la source XML TEI. Le même type de traitement est assuré pour l'ensemble des informations données par la transcription XML TEI.

Précisons que l'identification de l'item dans l'inventaire contemporain conserve la trace de la responsabilité scientifique avec l'élément `note` et l'attribut `resp` (ligne 9, figure 9.3) qui fait lui aussi l'objet d'un traitement spécifique au moment de l'extension de l'ontologie et de son peuplement.

Les figures 9.5 et 9.6 présentent des résultats des opérations d'exploitation des flux encodés en XML TEI par les chercheurs.

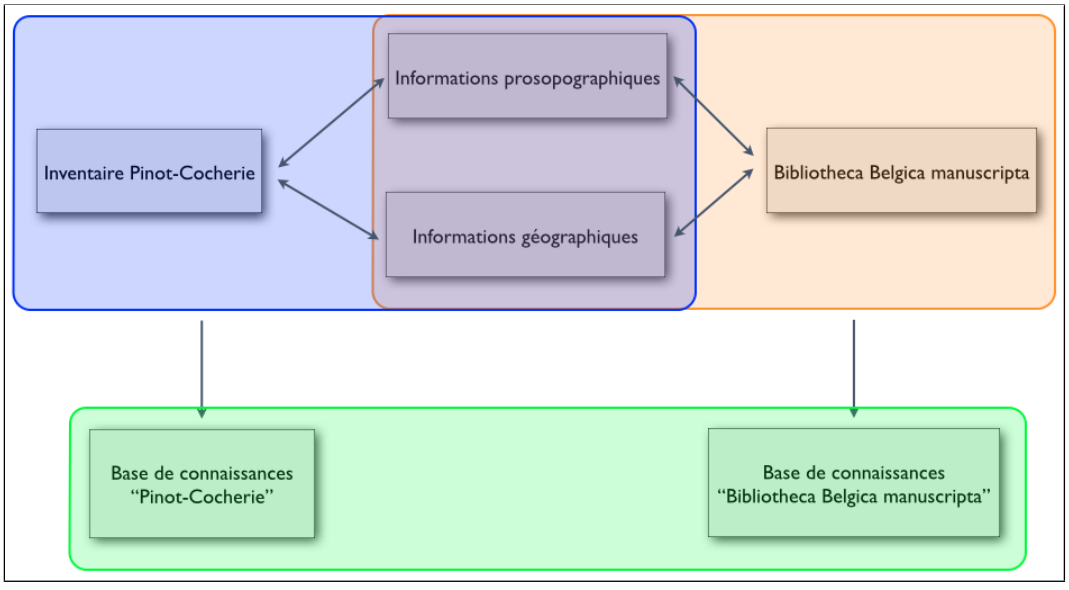

FIGURE 9.5 – Constitution de bases de connaissances à partir de flux de textes.

114. Dans les deux cas, "PC" signifie bien entendu PINOT-COCHERIE.

La figure 9.5 donne une vue générale du périmètre de travail et de la logique à l'œuvre. Chaque source, dont nous donnons ici deux exemples avec l'inventaire PINOT-COCHERIE et l'inventaire de la *Bibliotheca Belgica Manuscripta*, articulée avec des ressources externes partagées, l'une prosopographique et l'autre géographique, constitue une base de connaissances interrogeable en tant que telle. De plus dans la mesure où les sources partagent un même modèle ontologique, construit à partir de solutions de structuration identiques et les mêmes ressources externes, elles peuvent aussi être interrogées simultanément comme une base de connaissances unique.

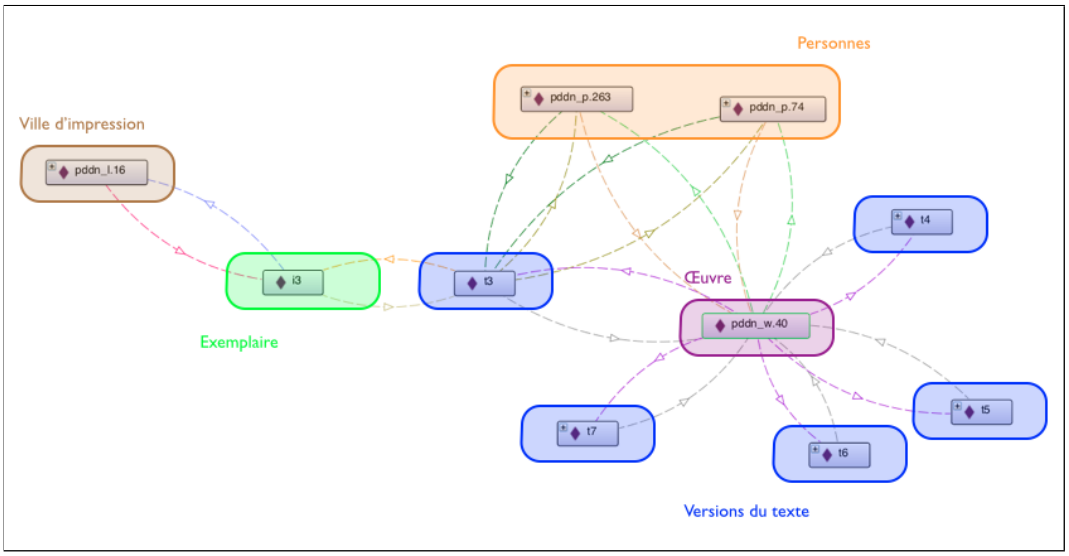

FIGURE 9.6 – Les œuvres communes de Luc D'ACHERY et de Jean MABILLON dans l'inventaire de PINOT-COCHERIE du Mont Saint-Michel.

La figure 9.6 propose une visualisation d'un graphe résultat de requête sur la base de l'inventaire PINOT-COCHERIE. Il s'agit de rassembler dans un graphe RDF unique toutes les informations, et leurs relations, disponibles dans l'inventaire et les ressources externes, sur les œuvres communes de Luc D'ACHERY et Jean MABILLON.

Si ces solutions ajoutent une expressivité considérable et précieuse aux textes, elles ne constituent cependant pas une solution de substitution de notre modèle de flux et de fragments. En effet, comme nous l'avons vu, il est tout à fait possible de produire des graphes en profitant de la simplicité de manipulation du langage XML pour le balisage et l'annotation des textes et en exploitant les catégories ajoutées par les spécialistes. Il s'agit bien de tirer le meilleur parti possible de ces deux technologies en articulant la commodité d'usage du XML et l'expressivité des graphes.

Quatrième partie

Expérimentation et validation

Introduction

Dans cette partie, nous présentons des expérimentations qui nous ont permis à la fois d'établir et de valider le modèle de flux et de fragments que nous proposons en appliquant les méthodes de construction itératives présentées. Les deux projets s'inscrivent tous deux dans une logique claire de diffusion des résultats : il s'agit de projets d'édition multimodales.

Parmi l'ensemble des projets de ce type ayant participé à l'établissement et à la validation de notre modèle, nous ne présentons que le plus ancien avec les *Sources du Mont Saint-Michel* [Bouet et Desbordes, 2009] et [Bougy, 2009] et le plus récent avec l'*Hortus Sanitatis* [Jacquemard *et al.*, 2013]. Nous avons choisi ces projets en raison de leurs apports spécifiques et de leur position chronologique soit au tout début des opérations de ce type aux Presses universitaires de Caen, soit très récemment et profitant donc des dernières avancées. C'est en particulier le cas sur le plan de l'outillage et des méthodes nécessaires pour permettre aux chercheurs de travailler eux-mêmes sur des versions structurées de leurs textes et de réaliser leurs expérimentations au sein d'un laboratoire de texte. En effet, au moment de la production des *Sources du Mont Saint-Michel*, les travaux sur les interfaces de travail ergonomiques sur des flux XML n'étaient pas encore assez avancés pour permettre aux chercheurs d'en disposer dans des conditions satisfaisantes alors que c'était le cas pendant la réalisation de l'*Hortus Sanitatis*.

D'autres projets ont ainsi directement participé à la construction de notre modèle et à sa validation expérimentale. Citons ici le projet *Montedite*[115], dirigé par Carole Dornier, qui propose de diffuser les trois volumes de notes de travail de Montesquieu qui n'avaient pas vertu à être édités en tant que tels, ce qui ne va pas sans poser problème et imposer un certain nombre de contraintes, [Buard et Dornier, 2008], [Dornier et Buard, 2010], comme la nécessité de restituer plusieurs états du texte en tenant compte des ajouts, biffures et corrections. Ces cahiers se composent de

115. https://www.unicaen.fr/services/puc/sources/Montesquieu/

fragments de textes écrits par Montesquieu lui-même ou par ses secrétaires. Certains de ces passages se retrouvent dans des textes édités. Il s'agit donc véritablement d'un recueil de notes dans lequel l'auteur venait puiser pour alimenter ses œuvres. Le travail dans le cadre du projet *Montedite* a permis de contrôler la souplesse du modèle de flux et de fragments et la possibilité de manipuler des sources plus récentes que les textes médiévaux.

Rappelons, avant d'entrer dans les détails des expérimentations retenues, que les flux XML sont toujours évolutifs. En effet, ils peuvent faire l'objet de nouvelles campagnes d'annotation et d'études dans le cadre de nouveaux projets dans la continuité ou non des éditions qui leur ont donné naissance initialement. Ainsi, le fait de figer le travail à instant t dans une édition n'implique en aucune manière de cesser le travail dans un *laboratoire de texte*.

10

Les sources du Mont Saint-Michel

Nous présentons ici la première mise en œuvre du modèle que nous proposons, à travers l'exemple des sources du Mont Saint-Michel [Bouet et Desbordes, 2009] et [Bougy, 2009]. Notons que cette première expérimentation permet de mettre à l'épreuve un certain nombre de points nodaux de notre proposition.

L'objectif est ici de tester la validité de la démarche du *single source publishing* et du modèle de flux de texte que nous proposons pour traiter des sources anciennes associées à un appareil scientifique complexe. Il s'agit en particulier, en suivant les principes proposés plus haut, de mettre en place un système de mise en page semi-automatique pour le papier et automatique dans l'édition en ligne à partir d'un flux de texte unique.

Notons également, pour être tout à fait précis, que nous avions déjà aux Presses universitaires de Caen, où cette expérimentation a été réalisée, mis en œuvre une chaîne de traitement exploitant les langages à balises pour d'autres publications plus classiques telles que des revues scientifiques, des actes de colloques ou encore des monographies.

D'un point de vue éditorial, l'objectif de cette première expérimentation est de poser les bases de la diffusion du texte d'une source, accompagné des commentaires scientifiques permettant sa bonne compréhension, sur plusieurs supports différents sans chercher nécessairement à optimiser tous les usages possibles de l'ensemble des supports concernés et de la diffusion en ligne en particulier. Ainsi, nous ne cherchons pas ici à tirer parti de toute la complexité des catégories textuelles manipulées par les chercheurs. Le balisage produit pour les sources du Mont Saint-Michel reste à un niveau éditorial et est l'œuvre de l'éditeur matériel. C'est donc ce dernier qui assure la tâche de structuration XML TEI. Les éditeurs scientifiques ont, dans le cadre de cette expérimentation, travaillé de manière traditionnelle, avec un logiciel

de traitement de texte, avec tout ce que cela implique, notamment du point de vue de la récupération des textes traités avec les toutes premières versions de ces logiciels.

Les sources du Mont Saint-Michel constituent donc une première couche d'expérimentation qui permet de poser les bases opérationnelles du modèle et de vérifier qu'un flux unique de texte, avec un niveau de balisage relativement simple discriminant les fragments de haut niveau comme les chapitres, titres, paragraphes et notes pour l'essentiel et surtout sans gestion des variantes en tant que telles, permet d'instancier de manière satisfaisante d'un point de vue éditorial et scientifique plusieurs supports différents. Le niveau d'encodage ne permet donc pas ici de restituer le texte de chacun des témoins.

10.1 Principes de structuration

10.1.1 Identification des fragments

Être en mesure de nommer les éléments constitutifs des textes manipulés de manière univoque est fondamental de plusieurs points de vue.

Tout d'abord éditorialement, il est capital, comme nous l'avons déjà évoqué, de pouvoir faire référence précisément aux textes. C'est particulièrement le cas dans le domaine de l'édition de sources anciennes, puisque les chercheurs doivent pouvoir pointer un passage précis, parfois une ligne d'un texte.

Ensuite, techniquement, pour construire des dispositifs efficaces, qu'il s'agisse d'outils de lecture ou de recherche, il est nécessaire de pouvoir manipuler les textes avec un niveau de précision adapté et avec un degré de certitude absolu.

L'attribution d'un identifiant unique à chaque fragment distingué lors du processus de structuration, quel que soit son niveau, éditorial ou scientifique, est indiscutablement la meilleure des solutions pour répondre à ces exigences techniques et éditoriales.

De plus, dans ses principes, l'identification des fragments doit être réalisée d'une manière rationnelle, directement liée à la source traitée, c'est-à-dire en rapport étroit avec l'organisation du texte, et *human understandable*. En effet, les chercheurs doivent pouvoir explicitement désigner un fragment sans être totalement dépendant d'un outil d'assistance. Il s'agit, ni plus ni moins, que de permettre l'échange autour du texte dans les meilleures conditions possibles.

La méthode de calcul des identifiants que nous utilisons s'inspire fortement du système de références absolues utilisées dans le domaine de l'édition de textes clas-

siques dans lequel chaque portion de texte, du chapitre à la ligne si nécessaire, est numérotée. Il s'agit de mettre en place une solution pour pouvoir faire référence à un texte indépendamment de ses formes d'édition, bref d'identifier de manière univoque chaque fragment constitutif d'un texte. Les solutions aléatoires et l'utilisation de *timestamps* sont écartées pour l'ensemble de ces raisons.

Chaque fragment est identifié en fonction de sa position dans le flux structuré. Autrement dit, l'arborescence est exploitée pour le calcul auquel elle sert même de base.

Les identifiants se composent d'une base, utilisée comme un préfixe stable à l'échelle du fragment de plus grande ampleur (le chapitre par exemple), suivie d'une série de chiffres séparés par des points.

Le préfixe se compose lui-même d'une chaîne de caractères comprenant la langue et éventuellement un fragment du titre, contenu dans l'élément `head` principal, c'est-à-dire le plus proche de l'élément `body` de l'arbre TEI ou caractérisé par un attribut d'un type particulier comme les chapitres par exemple. Le présupposé pour établir ce préfixe est la connaissance de la structure textuelle à l'identification de laquelle on souhaite procéder.

Sont ajoutées à cette base des séquences du type `.{1-9}*`. Le nombre de séquences ajouté est directement dépendant du niveau de profondeur d'identification auquel on souhaite parvenir et qui, bien entendu, est lui-même lié au niveau de balisage. Dans chacune de ces séquences, le chiffre correspond au numéro d'occurrence de l'élément courant dans l'arbre XML. Autrement dit, chaque séquence est en réalité un simple compteur d'éléments d'un type précis : `div`, `p`, `quote`, `seg`, etc.

L'identification peut être réalisée de manière continue pendant la saisie et la structuration en intégrant les dispositifs de calcul aux opérations d'insertion d'éléments, ou faire l'objet d'un traitement dédié global. Notons qu'une fois l'édition finalisée et publiée, ces identifiants ne devront plus être modifiés pour pouvoir être exploités pour le référencement et la citation.

Cette logique de construction systématique des identifiants permet également de simplifer la mise en relation d'éléments intégrés dans des flux textuels différents mais participant d'une même œuvre. Ainsi, une fois deux flux présentant des organisations logiques identiques identifiés avec cette méthode, les exploitations seront simples à réaliser.

Ainsi, la figure 10.1 donne un exemple du résultat de l'application de cette identification sur deux flux parallèles : une traduction à gauche et une transcription à droite. On peut observer la similarité de construction entre les identifiants des deux

flux. Dans la mesure où il s'agit de la transcription d'un texte et de sa traduction, on retrouve la même organisation dans les deux cas : il s'agit de s'appuyer sur cette identité de construction pour le calcul des identifiants qui vont simplifier la construction d'interfaces de lecture bilingues.

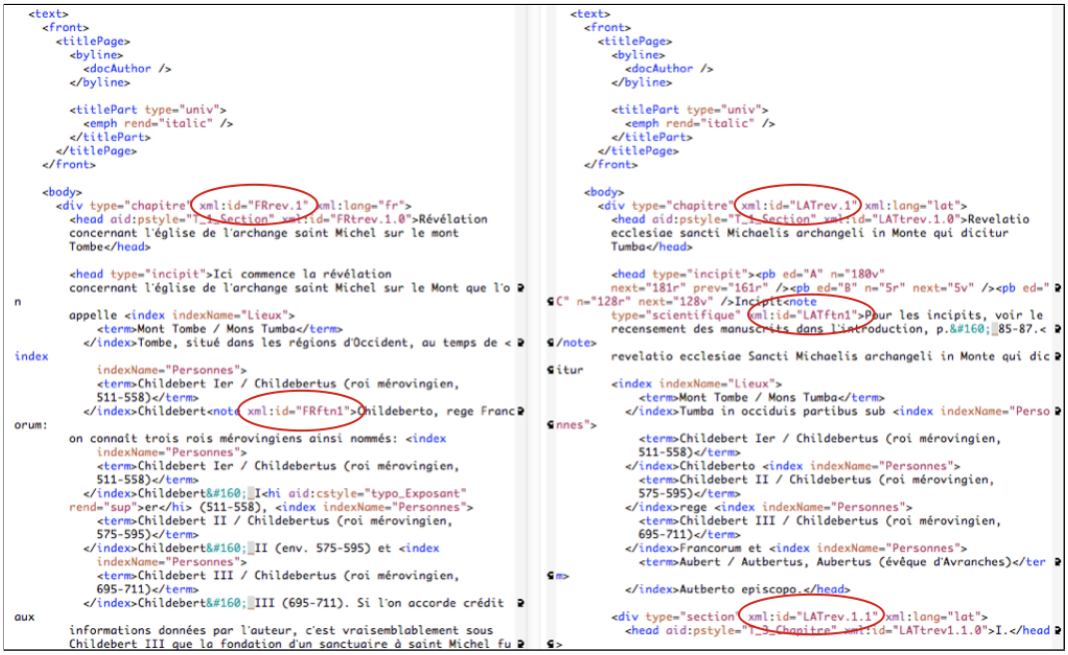

FIGURE 10.1 – Identification des fragments dans des flux XML TEI.

L'identification des fragments est aussi une étape indispensable pour beaucoup d'opérations relevant de l'exploitation des flux mis en place. Ainsi, pour permettre une annotation scientifique des textes, il est indispensable de disposer d'un moyen d'ancrer les notes dans l'arbre. Identifier chaque fragment permet de simplifier le travail en appliquant des solutions du type : ancrer une note dans tel élément identifié, à telle position.

De la même façon, une étude poussée de l'histoire des textes exige des possibilités de mises en relation qui imposent elles-mêmes de pouvoir désigner les éléments de manière univoque.

10.1.2 Flux de texte et corrections d'auteur

Dans la mesure où le contexte de travail est celui du *single source publishing*, la question des corrections d'auteur prend une importance particulière dans la mesure où nombre d'entre eux ont pris l'habitude de relire et de contrôler le fond uniquement au moment des premières épreuves d'imprimerie. Cet effet pervers des solutions modernes de traitement de texte est aujourd'hui intenable dans un contexte de mul-

tiplication des supports de diffusion et de lecture. Comment par exemple relire un texte sur le fond quand il n'y a pas de diffusion papier prévue, et donc *a fortiori*, d'épreuves d'imprimerie ?

L'introduction des flux de texte est donc ici lourde de conséquence. Il est indispensable de proposer et de mettre en place des solutions permettant aux auteurs de relire, contrôler et corriger les textes sur le fond sans bouleverser totalement leurs méthodes et habitudes de travail, même s'il n'est pas complètement illégitime de s'interroger sur leur pertinence.

Par ailleurs, cette question est l'une des plus récurrentes chez les éditeurs formés à ces techniques de production.

L'expérience des *Chroniques latines du Mont Saint-Michel* nous a permis d'aborder ce problème de manière totalement frontale. En effet, les auteurs ont relu l'ensemble des textes (transcriptions et traductions) au terme du travail de mise en forme de la version papier, qui, comme nous l'avons vu, exige tout de même un certain nombre de manipulations, automatisées ou non [116]. Autrement dit, c'est seulement à la fin de l'opération la plus chronophage que les auteurs ont été en mesure de relire leurs textes dans des conditions proches de celles d'un lecteur. Et, bien entendu, des problèmes de toute nature sont apparus : qualité de l'organisation des textes entre eux, cohérence de traduction d'un texte à l'autre, etc. Au final, ce sont de nombreuses corrections par page qu'il fallait traiter… Le problème réside dans le fait que les délais étaient très courts : l'impression devait débuter sans tarder pour respecter le planning. La figure 10.2 présente l'organisation du travail telle qu'elle était en 2009 au moment de l'édition des *Chroniques latines du Mont Saint-Michel*. Pris par le temps, nous avons privilégié la méthode qui consiste à exporter une version de traitement de texte [117] à partir du logiciel de PAO que nous avons pu de nouveau étiqueter correctement pour produire un nouveau flux balisé en XML TEI à partir d'OpenOffice dans des conditions acceptables. C'est la solution 2 sur la figure 10.2. Cette option n'est envisageable qu'avec un niveau de balisage simple, ne s'intéressant qu'aux catégories textuelles indispensables pour l'édition, et ne peut être considérée que comme une sorte d'ultime recours.

Si cette première expérience a permis de fixer les bornes d'intervention des différents acteurs dans des conditions de production réelles, il est totalement déraisonnable de construire des bonnes pratiques sur cette solution.

116. Voir p. 181 et suivantes.

117. Le logiciel de PAO utilisé, Indesign, est incapable de gérer des fragments de textes structurés dans les notes de bas de page ; il était donc impossible de récupérer un flux XML complet.

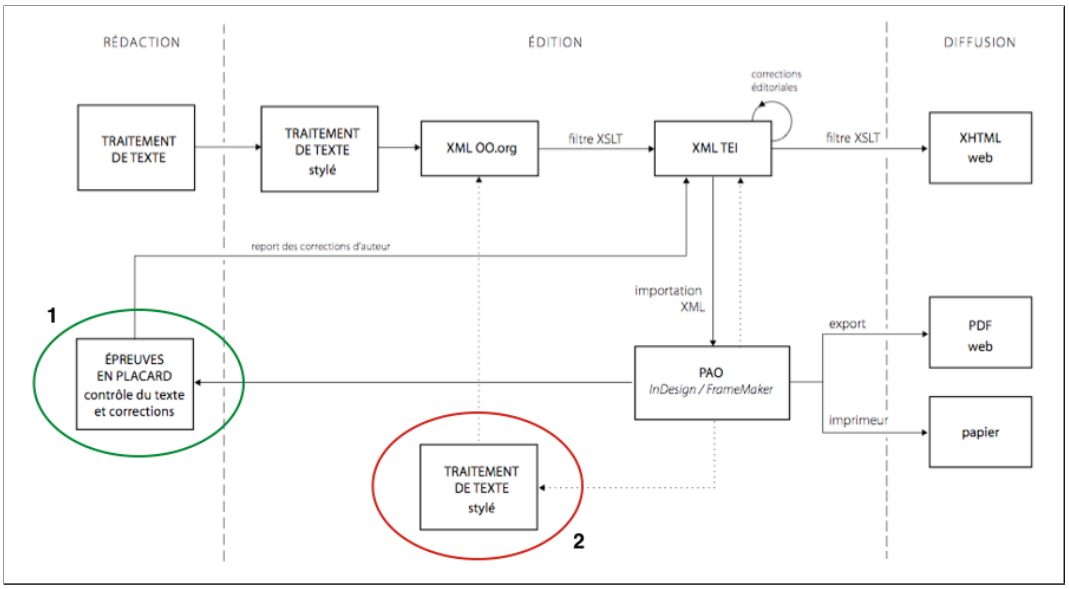

FIGURE 10.2 – La chaîne de production au moment de la production des *Chroniques latines du Mont Saint-Michel*.

Ainsi, dans le cadre de l'édition sur le modèle du *single source publishing*, la méthode à privilégier pour la gestion des corrections d'auteur est la solution 1 de la figure 10.2 qui consiste à importer le flux de texte une première fois dans une maquette de travail spécifiquement destinée à la relecture du fond. Le document ainsi produit doit respecter la longueur de ligne finale mais proposer une hauteur de bloc beaucoup plus importante et laisser des marges latérales confortables pour les commentaires. Il s'agit d'un document d'étape qui doit pouvoir être produit rapidement et sur lequel le travail de mise en forme doit être entièrement automatisé avec un minimum d'intervention humaine. Autrement dit, aucun travail de calage précis ne doit être entrepris : ce serait une perte de temps. Il s'agit d'être le plus efficace possible. Le fait d'étendre le texte sur une part importante de la hauteur de page permet aussi de simuler au mieux le flux de texte et de se rapprocher de certaines formes de diffusion. Ces épreuves de correction constituent donc une sorte de synthèse entre plusieurs modes de consultation, à la fois proche de la page et proche du flux…

Si le principe de ne commencer le travail de préparation et de correction de copie qu'une fois le manuscrit d'auteur complet est bien connu dans le monde de l'édition matérielle, il n'est pas toujours respecté, pour des raisons de calendrier le plus souvent. Cependant, dans le contexte de la convergence numérique, il est indispensable de respecter ce principe car, si le texte évolue trop lourdement, le risque de devoir reporter les corrections d'auteur sur chacun des supports de diffusion est réel.

10.2 Modèle éditorial

Les Chroniques latines du Mont Saint-Michel et *Le Roman du Mont Saint-Michel* constituent les deux premiers tomes d'une collection et présentent donc un certain nombre de traits communs, tous supports confondus. Nous présentons donc ici le modèle éditorial de la collection, c'est-à-dire l'ensemble de ses caractéristiques et principes communs en faisant éventuellement mention des spécificités de tel ou tel tome le cas échéant.

L'objectif, du point de vue éditorial, est d'articuler les supports les uns avec les autres et pas de les opposer. Tous les supports ne proposent pas le même contenu. En effet, seule la version papier de chaque tome, et le cédérom qui l'accompagne, propose la lecture des introductions scientifiques qui replacent les textes dans leurs contextes. La version électronique diffusée en ligne ne contient que les transcriptions et les traductions ainsi que l'intégralité des notes scientifiques qui les enrichissent et permettent de comprendre les règles d'établissement des textes.

Il s'agit donc de trouver des solutions de distribution des contenus sur chacun des supports pour faire en sorte qu'un même lecteur trouve un intérêt à utiliser tous les supports. Autrement dit, l'achat du volume papier ne doit pas rendre la consultation du site obsolète par exemple.

Enfin, les modalités de consultation ne sont pas les mêmes pour tous les supports. En dehors des spécificités évidentes inhérentes à la nature intrinsèque de chaque mode de diffusion, il est important de noter que la version électronique est totalement gratuite tandis que la version papier est payante et s'accompagne d'un cédérom qui reprend l'intégralité du contenu du volume imprimé.

10.2.1 Les versions papier

Le papier, dans les sources du Mont Saint-Michel, a un rôle éditorial fondamental puisque c'est ce support qui fixe le contenu afin de permettre un référencement stable. Il fournit également un document de travail profitant de toutes les qualités techniques du support papier : pas de besoin énergétique pour être consulté, confort de lecture, etc. Afin de tenir ce rôle, il porte l'intégralité des textes produits par les chercheurs dans le cadre de l'édition : de la transcription à la traduction, en passant par les notes et les introductions scientifiques. La seule limitation est le nombre d'images de manuscrits volontairement faible, pour des raisons évidentes de coûts d'impression. Ainsi, les volumes papier ne proposent qu'une sélection d'images et non l'ensemble des images des témoins principaux mobilisés par les chercheurs.

Les versions papier sont produites en utilisant le logiciel de publication assistée par ordinateur Indesign de la société Adobe. Le choix de cet outil est le résultat d'une observation pragmatique des situations de production chez les éditeurs institutionnels membres de l'AEDRES, contexte dans lequel les expérimentations ont été majoritairement menées, et qui utilisent très massivement, presque exclusivement en réalité, cet outil de publication.

Comme beaucoup de logiciels de sa génération, Indesign propose de travailler en construisant des blocs dont le contenu peut être déterminé précisément : texte, image ou les deux. Une fois ces blocs créés, ils peuvent être chaînés les uns aux autres pour permettre une circulation des informations à mettre en forme : dès qu'un bloc est plein, le contenu passe au bloc chaîné suivant. La proximité entre cette logique de chaînage et le modèle de flux de texte que nous proposons de placer au cœur de notre réflexion est évidente. Ainsi, à l'organisation des flux de textes balisés répond une logique de chaînage de blocs répartis sur une planche, c'est-à-dire le futur livre ouvert, auxquels on affecte des propriétés graphiques et typographiques pour l'instanciation de la forme papier.

Un autre point important pris en considération lors du choix du logiciel de PAO réside dans sa capacité à accepter les automatisations. Indesign propose ce type de fonctionnalités *via* des scripts qui permettent de réaliser toutes les opérations possibles dans l'interface graphique. Il s'agit donc de s'appuyer sur le flux de texte structuré pour automatiser, par des petits programmes, tout ce qui peut l'être. Ainsi les opérations répétitives telles que le placement des illustrations toujours présentes à la même place sur la page, l'extraction des notes de bas de page, la gestion des marqueurs d'index feront l'objet d'un traitement automatique sur la base de l'analyse des régularités exprimées dans la structure XML du flux textuel traité. Il convient cependant de garder à l'esprit que certaines opérations nécessitent toujours une intervention humaine, au moins pour faire les choix lorsque des règles deviennent contradictoires entre elles et que le logiciel n'est pas en mesure de faire un choix totalement pertinent et privilégie une règle plus ou moins aléatoirement. Un opérateur humain sera alors en mesure d'intervenir en amont sur les planches précédentes pour faire en sorte de régler le problème. En définitive le logiciel sera incapable de décider quelle règle il est préférable d'enfreindre.

L'exemple typique de ce cas de figure est la note composée de trois lignes qui ne tient pas en pied de page et dont l'appel est à la dernière ligne d'un paragraphe lui-même dernier sur la page. Bien entendu, il est impossible de séparer l'appel de note de la note, mais il est aussi impossible de couper un paragraphe de trois lignes.

Le plus souvent le logiciel choisit d'enfreindre la règle qui interdit de couper un paragraphe de trois lignes, autrement dit, de deux maux, il choisit le moindre. Pour autant, cette solution reste une infraction aux règles de composition : la seule solution est d'intervenir en amont pour, par exemple, réduire les espaces existants autour des intertitres ou des citations pour permettre de gagner de l'espace sur la page problématique.

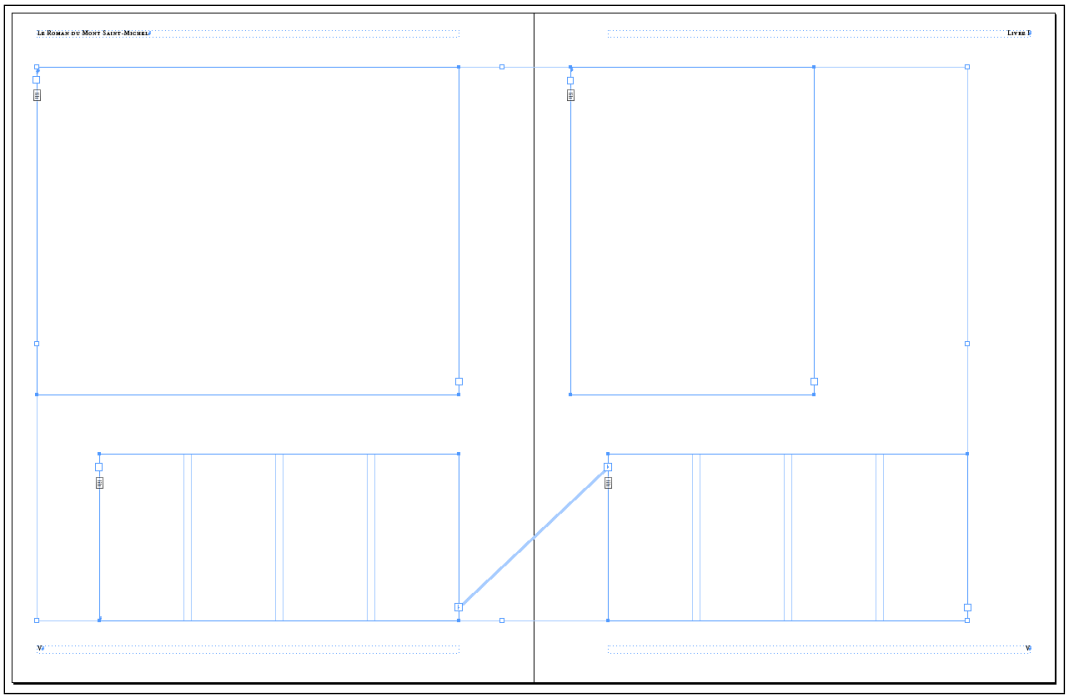

FIGURE 10.3 – Maquette Indesign pour l'accueil de flux de texte.

Pour l'instanciation de livres complexes tels que les éditions de sources anciennes, dans laquelle il s'agit d'articuler plusieurs types de textes entre eux, comme une transcription, une traduction et plusieurs systèmes de notes, le placement des blocs et la définition de leurs propriétés graphiques est une étape centrale. Pour cette étape capitale du travail, disposer d'un flux de texte typant précisément les différentes catégories textuelles est un atout important puisqu'il permet d'évaluer précisément le volume de chacun des types de textes. Ainsi, le maquettiste dispose d'un maximum d'informations pour faire les choix les plus efficaces lors de l'établissement et de la répartition des blocs sur les planches de la maquette. En fonction du nombre de caractères de chaque flux de texte, il pourra proposer les paramètres les plus appropriés pour chacun des types de blocs : taille de bloc, police et taille de caractère, nombre de colonnes, interlignage, espace intermots et intercaractères, etc. Dans tous les cas, la maquette proposera une solution de base qu'il s'agira d'adapter aux cas particuliers

rencontrés lors du travail de composition final. La figure 10.3 donne l'exemple d'une planche de maquette préparée pour le *Roman du Mont Saint-Michel* [Bougy, 2009] avant l'importation des textes définitifs.

La maquette porte donc un certain nombre de règles de placements et d'enchaînements des flux de texte sur les planches et entre les planches. Dans le cas précis des sources du Mont Saint-Michel, il y a deux flux principaux autour desquels les autres viennent s'organiser : la transcription et la traduction. La première se place toujours en page de droite et la seconde en regard, sur la page de gauche. Les flux périphériques, tels que les notes, sont importés dans un second temps.

Il s'agit donc dans un premier temps de séparer les flux de texte en fonction de leur nature : extraire les flux de notes scientifiques par exemple. La figure 10.4 donne une représentation générale de la séparation des flux et de leurs positionnements sur la planche. À partir du flux complet, trois flux différents sont produits :

— le premier contient uniquement la traduction ;
— le second contient la transcription (leçons retenues et variantes) ;
— le troisième contient les notes de commentaires scientifiques.

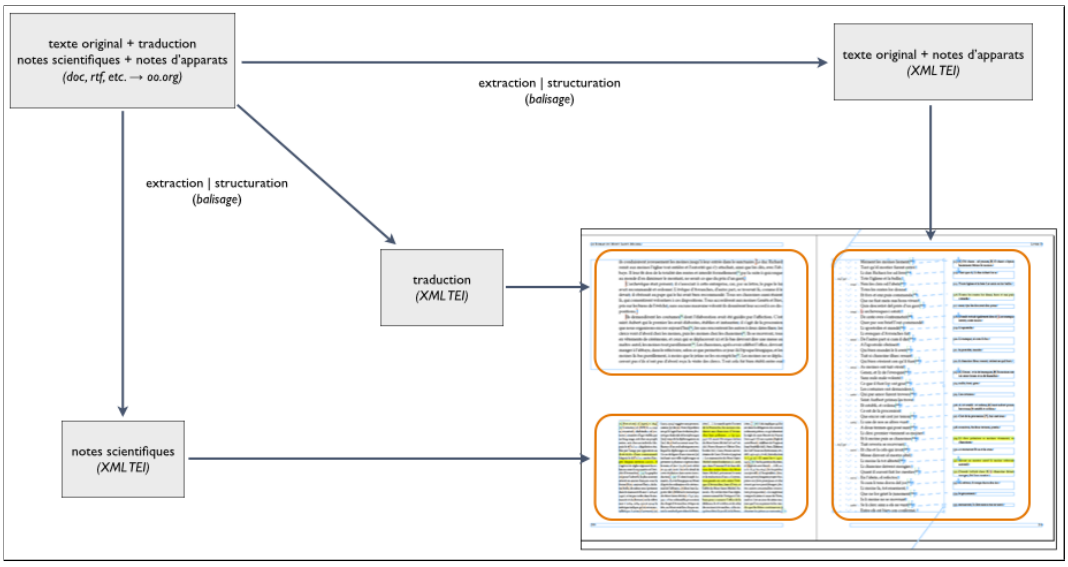

Figure 10.4 – Séparation des flux de texte.

Une fois tous les flux de texte importés, le travail consiste à les synchroniser ensemble de telle manière que le texte traduit sur la page de gauche corresponde exactement au texte transcris sur la page de droite et que tous les flux de notes apparaissent en relation avec leurs appels. Ce travail est réalisé de manière semi-automatique. En effet, la nature de certaines catégories textuelles permet d'automatiser leur placement tandis que certaines autres nécessitent une intervention manuelle pour le calage final

et la synchronisation des flux entre eux sur la planche, par exemple en redimensionnant certains blocs pour adapter la quantité de texte affichée sur une page. La logique qui préside est ici la suivante : tout ce qui peut être automatisé avec un taux d'efficacité suffisant doit l'être, tout ce qui va de toute façon imposer une intervention manuelle massive ne fait pas l'objet d'une tentative d'automatisation. Autrement dit, si les informations disponibles dans les différents flux de texte ne sont pas suffisantes pour permettre de placer avec certitude les éléments sur les planches, nous nous abstenons de faire des tentatives de déduction dont on sait qu'elles risquent fort de ne pas donner entière satisfaction et d'imposer finalement de refaire le travail. Il s'agit en définitive d'une position de prudence : mieux vaut ne pas régler un problème clairement que de prétendre le régler sans y parvenir réellement.

Le placement des notes de variantes en marge dans le *Roman du Mont Saint-Michel* est un bon exemple d'automatisation et donne une bonne perception du principe que nous venons de décrire.

Dans le flux contenant la transcription, avec les leçons et les variantes, nous disposons de toutes les informations nécessaires et correctement organisées. En effet, chaque vers est balisé par un élément `l` qui contient aussi, pour les vers concernés par une variante, un élément `note`, caractérisé par le couple attribut/valeur `type="variante"`, contenant le vers rejeté par l'éditeur scientifique.

Par ailleurs, l'objectif est de placer chaque vers rejeté en marge, en face de la leçon retenue dans le texte principal de la transcription.

Les informations sont donc claires, tant du point de vue de la structure d'entrée que de la forme souhaitée au terme des traitements. Il est donc tout à fait possible d'automatiser le placement des variantes en marge.

Techniquement il s'agit tout simplement de parcourir l'arborescence du flux de texte encodé en XML pour rechercher les éléments concernés : les éléments `note` avec l'attribut `type="variante"`. À chaque nœud rencontré correspondant aux critères de recherche, il suffit de créer un cadre ancré avec le bon style de bloc (pour attribuer les bonnes caractéristiques graphiques définies par le graphiste et enregistrées dans la maquette et le placement sur la page en particulier), d'extraire ensuite le texte du nœud, puis de le placer dans le bloc créé.

Une dimension importante de cette approche réside aussi dans le temps de traitement considérablement réduit. En effet, le placement des notes marginales manuellement n'exige pas moins de cinq opérations pour chaque note, toutes coûteuses en termes de temps :

— repérer la note ;

— couper le texte de la note de variante ;
— créer le bloc ;
— appliquer le style d'objet approprié (pour l'attribution des propriétés graphiques et le placement) ;
— coller le texte de la note de variante.

Même sans tenir compte des éventuelles erreurs humaines, qui ne manqueront pas de se produire si l'on tient compte du nombre de variantes [118], on compte par exemple 677 notes marginales contenant les variantes dans le *Roman du Mont Saint-Michel*, il est incontestable que prendre le temps d'automatiser cette série d'opérations est extrêmement rentable.

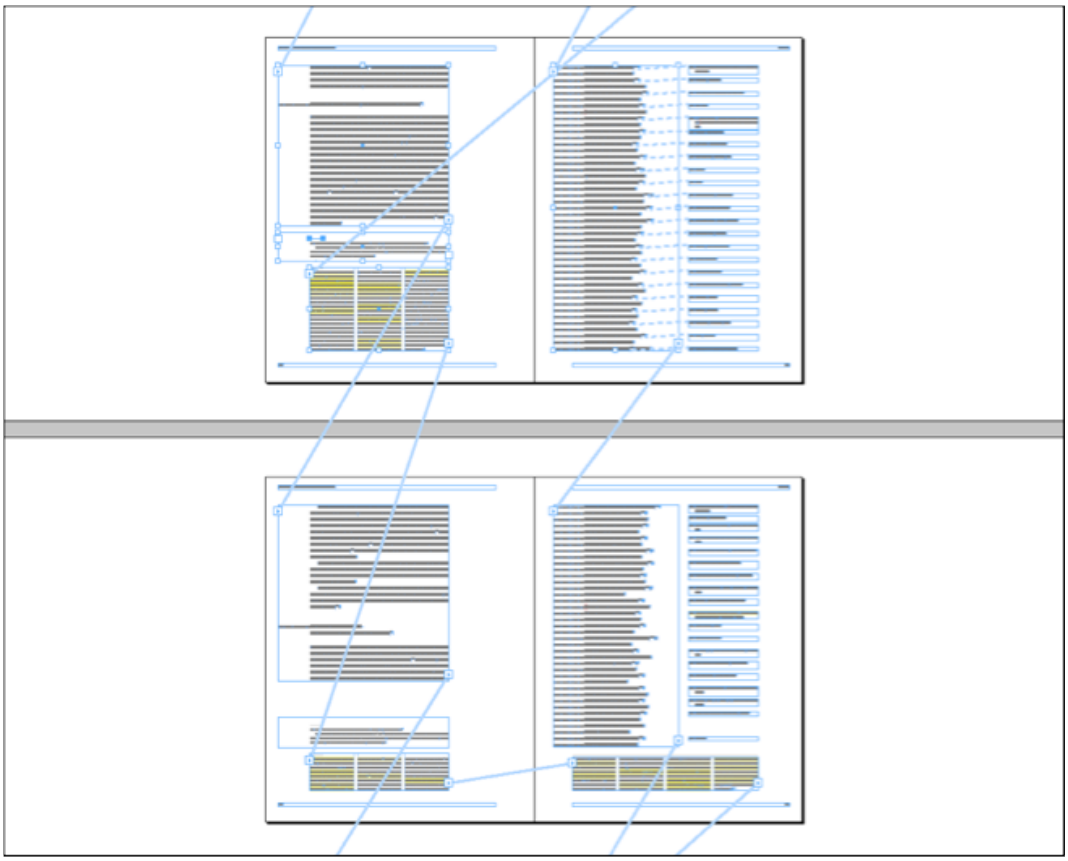

FIGURE 10.5 – Exploitation des flux XML TEI en PAO.

La figure 10.5 présente une vue de planches à l'issue du travail de placement des différents flux de texte. Les notes de variantes présentes en pages de droite sont dans des cadres ancrés (le lien vers l'ancre est symbolisé par des pointillés). Elles suivent donc le vers de la leçon en face de laquelle elles sont placées. Le système de

118. L'impact de ces erreurs peut être très lourd. Par exemple, si une note est oubliée, c'est l'ensemble de la numérotation qui devra être reprise.

chaînage de blocs sur les planches apparaît aussi clairement avec, en particulier, les liens entre les blocs de transcription en page de droite et la traduction en page de gauche. L'équilibre entre ces blocs, et donc entre les volumes de texte, est obtenu par des interventions manuelles de l'opérateur de PAO.

10.2.2 Les versions cédérom

Les versions cédérom se composent de fichiers PDF reprenant l'ensemble des contenus de la version papier et sont commercialisées avec ce support. La seule différence notable réalisée à la toute fin du processus de mise en page est l'ajout des liens hypertextes pour rendre actif tout le système de renvois : table des matières, index, glossaire, etc. Ces opérations se basent sur les numéros de pages pour assurer le système de circulation entre les points de départ et d'arrivée, c'est pourquoi ces opérations ne peuvent être réalisées définitivement qu'une fois la mise en page achevée.

Techniquement il s'agit simplement de fichiers PDF produits à partir d'Indesign une fois toutes les opérations de mise en page terminées. Il est en effet capital de réaliser les opérations de production des fichiers PDF à l'issue de l'ensemble des tâches de PAO car ces fichiers sont des images exactes de ce que sera la version imprimée et la pagination en est aussi l'unité de lecture.

La figure 10.6 donne une vue de la version cédérom du premier tome des *Sources du Mont Saint-Michel*. La table des matières dynamique, à droite, permet de circuler simplement et rapidement dans le texte. Cette table des matières a été produite automatiquement à partir de l'interprétation du flux de texte complet importé dans le logiciel de PAO.

Toutefois, il est possible de mettre en place des systèmes de liens hypertextes entre différentes parties d'un ou de plusieurs fichiers PDF. La table des matières dynamique de la figure 10.6 en est un exemple, mais examinons ici la méthode choisie pour faciliter l'utilisation du glossaire du *Roman du Mont Saint-Michel* dans sa version PDF.

La méthode, sur laquelle nous reviendrons en détail plus loin, repose sur le principe simple et parfaitement maîtrisé de la mise en place de sources et de destinations. Il s'agit de faire en sorte que chaque vers devienne une cible de lien potentiel, sans présumé *a priori* de son utilisation. Les opérations se déroulent en deux temps :

— pose des ancres sur chaque vers pour permettre de pointer vers chacun d'eux. Il s'agit en fait d'un point de référence interne à Indesign qui va servir de destination aux liens. Les ancres sont toutes formatées de la même manière ;

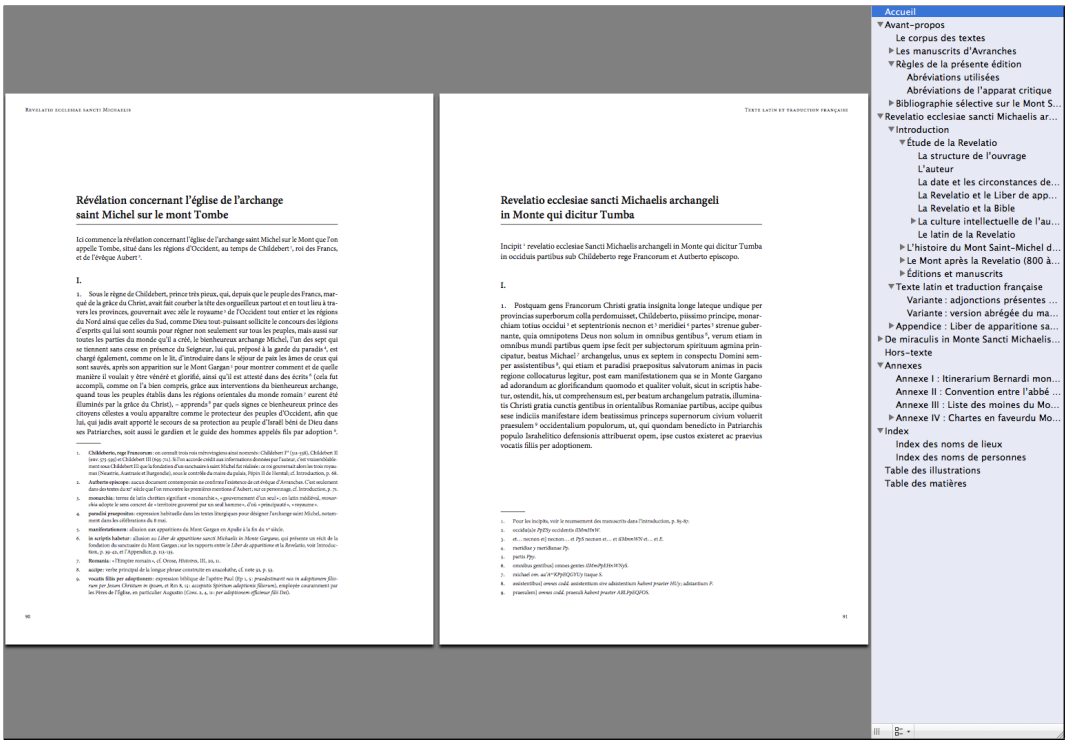

FIGURE 10.6 – Capture de la version cédérom des *Chroniques latines du Mont Saint-Michel* (consultée avec le logiciel *Aperçu* sous Mac OS).

— parcours du glossaire, repérage des vers (il s'agit d'une suite numérique de chiffres, chaque suite est séparée des autres par des virgules ou des points). Une simple expression régulière permet de repérer chaque occurrence. Pour chaque suite, un lien est posé sur la séquence de caractères pointant vers l'ancre correspondante, c'est-à-dire vers le vers repéré à l'étape précédente.

Pour ce glossaire, comme pour les autres exemples présentés, nous retrouvons ici la relation étroite entre la *structure logique* des fragments manipulés, nous travaillons toujours à l'échelle des chapitres définis soit dans les sources anciennes, soit par les éditeurs scientifiques, et leurs *structures physiques asynchrones* et dans ce cas précis, le fichier PDF (enrichi de liens hypertextes).

10.2.3 Les versions web

Les éditions en ligne [119] proposent un accès aux sources proprement dites, éventuellement enrichies, quand les droits ne sont pas si élevés qu'ils en deviennent rédhibitoires, de toutes les images des témoins principaux. Elles ne proposent pas, en

119. *Les Chroniques latines du Mont Saint-Michel* sont consultables en ligne à cette adresse : http://www.unicaen.fr/services/puc/sources/chroniqueslatines/. Et *Le roman du Mont Saint-Michel* à celle-ci : http://www.unicaen.fr/services/puc/sources/gsp/.

revanche, les introductions scientifiques qui sont réservées aux versions payantes, c'est-à-dire le papier accompagné du cédérom. Les éditions en ligne donnent donc "seulement" les textes portés par les manuscrits, édités par les chercheurs, la traduction de ces mêmes textes, ainsi que l'ensemble des systèmes d'annotations permettant leur bonne compréhension et détaillant des points de vocabulaire ou des choix d'établissement du texte. Nous verrons aussi que les modes d'accès à ces textes peuvent être multiples, avec par exemple, plusieurs parcours de lecture : contrairement au papier, cette multiplication des modalités d'accès et de lecture n'augmente pas significativement le coût de production.

L'échelle de consultation proposée par les éditions en ligne est celle du chapitre. C'est le fragment de base par lequel on accède au flux de texte des sources du Mont Saint-Michel. Le sommaire général liste l'ensemble des fragments correspondant aux chapitres, qu'ils soient présents dans la source primaire ou construits par les éditeurs scientifiques. Le lecteur choisit le chapitre qu'il souhaite consulter et le moteur d'affichage extrait les fragments correspondant dans les deux langues, latin ou ancien français et français contemporain.

Une fois un chapitre choisi depuis la page de sommaire, l'interface de lecture donne donc les deux versions du texte, la transcription latine ou en ancien français et la traduction en français contemporain, sur la même page écran, alignées en regard l'une de l'autre à l'échelle du paragraphe. La figure 10.7 présente une capture d'écran de l'édition électronique des *Chroniques latines du Mont Saint-Michel* qui permet de voir le système d'alignement des titres et des paragraphes de traduction avec les titres et les paragraphes de transcription. La composition de ce type de page de lecture repose sur la stricte application des principes d'identification évoqués plus haut. Le texte en colonne de gauche est placé sur la page à partir du flux de traduction et le moteur d'affichage va chercher le fragment de transcription correspondant, ici le paragraphe, pour le placer dans la colonne de droite.

Les versions en ligne des sources du Mont Saint-Michel constituent donc la première phase d'expérimentation et présentent certaines spécificités d'architecture logicielle dont nous devons dire un mot ici. En effet, pour ces deux tomes, c'est le système de fichiers qui a servi de solution de stockage des données. Autrement dit, les fichiers XML TEI contenant l'ensemble des informations sont directement stockés sur le serveur web dans l'arborescence locale. Chaque fragment correspondant à un chapitre est enregistré dans un fichier autonome. La gestion des fichiers et surtout la correspondance entre la transcription et la traduction sont gérées par une table de correspondance instanciée avant toute autre opération.

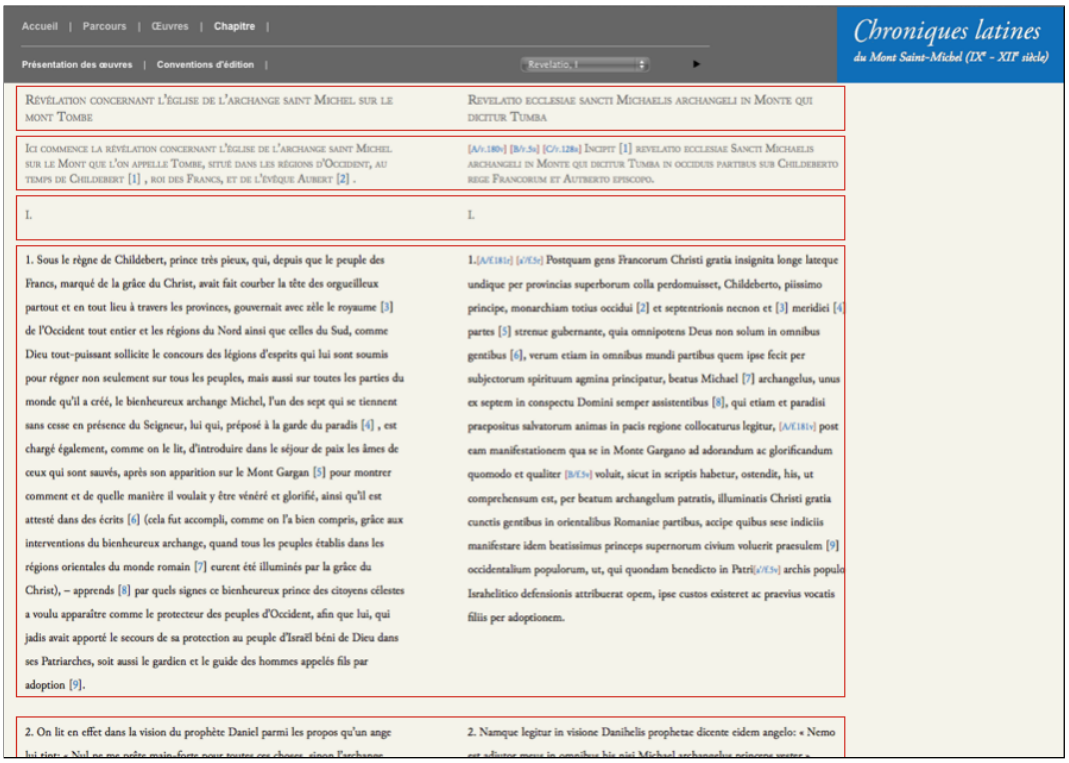

FIGURE 10.7 – Alignement des fragments de texte dans l'édition en ligne.

Les solutions de base XML native comme eXist [120] ou BaseX [121] n'étaient pas encore, en 2009, assez stables pour être intégrées au cœur d'éditions en ligne dont la stabilité est centrale.

Toutes les opérations de traitement comme la sélection de type de texte (notes marginales, marqueur d'index, etc.) ou la production d'outils de circulation (menus déroulants, boutons de changement de chapitre, etc.) sont produits par des feuilles de transformations XSLT sans recours à un langage de requête de type Xquery.

S'il est possible de considérer qu'il s'agit là d'une limitation fonctionnelle, rappelons que l'objectif est ici de valider les principes de base de construction d'une édition multimodale en exploitant des flux de textes encodés en XML dans le respect des recommandations de la TEI avec un niveau de balisage éditorial relativement simple. Les besoins en terme d'extraction sont donc plus limités par la profondeur d'étiquetage que par le choix des outils d'exploitation.

Enfin, ces expérimentations sont aussi, et avant tout, des productions réelles des Presses universitaires de Caen : l'impératif catégorique de proposer des solutions fiables et pérennes commande donc une certaine retenue.

120. http://exist-db.org
121. http://basex.org

10.3 *Chroniques latines du Mont Saint-Michel*

10.3.1 Présentation des sources et de l'édition

Les *Chroniques latines du Mont Saint-Michel* s'appuient sur quatre manuscrits principaux : Avranches, BM 210 ; Avranches, BM 211 ; Avranches, BM 212 ; et Avranches, BM 213. Cependant, beaucoup d'autres manuscrits ont été utilisés par les chercheurs pour l'établissement des textes édités [122]. La figure 10.8 présente des images numériques de ces manuscrits de la bibliothèque municipale d'Avranches.

Ces manuscrits proposent la lecture de six œuvres :

— *Revelatio ecclesiae sancti Michaelis archangeli in Monte Tumba* ;
— *Introductio monachorum* ;
— *De translatione et miraculis beati Autberti* ;
— *Miracula sancti Michaelis* ;
— *Baudri de Dol, De scuto et gladio sancti Michaelis* ;
— *Liber de apparitione sancti Michaelis in Monte Gargano.*

Ces textes racontent la fondation du premier sanctuaire du Mont par l'évêque Aubert, vers le début du VIII^{e} siècle, la manière dont le duc Richard I^{er} établit des moines bénédictins sur le Mont vers 965-966, la redécouverte des ossements et du crâne d'Aubert et présentent les prodiges attribués à l'archange des origines jusqu'en 1050.

L'édition, indépendamment de chacune de ses formes, présente des caractéristiques générales. L'ensemble de l'opération est réalisé en co-édition avec le Scriptorial d'Avranches, musée des manuscrits du Mont Saint-Michel.

L'édition est bilingue et articule les textes établis en latin par les chercheurs avec leur traduction en français contemporain.

Le programme éditorial met en œuvre trois supports de diffusion, en jouant sur leur complémentarité :

— qualité du support papier pour la lecture immersive et la stabilité de référencement ;
— souplesse, dynamisme et extensibilité volumétrique du web ;
— possibilités de consultation dynamique hors-connexion avec le cédérom [123].

122. Pour plus de précisions sur l'établissement des textes, se reporter aux introductions aux éditions dans [Bouet et Desbordes, 2009].

123. Depuis 2009, il est de plus en plus rare de ne pas disposer de connexion réseau, ce qui rend les qualités de ce support peut-être plus discutable qu'à l'époque. Ajoutons enfin, que de plus en plus de machines ne disposent plus aujourd'hui du périphérique de lecture nécessaire...

FIGURE 10.8 – Les principaux manuscrits de la bibliothèque municipale d'Avranches. *A* et *a′* : Avranches, BM 211 ; *B* : Avranches, BM 210 ; *C* : Avranches, BM 212.

Du point de vue commercial, les versions papier et cédérom, proposant l'intégralité du texte (introductions scientifiques, systèmes de notes, index, etc), sont distri-

buées dans un produit unique payant dans le réseau habituel de diffusion/distribution des Presses universitaires de Caen.

La version en ligne permet au lecteur de consulter l'ensemble des images des folios concernés des principaux manuscrits conservés au fonds ancien de la bibliothèque municipale d'Avranches. En outre, l'édition en ligne permet au lecteur de choisir un mode d'accès aux textes [124], de consulter, au fil de la lecture, les 275 images de manuscrits avec des niveaux de zoom progressifs pour être au plus près des sources.

Enfin, on pourrait imaginer produire d'autres versions en ligne des *Chroniques latines*, basées sur des sélections et des extractions de contenu pour toucher un public particulier. Par exemple, conserver uniquement la transcription latine et sa traduction, ainsi que les images, sans aucune note, pour faciliter une lecture par le grand public. Un des intérêts de ces méthodes de travail est justement de permettre, même bien après les productions initiales, de faciliter la création de nouvelles formes de diffusion à partir du même flux de texte, éventuellement en opérant des sélections sur ce dernier.

10.3.2 Modélisation : Textes et artefacts

L'une des caractéristiques importantes des *Chroniques latines* réside dans la multiplicité des versions de textes qui impose de clarifier la nature des rapports entre les textes et leurs supports dans le cadre d'une édition multimodale.

La figure 10.9 donne une représentation de la collection de textes des *Chroniques latines* dans les principaux manuscrits [125]. Les chercheurs bornent des fragments de texte correspondant à des récits dans chacun des manuscrits. Ils tracent en quelque sorte les zones textuelles de correspondance tout en identifiant les limites de ces zones indépendamment des supports matériels.

Une fois les éléments du flux central correspondant aux récits identifiés, c'est-à-dire les fragments de plus haut niveau, il reste aux chercheurs à les articuler les uns avec les autres. Bien entendu, il est tout à fait possible de tester plusieurs organisations des textes. C'est l'un des aspects fondamentaux de la notion de laboratoire de texte que nous défendons.

Chaque artefact porte donc des textes et ces textes sont organisés dans l'édition par les chercheurs pour en faciliter l'accès et la compréhension. Le flux central de la

124. Voir p. 197.

125. Pour des raisons de lisibilité, le cas particulier du manuscrit hétérogène Avranches BM 211 a été divisé sur la figure 10.9 en deux unités identifiées `ms211a` et `ms211b`, même s'il s'agit bien d'un seul manuscrit recomposé.

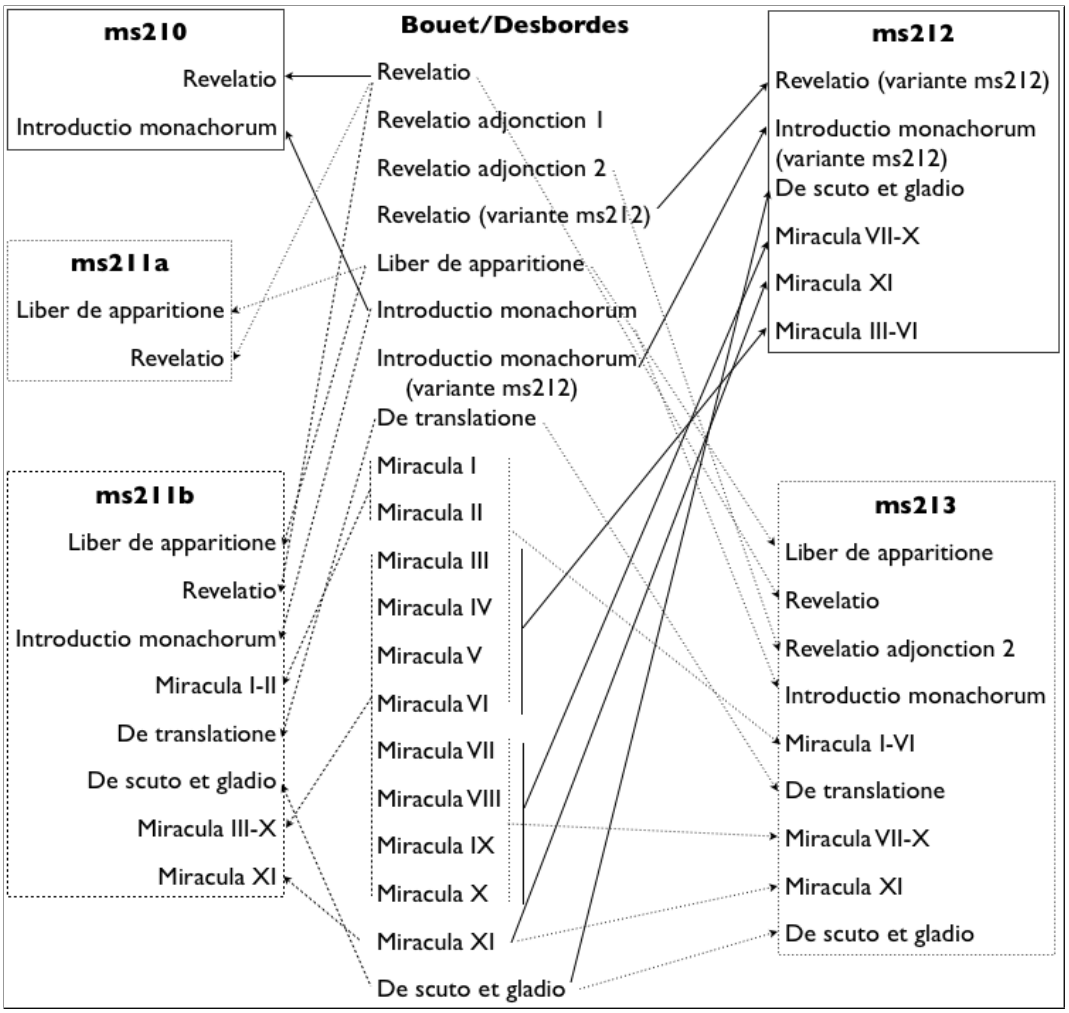

FIGURE 10.9 – L'organisation des textes dans les *Chroniques latines du Mont Saint-Michel*.

figure 10.9 correspond donc au travail des chercheurs et rend compte de leur analyse des textes. Ainsi, les artefacts proposent des versions différentes des mêmes textes que les chercheurs organisent en un système logique, c'est-à-dire en flux pour la lecture ou toute autre exploitation.

La figure 10.10 présente l'organisation des structures[126] du fragment de la *Revelatio ecclesiae sancti Michaelis archangeli in Monte Tumba* avec l'ensemble des instances qui en proposent la lecture, des manuscrits aux éditions, en passant par la vue d'édition. C'est l'articulation de sa structure logique et d'une structure physique qui permet d'accéder à un fragment. En effet, une fois un fragment logique identifié dans les manuscrits par les chercheurs, on retrouve la situation dans laquelle nous avons, en germe, une *structure logique*, l'arbre XML qui lui correspond, et plusieurs

126. Voir p. 131 et suivantes.

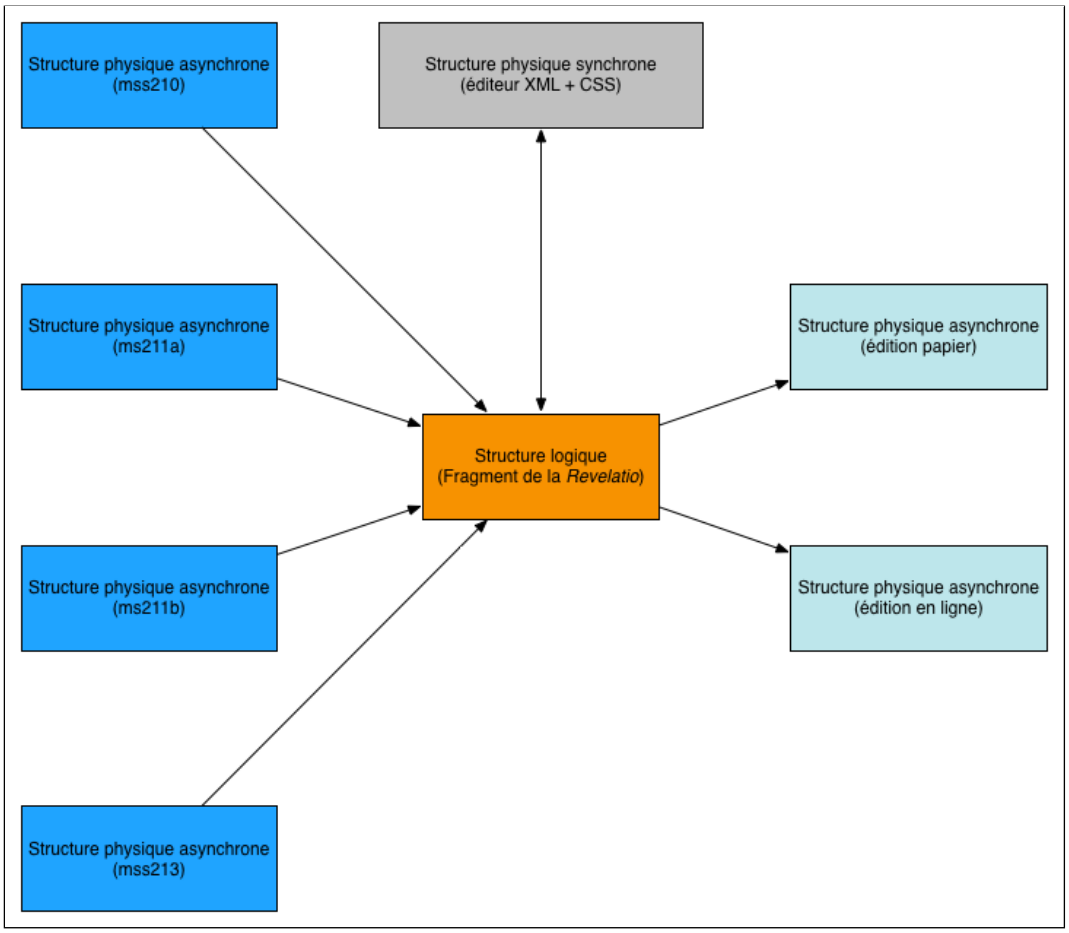

FIGURE 10.10 – Structure logique et structures physiques du fragment de texte de la *Revelatio*.

structures physiques asynchrones, une par manuscrit ainsi que, lors des travaux d'édition, une *structure physique synchrone*. Rappelons ici que les variantes textuelles ne sont pas encodées, le travail de balisage étant l'œuvre de l'éditeur matériel et non des éditeurs scientifiques, la restitution du texte exact de chaque témoin est donc impossible en intégralité. En réalité, l'articulation entre les structures physiques et la structure logique ne peut se faire qu'au niveau des fragments de niveau paragraphe, mais pas à des niveaux plus profonds. Le texte encodé rend donc compte du résultat du travail des chercheurs mais ne fournit pas l'ensemble des éléments sur lesquels ils se sont appuyés. Comme nous l'avons déjà évoqué, c'est bien le niveau de balisage qui limite les possibilités d'exploitation. Aller au-delà demande d'intégrer les chercheurs à la construction même du flux structuré.

Du point de vue théorique il s'agit de libérer les textes de leurs supports d'origine sans perdre le lien que ces textes entretiennent avec ce support. Nous sommes bien

dans une logique de factorisation des textes portés par plusieurs manuscrits. Il s'agit bien en effet, d'établir un texte édité, indépendant de chacune des versions proposées par les différents témoins. Autrement dit, le simple fait de retenir certaines formes textuelles et d'en rejeter d'autres, très exactement le travail d'édition scientifique des chercheurs, s'inscrit tout à fait dans ce modèle de factorisation et de mesure d'écarts. Mais il s'agit aussi pour nous, d'intégrer l'activité d'édition matérielle dans la même démarche. Dans le cas des *Chroniques latines*, le travail des chercheurs s'est déroulé de manière traditionnelle, cependant l'approche générale reste tout à fait similaire. Par ailleurs, dans le cadre de cette première expérimentation, l'enjeu était de vérifier la possibilité d'intégrer le travail d'édition matérielle.

L'objet du travail courant est donc le texte comme expression du sens, indépendamment de tout support, presque au sens FRBR : le travail d'édition consiste alors en quelque sorte à séparer le texte de ses manifestations. Le travail est donc réalisé sur un flux d'informations textuelles, finalement assez proche, dans sa nature, d'un *verbatim* de conversation dans lequel sont données les propositions de chaque témoin dont certaines sont retenues comme élément du texte édité, le résultat du travail, et d'autres écartées de ce résultat.

Les fragments qui composent le flux de texte sont de plusieurs types ici : chapitres, paragraphes, titres et notes d'apparats critiques et scientifiques. Si la définition des fragments de haut niveau ne présente pas de difficulté dans le cas des *Chroniques latines*, puisque les récits sont facilement identifiables en tant que tels par les spécialistes, il n'en va pas de même pour les fragments internes qui les constituent. Il s'agit donc ici d'examiner les principes et les méthodes pour redéfinir et/ou identifier des unités.

Tout comme les chapitres, les notes scientifiques ne posent bien entendu aucun problème puisqu'elles sont ajoutées par les chercheurs eux-mêmes et ne sont évidemment pas présentes dans les textes portés par les manuscrits auquel elles apportent justement des informations supplémentaires pour en faciliter la compréhension. De la même manière, les titres sont attribués par les chercheurs. Leurs positionnements dans le flux et la définition de leurs limites se déroule donc naturellement.

En revanche, si les paragraphes n'existent pas non plus dans les textes portés par les manuscrits, il s'agit tout de même d'intervenir sur l'organisation du texte d'une toute autre manière. En effet, il ne s'agit pas seulement d'une rupture formelle, ou encore d'une intervention sur une structure physique synchrone ou asynchrone, mais bien d'une modification des éléments logiques du texte porté initialement par les témoins. La figure 10.11 présente deux portions de structures physiques asynchrones

du texte de la *Revelatio*. on y voit clairement que l'un des critères retenus par les chercheurs pour insérer un nouveau paragraphe, en plus du sens du texte, est ce que nous appellerons ici la capitale. Mais à vrai dire, toutes ces capitales n'ont pas le même rôle : certaines marquent le début d'un nom propre, d'autres le début d'un discours rapporté, etc.

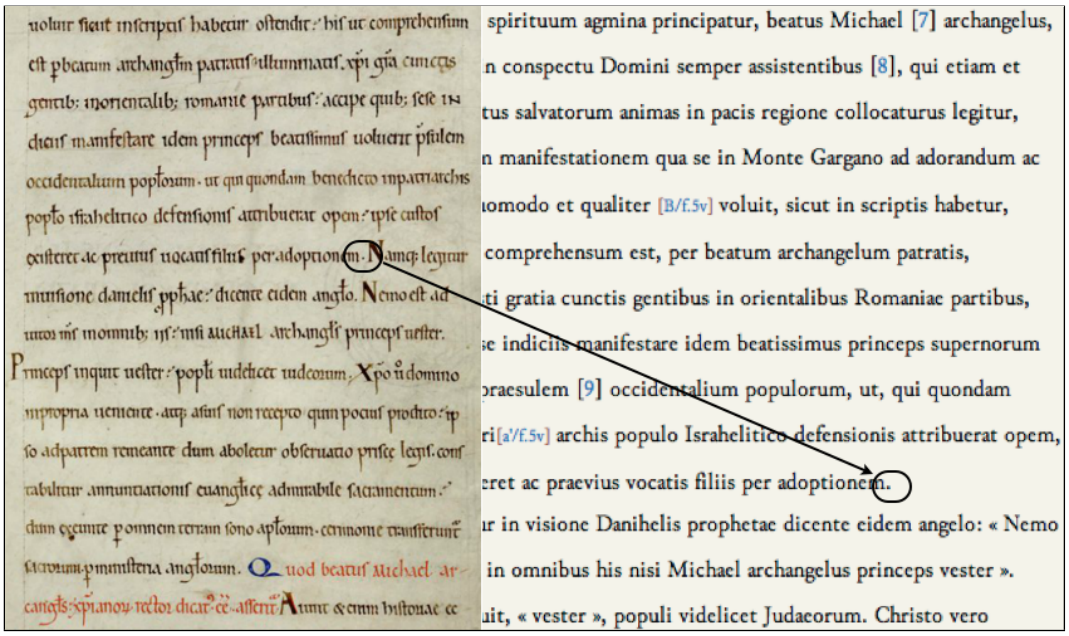

FIGURE 10.11 – Identification des fragments de texte : vue d'un témoin et de l'édition en ligne.

Les chercheurs estiment la force de la rupture textuelle en s'appuyant sur le sens et sur des indications formelles puis ils lui assignent une nature plus moderne ; là encore pour en faciliter la transmission, la compréhension et l'intelligibilité par un lecteur aujourd'hui. Concrètement, dans les *Chroniques latines*, les éditeurs scientifiques ont reconstitué des paragraphes dans leur logiciel de traitement de texte au niveau de la rupture indiquée sur la figure 10.11, paragraphes qui sont devenus des éléments `p` dans le flux XML après exportation et qui sont traités en tant que tels lors de la phase de production de la structure physique asynchrone de lecture en ligne.

10.3.3 Parcours de lecture

La notion de parcours de lecture est intimement liée à la multiplicité des manuscrits. En effet, aucun des manuscrits ne porte l'intégralité des textes édités et, en conséquence, aucun ne correspond à l'organisation des textes établie par les chercheurs. La séquence de lecture proposée par les éditeurs scientifiques peut donc, dans une

certaine mesure, être considérée comme une manière spécifique d'aborder les textes. En réalité, chacun des manuscrits propose sa propre organisation qu'il s'agit de considérer et de présenter en tant que telle.

Le principe de parcours de lecture proposé par l'édition en ligne est en lien direct avec la collection de textes et sa répartition dans les différents manuscrits principaux présenté par la figure 10.9 [127]. Chaque page de sommaire propose la liste des textes lisibles sur l'édition contemporaine ou sur le témoin choisi ainsi qu'une présentation des spécificités du support : contexte de production, répartition des folios, etc.

La figure 10.12 donne les six sommaires de l'édition en ligne pour chaque organisation des textes : l'édition contemporaine de Pierre Bouet et Olivier Desbordes et les quatre manuscrits principaux. Notons qu'il y a en tout six parcours et non cinq car les deux sommaires du manuscrit Avranches BM 211 correspondent à deux cahiers différents assemblés en un recueil unique au XVII^{e} siècle et qui correspondent en réalité à deux organisations des textes différentes.

Le parcours principal, le premier présenté en haut à gauche sur la figure 10.12, et mis en avant est celui qui repose sur l'analyse et l'interprétation des éditeurs scientifiques, mais l'édition en ligne offre tout de même la possibilité au lecteur de lire les textes en fonction de leur organisation sur chacun des témoins. On considère qu'un lecteur qui s'intéresse à un témoin particulier est *a priori* un latiniste qui n'a donc pas besoin de la traduction. L'exploitation des balises marquant les sauts de page est particulière : dans la mesure où l'on s'intéresse à un témoin particulier, il est possible de donner une miniature de l'image du manuscrit concerné en marge. La figure 10.13 présente une capture d'écran de l'interface du texte édité de lecture d'un témoin unique.

Mais la lecture du texte édité en fonction de l'organisation d'un témoin donné impose aussi une autre opération de sélection des fragments : la sélection des notes d'apparat pertinentes, c'est-à-dire uniquement celles qui concernent le témoin sélectionné par le lecteur. Cette opération consiste, une fois le témoin choisi, à examiner le contenu de chaque note afin de déterminer si elle doit être retenue ou non.

Techniquement, il s'agit en réalité de mettre en mémoire l'identifiant du témoin choisi par le lecteur et de tester sa présence parmi la liste d'identifiants qui constituent les valeurs des attributs `resp` des éléments `note`.

La figure 10.14 donne un exemple d'encodage permettant cette comparaison dans le cadre de la sélection des notes à retenir. Cette dernière s'opère en cherchant l'iden-

127. Voir p. 194.

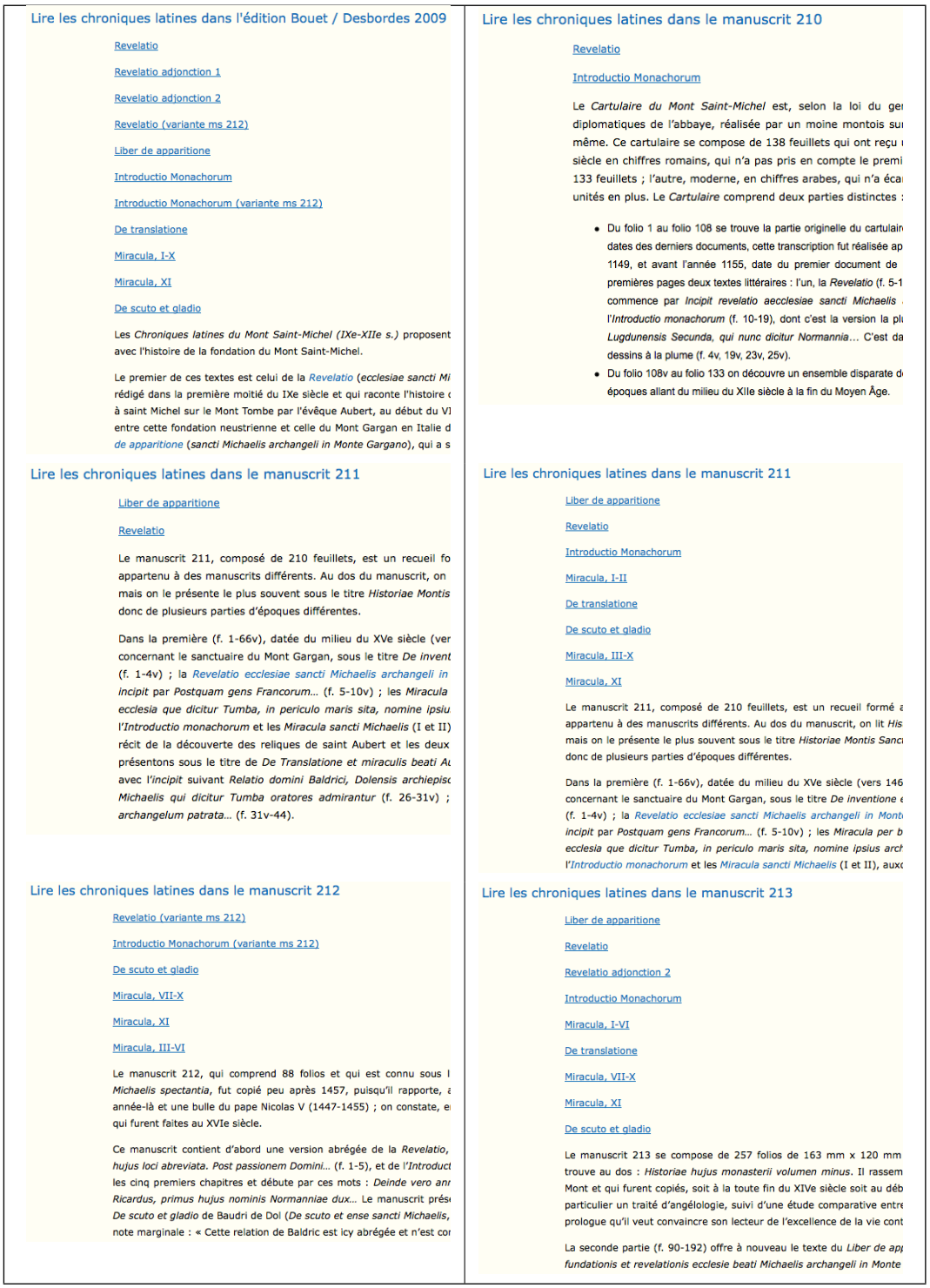

FIGURE 10.12 – Parcours de lecture des *Chroniques latines*.

tifiant du témoin courant dans les valeurs des attributs `resp` des lignes 7, 9 et 12 des éléments `note` des lignes 7, 10 et 13.

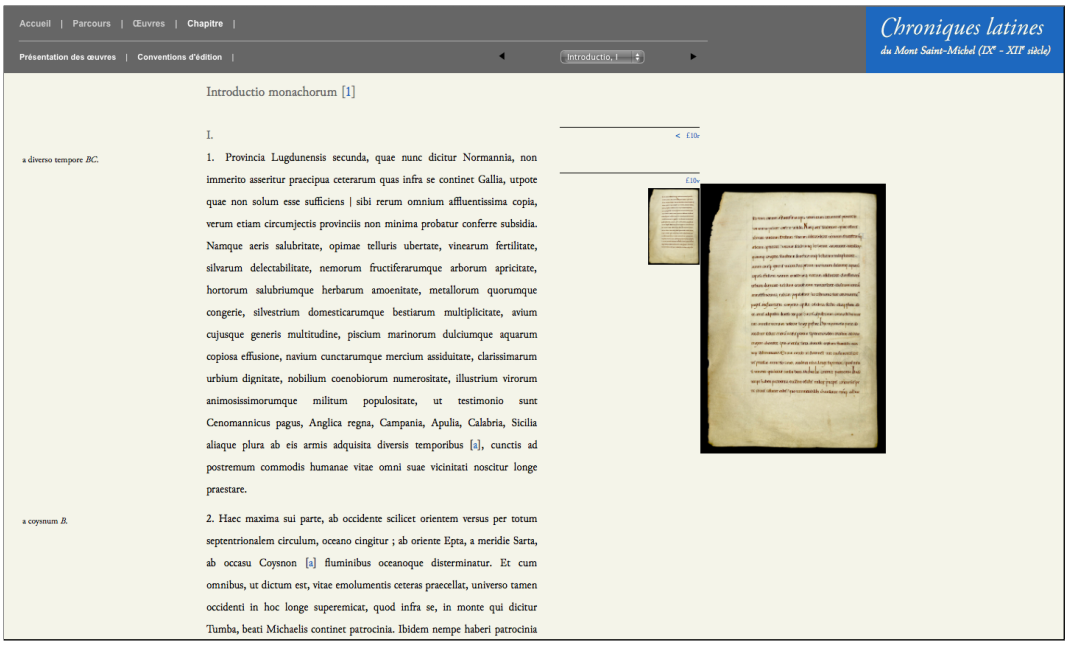

FIGURE 10.13 – Interface de lecture du texte d'un témoin spécifique.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  [...]
3  <p aid:pstyle="txt_Original_Prose_Numero"
4  xml:id="LATrev.1.1.1">1.
5    <pb ed="A" n="181r" next="181v" prev="180v" />
6    <pb ed="aprime" n="5r" next="5v" prev="4v" />Postquam gens
       Francorum Christi gratia insignita longe lateque undique
       per provincias superborum colla perdomuisset, Childeberto,
        piissimo principe, monarchiam totius occidui
7    <note resp="#P #p #E #S #y #i #I #M #m #H #n #W" type="
       apparat" xml:id="LATftn2">occidu(a)e <emph>PpESy</emph>
       occidentis <emph>iIMmHnW</emph>.</note>
8  et septentrionis necnon et
9    <note resp="#P #p #S #i #I #M #m #n #W #N #E" type="apparat"
       xml:id="LATftn3">
10 et... necnon et] necnon... et <emph>PpS</emph> necnon et... et
     <emph>iIMmnWN</emph> et... et <emph>E</emph>.</note>
11 meridiei
12   <note resp="#y #P #p" type="apparat" xml:id="LATftn4">
13 meridiae <emph>y</emph> meridianae <emph>Pp</emph>.</note>
14 partes
15 [...]
16 </p>
```

FIGURE 10.14 – Extrait de code XML TEI d'un fragment des *Chroniques latines*.

Cette structure est obtenue automatiquement sur la base de l'analyse des chaînes de caractères qui composent la note. En effet, rappelons que l'encodage XML des *Chroniques latines* est réalisé par l'éditeur matériel sans aucune intervention des cher-

cheurs. Ainsi, les valeurs des attributs `resp` sont renseignées à partir des séquences saisies en note de bas de page par les chercheurs dans leur logiciel de traitement de texte. À l'issue de l'exportation, le flux XML ne contient que des éléments `note` avec des attributs `type` et `xml:id` enrichis d'une sous-structure composée de caractères et d'éléments `emph` correspondants aux chaînes en italiques sous le logiciel de traitement de texte.

Pour ajouter les attributs `resp` avec les bonnes valeurs, il faut repérer et exploiter les éléments `emph` composés de chaînes de caractères contenant uniquement une série de lettres capitales ou minuscules et sans aucune espace typographique. Une fois ces séquences en mémoire, il est nécessaire d'ajouter un traitement qui ajoute le caractère `#` avant chaque lettre et un espace de séparation après pour produire une valeur exploitable simplement et stockable dans l'attribut `resp` qui ne doit contenir, d'après les recommandations de la TEI, que des pointeurs (`URI`).

Enfin, du point de vue du modèle, les parcours de lecture imposent bien de produire de nouvelles structures physiques asynchrones associées aux structures logiques des textes portés par le témoin choisi par le lecteur.

Notons enfin que l'ensemble des possibilités que nous venons d'évoquer reposent sur un balisage simple et destiné à l'édition : il n'est donc pas nécessaire de mettre en place des solutions de structuration extrêmement poussées pour être en mesure de proposer des systèmes de diffusion offrant de multiples options de consultation et de lecture.

10.3.4 Alignement des versions des textes

L'alignement des versions latine et française des textes édités dans le cadre des *Chroniques latines* repose sur l'exploitation du système d'identification de fragments que nous avons présenté plus haut [128].

Ainsi, la construction de la page web finale, avec la mise en regard des fragments latins qui rythment le placement des passages de traduction en français, repose sur la circulation de ces identifiants dans les requêtes formulées par le lecteur depuis l'interface de consultation. La figure 10.15 illustre le résultat obtenu à l'issue de l'ensemble des opérations. On y voit clairement comment les espaces verticaux sont ménagés entre les fragments de niveau paragraphe pour permettre l'alignement de chacun d'eux dans les deux langues.

128. Voir p. 176.

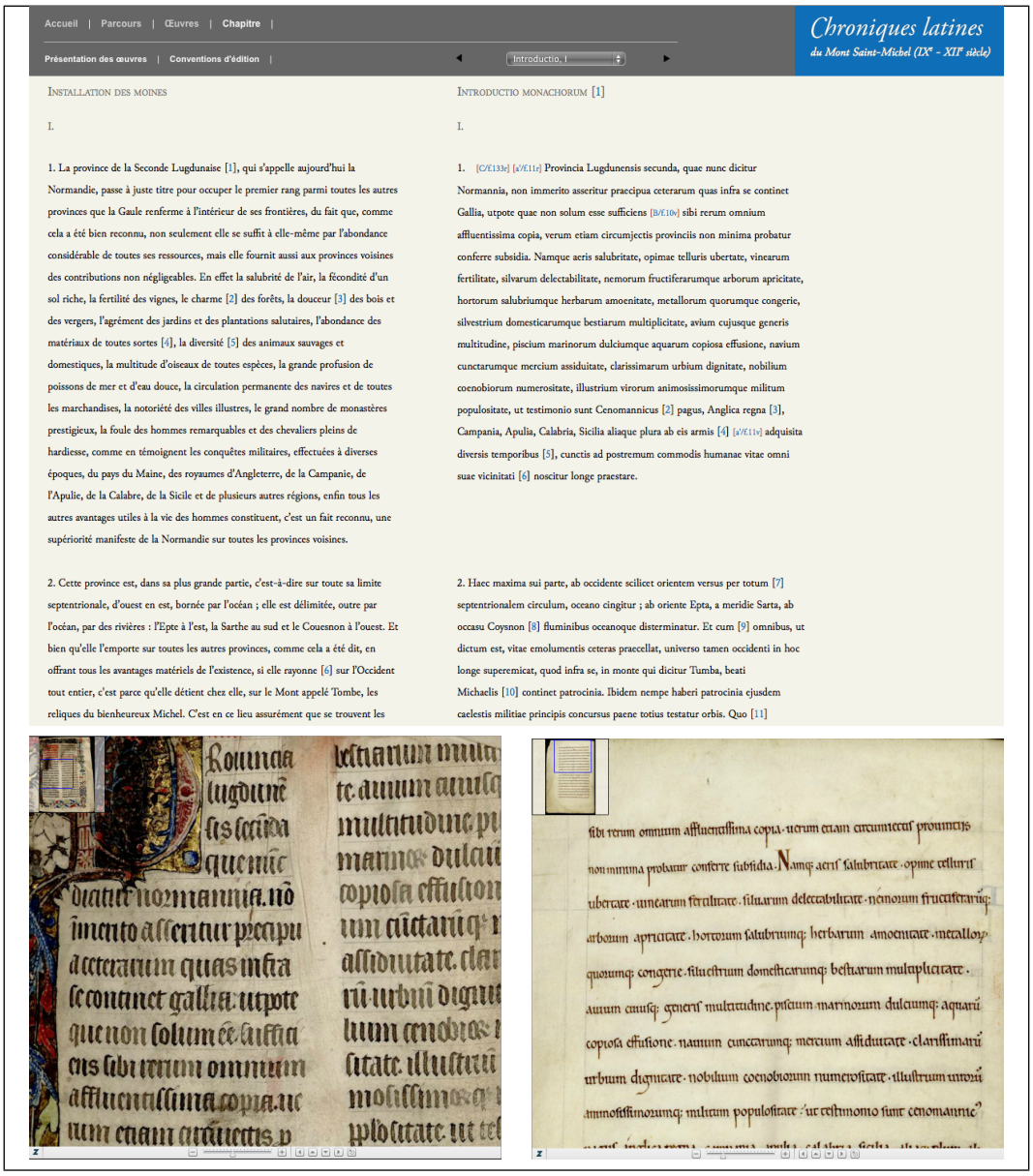

FIGURE 10.15 – L'interface de lecture en ligne bilingue des *Chroniques latines du Mont Saint-Michel*.

La figure 10.16 présente le mécanisme de production de cette page de consultation bilingue depuis la page présentant l'organisation des textes proposée par Pierre BOUET et Olivier DESBORDES. Le principe est simple : le lecteur demande à lire un fragment depuis le parcours de lecture Bouet/Desbordes [129] en cliquant sur l'un des textes proposés. L'identifiant du fragment français est envoyé au système qui détermine le fragment latin correspondant, grâce au système de construction des

129. Voir figure 10.12.

identifiants, puis extrait les deux fragments du flux central. Ensuite, les opérations de production de la page web demandée sont exécutées.

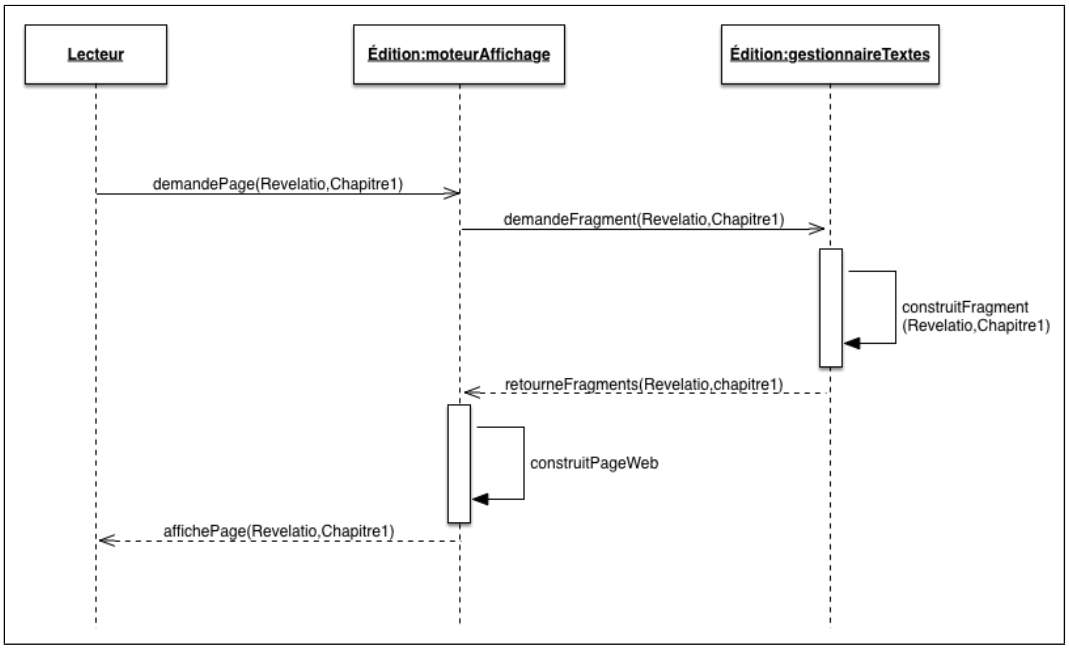

FIGURE 10.16 – Consultation du premier chapitre de la *Revelatio*.

Du point de vue du modèle de flux, il s'agit donc bien de manipuler deux flux distincts, la transcription latine et la traduction française, pour en extraire et unir deux fragments. Nous croisons donc deux structures logiques pour en produire une nouvelle contenant les deux versions de chaque fragment et c'est au résultat de ce croisement qu'est associée une structure physique asynchrone pour composer la page web de lecture.

Dans ce cas des *Chroniques latines*, le niveau de balisage éditorial ne permet pas de mettre en place un alignement plus fin que le niveau de paragraphe. Cependant, rien n'interdit de reprendre la structure pour la préciser davantage et ainsi donner la possibilité de mettre en place une interface assurant un alignement au niveau de fragments plus précis.

L'importance de l'architecture des données nous oblige ici à revenir sur l'organisation utilisée pour les *Chroniques latines*. Si la sélection des fragments inférieurs, à partir de la structure des textes, a été décrite, le choix du texte ou du récit (la *Revelatio*, l'*Introductio monachorum* par exemple) est en effet très dépendant de la manière de sérialiser les données.

Nous avons ici, dans le cas des *Chroniques latines*, choisi de sérialiser chaque texte dans des fichiers distincts les uns des autres et stockés directement sur le système de

fichiers du serveur sans utiliser de base de données[130]. Les parcours de lecture sont définis dans des tables d'association dans l'édition en ligne. Lorsque le lecteur choisi un texte sur la page du parcours de lecture, il permet en réalité de sélectionner les fichiers contenant les textes latins et français sur lesquels travailler.

10.3.5 Rapports texte/image

Le cas d'usage le plus évident des images numériques dans le cadre d'une édition de sources telle que *Les chroniques latines du Mont Saint-Michel* consiste à permettre au lecteur de convoquer les images des principaux manuscrits pendant sa lecture du texte édité et de sa traduction.

L'intégration qui nous occupe ici est donc la plus simple possible. Il s'agit de marquer dans les flux les ruptures de pages pour permettre de connaître le rythme de changement pour les manuscrits principaux. Autrement dit, il faut simplement insérer des éléments de type `milestone` dans le flux de texte aux endroits où se terminent les rectos et les versos des folios pour les quatre manuscrits de base. En définitive, l'objectif est bien de signaler un événement survenant sur l'un des témoins directement dans le flux de texte structuré. En fait, l'insertion du lien vers une nouvelle image doit correspondre au moment où un lecteur consultant le manuscrit doit tourner une page du codex. C'est bien la rupture de page physique qui commande l'action. Dans le cas d'une lecture directe sur le manuscrit, l'action consiste à manipuler l'objet porteur du texte pour poursuivre la lecture, mais dans le cadre de l'édition en ligne, bien entendu la rupture de page ne peut pas entraîner la même action. L'événement de rupture doit être signalé dans le flux de texte pour permettre au lecteur d'agir, par exemple en activant l'affichage de l'image numérique.

La figure 10.14, p. 200, donne des exemples d'encodage, aux lignes 5 et 6, d'un traitement XSLT permettant la transformation d'éléments TEI `pb` en liens hypertextes pointant vers les images dans l'édition en ligne, en exploitant l'élément XHTML `a`. Le flux XML TEI contient des éléments `pb` caractérisés avec les attributs `ed` pour le témoin, `n` pour le numéros du folio, `next` pour le folio suivant, `prev` pour le folio précédent. L'ensemble de ces informations sera mobilisé pour construire les liens dans l'un ou l'autre des modes de lecture proposés (latin ou bilingue).

La figure 10.17 présente un `template` XSLT qui prend en charge la création des liens vers le système de visualisation des images haute résolution des manuscrits à partir des flux encodés XML TEI. Les lignes 5, 7, 8 et 9 assurent l'insertion d'une

130. Voir p. 188 et suivantes.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  [...]
3  <xsl:template match="pb">
4    <span class="pagebreak">
5      <xsl:text>[</xsl:text>
6      <a target="_blank" href="javascript:void(0)"><xsl:attribute
           name="onclick">myWindow=window.open('images/Template.
         php?parcours=ms<xsl:value-of select="@ed"/>\&amp;folio=<
         xsl:value-of select="@n"/>','Manuscrit_<xsl:value-of
         select="@ed"/>','width=850,height=750,top=50,left=50,
         toolbar=no,resizable=1,scrollbars=yes');myWindow.focus()
         ;return false</xsl:attribute>
7        <xsl:value-of select="@ed"/>
8        <xsl:text>/f.</xsl:text>
9        <xsl:value-of select="@n"/>
10     </a>
11     <xsl:text>]</xsl:text>
12   </span>
13 </xsl:template>
14 [...]
```

FIGURE 10.17 – Extrait de code XSLT de création des liens vers les images de manuscrits.

chaîne de caractères spécifique directement dans le texte affiché dans l'édition en ligne. La création du lien *stricto sensu* se situe à la ligne 6 avec l'insertion de l'appel à l'outil de visualisation des images haute résolution avec des fonctionnalités de zoom.

La figure 10.15, p. 202, donne des exemples de liens du texte vers les images et de fenêtres de consultation des images.

10.4 *Le Roman du Mont Saint-Michel*

Les apports de l'expérience des *Chroniques latines* sont intégrés ici. Nous ne revenons donc pas sur la base de modélisation qui est identique (un flux de texte par langue, identification des fragments, rapports entre structures logiques et structures physiques, etc.), mais nous nous focalisons sur la capacité du modèle à rendre compte des textes du *Roman du Mont Saint-Michel* d'une part, ainsi que sur les apports spécifiques de cette source pour notre modèle d'autre part.

10.4.1 Présentation des sources et de l'édition

Le *Roman du Mont Saint-Michel* [BOUGY, 2009] se lit dans deux manuscrits conservés à la *British Library* :

— BL Additional 10289 : « manuscrit A » dans l'édition. Ce témoin a probablement été réalisé par un copiste du Mont. Ce manuscrit est incomplet et 7 passages sont manquants ;
— BL Additional 26876 : « manuscrit B » dans l'édition. Ce témoin présente l'avantage de proposer un texte complet mais l'inconvénient d'être peu soigné dans sa réalisation (parchemin de piètre qualité, erreurs, lacunes, etc.).

Les conditions d'acquisition de ces deux manuscrits sont inconnues, mais celles-ci datent probablement d'avant la Révolution française.

L'édition bilingue propose le texte édité en ancien français versifié en regard de la traduction en français contemporain en prose sur trois supports différents : papier, en ligne et cédérom. À cause des variations importantes du volume occupé par les deux versions du texte, c'est la version en ancien français qui rythme l'apparition des passages de traduction correspondants en français contemporain, laissant donc, dans certains cas, de larges espaces entre les paragraphes. La figure 10.21 permet de voir le rendu de cette solution d'alignement dans l'édition en ligne. De plus, dans les versions en ligne et cédérom, le glossaire doit être dynamique, ce qui représente la création de plus 15 000 liens en tout pour chacun de ces deux supports de lecture. Enfin, une version grand public du texte doit être préparée pour être diffusée sur le portail régional *Normannia*.

L'auteur du *Roman du Mont Saint-Michel*, Guillaume de Saint-Pair, un jeune moine de l'abbaye du Mont Saint-Michel, s'est appuyé sur les textes latins qui rendent compte de l'histoire du Mont Saint-Michel, les mêmes que ceux qui ont été édités par Pierre Bouet et Olivier Desbordes dans les *Chroniques latines*[131] :

> Guillaume de Saint-Pair a donc, pour ceux qui ne connaissent pas le latin et ne peuvent consulter les documents originels conservés à l'abbaye, *torné*, c'est-à-dire « traduit » ou plutôt « transposé » l'histoire du Mont Saint-Michel du latin en français.
> L'ouvrage a en effet pour principales sources les textes fondateurs du Mont Saint-Michel, rédigés en latin, et dont une copie partielle avait été récemment réalisée, entre 1149 et 1155, peut-être sur ordre de Geoffroy, abbé de mai 1149 à décembre 1150, ou de Robert de Torigni, élu en 1154, dans le *Cartulaire du Mont Saint-Michel*, recueil de textes et de chartes concernant l'abbaye, conservé à la bibliothèque d'Avranches sous la cote 210.

131. Pour plus de détails sur l'auteur et son œuvre, voir la rubrique *présentation de l'œuvre* dans l'édition en ligne : http://www.unicaen.fr/services/puc/sources/gsp/index.php?page=presentation.

Ainsi, le *Roman du Mont Saint-Michel* constitue en définitive une première exploitation des sources du Mont Saint-Michel tout en faisant partie intégrante de ces mêmes sources. En effet, Guillaume de Saint-Pair, soucieux de transmettre des informations vérifiées aux pèlerins qui ne connaissaient pas le latin, propose une version dans la langue de l'époque et destinée à les accompagner pendant leur pèlerinage.

Cependant, il ne s'agit pas simplement d'une transmission d'informations, mais aussi d'un texte avec une dimension politique :

> Le « guide touristique » qu'a composé Guillaume de Saint-Pair à l'usage des pèlerins est en réalité une œuvre polémique qui rappelle et défend avec vigueur les privilèges de sa communauté.

Enfin, le texte est aussi le résultat d'une activité créatrice exploitant les sources latines du Mont Saint-Michel, car Guillaume de Saint-Pair

> [...] a su adapter les chroniques latines du sanctuaire pour en faire une œuvre originale. Ses précisions et ses ajouts aux ouvrages qui lui ont servi de modèle, grâce à son érudition et à sa connaissance des chartes de l'abbaye, renforcent l'aspect didactique de son texte, mais lui permettent aussi de donner sa vision personnelle des faits. Son goût pour les digressions entraîne le lecteur à la découverte des espèces marines de la baie, le guide sur les chemins entre deux sanctuaires dédiés à l'archange, lui offre une description poétique du Mont contemplé depuis les hauteurs de l'Avranchin...

10.4.2 Alignement texte versifié / traduction en prose

Rappelons pour commencer que, dans le cadre des sources du Mont Saint-Michel en général et du *Roman du Mont Saint-Michel* en particulier, c'est l'éditeur matériel qui se charge de l'ensemble des opérations de balisage du texte. L'éditeur scientifique a remis un document de traitement de texte réunissant les deux flux, texte édité et traduction, "alignés" autant que possible, dans les limites de l'outil [132].

La première tâche consiste donc à extraire de ce document unique les deux flux à manipuler pour la construction des différents systèmes de lecture. Pour cela, le document est stylé en ajoutant des points de rupture qui permettent de cibler précisément les changements de langues. Ensuite, deux transformations sont appliquées en série : la première récupère les éléments constitutifs du flux en ancien français et la seconde les éléments du flux de traduction en français contemporain.

132. Voir p.83.

Il s'agit donc de développer une feuille de style spécifique donnant les renseignement sur la langue du texte étiqueté puis d'appliquer des transformations, ou plus exactement dans le cas présent, des extractions s'appuyant sur cette information de langue. C'est l'information de même nature, ici la langue, qui permet de regrouper dans un flux de texte unique l'ensemble des éléments discriminés par l'éditeur au moment de l'application des styles. Il est ainsi possible, à partir d'un document unique, de produire les deux flux de texte encodés XML nécessaires à la production de l'ensemble des formes de diffusion.

Cette souplesse et cette adaptabilité des opérations de production qui s'appuient sur la libération du texte hors des contraintes du document dans lequel il est enfermé, est l'une des forces du modèle de flux tel que nous le proposons.

Une fois la méthode de production des deux flux fixée, il s'agit de s'assurer que les deux structures logiques produites possèdent toutes les informations nécessaires à leur alignement.

Mais il y a une différence majeure entre la traduction et le texte édité en ancien français. En effet, si le texte édité est versifié, ce n'est pas le cas de la traduction, qui elle, est en prose. Pour permettre l'alignement d'un ensemble de vers avec le paragraphe de prose lui correspondant, il est indispensable d'ajouter un niveau de structuration, c'est-à-dire, du point de vue du modèle un nouveau type de fragment : celui du groupe de vers.

Concrètement, cela se traduit par la création d'un style particulier et l'insertion d'une séquence de texte totalement étrangère aux textes (texte édité et traduction) qui ne sera pas, bien entendu, incluse dans les flux résultants des opérations d'extraction. La figure 10.18 montre deux de ces séquences, *prose1* et *vers1*, au début d'un fichier de travail de l'éditeur matériel.

La première passe du traitement de transformation va produire un flux XML unique dans lequel ces deux séquences seront conservées, mais seront explicitement constituées en fragments cohérents et limités puisque le style défini est du même niveau hiérarchique. Les notes sont elles aussi incluses dans ce nouveau fragment, discriminant ainsi dans le même mouvement, le texte édité avec son système de notes et la traduction avec le sien. Au terme de la première passe, nous disposons donc d'un flux de texte unique encodé en XML, dans lequel tous les niveaux sont explicitement marqués.

La seconde passe se charge de séparer les deux flux de texte sur la base des éléments discriminés lors que premier passage. Ces deux flux présentent des structures

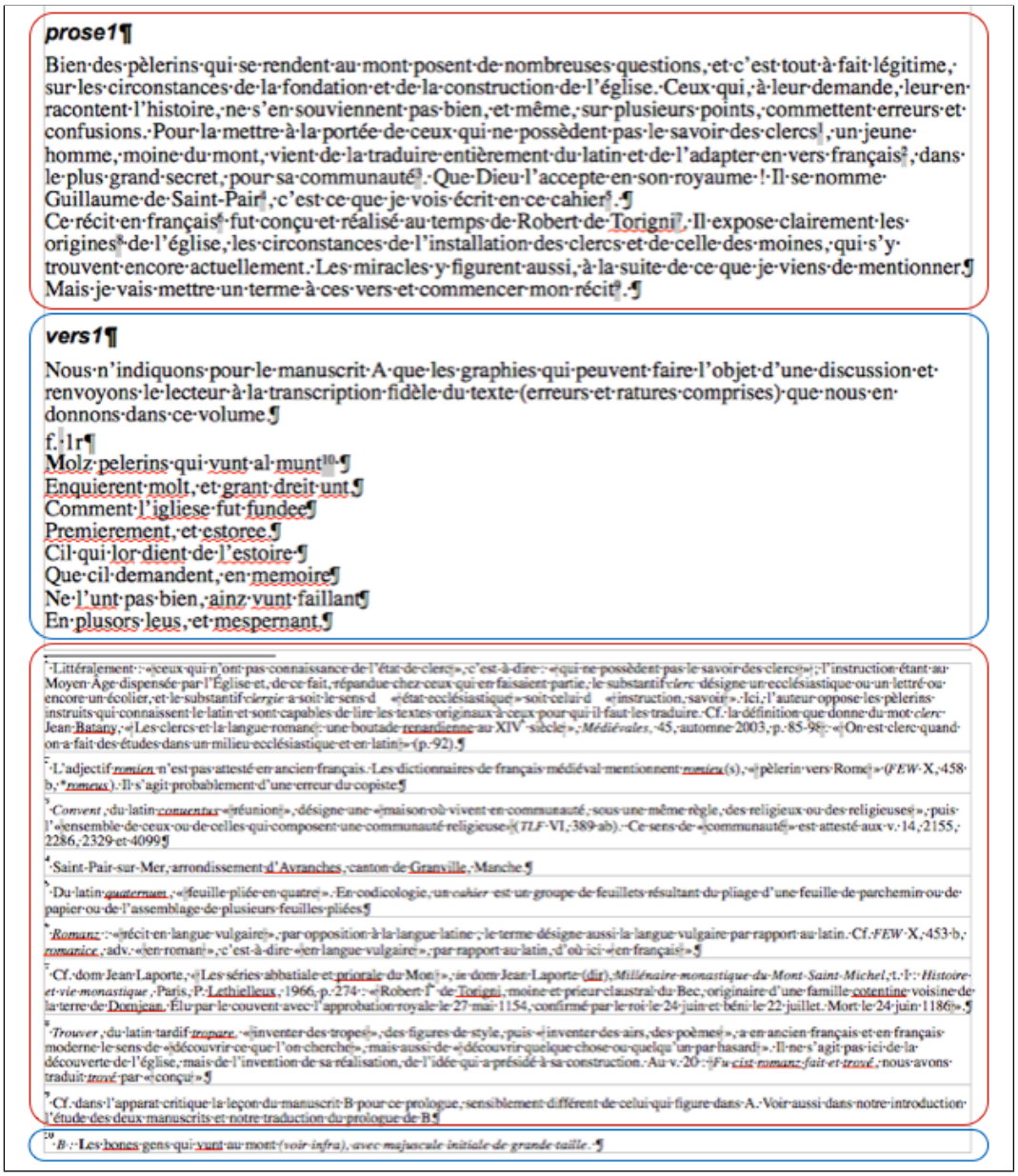

FIGURE 10.18 – Insertion d'un nouveau type de fragment dans le *Roman du Mont Saint-Michel*.

logiques d'une complexité similaire jusqu'au fragment indispensable pour construire l'ensemble des formes de diffusion.

Les figures 10.19 et 10.20 donnent des extraits des résultats obtenus au terme du traitement. Sur la figure 10.19, l'élément `lg`, ligne 5 est caractérisé par l'identifiant `xml:id=AFR.1.1` qui permet son alignement avec l'élément `p`, ligne 7 de la figure 10.20 identifié par le couple attribut/valeur `xml:id=FR.1.1`. Nous retrouvons

```
<?xml version="1.0" encoding="UTF-8"?>
[...]
<div type="chapitre">
  <div type="section" xml:id="AFR.1" xml:lang="afr">
    <lg xml:id="AFR.1.1">
      <l n="1" aid:pstyle="txt_Original_Vers" xml:id="vers1"
            xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">
               <pb ed="A"
            n="1r" />Molz pelerins qui vunt al munt<note type="
               marginal"
            xml:id="AFRftn1"> <emph aid:cstyle="typo_Italique">
                B  </emph>: Les
            bones gens qui vunt au mont <emph aid:cstyle="
               typo_Italique">(voir infra), avec majuscule
               initiale de grande taille.</emph>
            </note></l>
      <l n="2" aid:pstyle="txt_Original_Vers" xml:id="vers2"
            xmlns:aid="http://ns.adobe.com/AdobeInDesign/4.0/">
               Enquierent molt, et grant dreit unt,</l>
       [...]
   </lg>
[...]
```

FIGURE 10.19 – Extrait de code XML TEI d'un fragment du texte édité du *Roman du Mont Saint-Michel*.

```
<?xml version="1.0" encoding="UTF-8"?>
[...]
<div type="chapitre">
  <head>Livre I <num>vers 1 à 1378</num></head>
    <div type="section" xml:id="FR.1" xml:lang="fr">
      <head>
        <num>v. 1 -28</num>Introduction, l'auteur et son œuvre
      </head>
        <p xml:id="FR.1.1">Bien des pèlerins qui se rendent au
           mont posent de nombreuses questions, et  c est  tout
            à fait légitime, sur les circonstances de la
           fondation et de la construction de l'église. Ceux
           qui, à leur demande, leur en racontent l'histoire,
           ne  s en  souviennent pas bien, et même, sur
           plusieurs points, commettent erreurs et confusions.
           Pour la mettre à la portée de ceux qui ne possèdent
           [...]
        </p>
[...]
```

FIGURE 10.20 – Extrait de code XML TEI d'un fragment de la traduction du *Roman du Mont Saint-Michel*.

ici les règles d'identification présentées plus haut qui assurent la mise en place des solutions d'alignement et de mise en correspondance de fragments d'un flux à l'autre.

En définitive, il est indispensable de fixer les principes éditoriaux pour être mesure de mettre en place des flux de texte permettant la production des formes de diffusion. Cependant, la souplesse de manipulation favorise aussi l'expérimentation : il est toujours possible de revenir à une étape antérieure pour ajouter un nouveau type de fragment logique de manière à répondre au mieux aux objectifs éditoriaux.

La figure 10.21 donne une vue de l'édition en ligne. La page web présentée est produite à partir des deux flux structurés selon les principes que nous venons de présenter. Nous pouvons clairement voir la manière dont les fragments de groupes de vers (`lg`) permettent de bâtir l'alignement avec les paragraphes de prose.

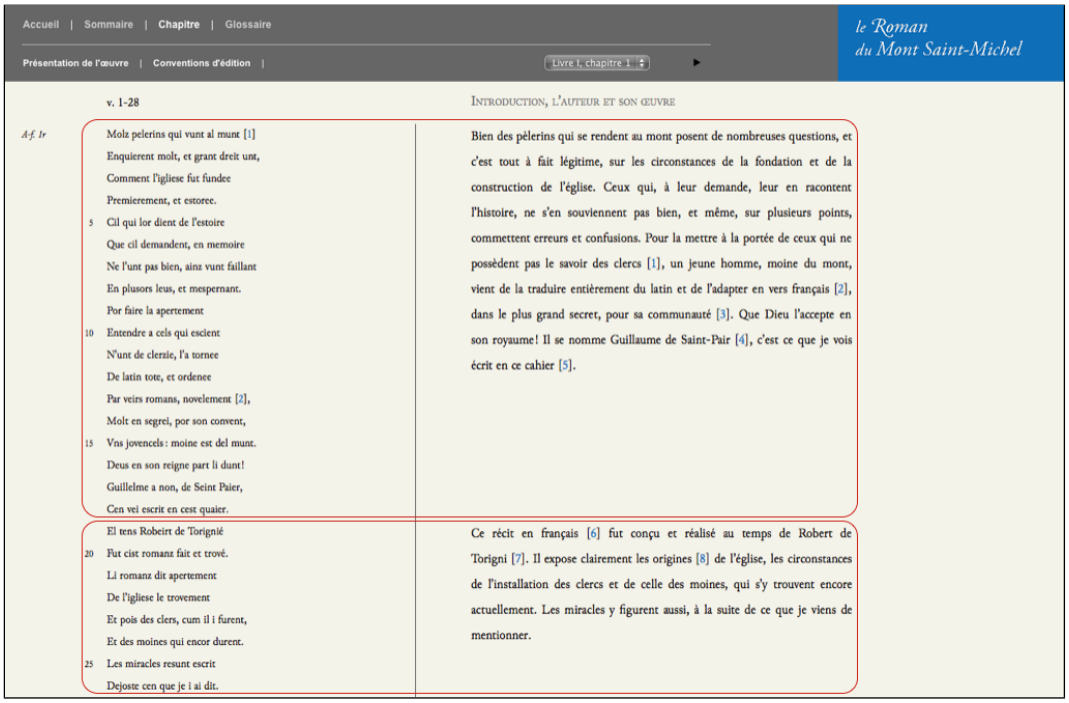

FIGURE 10.21 – Interface de lecture de l'édition en ligne bilingue du *Roman du Mont Saint-Michel*.

10.4.3 Glossaire et flux

Le glossaire, que nous avons déjà évoqué un peu plus haut, regroupe un ensemble de termes rencontrés dans le *Roman du Mont Saint-Michel* présentés par ordre alphabétique et accompagnés de la liste des numéros de vers où le terme concerné apparaît. Il donne les écarts de formes ainsi que leur traduction et des précisions grammaticales. Enfin, pour chaque forme le glossaire donne la liste des numéros de vers auxquels le lecteur pourra se reporter. La figure 10.22 présente la version en ligne du glossaire

dans laquelle chaque numéro de vers est un lien permettant de revenir au texte pour consulter les termes dans leurs contextes.

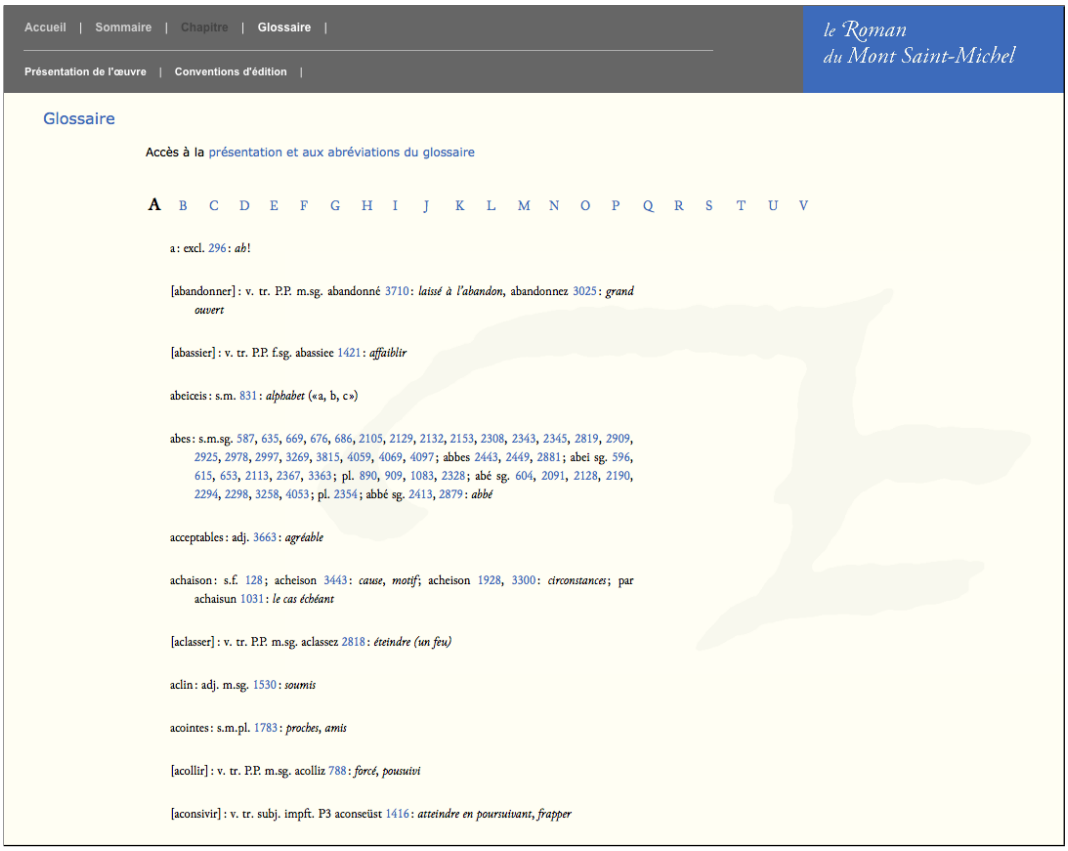

FIGURE 10.22 – Le glossaire du *Roman du Mont Saint-Michel* dans l'édition en ligne.

La gestion du glossaire ne présente pas de différence radicale sur le fond avec celle que nous avons présentée pour les textes édités et les traductions. Il s'agit d'un flux d'information particulier mais qui est géré de la même manière. La première étape pour l'éditeur matériel consiste donc à styler le document remis par l'auteur puis à produire, à partir de cette phase initiale d'étiquetage, une version encodée en XML.

Mais le glossaire présente une particularité qui le distingue des autres types de textes traités jusqu'à maintenant et qu'il s'agit de prendre en compte : il est, par nature, un outil d'accès spécifique aux textes édités qui repose en grande partie sur un système de renvois. Il s'agit bien évidemment d'exploiter ce système de renvois pour le rendre dynamique sur tous les supports qui le permettent, c'est-à-dire sur la version en ligne et sur la version cédérom.

Les différences techniques sont trop grandes entre ces deux supports pour permettre une seule méthode de traitement pour l'activation des liens du glossaire, même si, en réalité, les principes sont très voisins. En effet, dans la version en ligne, il s'agit

de pointer vers une zone, identifiée par le numéro de vers, d'une page construite à la volée tandis que, dans la version cédérom, il s'agit de renvoyer le lecteur à une page numérique d'un fichier PDF.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  [...]
3  <div type="lettre-glossaire">
4    <head>A</head>
5    [...]
6    <p rend="mot-glossaire">a l e i r : v. intr. et pron. 1541,
       3207, 3920; subj. prést P3 aut 1331, 3711, P5 (pron.)
       augiez 1896, P6 augent 1881, 3279, 3656, 3827, algent 306;
        P.P. m.sg. alei 670, 711, 130, m.pl. 571, 613, 950, 3355;
        m.pl. alé 278, 548, 562, 1619, 1922, 2031, 2337, 3257; f.
       sg. alee 999, 3029, 3205, 3205, 3353, 3484, 3922, aleie
       3191: <hi rend="italic">aller</hi>
7    </p>
8    <p rend="mot-glossaire">aleuz: s.m.pl. 1672   : <hi rend="
       italic">domaines</hi>
9    </p>
10   [...]
11 </div>
12 [...]
```

FIGURE 10.23 – Extrait de code XML du glossaire du *Roman du Mont Saint-Michel* après export depuis le logiciel de traitement de texte.

La version en ligne du glossaire est produite à partir du flux XML exporté depuis la version stylée du document de traitement de texte de l'auteur. La figure 10.23 présente un extrait du code XML obtenu à l'issue de cette exportation. La structure est simple avec une division, `div` pour chaque lettre et un élément `p` pour chaque terme contenant également la liste des numéros de vers où se reporter pour consulter le contexte d'occurrence.

Il s'agit d'enrichir cette version pour ajouter des pointeurs pour l'ensemble des vers rencontrés dans le glossaire. Cet enrichissement est réalisé au moyen d'une expression régulière qui repère chaque numéro de vers, c'est-à-dire une suite composée exclusivement de 1 à 4 chiffres sans aucun autre caractère, et ajoute l'élément XML approprié `ref` caractérisé par l'attribut `target` contenant la destination, donc le numéro du vers. Avec cette méthode, les liens vont pointer vers chaque vers à condition que l'ensemble de ces derniers soient identifiés de manière rationnelle comme on peut le voir à la ligne 12 de la figure 10.19, p. 210. La figure 10.24 présente le résultat de l'opération d'enrichissement. Chaque numéro de vers est maintenant encapsulé dans

un élément **ref** avec l'attribut **target** contenant le pointeur vers le vers correspondant.

```
<?xml version="1.0" encoding="UTF-8"?>
[...]
<div type="lettre-glossaire"><head>A</head>
  [...]
  <p rend="mot-glossaire">aleir : v. intr. et pron. <ref
     target="#vers1541">1541</ref>, <ref target="#vers3207">
     3207</ref>, <ref target="#vers3920">3920</ref>   ; subj.
     prést P3 aut <ref target="#vers1331">1331</ref>, <ref
     target="#vers3711">3711</ref>, P5 (pron.) augiez <ref
     target="#vers1896">1896</ref>, P6 augent <ref target="#
     vers1881">1881</ref>, <ref target="#vers3279">3279</ref>,
     <ref target="#vers3656">3656</ref>, <ref target="#vers3827
     ">3827</ref>, algent <ref target="#vers306">306</ref>   ;
     P.P. m.sg. alei <ref target="#vers670">670</ref>, <ref
     target="#vers711">711</ref>, <ref target="#vers130">130</
     ref>, m.pl. <ref target="#vers571">571</ref>, <ref target=
     "#vers613">613</ref>, [...] aleie <ref target="#vers3191">
     3191</ref>   : <hi rend="italic">aller</hi></p>
  <p rend="mot-glossaire">aleuz : s.m.pl. <ref target="#
     vers1672">1672</ref>   : <hi rend="italic">domaines</hi></
     p>
  [...]
</div>
[...]
```

FIGURE 10.24 – Extrait de code XML du glossaire du *Roman du Mont Saint-Michel* après enrichissement.

La version cédérom est produite intégralement avec Indesign dans lequel est importée la version XML du glossaire exportée depuis le logiciel de traitement de texte. Rappelons ici que l'ensemble des textes proposés sur le cédérom sont des fichiers PDF produits avec Indesign. Une fois les documents Indesign créés, le principe est très similaire à celui appliqué pour la version en ligne : il s'agit de poser des ancres sur chaque vers, car ils sont tous susceptibles de faire l'objet d'un lien depuis le sommaire. Ensuite, un parcours du glossaire à la recherche de toutes les mentions de numéro de vers, avec exactement les mêmes critères de recherche que pour la version en ligne, pour ajouter à chaque occurrence, un lien vers l'ancre correspondante.

Dans tous les cas, on retrouve l'importance de disposer d'un système d'identification rationnel des éléments constitutifs du texte. Avec un système aléatoire, ce type de procédures serait totalement impossible à mettre en place : il faudrait systématiquement récupérer l'identifiant sur le vers concerné ce qui serait beaucoup

moins rapide. Construire un système rationnel d'identification permet de simplifier considérablement les exploitations.

La création des liens permettant de circuler du glossaire vers le texte édité, pour la lecture des termes dans leurs contextes, se règle donc selon un principe d'exploitation de la structure versifiée du *Roman du Mont Saint-Michel*. Cependant, il faut aussi que le lecteur soit en mesure de revenir au glossaire, au terme précis qu'il était en train d'examiner et la question de ce retour se pose aussi bien dans le cas de l'édition en ligne que dans celui du cédérom.

Dans le cas de l'édition en ligne, il faut permettre au lecteur de revenir au glossaire au mot précis où il se trouvait avant de venir consulter ce terme dans son contexte. Pour cela, un lien de retour est ajouté à la page web produite. Du point de vue du modèle, il s'agit d'intervenir sur la structure physique asynchrone pour ajouter des informations qui ne sont pas directement incluses dans les textes pour simplifier les manipulations et donc l'activité de lecture du texte. La figure 10.25 donne une vue du résultat de l'insertion du lien de retour au glossaire ainsi que de la mise en surbrillance du vers contenant le terme concerné.

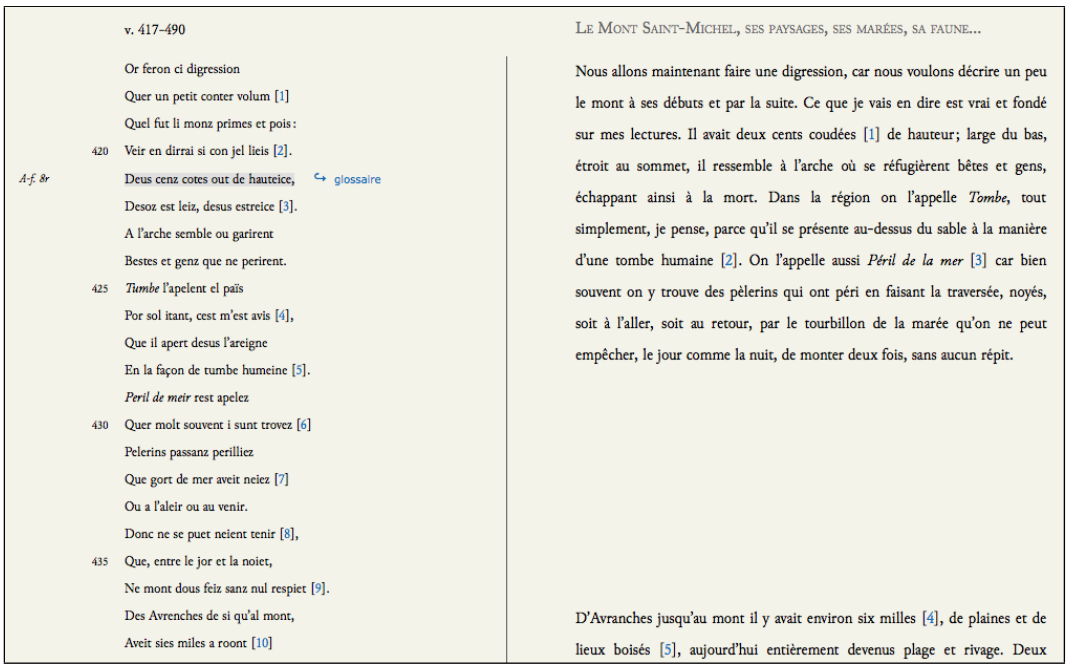

FIGURE 10.25 – Lien de retour au glossaire dans l'édition en ligne du *Roman du Mont Saint-Michel*.

Dans le cas du cédérom, dans la mesure où il est impossible d'intervenir sur le contenu des fichiers affichés, la solution retenue consiste simplement à s'appuyer sur

les systèmes de circulation internes des lecteurs PDF en activant un lien de retour générique vers la vue précédente directement depuis l'interface du logiciel.

10.4.4 Rapports texte/image

Idéalement, et comme pour les *Chroniques latines*, les images numériques des deux manuscrits du *Roman du Mont Saint-Michel* devraient être consultables en haute résolution depuis l'édition en ligne. Mais, comme nous l'avons déjà signalé, les deux manuscrits sont conservés à la *British Library* qui n'accorde, ou qui n'accordait en 2009, que des droits d'exploitation pour un temps donné. Il était impossible pour les Presses universitaires de Caen de payer les loyers demandés. Les images ne sont donc pas consultables dans l'édition en ligne du *Roman du Mont Saint-Michel* comme c'est le cas pour les manuscrits principaux des *Chroniques latines*.

Cependant, l'ensemble des flux de texte est prêt pour ce type d'exploitation et la préparation des liens est réalisée de la même manière que pour les *Chroniques latines*, avec un élément `pb`, mais l'absence des images interdit toute autre exploitation que le simple signalement. Ainsi, chaque rupture de folio dans l'un ou l'autre des manuscrits est balisée dans le flux XML TEI du texte édité en ancien français, mais ces éléments ne sont tout simplement pas activés dans l'édition en ligne et permettent juste de signaler les changements de folio au lecteur pendant sa lecture.

Nous sommes ici confrontés à l'une des limitations importantes concernant l'exploitation des nombreuses initiatives de numérisation et de constitution de bibliothèques numériques : l'accès aux images. En effet, à l'heure actuelle, chaque initiative en la matière conduit à la mise en place d'une nouvelle interface à consulter sans possibilité de croiser les recherches et de comparer les résultats dans une interface unique. Il existe cependant des solutions qui sont en cours de développement et d'expérimentation. Ainsi, le pool de l'équipement d'excellence Biblissima [133] réalise de nombreuses expérimentations avec le modèle *shared canvas* que nous avons déjà évoqué plus haut [134].

Le principe général de *shared canvas* consiste à définir des *canevas* qui décrivent des zones auxquelles sont affectés des contenus iconographiques ou textuels. Le canevas ainsi mis en place agit donc comme un système de référence pour l'ensemble des informations et permet de les faire interagir.

133. Pour plus de précisions, voir les pages http://doc.biblissima-condorcet.fr/entrepots-dimages et http://doc.biblissima-condorcet.fr/visualiseur-mirador.

134. Pour des précisions du modèle *Shared Canvas*, voir p. 84 et http://www.shared-canvas.org.

Du point de vue de l'interopérabilité des images, il est indispensable de construire une architecture logicielle proposant un serveur capable de fournir les informations respectant le modèle de données *shared canvas* et un visualiseur capable de les interpréter.

L'un des avantages les plus significatifs de cette architecture et de ce modèle de données réside dans le fait que les images restent stockées uniquement sur les serveurs de l'institution de conservation qui les expose avec son propre serveur. Les visualiseurs se connectent à ces serveurs pour permettre l'affichage. Ainsi, un même visualiseur peut afficher des images stockées sur plusieurs serveurs d'images différents.

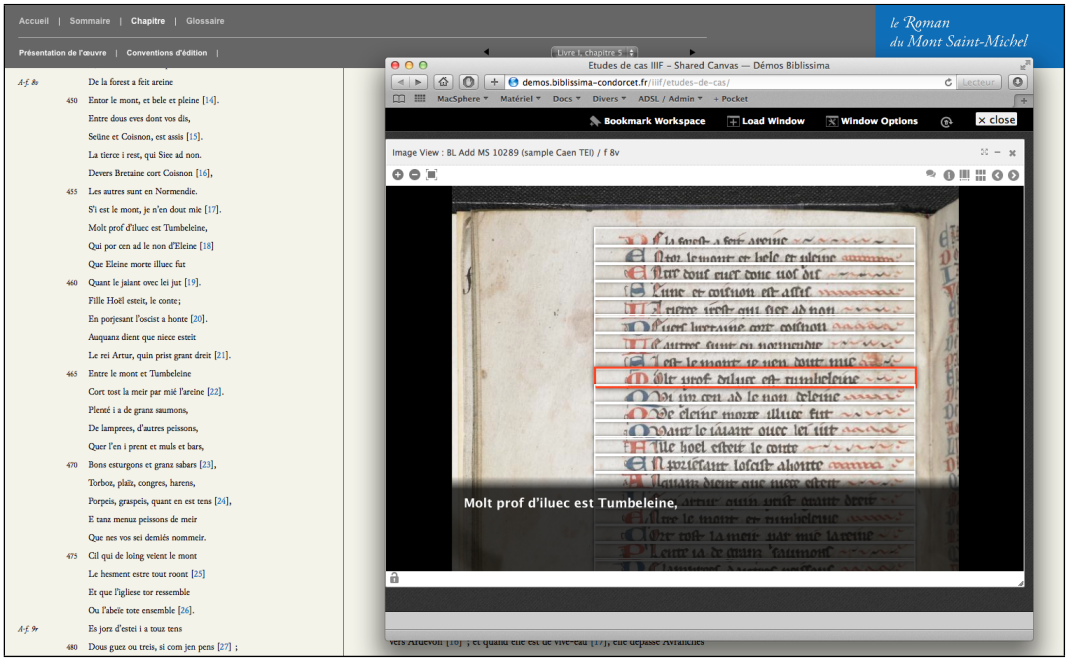

FIGURE 10.26 – Rapports texte/image dans le *Roman du Mont Saint-Michel*.

De plus, le modèle de données permet d'associer tout type d'annotation à une zone du canevas, et, en l'occurrence pour ce qui nous intéresse ici, la transcription d'une zone. Le pool Biblissima, à partir du texte édité en XML TEI, a produit les *manifests* contenant les annotations de transcriptions de chaque vers, avec les coordonnées des zones, pour certains folios du manuscrit Add MS 10289 conservé à la *British Library*[135].

La figure 10.26 propose une capture d'écran du résultat obtenu. Un clic sur les zones encadrées en blanc dans la fenêtre au premier plan permet d'afficher la trans-

135. Le résultat de ce travail du pool Biblissima est consultable en ligne à cette adresse : http://demos.biblissima-condorcet.fr/iiif/etudes-de-cas/.

cription issue du travail d'édition du *Roman du Mont Saint-Michel*, dans la partie basse de cette même fenêtre.

10.5 Apports et influence sur le modèle

L'expérience du travail sur les textes des *Chroniques latines du Mont Saint-Michel* a permis de valider la possibilité de manipuler les textes portés par les sources primaires comme des flux bornés et fragmentés. L'apport principal du travail sur l'édition du *Roman du Mont Saint-Michel* réside dans le fait qu'il a permis de mettre en évidence les qualités de souplesse et d'adaptabilité du modèle. Ainsi, il est tout à fait possible de définir, si c'est nécessaire, un niveau de fragment, comme nous l'avons vu avec l'ajout des groupes de vers pour l'alignement du texte édité en ancien français et de sa traduction en prose dans la version en ligne. Ainsi, si l'enrichissement des flux est nécessaire pour une exploitation donnée, il est tout à fait possible de le réaliser sans que cela perturbe les autres exploitations.

Ces enrichissements peuvent être réalisés bien après la mise en place initiale des flux structurés comme nous l'avons vu avec les expérimentations réalisées avec *shared canvas* dans le cadre de l'équipement d'excellence Biblissima. La constitution de flux de texte encodés dans le respect des standards internationaux assure une évolution et une intégration efficace des nouvelles solutions d'exploitation et des nouveaux modèles de données.

Enfin, l'expérimentation de l'édition du *Roman du Mont Saint-Michel* apporte aussi des certitudes sur la capacité à gérer les outils d'accès aux textes comme les glossaires en simplifiant aussi considérablement les opérations de traitement.

11

L'*Hortus Sanitatis*

Comme nous l'avons vu avec les sources du Mont Saint-Michel, le modèle de flux et de fragments que nous proposons a fait ses preuves dans le cadre d'un niveau de balisage éditorial. Il reste cependant à savoir si ce modèle de flux fonctionne avec un niveau de structuration plus fin et des systèmes d'annotation plus avancés et plus complexes.

Les bases fixées plus haut restent tout à fait valides ici et nous allons examiner la manière dont le modèle se comporte, ou doit évoluer, quand le niveau de description est plus fin. Il s'agit en réalité d'étudier l'influence de la multiplication des flux d'information et de l'augmentation de la finesse des fragments sur les possibilités d'exploitation tant du point de vue de la recherche que des solutions de production des supports de diffusion.

Pour valider le modèle et le mettre à l'épreuve sur un projet plus complexe dans lequel les chercheurs se sont investis dans le travail de balisage, le cas de l'*Hortus Sanitatis* est idéal. En effet, la complexité des textes traités et l'investissement de l'équipe de chercheurs en font un terrain d'expérimentation parfaitement adapté.

11.1 Présentation des sources

L'*Hortus Sanitatis* [Jacquemard *et al.*, 2013] est une édition d'un traité latin d'ichtyologie du XV^e^ siècle qui a été plusieurs fois réédité et traduit dont des exemplaires sont conservés dans plusieurs bibliothèques[136]. Il se caractérise en particulier par le fait qu'il s'agit en réalité d'une compilation de sources avec très peu de passages originaux. On trouve ainsi dans l'*Hortus Sanitatis* des passages de Vincent de Beau-

136. Voir la rubrique "Éditions et catalogues" de la bibliographie sur l'édition en ligne de l'*Hortus Sanitatis* : http://www.unicaen.fr/puc/sources/depiscibus/bibliographie.

vais, Pline l'ancien, Thomas de Cantimpré, Isidore de Séville et d'autres. Le projet a consisté pour une large part à identifier des sources de l'*Hortus Sanitatis* ainsi qu'à reconstituer de l'histoire de ces textes.

11.2 Flux et fragments dans l'*Hortus Sanitatis*

L'*Hortus Sanitatis* est donc une compilation de sources, c'est-à-dire, du point de vue de notre modèle, un ensemble constitué de plusieurs flux de texte entrecroisés, repris et transformés, les uns avec les autres : un flux de texte d'Aristote, un flux de Thomas de Cantimpré, un flux d'Avicenne, etc. Chacun de ces flux se compose de fragments pouvant ou non proposer une cohérence autonome.

La question centrale dans le projet d'édition de l'*Hortus Sanitatis* est alors de déterminer par où commencer le travail d'édition. En effet, dans la mesure où très peu de passages sont originaux, est-il nécessaire pour rendre compte de l'établissement et de l'histoire du texte, d'éditer l'ensemble des sources de l'*Hortus Sanitatis* ? Bien entendu, devant l'ampleur de la tâche, cette dernière option est intenable ; il est donc indispensable de trouver une autre solution, plus réaliste sur le plan organisationnel.

Là encore, examiner le problème sous l'angle de la fluidité des textes permet d'établir une position qui, si elle ne règle pas le problème au sens strict, permet en quelque sorte d'en annuler la portée. En effet, si l'on considère l'*Hortus Sanitatis* et l'ensemble de ses sources comme un réseau de textes connectés les uns aux autres, il s'agit simplement de choisir un point d'entrée pour débuter le travail sur l'ensemble du réseau et non plus de trouver une solution pour traiter, dans un seul et même mouvement, l'intégralité de l'énorme masse textuelle que l'*Hortus Sanitatis* et ses sources constituent.

Ainsi, l'approche du flux de texte qu'est l'*Hortus Sanitatis* permet de le considérer en tant que tel, les exemplaires sont bien réels, et pour chaque fragment identifié par les chercheurs, il sera possible d'ajouter les informations de provenance, sur autant de niveaux que nécessaires sur le plan scientifique pour retracer l'histoire du texte discriminé. Il s'agit donc de traiter le texte de l'*Hortus Sanitatis* comme un flux de texte composé de fragments disposant tous de leur propre histoire qu'il s'agit de retracer. Ce qui, à vrai dire, est exactement la réalité textuelle à laquelle les chercheurs sont confrontés.

La complexité du flux de texte de l'*Hortus Sanitatis* est donc au départ totalement liée à la source matérielle étudiée par les chercheurs. On retrouve ainsi les niveaux

de fragments déjà manipulés dans les sources du Mont Saint-Michel : les chapitres et les paragraphes.

```
<?xml version="1.0" encoding="UTF-8"?>
[...]
<div type="section1" xml:id="LAT.hs.4.16.section" xml:lang="la">
  <figure>
    <graphic url="../IMG/hr/16.tif"/>
    <head>Cancer (Épernay, BM, Inc. 3017)</head>
  </figure>
  <p rend="renvoi" aid:pstyle="txt_Renvoi_Interne" xml:id="LAT.hs.4.16.ri"/>
  <p rend="Lieux" aid:pstyle="txt_Lieux_Parallele" xml:id="LAT.hs.4.16.lp"/>
  <p n="1" aid:pstyle="txt_Normal" xml:id="LAT.hs.4.16.1">
    [...]
  </p>
  <p n="2" aid:pstyle="txt_Normal" xml:id="LAT.hs.4.16.2">
    <bibl subtype="paragraphe" type="source">
      <author>VB</author>
      <biblScope>17, 37, 1</biblScope>
    </bibl>
    <index indexName="marqueurCitation">
      <term type="orig">Aristoteles</term>
      <term type="norm">Arist.</term>
      <term type="identifie">Arist.</term>
    </index>.
    <seg type="citation">
      <bibl subtype="fragment" type="source">
        <author>Arist.</author>
        <title type="oeuvre">HA</title>
        <biblScope>490 b 6 MS</biblScope>
        <note type="sources" xml:id="LATnote252"><hi rend="italic">Cancer vero habet octo pedes, per quos movetur motu equali</hi>.</note>
      </bibl>
      <index indexName="piscibus">
        <term>cancer</term>
      </index>
      Cancer habet octo pedes per quos<note type="apparat" xml:id="LATnote253">quas <hi rend="italic">1491 Prüss<hi rend="sup">1</hi> 1536.</hi></note> movetur aequali motu. Caret sanguine.
    </seg>
  [...]
  </p>
[...]
```

FIGURE 11.1 – Extrait du code XML du texte latin de l'*Hortus Sanitatis*.

Mais l'*Hortus Sanitatis* a fait l'objet d'un travail supplémentaire de la part des chercheurs. En effet, chaque paragraphe provient d'une source primaire : Vincent de Beauvais. Ce premier niveau d'identification fait l'objet d'un marquage spécifique au niveau du fragment de type paragraphe. La figure 11.1 présente un extrait du code XML TEI du texte latin de l'*Hortus Sanitatis*. Les lignes 14 à 17 offrent un exemple du balisage utilisé pour marquer la source du paragraphe en reprenant sa référence absolue, c'est-à-dire en l'occurrence `VB 17, 37, 1`. L'attribut `subtype` de l'élément `bibl` renseigne sur la portée de la référence dans le texte, c'est-à-dire en réalité, sur le fragment concerné par la référence. De cette manière, lorsque le texte de Vincent de Beauvais sera à son tour édité, il sera très aisé d'activer un lien dans l'édition en ligne depuis cette référence bibliographique absolue, par exemple en ajoutant une simple transformation.

De plus, à l'intérieur de chaque paragraphe, chacun des fragments de citation est borné par l'élément `seg`, et lui aussi identifié par une référence bibliographique absolue selon un schéma de structuration tout à fait similaire bien entendu. Les lignes 24 à 29 de la figure 11.1 donnent un exemple d'une telle identification. La portée de cette identification concerne le fragment, comme l'indique l'attribut `subtype` de l'élément `bibl` de la ligne 24. Le fragment en question est donc celui qui contient la référence, qui débute à la ligne 23 et se termine à la ligne 34. Les éléments `author`, `title` et `biblScope` donnent les précisions sur la référence. De plus, les chercheurs donnent le texte original de la source secondaire, ici Aristote, dans l'élément `note`, ligne 28, à l'échelle de l'ensemble du texte du segment.

Ce dernier point est nécessaire pour permettre au lecteur de lire le texte d'origine et d'évaluer le degré de modification subi par le fragment sans être contraint de disposer du texte d'Aristote. Il s'agit d'une facilité proposée aux lecteurs. On pourrait envisager, une fois le flux de texte d'Aristote constitué, de réaliser des extractions pour convoquer les fragments concernés et alimenter le flux de l'*Hortus Sanitatis*.

Ainsi, si la définition des fragments constitutifs du flux de l'*Hortus Sanitatis* repose pour commencer sur les exemplaires pour les chapitres et les paragraphes, elle s'appuie à partir des segments sur l'histoire des textes et se détache par là même, dans une certaine mesure, de la matérialité de la source primaire.

La figure 11.2 donne un exemple d'exploitation de cette profondeur de fragmentation des flux de texte latin et de la traduction en français. Les deux flux faisant l'objet du même niveau de découpage fragmentaire, avec le même système d'identi-

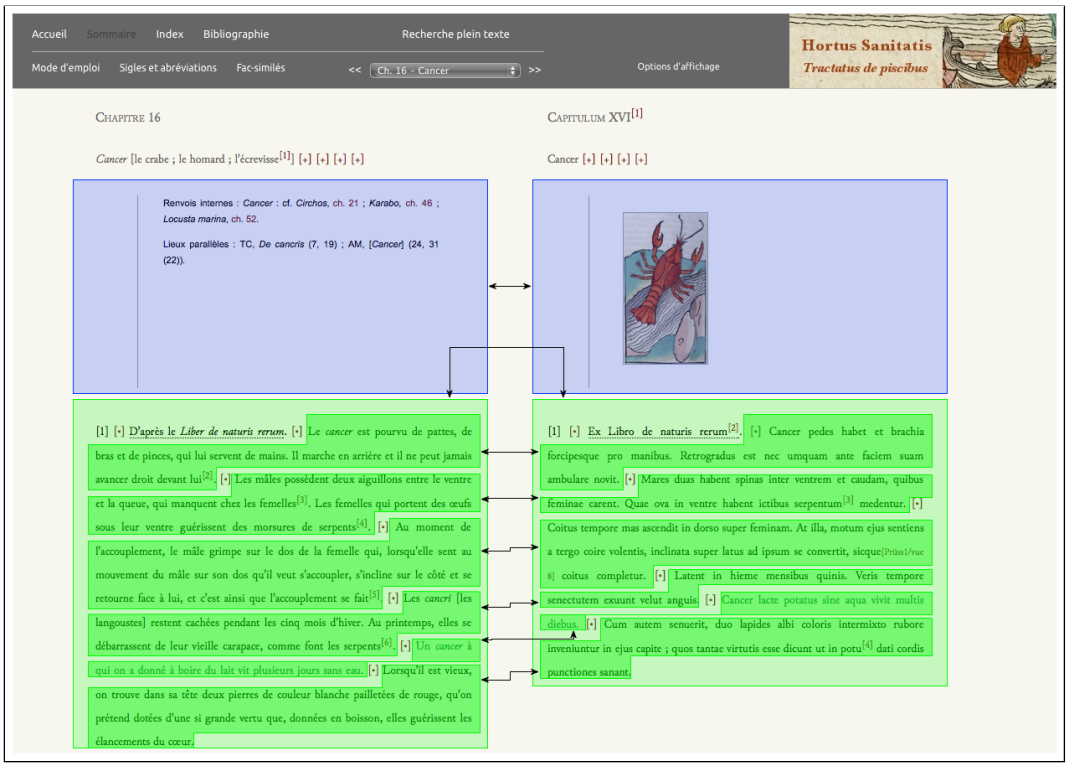

FIGURE 11.2 – Vue des niveaux de fragments dans la version en ligne de l'*Hortus Sanitatis*.

fication décrit pour les sources du Mont Saint-Michel [137], il est possible d'aligner les segments d'identification des sources secondaires les uns avec les autres. Dans l'édition en ligne de l'*Hortus Sanitatis*, une solution de mise en surbrillance a été choisie pour permettre au lecteur de profiter de cet alignement. Ainsi, lorsque le lecteur passe le pointeur sur un segment identifié, sur le latin ou sur le français, la couleur des deux segments en correspondance change. Sur la figure 11.2 les paragraphes et les segments internes sont mis en évidence avec leurs correspondances dans les deux flux de texte. Rappelons que les fragments de niveau "paragraphe" servent de base et constituent la brique centrale de l'alignement vertical des deux langues lors de la construction dynamique de la page web.

11.3 Une édition multimodale

Le projet d'édition vise à donner au lecteur la possibilité d'accéder au texte en latin et en français ainsi qu'à l'histoire de ce texte dans les conditions les plus efficaces possible. Il s'agit bien de fournir au lecteur des textes directement utilisables et in-

137. Voir p. 176 et suivantes.

telligibles avec l'ensemble des informations nécessaires à une bonne compréhension. Ainsi chaque fragment de source ou segment de citation est enrichi des informations d'identification retraçant son histoire. Comme pour les éditions déjà décrites précédemment, les deux supports de diffusion, papier et web, sont alimentés à partir d'un seul flux de données encodées en XML TEI.

L'*Hortus Sanitatis* présente toutefois la particularité de proposer un grand nombre de solutions d'accès aux textes quel que soit le support considéré. Il s'agit d'une édition multimodale au sens plein du terme puisque le lecteur peut accéder au texte par le sommaire, en suivant l'organisation des témoins, par œuvre ou par auteur secondaires, par les index des noms de poissons (latins ou français), par le moteur de recherche, etc.

11.3.1 La version papier

La version papier entre, comme les sources du Mont Saint-Michel, dans les canons habituels de l'édition critique de textes anciens avec notes d'apparats critiques, commentaires philologiques, etc.

La logique de manipulation de flux de données structurées dans le but de produire un volume papier de l'édition de l'*Hortus Sanitatis* est similaire à celle que nous avons déjà présentée pour les sources du Mont Saint-Michel. Cependant, le niveau de balisage scientifique mis en place par les chercheurs contraint à certaines adaptations : il est indispensable de déterminer les fragments qui feront l'objet d'une mise en forme graphique dans le volume papier produit.

Le même flux de texte doit permettre d'alimenter une version en ligne et une version papier. À partir du flux scientifique contenant l'intégralité des informations mises en place et manipulées par les chercheurs, il s'agit de définir des flux éditoriaux qui permettront de construire les planches de l'édition papier.

La technique utilisée consiste à reconstituer des flux de textes en fonction de leur type. Cette solution de reconstitution permet d'évaluer les volumes de chacun des flux de texte (transcription ou traduction) et de commentaire, ce qui permet de travailler avec un maximum d'efficacité sur leur distribution dans la page.

La figure 11.3 donne une représentation de l'extraction des différents flux de texte et de leur répartition sur une page de l'édition papier. Il convient ici de distinguer pour commencer deux grands types de flux : le texte édité latin, avec un certain nombre d'éléments périphériques sur lesquels nous reviendrons plus loin, et les flux des différents systèmes de notes scientifiques. Ainsi, tout ce que les chercheurs ont ba-

lisé en utilisant l'élément `note` fait l'objet d'une extraction, les notes sont rassemblées dans un flux distinct, lui-même défini sur la base de la valeur affectée à l'attribut `type` de chacune d'elle [138]. Précisons qu'un marqueur est inséré dans le flux de texte principal pour signaler la présence de la note sous la forme d'un simple appel. Un exemple d'un tel marqueur est donné à la ligne 12 de la figure 11.4 avec l'élément `hi`. Ces appels serviront aussi bien au lecteur, de manière tout à fait traditionnelle, qu'à l'opérateur de PAO qui va s'appuyer sur ces marqueurs pour rythmer l'apparition des notes sur la planche en fonction de la présence des appels.

Cette distinction élémentaire entre flux de texte et flux de notes doit cependant être précisée car certaines informations qui relèvent pourtant de l'annotation sont conservées dans le flux de texte principal et ce, pour deux raisons.

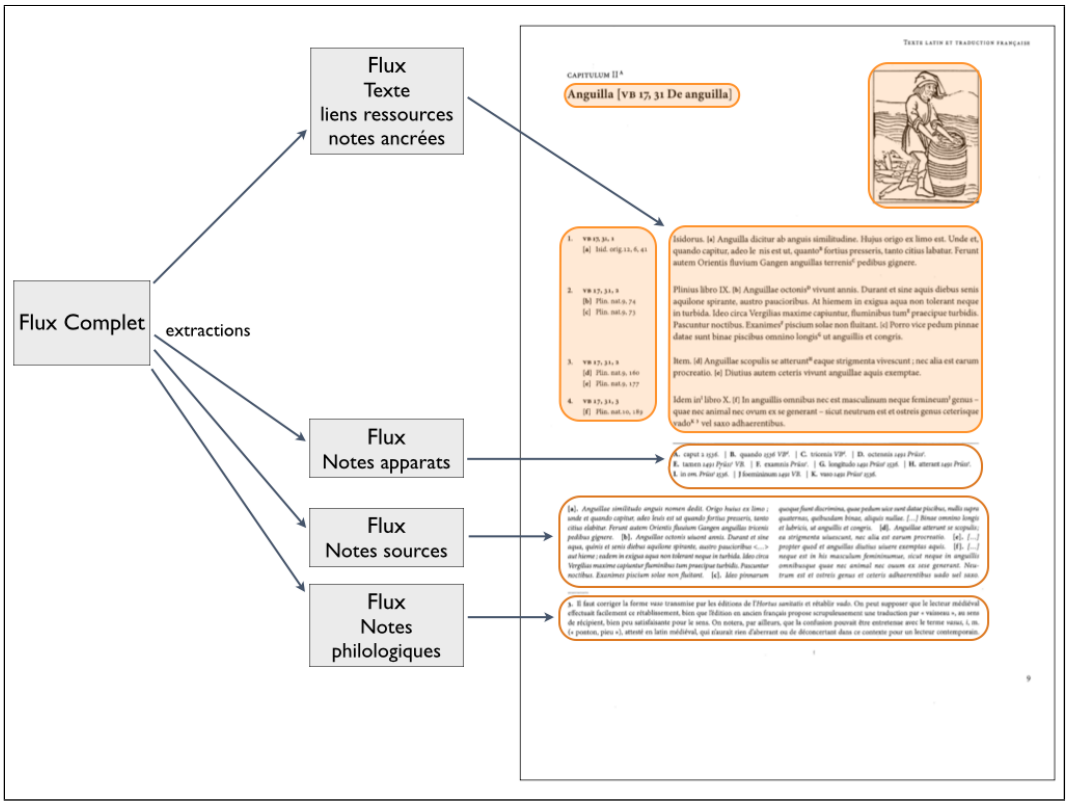

FIGURE 11.3 – Organisation des flux de textes dans l'*Hortus Sanitatis*.

La première explication est d'ordre tout à fait pragmatique. Les notes marginales font l'objet d'un traitement particulier. En effet, ce type de contenu, qu'il soit textuel ou iconographique, doit être placé relativement à son point de référence dans le texte principal. Pour définir les points d'ancrage, la meilleure solution est de conserver les informations dans le flux qui doit contenir ces points d'ancrage au moment de

138. Voir la ligne 28 de la figure 11.1, page 221.

l'importation dans Indesign et de réaliser les opérations d'extraction du flux principal et de placement dans le même mouvement. Notons que l'illustration, qui doit aussi être placée relativement au chapitre qu'elle concerne, fait l'objet d'un traitement équivalent, et d'un placement dans un bloc dédié, aux dimensions fixes, en haut à droite de la page et non en marge.

Mais au-delà de cette explication pragmatique, certaines informations restent aussi intégrées dans un des flux de texte principaux en fonction de leur nature. Ainsi, les informations d'identification des segments de texte doivent, bien entendu, apparaître le plus près possible du fragment qu'ils concernent. Ces informations sont donc laissées dans le flux de texte latin et feront l'objet du traitement spécifique décrit plus haut. En réalité, les identifications sont conservées dans le flux principal parce qu'il s'agit d'une information donnée au lecteur et non d'un commentaire scientifique au sens strict, l'explication scientifique de cette identification est, elle, donnée dans le flux des notes sources dont le placement est indiqué à la figure 11.3.

Par ailleurs, le nombre d'identifications [139] interdit de les traiter unité par unité sans imposer trop de temps de calcul et de manipulation au moment de l'ajustement de la mise en page. En plus de la reconstitution des flux, il s'agit donc d'opérer un certain nombre de regroupement d'éléments pour obtenir un grain de description plus facile à exploiter du point de vue éditorial. C'est le cas de l'ensemble des informations d'identification qui sont regroupées dans un élément unique à l'échelle du paragraphe.

Comme nous l'avons vu plus haut, les chercheurs ont identifié et balisé les sources de chacun des segments de texte qui composent l'*Hortus Sanitatis* et ajouté toutes les informations nécessaires à la compréhension de l'histoire de ces fragments. Il y a 677 marques d'identification pour l'ensemble des 106 chapitres. Rappelons que ces informations ne sont pas les seules à composer l'appareil scientifique produit par les chercheurs. Celui-ci se compose également des flux de commentaires philologiques, zoologiques, etc. Ce sont ces flux qui font l'objet des extractions pour permettre la mise en place de solutions de distribution harmonieuses et intelligibles de l'ensemble des flux de texte sur la planche du futur livre, mais les informations d'identification proprement dites sont restées dans le flux latin principal comme nous l'avons vu.

La figure 11.4 donne un extrait de code XML incluant le regroupement des informations d'identification. Cet exemple est extrait du nouveau flux de texte latin, dépouillé des systèmes de notes et avec un niveau de balisage différent, et qui corres-

139. Rappelons que l'identification est réalisée au niveau du paragraphe pour la citation primaire et au niveau des segments pour les citations secondaires.

```
<?xml version="1.0" encoding="UTF-8"?>
[...]
<head aid:pstyle="LAT_T_Chapitre"><index indexName="piscibus"><
   term>cancer</term></index>Cancer</head>
<bibl type="source_chap" aid:pstyle="LAT_txt_SO"><author aid:
   cstyle="typo_SC">VB </author><biblScope>17, 36 </biblScope><
   title>De cancro </title><author aid:cstyle="typo_SC">VB </
   author><biblScope>17, 37 </biblScope><title>De eodem </title
   ><author aid:cstyle="typo_SC">VB </author><biblScope>17, 38
   </biblScope><title>De operatione cancrorum in cibo, vel
   medicina </title><author aid:cstyle="typo_SC">VB </author><
   biblScope>17, 39 </biblScope><title>De eodem </title></bibl>
<div type="section1" id="LAT.hs.4.16.section" xml:lang="la"><
   graphic href="../IMG/hr/16.tif"/>
[...]
<p aid:pstyle="LAT_txt_Normal">
  <note type="source_para" aid:pstyle="LAT_renvoi_SO_1"><c aid:
     cstyle="reference_Paragraphe">2</c>.     <ab><author aid:
     cstyle="typo_SC">VB </author><biblScope>17, 37, 1 </
     biblScope></ab><c>  </c><ab aid:cstyle="reference_SO">[g]<
     /ab><c>   </c><author aid:cstyle="typo_SC">Arist. </author
     ><title>HA </title><biblScope>490 b 6 MS </biblScope><c>
             </c><ab aid:cstyle="reference_SO">[h]</ab><c>   </
     c><author aid:cstyle="typo_SC">Arist. </author><title>HA <
     /title><biblScope>523 b 5-8 MS </biblScope><c>      </c><
     ab aid:cstyle="reference_SO">[i]</ab><c>   </c><author aid
     :cstyle="typo_SC">Arist. </author><title>HA </title><
     biblScope>525 b 7-9 MS </biblScope>
  </note>
  <ab type="marqueurCitation"><term type="orig">Aristoteles</
     term></ab>.
  [...]
  <seg type="citation">
    <index indexName="piscibus">
      <term>cancer</term>
    </index>
    <hi type="source" aid:cstyle="appel_SO">g</hi>
    Cancer habet octo pedes per quos movetur aequali motu.
       Caret sanguine.
  </seg>
  <seg type="citation">
  <hi type="source" aid:cstyle="appel_SO">h</hi>Durum extra et
     intus molle nec est illud durum conteribile
  [...]
  </seg>
</p>
[...]
```

FIGURE 11.4 – Extrait de code XML du texte latin de l'*Hortus Sanitatis* préparé pour l'importation dans Indesign.

pond en définitive à un état de balisage éditorial. En réalité, il s'agit bien de produire à partir du flux encodé à un niveau scientifique, un flux avec un balisage éditorial plus simple à exploiter du texte principal en plus des flux de notes scientifiques. C'est l'élément `note` de la ligne 8 de la figure 11.4 qui doit ici retenir notre attention. L'ensemble des informations d'identification, dans le flux scientifique complet, est balisé en utilisant les éléments `bibl` pour l'ensemble de la source, `title` pour son titre, `author` pour son auteur et `biblScope` pour son étendue [140].

C'est à partir de cette structuration que les éléments `note` créés dans le flux éditorial sont alimentés en utilisant une transformation XSL. Celle-ci, en plus d'ajouter l'élément `note` correspondant au bloc marginal, remplace les éléments `bibl` par des éléments `ab` pour les sources de paragraphe et les supprime pour les segments dans la mesure où ils ne sont d'aucune utilité pour la mise en forme. Enfin cette transformation ajoute aussi des éléments `c` pour contenir des espaces typographiques nécessaires pour la mise en forme ainsi que des éléments `ab` pour baliser les appels de note concernant les sources. Ces appels sont insérés au début des segments concernés puisque leur portée s'étend à l'intégralité de leur texte.

L'élément `note` ainsi produit, et respectant les règles d'indentation pour une bonne interprétation par Indesign [141], contient l'intégralité des informations nécessaires à la création d'un bloc marginal qui fera l'objet d'attribution d'un style au moment de sa création dans Indesign pour automatiser son placement. En effet, un traitement lancé directement dans Indesign se chargera de repérer toutes les notes de type `source_para`, ligne 8 de la figure 11.4, et assurera la création d'un bloc préalablement défini dans la maquette Indesign disposant de l'ensemble des propriétés graphiques adaptées.

Rappelons pour terminer que, comme dans le cas des sources du Mont Saint-Michel, la synchronisation des flux les uns par rapport aux autres sur la planche de travail se fait manuellement sous Indesign même si le travail est préparé en intégrant dans les flux les points de repères qui permettent de simplifier ce travail au maximum.

11.3.2 La version web

L'architecture logicielle est très similaire à celle utilisée pour les sources du Mont Saint-Michel à une différence majeure près. Pour l'édition en ligne de l'*Hortus Sanitatis*, les fichiers XML TEI ne sont plus stockés sur le système de fichiers du serveur

140. Voir les lignes 24 à 27 de la figure 11.1, p. 221.

141. Ces règles, détaillées p. 120 et suivantes, sont impossibles à reproduire dans les exemples de code XML que nous donnons.

mais dans une base XML native. Cette dernière permet d'entrer véritablement dans la logique de manipulation de flux, car dans une base XML les documents initiaux contenant les données ne sont plus que des niveaux d'interrogation au même titre que les éléments de structuration de l'arborescence XML.

L'édition électronique de l'*Hortus Sanitatis* propose quatre solutions d'accès aux textes qui sont autant de solutions d'exploitation des flux de textes constitués.

La première solution consiste à lire les textes en respectant l'organisation originale de la source via un sommaire traditionnel donnant le numéro et le titre des chapitres. L'interface de lecture donne ensuite, comme nous l'avons vu plus haut, accès aux textes dans deux langues (latin et français) alignées l'une sur l'autre à l'échelle du segment de citation, le lecteur peut ainsi très rapidement faire la correspondance entre le fragment latin original et sa traduction française. En plus de ces mises en forme classiques, la version électronique propose au lecteur d'extraire l'ensemble des fragments d'une même source pour les rassembler sur une même page à la volée. Autrement dit, le lecteur peut instancier un nouveau document, certes totalement inadapté à une lecture immersive et ne correspondant à aucune réalité historique, rassemblant tous les segments provenant d'un auteur spécifique. Au cours de cette opération d'instanciation, ce sont les mêmes données qui sont exploitées d'une autre manière, aucun balisage supplémentaire n'est nécessaire. Il s'agit en définitive d'un premier pas vers une analyse statistique des sources de l'*Hortus Sanitatis*.

La seconde méthode d'accès aux textes proposée au lecteur consiste à utiliser les index raisonnés. L'édition fournit deux index des noms de poissons, l'un en français et l'autre en latin. Le lecteur peut ainsi accéder au chapitre traitant d'un poisson spécifique rapidement. Ces ressources sont produites à partir des balises `index`, avec l'attribut `indexName`, et `term` contenues dans les flux de transcription et de traduction.

La troisième solution d'accès aux textes prend la forme d'un moteur de recherche. Il propose de manière traditionnelle de rechercher une séquence de caractères saisie par l'utilisateur qui peut aussi limiter son investigation à certains contextes. Ainsi, le lecteur peut chercher dans tout le texte ou seulement dans l'un des deux flux français ou latin, ou encore seulement dans les notes. Bien entendu, il s'agit là de possibilités offertes par les structures XML mises en place.

Enfin, le lecteur peut également accéder aux textes par les images numériques de deux incunables, les exemplaires de Valognes et d'Épernay [142] ayant fait l'objet

142. Valognes, BM, R 99 et Épernay, BM, Inc. 3017.

d'une numérisation en mode image. Pour produire la page de ce mode d'accès, seuls les éléments `pb` du flux XML de transcription latine sont retenus.

11.4 Principes éditoriaux

L'édition de l'*Hortus Sanitatis* n'est pas un projet isolé et clos, mais prend place dans le cadre d'un programme de recherche sur la constitution des savoirs dans le domaine de l'ichtyologie médiévale. Il s'agit dans ce programme d'étudier la manière dont ces savoirs se sont constitués par accumulation.

Le modèle de flux et de fragments est parfaitement adapté à ce type d'approche car les savoirs dont il est question ici vont trouver de multiple projections dans des sources différentes avec des altérations et des transformations variables.

Les formes éditoriales de l'*Hortus Sanitatis* offrent ainsi la possibilité de circuler entre ces savoirs en proposant les lieux parallèles où ces savoirs peuvent être lus dans des contextes différents chez d'autres auteurs ou compilateurs. Selon l'état d'avancement des travaux, les références, données sous leurs formes absolues lorsque les éditions existent, sont actives sous forme de liens hypertextes, quand les textes sont disponibles en ligne, dans le laboratoire *Ichtya*, qui contient les textes en ligne rassemblés dans le cadre du programme.

Les solutions mises en place, et en particulier la méthode d'expérimentation/validation, dans le cadre d'édition de l'*Hortus Sanitatis* sont appliquées dans les mêmes conditions dans le cadre du laboratoire *Ichtya*.

11.5 Compilation de sources et base scientifique

Le caractère compilatoire de l'*Hortus Sanitatis* est donc l'une de ses caractéristiques essentielles. Comme nous l'avons vu, l'édition en ligne restitue le contenu du discours savant tel qu'il est véhiculé par la source, en respectant ses incohérences et ses défaillances, tout en permettant au lecteur de les comprendre à travers les notes scientifiques et les passages originaux dont l'*Hortus Sanitatis* donne des formes rapportées. Les données manipulées sont donc de nature profondément fragmentaire.

La figure 11.5 donne une capture d'écran de la page du répertoire de citations rassemblant l'ensemble des fragments provenant d'Aristote. Ces segments de discours rapportés sont dispersés dans l'ensemble des 106 chapitres et la page qui les regroupe ne se prête pas à une lecture immersive, il s'agit bien d'étudier les fragments de l'*Hortus Sanitatis* qui proviennent d'Aristote. Pour faciliter la tâche du lecteur, les

chercheurs ont ajouté à chaque segment d'identification une note restituant le texte tel qu'il apparaît chez Aristote. Ainsi, il est possible de construire la page de consultation dans l'édition en ligne en juxtaposant le fragment de discours dans l'*Hortus Sanitatis* et tel qu'il apparaît dans l'œuvre d'Aristote.

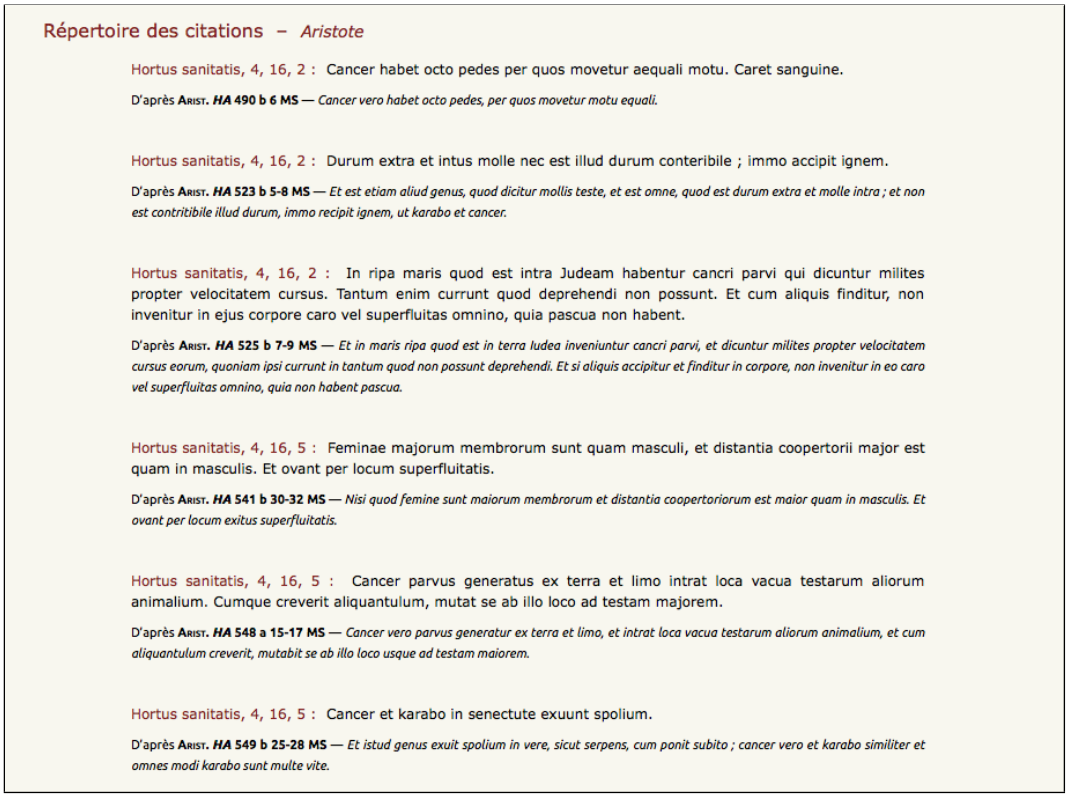

FIGURE 11.5 – L'ensemble des segments provenant de l'œuvre d'Aristote dans l'*Hortus Sanitatis*.

Nous retrouvons ici, les éléments de notre modèle. Tout d'abord, cette application repose entièrement sur la possibilité de distinguer le texte de son support. D'autre part, l'unité de circulation est bien celle du fragment et non celle d'une unité documentaire physique. Enfin, c'est bien le fragment de discours qui circule, sa structure logique, avec d'éventuelles altérations, et non sa forme physique.

Cette utilisation constitue une exploitation du modèle de flux et de fragments tendant vers la base de recherche. Il existe cependant un point de faiblesse : la copie par les chercheurs du fragment d'origine dans les flux de l'*Hortus Sanitatis*. Cette opération est potentiellement porteuse d'erreur humaine, ce qui serait un comble étant donné la nature des objectifs scientifiques. La solution idéale, beaucoup trop lourde à mettre en place avec les moyens dont nous disposions mais parfaitement possible avec notre modèle, aurait été d'intégrer l'œuvre d'Aristote au serveur de

fragments pour permettre l'extraction dynamique des fragments des autorités au moment de la requête.

Une autre exploitation de recherche a consisté à alimenter la base de données du projet Sourcencyme (ANR 2007-2011) à partir de l'encodage mis en place dans le cadre de l'édition de l'*Hortus Sanitatis*. L'objectif du projet Sourcencyme, porté par l'atelier Vincent de Beauvais[143], est d'étudier la circulation du savoir chez les encyclopédistes. Comme nous l'avons vu, le compilateur de l'*Hortus Sanitatis* s'est massivement appuyé sur les textes de Vincent de Beauvais, qui constitue sa source principale. Une grande part du travail d'identification des sources secondaires entrait directement dans le champs du projet Sourcecyme. L'équipe de l'*Hortus Sanitatis* a donc très naturellement intégré le programme. Pour éviter de réaliser le travail deux fois, l'ensemble des recherches réalisé par les chercheurs de Caen dans le cadre de l'*Hortus Sanitatis* a été intégré dans la base de données en utilisant une feuille de transformation pour ajuster les niveaux d'annotation et surtout supprimer les informations directement liées à l'édition.

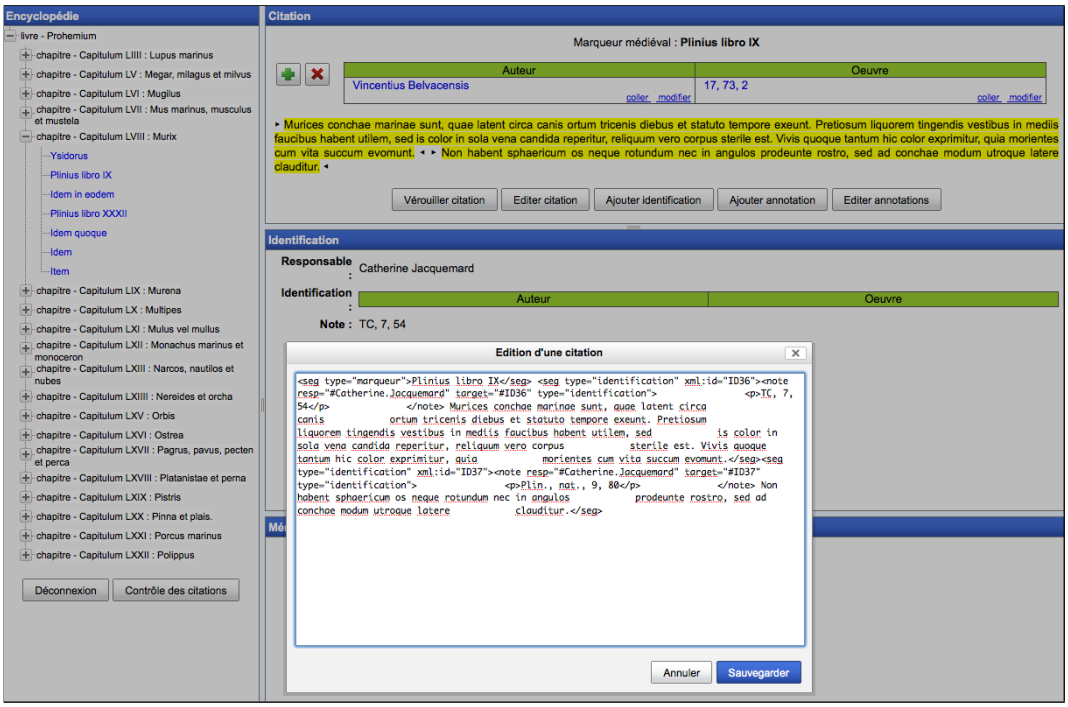

FIGURE 11.6 – Un fragment de l'*Hortus Sanitatis* dans la base de recherche du projet ANR Sourcencyme.

143. Cet atelier n'existe plus aujourd'hui, mais le programme est maintenant accueilli par l'Institut de recherche et d'histoire des textes (IRHT).

La figure 11.6 donne un fragment provenant des flux mis en place pour l'édition de *Hortus Sanitatis* visualisé dans la base de données du projet Sourcencyme et édité avec les outils développés dans ce cadre.

11.6 Apports et influences sur le modèle

La nature compilatoire de la source primaire, et ses caractéristiques fragmentaires et régulières (les chapitres sont toujours construits selon un modèle unique) sont particulièrement intéressantes pour notre modèle et font de l'*Hortus Sanitatis* un excellent terrain d'expérimentation.

Contrairement au cas des *Sources du Mont Saint-Michel* les spécialistes se sont pleinement investis dans le travail de mise en place des flux XML de l'*Hortus Sanitatis*. Ce travail a permis de mettre à l'épreuve les solutions d'articulation des niveaux d'annotation et de structuration dans le cadre de projets d'éditions multimodales. C'est en particulier le cas de la possibilité d'opérer une réduction du niveau d'encodage pour ramener le balisage scientifique à un niveau éditorial adapté à la production de formes de diffusion stables et référençables.

Enfin, le travail sur l'*Hortus Sanitatis* a permis de mettre à l'épreuve la possibilité, une fois les flux mis en place, de les manipuler pour en proposer des modes d'accès et d'utilisation radicalement différents de l'organisation fournie par la source primaire.

Conclusion

Bilan

Le numérique impacte l'ensemble des acteurs concernés, de près ou de loin, par les projets d'édition de sources anciennes. Les techniques utilisées, en particulier la structuration de données avec les « langages à balises » de type XML, apporte une base technologique commune favorisant l'échange et la circulation d'informations.

De plus, l'existence de ce socle technique partagé, qui permet de donner aux échanges entre les acteurs une base solide dans le cadre de projets de collaboration, est renforcée par l'omniprésence du réseau qui tisse des liens entre tous les acteurs concernés par l'édition de sources anciennes : de l'archiviste ou conservateur à l'éditeur en passant par le chercheur.

Par ailleurs, la multiplication des supports de lecture, dernière étape de la convergence numérique, est une phase importante que nous avons pris en compte dans le cadre de notre réflexion sur l'édition numérique. En effet, la nécessité de produire des formes adaptées à l'ensemble des nouveaux supports de lecture (web, tablettes, *smartphones*, liseuses) n'est pas sans impacter l'organisation du travail éditorial. Sur ce plan, les éditeurs sont de plus en plus amenés à mettre en place des méthodes de travail intégrant la logique du *single source publishing*, distinguant *le* fond *des* formes, pour répondre à l'attente des lecteurs qui souhaitent lire les mêmes textes sur plusieurs supports. Ainsi, la structuration de données, conçue indépendemment de toute forme de lecture, revêt, dans le monde numérique, une importance centrale dans l'organisation du travail [Vitali-Rosati et E. Sinatra, 2014]. Si cette activité d'explicitation des structures des textes a toujours fait partie des grandes fonctions de l'éditeur [144], le numérique en accroît l'importance, mais il ne s'agit pas véritablement d'une compétence nouvelle à acquérir pour les éditeurs. Il ne s'agit pas de dire qu'il n'y a pas d'innovation sur ce plan ; il y en a, en particulier sur le plan de l'intégration des langages de structuration de type XML au cœur des chaînes de production. Mais sur le fond, la nécessité de structurer les contenus s'est systématisée depuis l'utilisation des dispositifs numériques dans les maisons d'édition. En définitive, c'est sans doute sur la sphère de diffusion et de distribution que l'impact de la multiplication des supports de lecture est la plus forte : on ne diffuse pas un livre numérique comme un livre papier...

La structuration de données est donc un aspect central tant du point de vue de la conservation, que de l'étude ou de l'édition matérielle des textes anciens.

144. Les livres sont structurés : il suffit d'examiner une table des matières pour s'en convaincre, et les éditeurs utilisent depuis plusieurs décennies des outils informatiques pour simplifier les opérations de production de ce type de solution d'accès au texte.

Ainsi, dans ce dernier domaine de l'édition, plus précisément sur le plan de la fabrication, un même *texte* structuré doit pouvoir permettre la production de plusieurs *documents* mis en forme en fonction des contraintes de chaque support de diffusion : ePub, PDF imprimeur, pages web, etc. éventuellement avec des sélections de contenus pour tel ou tel support.

Nous proposons donc un modèle centré sur la place prédominante du texte en replaçant la notion trop restrictive de *document* à la périphérie. Dans notre proposition le *document* doit être considéré comme un état du texte à instant donné.

Il s'agit donc d'entrer dans une logique de manipulation de flux d'informations textuelles structurées. Ces flux composés de fragments de textes, de taille et de complexité variables en fonction des besoins, sont structurés et annotés par les spécialistes à chaque étape de la réalisation d'un projet. Chaque acteur mobilise les vocabulaires de son domaine pour s'acquitter de sa tâche. Le socle technique commun, c'est-à-dire les langages XML et les technologies associées, assure la possibilité de passer d'un vocabulaire de description à un autre simplement.

Mais la manipulation d'un flux de texte impose également une réflexion sur l'ergonomie des outils et sur les systèmes d'interaction entre les acteurs et les textes indépendamment de la forme qui leur sera affectée pour leur annotation ou leur diffusion. Nous proposons de distinguer la *structure logique* du texte, qui en décrit l'ensemble des éléments constitutifs, et les *structures physiques synchrone* ou *asynchrone*, qui donnent des formes affectées aux éléments de texte, et qui sont associées à une *structure logique* en fonction des opérations à réaliser.

Ainsi, l'association d'une *structure logique* et d'une *structure physique synchrone* permet d'interagir en temps réel avec le texte : il s'agit d'une solution d'édition et de modification du texte. Cette association revient en réalité à produire un *document* temporaire construit pour un usage spécifique du texte.

De la même manière, l'association d'une *structure logique* et d'une *structure physique asynchrone* permet d'obtenir une vue éventuellement interactive du texte mais sans autoriser de modifications lourdes. Si une *structure physique synchrone* permet l'interaction et la modification de la *structure logique* du texte, une *structure physique asynchrone* permet simplement d'en prendre connaissance et éventuellement de la commenter, par exemple en utilisant un outil d'annotation, mais pas de la bouleverser.

L'importance de la conception des interfaces d'édition est capitale. En effet, elles doivent permettre aux spécialistes d'interagir avec des flux organisés en arbres XML. L'utilisateur doit donc pouvoir se concentrer sur les choix sémantiques des éléments et non sur les aspects techniques imposés par les solutions informatiques. Pour Andrea Bozzi [Bozzi, 2014] un point

> [...] considéré comme inévitable dans le développement de notre système est de mettre à disposition de l'utilisateur final une interface qui facilite le choix des balises et des opérations de codage. Un éditeur occupé par des problèmes scientifiques parfois très complexes ne doit pas être distrait par la consultation d'un catalogue de balises pour choisir celle qui convient à chaque situation spécifique.
> Tout en réaffirmant que la machine pourra fournir des résultats significatifs, si et seulement si on y entre une information bien structurée ; il est aussi essentiel que le spécialiste travaille avec une méthode simple, et intuitive, qui l'aide à formaliser son questionnement avec des balises que la machine peut compter et analyser.

Nous proposons de changer de paradigme théorique, mais pas forcément technique. Comme nous l'avons vu, il s'agit, comme cela est fait très couramment aujourd'hui, d'exploiter les solutions de structuration habituelles comme XML et autres "langages à balises". L'idée est plutôt de replacer au cœur des méthodes et des pratiques la nature profondément fluide du texte.

Notre modèle permet de fournir aux acteurs les moyens de mettre en place des flux de textes richement annotés qui pourront faire l'objet d'exploitation d'autant plus efficaces et fines que le niveau d'annotation sera précis.

De ce point de vue, il s'agit d'une évolution des *humanités numériques*. En effet, la logique majoritaire dans ce champs consiste à numériser des masses importantes de données avec une annotation minimale puis à s'appuyer sur des techniques de fouille et de traitement automatique pour extraire des connaissances des textes. À cette logique *top/down* nous proposons d'ajouter une approche *bottom/up* en partant des pratiques des spécialistes des domaines concernés pour mettre en place des entrepôts de données textuelles riches avec des annotations contrôlées. De ces ressources textuelles, fabriquées au sein de *laboratoires de textes*, peuvent être extraits des fragments pouvant alimenter des éditions ou d'autres outils de diffusion ou d'exploitation scientifiques. La complexité et la richesse des niveaux de structuration et d'annotation des flux mis en place dans les laboratoires de texte constituent le facteur limitant principal des solutions d'extractions possibles : plus les flux sont finement structurés et annotés, plus les solutions d'extractions seront riches et précises.

Il s'agit donc aussi de proposer une approche différente des humanités numériques qui ne favorise pas une approche de masse, mais qui parte des pratiques métiers pour constituer des systèmes documentaires riches en apportant aux SHS une dimension expérimentale, avec une méthode de constitution itérative, et d'évaluation des résultats. En étudiant les pratiques des différents acteurs mobilisés dans le cadre d'une édition de sources, nous avons pu dégager un certain nombre de bonnes pratiques et montrer comment elles peuvent constituer la base d'un modèle de travail sur les textes en tant que tel.

Mais l'évolution des textes ne prend pas forcément fin avec la production d'une édition. En effet, tout ou partie du flux peut être récupéré par d'autres chercheurs pour faire l'objet d'une nouvelle campagne d'annotation et d'enrichissement, avec des objectifs scientifiques similaires ou totalement différents. Ainsi, les textes diffusés dans le cadre d'une édition donnée, c'est-à-dire une association entre une *structure logique* et des *structures physiques asynchrones*, peuvent connaître un nouveau cycle de vie dans le cadre d'un nouveau laboratoire de texte dans lequel la même *structure logique* sera associée à de nouvelles *structures physiques synchrones* pour être enrichie ou améliorée.

Perspectives

Outre les perspectives directes liées à l'utilisation des technologies du web sémantique déjà évoquées plus haut [145], le modèle proposé offre d'autres pistes de recherche et d'étude.

Ainsi sur le plan juridique, le modèle de flux et de fragments que nous proposons pour l'organisation du travail sur les sources anciennes a permis d'identifier des objets dans une optique opérationnelle mais il serait particulièrement intéressant d'étudier la possibilité de développer un discours juridique sur ces objets. La question est particulièrement sensible dans le monde de l'édition de textes anciens [146].

Il serait donc intéressant de travailler avec des juristes afin d'évaluer la capacité de notre modèle à fournir des bases pour identifier le travail de chacun sur les différents états des textes. Dans la mesure où toutes les étapes dans le travail de structuration

145. Voir p. 163 et suivantes.

146. Après le procès opposant les maisons d'édition Libraire Droz et Classiques Garnier en mars 2014, concluant que l'édition d'un texte dépourvu d'apparat critique ne pouvait fonder un droit d'auteur, le colloque *L'éditeur de textes est-il un auteur ? Questions juridiques et scientifiques à propos de l'édition critique* organisé à l'IRHT en février 2015 montre l'intérêt que suscitent ces questions parmi les chercheurs impliqués dans des travaux d'éditions de textes anciens.

peuvent être attribuées, l'établissement de la *structure logique* du texte peut peut-être faire l'objet d'une autorité individuelle ou collective. De la même manière, la mise en place de *structures physiques asynchrones* peut être considérée comme une valeur ajoutée et le couple *structure logique*/*structure physique asynchrone* peut peut-être être considéré comme une création pouvant faire l'objet de droits [Roux, 2015]. Ainsi, disposer de droits sur une forme papier pourrait ne pas nécessairement aller de pair avec des droits sur une quelconque autre forme de diffusion et encore moins sur la *structure logique* du texte.

La clarification des objets manipulés pourrait donc peut-être éclaircir les interventions des différents acteurs pour permettre aux juristes d'identifier ou pas des zones de droits sur tel ou tel état du texte. Une avancée sur ce point pourrait favoriser la circulation des textes et simplifier la constitution ou l'accès aux ressources textuelles.

Enfin, même si notre modèle s'appuie fortement sur les pratiques des acteurs concernés, il serait intéressant de mesurer son seuil d'acceptation dans ces mêmes communautés. Si les outils mis en place font l'objet de formation auprès d'un nombre toujours croissant d'éditeurs institutionnels, d'abord dans le cadre de l'AEDRES puis maintenant dans le cadre de BSN 7, une enquête à la fois quantitative et qualitative à l'échelle nationale portant spécifiquement sur les points abordés dans cette étude pourraient permettre de critiquer le modèle et de l'améliorer. Bien entendu, cette enquête devrait englober tous les acteurs concernés et pas seulement les éditeurs matériels pour constituer un panel représentatif et apporter des résultats exploitables.

Bibliographie

[BSN, 2012] BSN (2012). Bibliothèque Scientifique Numérique. *Bibliothèque Scientifique Numérique. http://www.bibliothequescientifiquenumerique.fr/. (consulté le 15 décembre 2015).*

[ECMA, 2008] ECMA (2008). Office Open XML File Formats. Rapport technique, ECMA.

[EDItEUR, 2014] EDItEUR (2014). ONIX for Books, Product Information Format Specification. Rapport technique, EDItEUR.

[ICA, 2000] ICA (2000). ISAD(G) : Norme générale et internationale de description archivistique. Rapport technique, International council on archives, Ottawa.

[IFLA, 1999] IFLA (1999). Functional requirements for bibliographic records. Rapport technique, International Federation of Library Associations and Institutions.

[METS Editorial Board, 2010] METS Editorial Board (2010). Metadata Encoding and Transmission Standard. Rapport technique, Digital Library Federation.

[NISO, 2004] NISO (éd) (2004). *Understanding metadata.* NISO Press, Bethesda, MD.

[OCDE, 2000] OCDE (2000). La littératie à l'ère de l'information. Rapport technique, Organisation de Coopération et de Développement Économiques, Paris.

[TEI, 2012a] TEI (2012a). 12 Critical Apparatus - TEI P5 : Guidelines for Electronic Text Encoding and Interchange. Rapport technique, Text Encoding Initiative Consortium.

[TEI, 2012b] TEI (2012b). TEI P5 : Guidelines for Electronic Text Encoding and Interchange. Rapport technique, Text Encoding Initiative Consortium.

[Baechler et Ingold, 2010] Baechler, M. et Ingold, R. (2010). Medieval manuscript layout model. *In ACM Symposium on Document Engineering*, p. 275–278.

[Barbier, 2000] Barbier, F. (2000). *Histoire du livre.* Armand Colin, Paris.

[Berglund, 2006] Berglund, A. (2006). Extensible Stylesheet Language (XSL) version 1.1. Rapport technique, W3C.

[Berjon *et al.*, 2014] Berjon, R., Faulkner, S., Leithead, T., Navara, E. D., O'Connor, E., Pfeiffer, S. et Hickson, I. (2014). HTML5, A vocabulary and associated APIs for HTML and XHTML. Rapport technique, W3C.

[Berra, 2012] Berra, A. (2012). Faire des humanités numériques. *In* Mounier, P. (éd) : *READ/WRITE BOOK 2*, p. 25–43. OpenEdition Press, Marseille.

[Berti *et al.*, 2009] Berti, M., Romanello, M., Babeu, A. et Crane, G. (2009). Collecting fragmentary authors in a digital library. *In Joint International Conference on Digital Libraries*, p. 259–262, New York.

[Bouet et Desbordes, 2009] Bouet, P. et Desbordes, O. (2009). *Chroniques latines du Mont Saint-Michel (IX^{e}-XII^{e} siècle)*. Presses universitaires de Caen ; Scriptorial-Ville d'Avranches, Caen ; Avranches.

[Bougy, 2009] Bougy, C. (éd) (2009). *Le Roman du Mont Saint-Michel (XII^{e} siècle)*. Presses universitaires de Caen, Caen.

[Bozzi, 2014] Bozzi, A. (2014). Édition numérique de documents textuels. *Labex OBVIL, Carnet de recherche, http://obvil.paris-sorbonne.fr/carnet-de-recherche/presentations/edition-numerique-de-documents-textuels*.

[Bray *et al.*, 2008] Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E. et Yergeau, F. (2008). Extensible Markup Language (XML) 1.0 (Fifth Edition). Rapport technique, W3C.

[Briet, 1951] Briet, S. (1951). *Qu'est-ce que la documentation ?* EDIT, Paris.

[Bryan, 1992] Bryan, M. (1992). An Introduction to the Standard Generalized Markup Language (SGML). *The SGML Centre*.

[Buard et Dornier, 2008] Buard, P.-Y. et Dornier, C. (2008). Éditer un cahier de travail de montesquieu : les apports du numérique. *Recherches et travaux*, (72):139–156.

[Buckland, 1997] Buckland, M. K. (1997). What is a "document" ? *Journal of the American Society of Information Science*, 48(9):804–809.

[Burghart, 2013] Burghart, M. (2013). Les Trois Ordres ou l'Imaginaire des Digital Humanities #dhiha5. *Digital Humanities à l'IHA. http://dhiha.hypotheses.org/804. (consulté le 15 janvier 2014)*.

[Burnard, 2012] Burnard, L. (2012). Du *literary and linguistic computing* aux *digital humanities* : retour sur 40 ans de relations entre sciences humaines et infor-

matique. *In* Mounier, P. (éd) : *READ/WRITE BOOK 2*, p. 45–58. OpenEdition Press, Marseille.

[Burnard, 2014] Burnard, L. (2014). *What is the Text Encoding Initiative ?* OpenEdition Press, Marseille.

[Clark et DeRose, 1999] Clark, J. et DeRose, S. (1999). XML Path Language (XPath). Rapport technique, W3C.

[Coombs *et al.*, 1987] Coombs, J. H., Renear, A. H. et DeRose, S. J. (1987). Markup systems and the future of scholarly text processing. *Communications of the ACM*, 30(11):933–947.

[Crofts *et al.*, 2011] Crofts, N., Doerr, M., Gill, T., Stead, S. et Stiff, M. (2011). Definition of the cidoc conceptuel reference model. *Documentation Standards Group – CIDOC CRM Special Interest Group*, (May).

[Dacos, 2010] Dacos, M. (2010). A webservice to convert from Office to XML TEI.

[Dacos et Mounier, 2010] Dacos, M. et Mounier, P. (2010). *L'édition électronique*. Repères. La Découverte, Paris.

[Dacos et Mounier, 2014] Dacos, M. et Mounier, P. (2014). Humanités numériques. Rapport technique, Institut francais.

[Darnton, 1999] Darnton, R. (1999). Le nouvel âge du livre. *Le Débat*, 105:176–184.

[Dornier et Buard, 2010] Dornier, C. et Buard, P.-Y. (2010). L'édition électronique de cahiers de travail : l'exemple de mes pensées de montesquieu. *In* Fischer, F., Fritze, C. et Vogeler, G. (éds) : *Codicology and Palaeography in the Digital Age 2*, p. 361– 374. Books on Demand, Norderstedt.

[Doumat *et al.*, 2008] Doumat, R., Egyed-Zsigmond, E., Pinon, J. M. et Csiszar, E. (2008). Online ancient documents : Armarius. *In Proceeding of the eighth ACM symposium on Document engineering*, p. 127–130. ACM.

[Driscoll, 2010] Driscoll, M. J. (2010). The words on the page : Thoughts on philology, old and new. *In* Lethbridge, E. et Quinn, J. (éds) : *Creating the medieval saga : Versions, variability, and editorial interpretations of Old Norse saga literature*, p. 85–102. University Press Of Southern Denmark, Odense.

[Duchein, 1977] Duchein, M. (1977). Le respect des fonds en archivistique : principes théoriques et problèmes pratiques. *La Gazette des archives*, 97:71–96.

[Durusau et Brauer, 2011] Durusau, P. et Brauer, M. (2011). Open Document Format for Office Applications. Rapport technique, OASIS.

[Greg, 1966] Greg, W. W. (1966). *The Rationale of Copy-Text*. Clarendon Press.

[Hulle, 2008] Hulle, D. v. (2008). Les manuscrits bilingues de beckett : la combinaison des approches "documentaire" et "textuelle" dans une édition numérique. *Recherches & Travaux*, 72:53–58.

[Imprimerie Nationale, 2002] Imprimerie Nationale (éd) (2002). *Lexique des règles typographiques en usage à l'imprimerie nationale*. Imprimerie nationale, Paris.

[Jacquemard *et al.*, 2013] Jacquemard, C., Gauvin, B. et Lucas-Avenel, M.-A. (éds) (2013). *Hortus sanitatis : Livre IV, Les Poissons*. Fontes & Paginæ. Presses universitaires de Caen, Caen.

[Jeanneret et Souchier, 2005] Jeanneret, Y. et Souchier, E. (2005). L'énonciation éditoriale dans les écrits d'écran. *Communication et langages*, 145(1):3–15.

[Kay, 2007] Kay, M. (2007). XSL Transformations (XSLT) Version 2.0. Rapport technique, W3C.

[Kummer, 2010] Kummer, R. (2010). Semantic Technologies for Manuscript Descriptions — Concepts and Visions. *In* Fischer, F., Fritze, C. et Vogeler, G. (éds) : *Codicology and Palaeography in the Digital Age 2*, p. 133–154. Books on Demand, Norderstedt.

[Le Hors *et al.*, 2004] Le Hors, A., Le Hégaret, P., Wood, L., Nicol, G., Robie, J., Champion, M. et Byrne, S. (2004). Document Object Model (DOM) Level 3 Core Specification. Rapport technique, W3C.

[Lebarbé *et al.*, 2008] Lebarbé, T., Blanchard, A. et Meynard, C. (2008). Manuscrits de Stendhal. *Recherches & Travaux*, 72:97–117.

[Lebarbé et Meynard, 2009] Lebarbé, T. et Meynard, C. (2009). Nouvelles pratiques éditoriales, nouvelles lectures : les enjeux de l'édition électronique de manuscrits littéraires. *Mémoires du livre*, 1(1).

[Lecarpentier, 2011] Lecarpentier, J.-M. (2011). *Sydonie : Architecture de Document et Ingénierie du Web*. Thèse de doctorat, Université de Caen Basse-Normandie, Caen.

[Lecarpentier *et al.*, 2010] Lecarpentier, J.-M., Bazin, C. et Le Crosnier, H. (2010). Multilingual composite document management framework for the internet : an frbr approach. *Document Engineering*, p. 13–16.

[Meynard *et al.*, 2013] Meynard, C., Jacquelot, H. et Corredor, M.-R. (éds) (2013). *Stendhal, Journaux & Papiers*, volume 1 – 1797-1804. ELLUG, Grenoble.

[Ore et Eide, 2009] Ore, C. E. et Eide, O. (2009). TEI and cultural heritage ontologies : Exchange of information ? *Literary and Linguistic Computing*, 24(2): 161–172.

[Pemberton, 2002] Pemberton, S. (2002). XHTML^{TM} 1.0 The Extensible HyperText Markup Language. Rapport technique, W3C.

[Polis et Stasse, 2009] Polis, S. et Stasse, B. (2009). Pratiques du document. entre tradition et renouvellement. *MethIS*, p. 7–12.

[Pritchett et Gylling, 2011] Pritchett, J. et Gylling, M. (2011). EPUB Open Container format (OCF) 3.0. Rapport technique, IDPF.

[Prost, 2007] Prost, B. (2007). Rapport d'étude sur l'édition numérique de livres scientifiques et techniques. Rapport technique, Édition QUÆ.

[Prost, 2011] Prost, B. (2011). *XML pour l'édition structurer saisir publier*. Eyrolles, Paris.

[Pédauque, 2003] Pédauque, R. T. (2003). Document : forme, signe et médium, les re-formulations du numérique.

[Pédauque, 2007] Pédauque, R. T. (2007). *La redocumentarisation du monde*. Cépaduès, Toulouse.

[Quint, 1987] Quint, V. (1987). *Une approche de l'édition structurée des documents*. Thèse de doctorat, Université Joseph-Fourier-Grenoble I.

[Quint *et al.*, 2010] Quint, V., Roisin, C., Sire, S. et Vanoirbeek, C. (2010). From templates to schemas : bridging the gap between free editing and safe data processing. *In ACM Symposium on Document Engineering*, p. 61–64.

[Raggett *et al.*, 1999] Raggett, D., Le Hors, A. et Jacobs, I. (1999). HTML 4.01 Specification. Rapport technique, W3C.

[Robinson et Meschini, 2010] Robinson, P. et Meschini, F. (2010). Works, documents, texts and related resources for everyone. *In Digital Humanities 2010 (DH2010)*.

[Romanello *et al.*, 2009a] Romanello, M., Berti, M., Babeu, A. et Crane, G. (2009a). When printed hypertexts go digital : Information extraction from the parsing of indices. *In HyperText*, p. 357–358, New York.

[Romanello *et al.*, 2009b] Romanello, M., Berti, M., Boschetti, F. et Crane, G. (2009b). Rethinking critical editions of fragmentary texts by ontologies. *In Electronic Publishing*, p. 155–174, Milan.

[Roux, 2015] Roux, D. (2015). Les tournants numériques de l'édition scientifique, (à paraître).

[Salaün, 2012] Salaün, J.-M. (2012). *Vu, lu, su : les architectes de l'information face à l'oligopole du Web*. La Découverte, Paris.

[Sanderson *et al.*, 2011] Sanderson, R., Albritton, B., Schwemmer, R. et Van de Sompel, H. (2011). Sharedcanvas : a collaborative model for medieval manuscript layout dissemination. *In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, p. 175–184. ACM.

[Schreibman *et al.*, 2004] Schreibman, S., Siemens, R. et Unsworth, J. (2004). *A companion to digital humanities*. Blackwell, Oxford.

[Schuwer, 1997] Schuwer, P. (1997). *Traité pratique d'édition*. Ed. du Cercle de la librairie, Paris.

[Smith, 2009] Smith, N. (2009). Citation in classical studies. *Digital Humanities Quarterly*, 3(1).

[Teorey, 1990] Teorey, T. J. (1990). *Database modeling and design :the entity-relationship approach*. Morgan Kaufman Publishers, San Mateo, Calif.

[UCP, 2010] UCP (2010). *The Chicago manual of style*. The University of Chicago Press, Chicago, 16e édition.

[Vertan et Reimers, 2012] Vertan, C. et Reimers, S. (2012). A TEI-based Application for Editing Manuscript Descriptions. *Journal of the Text Encoding Initiative*, (2).

[Virbel et Maignien, 1999] Virbel, J. et Maignien, Y. (1999). *Le livre électronique et le concept de station de lecture assistée par ordinateur*, p. 11–48. PULIM.

[Vitali-Rosati et E. Sinatra, 2014] Vitali-Rosati, M. et E. Sinatra, M. (2014). Introduction. *In* Vitali-Rosati, M. et Sinatra, M. E. (éds) : *Pratiques de l'édition numérique*, p. 7–10. Presses de l'Université de Montréal, Montréal.

[Walsh, 2010] Walsh, N. (2010). *DocBook 5 : the definitive guide*. O'Reilly, Beijing ; Sebastopol.

[TGIR-Huma-Num, 2013] TGIR-Huma-Num (2013). Dariah - digital research infrastructure for the arts and humanities. Rapport technique, TGIR Huma-Num.

Résumé

Les sources anciennes présentent une complexité d'organisation textuelle qui incite à la définition de modèles d'édition spécifiques adaptés. Sur la base de l'étude des pratiques des métiers de la conservation, de l'analyse et de l'édition des textes anciens, la thèse propose un modèle général d'organisation du travail permettant la circulation des informations sur les objets conservés ainsi que les textes portés par ces mêmes objets. La prise en compte des cultures métier confrontées au contexte de la convergence numérique et de l'omniprésence du réseau conduit à interroger la notion de document pour la replacer dans une logique plus vaste des flux de données et des fragments d'informations numériques. Nous proposons de considérer et d'organiser l'ensemble des informations manipulées en flux de textes structurés par les chercheurs en sciences en humaines et sociales. En nous appuyant sur de nombreuses expérimentations menées dans le monde de l'édition institutionnelle, nous proposons un modèle d'organisation des ressources textuelles finement structurées dans le but de faciliter les exploitations éditoriales et les programmes de recherche dans le domaine des humanités numériques.

Mots-clés: histoire – sources, patrimoine écrit, édition scientifique, édition électronique, structures de données (informatique), ontologies (informatique), XML (langage de balisage), informatique (normes).

Abstract

Ancient sources present a very complex text organization that guides to define dedicated patterns. Beginning with the study of practices in the fields of preservation, analyze and edition of ancient texts, this thesis gives a general work pattern to organize circulation of information on objects and on texts carried by these objects. Paying attention to the cultural dimension of skills envolved in ancient sources facing both digital convergence and global networking guides us to rethink the document in the larger fields of flows and fragments. We propose to organize all informations used in text flows marked up by scholars. Based on many experimentations lead in institutional publishing, we propose a general organisational pattern to manage highly annotated textual resources in order to easily build editorial or research exploitations.

Keywords: history – sources, science publishing, electronic publishing, data structures (computer science), ontologies (information retrival), XML (document markup language), computer science – standards.

Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen – UMR 6072
Université de Caen Basse-Normandie, Campus Côte de Nacre, Boulevard du Maréchal Juin
CS 14032 — 14032 CAEN cedex 5