Théia

1 | 2024
Retours d'expérience en édition numérique de sources en histoire et histoire de l'art

Des plumes aux pixels : actualité de l'édition scientifique numérique

From writing with quills to pixels: the latest in digital scientific publishing

Ariane Pinche

http://publications-prairial.fr/theia/index.php?id=102

DOI : 10.35562/theia.102

Référence électronique
Ariane Pinche, « Des plumes aux pixels : actualité de l'édition scientifique numérique », *Théia* [En ligne], 1 | 2024, mis en ligne le 14 avril 2025, consulté le 18 septembre 2025. URL : http://publications-prairial.fr/theia/index.php?id=102

Des plumes aux pixels : actualité de l'édition scientifique numérique

From writing with quills to pixels: the latest in digital scientific publishing

Ariane Pinche

PLAN

TEXTE

1 Les pratiques de l'édition savante [1] sont les héritières d'une méthodologie qui remonte aux premières tentatives de reconstitution d'un texte original ou du moins d'une version canonique pour des textes aux traditions complexes comme la Bible, les œuvres d'Homère, Cicéron, Virgile, Chrétien de Troyes ou encore Shakespeare [2]. Depuis les humanistes du XVI^e^ siècle jusqu'à nos jours, les éditeurs ont mis au point différentes méthodes pour rendre la pluralité textuelle. Ces méthodes vont de la restitution minutieuse d'un texte le plus proche possible de sa source – restitution de la mise en page, des abréviations, etc. – à la reconstitution d'un archétype le plus proche possible de la version de l'auteur. L'éditeur y retrace les relations entre les différentes réalisations du texte depuis le(s) texte(s) de l'auteur jusqu'aux réalisations accessibles aujourd'hui [3].

2 Depuis près d'un siècle persiste une tension méthodologique entre éditions « conservatrices », qui suivent un témoin issu de la tradition

manuscrite, et les éditions « reconstructionnistes », qui cherchent à retrouver le texte original de l'auteur. La méthode reconstructionniste, née, entre autres, des travaux de K. Lachmann [4], s'est établie comme la méthodologie de référence dans les études classiques. Ces éditions opèrent une nette distinction entre le texte conçu comme étant le texte original et les sources manuscrites regardées comme un véhicule imparfait du texte à reconstruire, là où l'éditeur d'une édition conservatrice envisagera le texte du témoin manuscrit comme une réalisation possible du texte [5]. La méthode conservatrice, initiée au début du XX^e^ siècle par J. Bédier, prône l'édition du « meilleur des manuscrits » (qualité du texte, place dans la tradition manuscrite). Elle se prête tout particulièrement à l'édition des textes médiévaux, en permettant de conserver les graphies d'un document qui a réellement circulé, là où la méthode reconstructionniste aboutit à la création d'un texte dont la langue ou les graphies ne sont plus en lien avec des attestations historiques [6]. Dans la continuité de cette approche, dans les années quatre-vingt, B. Cerquiglini défend, dans l'*Éloge de la variante* [7], la place du témoin manuscrit comme représentant d'un état de transmission du texte ayant sa valeur intrinsèque. Il donne alors naissance à la *New Philology* et influence également la philologie génétique [8]. Dans la continuité de ces réflexions, Peter Shillingsburg définit son travail davantage comme la présentation d'un processus textuel, plutôt que comme l'établissement d'un produit final immuable [9]. Enfin, J. Bryant considère que « the only "definitive text" is a multiplicity of texts, or rather, the fluid text » [10], proposant ainsi un point de vue sur le texte opposé à celui de l'école lachmanienne. S'ajoutent, aujourd'hui, au paysage scientifique, les éditions génétiques qui retracent le processus complet d'écriture de l'auteur à travers l'analyse de ses brouillons [11]. Toutefois, si les méthodes changent, les éditions scientifiques visent toutes l'établissement d'un texte fiable et contextualisé, quelle que soit l'alternative choisie.

3 Enfin, malgré le soin apporté à ces éditions, elles sont souvent difficiles à appréhender à cause des contraintes imposées par le format papier qui entraînent l'utilisation de règles de représentation et/ou de schématisation de l'information [12], en raison du nombre restreint d'informations transmissibles aux lecteurs. En outre, jamais un livre ne pourra contenir dans ces pages les illustrations,

l'arrangement des feuillets, les variations de la tradition manuscrite et rester consultable [13]. Enfin, ces productions, figées dans le papier, nécessitent d'être mises à jour, voire refaites au gré des évolutions méthodologiques ou des nouvelles avancées de la recherche. Les éditions numériques ont, pour partie, émergé pour essayer de dépasser ces limitations, permettant de consulter les numérisations des sources, d'offrir des parcours de lecture, de plonger le lecteur dans un réseau de multi-fenêtrages [14] et d'hyperliens pour proposer une lecture augmentée. Ainsi, après avoir dressé un état de l'art de l'édition numérique, nous présenterons ses standards, son lectorat, mais également ses nouveaux enjeux. Enfin, nous analyserons l'impact des récentes innovations technologiques sur les systèmes de production textuelle.

Édition numérique : définition et héritages méthodologiques

4 Le passage des éditions au format numérique n'est pas synonyme d'une révolution totale des méthodologies de l'édition [15]. Au contraire, elles semblent être solidement enracinées dans les débats philologiques du XX^e^ siècle et ont contribué à enrichir les multiples voies envisageables. Certains théoriciens de l'édition numérique considèrent le passage au numérique comme une évolution naturelle de la philologie traditionnelle. Selon H. W. Gabler, elles doivent respecter les mêmes critères d'érudition scientifique pour l'établissement du texte et être mises en œuvre à l'aide d'instruments qui renforcent l'analyse critique du texte (collation, concordance, stemma, etc.) [16]. En outre, la philologie matérialiste tire parti de la mise en ligne des numérisations des documents, de systèmes de multifenêtrage pour la comparaison, et de la contextualisation diachronique et synchronique des œuvres via des systèmes d'hyperliens [17].

Édition numérique scientifique : définition

5 Que signifie éditer numériquement ? Patrick Sahle propose ces quelques critères pour définir les « *digital scholarly editions* » (DSE)

que nous reprenons et traduisons pour partie ci-dessous [18] :

- **S pour Scholarly** : une édition doit offrir une critique textuelle, apportant ainsi une valeur supplémentaire par rapport à la simple mise en ligne d'un fac-similé. Une bibliothèque numérique, qui se contente de numériser des documents, sans offrir une analyse critique, ne répond pas à cette définition.
- **D pour Digital** : une édition numérique ne doit pas pouvoir être convertie en une édition imprimée sans une perte substantielle d'informations ou de fonctionnalités. De même, une conversion numérique d'une édition imprimée n'est pas une édition numérique, sauf si elle est enrichie de contenus ou de nouvelles fonctionnalités.
- **E pour Édition** : une édition numérique doit fournir un texte, qu'il s'agisse d'une simple transcription ou d'un texte plus élaboré. Ainsi, les catalogues, index ou bases de données ne sont pas inclus dans cette catégorie.

Enfin, selon J. Carlquist, une édition scientifique numérique de qualité doit s'appuyer sur un encodage complexe et riche, doit inclure un texte interrogeable et des images, et être enrichie de métadonnées appropriées et riches de sens, un apparat critique, des index et un glossaire [19].

Édition documentaire contre édition monumentale ?

6 Les nouvelles possibilités offertes par les éditions numériques ont ravivé le débat entre éditions documentaires et éditions monumentales, pour reprendre les termes de P. Robinson :

> *One cannot know the work without the documents – equally, one cannot understand the documents without a comprehension of the work they instance. From this, a principle appears: a scholarly edition must, so far as it can, illuminate both aspects of the text, both text-as-work and text-as-document. Traditional print editions have focused more on the first. An evident advantage of digital editions is that they might redress this balance, by including much richer materials for the study of text-as-document than can be achieved in the print medium.* [20]

7 L'objet numérique libéré des contraintes du papier permet une accumulation d'images, de textes et de représentations offrant la possibilité inédite de représenter le texte dans ses réalisations en tant que document, là où le papier représente le texte comme monument. Bien que nombre d'éditions en ligne soient tournées vers le texte comme document, comme les éditions du projet ELEC de l'École nationale des chartes [21], de nombreux critiques ont souligné les failles d'une pratique qui se limiterait uniquement à une représentation imitative de la source. Toutefois, M. Dahlström clame combien il est fallacieux de dire qu'une édition numérique imitative ne découle pas d'une démarche scientifique [22]. La conversion des signes manuscrits en caractères informatiques relève déjà d'une interprétation, car le passage à une représentation normalisée entraîne une réduction de la variété des formes de lettres de la source [23]. Transcrire et encoder sont le fruit d'un processus de sélection impliquant réflexion et méthode : choix de la granularité de l'imitation de la source, représentation du système abréviatif, niveaux de différenciation des allographes, etc. Toutefois, quoiqu'on puisse considérer qu'une édition documentaire est déjà le premier pas vers une édition critique en tant que source primaire [24], il est primordial de dépasser le stade de la mise en ligne d'archives ou d'une accumulation documentaire, sans quoi le risque est de perdre le lecteur dans un amas d'informations dont il ne parviendra pas à faire sens, amenant à une forme de refus d'éditer [25].

8 Certains projets comme le projet *Hyperdonat* ont expérimenté une voie médiane, essayant de concilier l'approche du texte-comme-document et du texte-comme-monument en proposant un texte de référence et la possibilité de parcourir la tradition manuscrite à travers une interface de comparaison des témoins [26]. Le projet d'édition de *Guiron le Courtois*, en s'appuyant sur la modularité des corpus numériques, propose d'établir le texte branche par branche avec des éditions intermédiaires pour chacune d'entre elles, afin d'établir un texte critique qui garde une trace de la surface linguistique pour chaque branche de la tradition. La réunion des différentes éditions permettra à terme d'arriver à l'édition critique finale de l'œuvre, tout en suivant les innovations du texte au fil de la transmission [27].

9 Ainsi, même si les premières éditions numériques ont surtout favorisé le texte comme document, tandis que les éditions papier ont favorisé le texte comme monument, les progrès techniques de ces dernières années, grâce, entre autres, à la reconnaissance automatique d'écriture (ATR), offrent la possibilité de traiter des corpus de plus en plus vastes, et donc de traiter des traditions manuscrites complètes, tout en établissant un texte de référence.

Les différentes formes d'édition numérique

10 L'une des premières applications du numérique pour la mise en ligne de textes a été la création de bibliothèques numériques, offrant la possibilité de mener des fouilles textuelles. En France, l'Observatoire de la vie littéraire (OBVIL, 2014-…) [28] a mis à disposition plus d'une trentaine de corpus littéraires français structurés en XML TEI [29] sur Github [30], accompagnés d'outils tels que *Dramagraph* pour faciliter l'analyse de la répartition des répliques dans les pièces de théâtre [31]. À l'échelle internationale, le projet *Perseus Digital Library* (1987 – …) [32] donne accès à des centaines d'éditions de textes, ainsi qu'à une interface de consultation, de recherche, et à des URN [33] pérennes pour citer les textes, tout en assurant l'accès aux fichiers sources encodés en XML TEI.

11 Certaines éditions numériques, tout en respectant les codes de l'édition traditionnelle, proposent des éditions enrichies. C'est le cas du projet la *Queste del saint Graal* (1999 – …) de Christiane Marchello-Nizia et Alexei Lavrentiev [34] à l'origine de la création du portail de la Base de français médiéval (BFM) [35] qui héberge des éditions en XML TEI de textes en ancien français. Grâce au média numérique, la *Queste* met en regard le texte avec sa source, tout en fournissant une transcription imitative avec les abréviations et une transcription normalisée. En outre, le texte est intégralement étiqueté : lemmes et POS (Part-of-Speech [36]) facilitant les recherches lexicales et des analyses textométriques.

12 Influencés par la *New Philology*, certains projets proposent des parcours de lecture où la matérialité des œuvres est centrale. Le projet *The Walt Whitman Archive* (1995– …) [37] ou encore *The Rossetti Archive* (1993-2008) [38] en sont des exemples. Le projet *The*

Complete Writings and Pictures of Dante Gabriel Rossetti offre un parcours enrichi de numérisations d'un corpus où l'auteur a non seulement écrit les textes, émis des productions picturales (souvent en amont du texte et en lien étroit avec lui)[39], mais aussi élaboré la conception matérielle des livres, leur donnant une importance cruciale à la compréhension du processus créatif de l'auteur[40]. Plus récemment, le projet *Woolf Online* retrace l'univers mental de Virginia Woolf en mettant à disposition des lecteurs ses brouillons et des collections d'images. Pour les éditions génétiques, le numérique permet d'explorer les brouillons d'auteurs. Dans l'édition de *Frankenstein* de Mary Shelley du projet the *ShelleyGodwin Archive* (2013– …)[41], on peut parcourir les différentes étapes de l'écriture et identifier les mains responsables des corrections manuscrites[42]. En France, le projet *Bovary* (20022009)[43] propose un plan de la création de l'œuvre de Gustave Flaubert, permettant de retracer le processus d'écriture du roman. Ces projets, à travers la création de réseaux d'hyperliens et une documentation abondante, mettent en scène de manière dynamique le processus créatif. Elles promeuvent une textualité numérique ouverte et interactive en opposition au livre statique et fermé[44]. Toutefois, une telle abondance de documents demande un investissement important du lecteur, avec une navigation complexe pouvant entraîner une perte du sens au profit d'une navigation compulsive, comme le décrit A. Mangen :

> *The urge to click can easily become too tempting to resist, if we are cognitively or perceptually stimulated with possibilities that seem more exciting than what we are presently focused on. Knowledge sites have a wealth of potentials that can risk disrupting our phenomenological preoccupation with them, thereby limiting the possibility of hermeneutical reflection.* [45]

13 En outre, les systèmes de renvoi dépendent des interfaces en ligne, ce qui les rend vulnérables aux mises à jour techniques, menaçant ainsi leur maintenabilité.

14 Le passage au format numérique offre donc des opportunités inédites aux éditeurs et permet d'aborder des corpus de manière exhaustive, tout en engageant le lecteur dans des expériences interactives. Toutefois, la vigilance reste de mise pour éviter la prolifération

d'informations sans valeur critique et assurer la pérennité des projets dans un environnement numérique en constante évolution.

Édition numérique : comment, pour qui, nouveaux enjeux ?

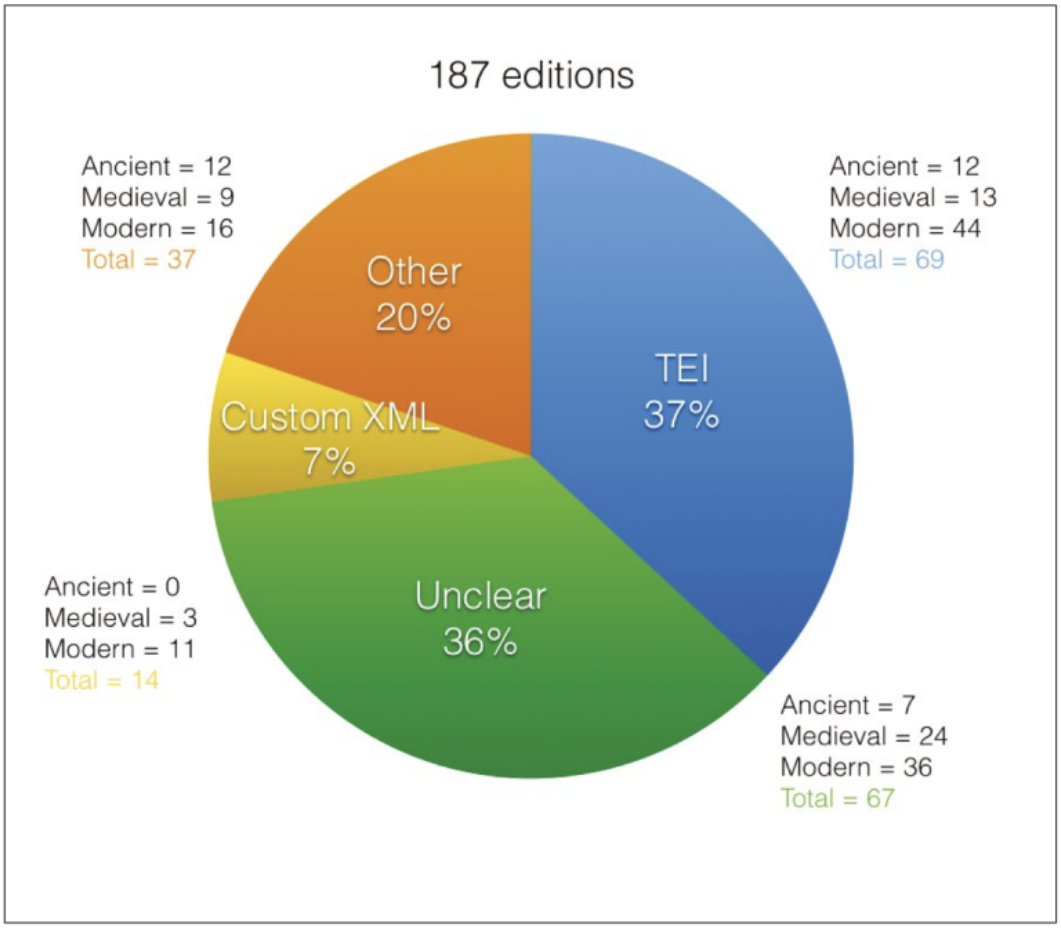
15 Les éditions numériques découlent d'une histoire décennale ayant forgé une communauté scientifique unie autour de la norme qu'est la TEI (*Text Encoding Initiative*). Non seulement elle a fourni les outils nécessaires à la création d'éditions numériques, mais elle a également été propice à la réflexion et aux échanges sur les questions relatives à l'édition des textes au sein d'une communauté dynamique et ouverte.

Un standard : XML TEI

16 La TEI a vu le jour en 1987 avec pour objectif de standardiser la représentation informatique des données textuelles et des sources historiques. La même année, lors d'une réunion réunissant des spécialistes des archives, des sciences computationnelles et des sciences des textes à Poughkeepsie (États-Unis), la TEI s'est dotée de *guidelines* [46]. Depuis 2003, elle repose exclusivement sur le langage XML (*eXtensible Markup Language*), qui fonctionne sur un système d'arborescence et d'imbrication strict. Son système de balisage permet d'enrichir le texte source d'informations aussi diverses que la hiérarchisation du texte, les entités nommées et les annotations linguistiques et plus encore, le tout dans un langage lisible à la fois par l'humain et par l'ordinateur. L'objectif de la TEI est de fournir une solution aussi bien aux débutants pour structurer un texte qu'aux experts cherchant une solution pour bâtir un corpus complexe. Elle offre suffisamment de liberté pour permettre à chacun de proposer un encodage adapté à sa source et à ses objectifs de recherche. L'esprit même de la TEI n'est pas de contraindre les éditeurs à un choix technique appliqué de manière mécanique, mais de permettre aux choix épistémologiques de rester centraux dans la modélisation des données. Cependant, la contrepartie de cette liberté se traduit par des pratiques éditoriales, tout comme la philologie traditionnelle, qui ne sont pas uniformisées.

17 Enfin, bien que la plupart des projets cités dans cet article reposent sur des corpus encodés en XML TEI, il est indéniable que le standard n'est pas encore appliqué de manière systématique. Comme le soulignent G. Franzini, S. Mahony et M. Terras, seulement 37 % des éditions numériques parmi celles répertoriées dans un catalogue de 187 éditions suivent les prescriptions de la TEI [47].

Figure 1 : Répartition des technologies utilisées pour les éditions numériques, illustration issue de G. Franzini, S. Mahony et M. Terras, « A Catalogue of Digital Editions », 2016



18 Ainsi, malgré l'établissement de standards documentés et partagés par une large communauté, les pratiques de l'édition numérique ne sont pas encore unifiées, ne serait-ce que d'un point de vue technologique.

Qui sont les « lecteurs » des éditions numériques ?

19 Il est relativement complexe de prévoir quels seront les usages et les lecteurs des éditions numériques. Non seulement le lectorat potentiel échappe à toute délimitation disciplinaire, mais la multiplicité de ses usages ouvre encore davantage les possibilités, entraînant même parfois une perte de contrôle de l'éditeur sur l'utilisation de son édition [48]. Bien que l'on qualifie souvent les « utilisateurs » d'éditions numériques sous le terme générique de « lecteurs », le terme même est contestable dans le cas d'une analyse textuelle basée sur du *distant reading* [49]. Voici une tentative de délimitation de trois types principaux d'utilisateurs/lecteurs en fonction du type d'usage du texte [50] :

- Le lecteur-consultant qualitatif, qui adopte une lecture linéaire du corpus similaire au support papier, soit pour lire une œuvre accessible en ligne, soit pour établir un corpus à analyser de manière qualitative.
- Le lecteur-explorateur pour les éditions basées sur du multifenêtrage ou des hyperliens, comme les projets The Rossetti Archive ou encore Woolf Online. Le lecteur y crée son propre parcours dont la complexité peut être déterminée en fonction de son degré d'expertise.
- Le lecteur-utilisateur de données. Ce sont des utilisateurs qui ne consultent pas les interfaces de lecture ou d'exploration, mais qui utilisent le texte brut, structuré, voire enrichi pour faciliter la fouille [51] via une API ou en récupérant directement les fichiers sources pour des études quantitatives ou en faire des données d'entraînement [52]. Le texte est alors traité comme des données [53].

20 L'éventail des utilisations possibles est très large, de la simple consultation à la fouille de texte, exigeant une donnée structurée et enrichie. Le traitement du texte comme donnée implique un passage d'une pratique individuelle visant à produire un objet fini à la production de données réexploitables dans un autre cadre que celui pour lequel elles ont été créées. Ainsi, l'éditeur numérique n'est plus simplement le producteur d'un texte clos sur lui-même, mais il devient également un modélisateur de données à partager, d'où l'importance de l'utilisation de standards (caractères Unicode, encodage, etc.) pour en faciliter la réutilisation.

Nouvelles problématiques

21 Les éditions numériques demeurent aujourd'hui un ensemble de pratiques multiples, ce qui les rend difficiles à évaluer. En effet, tandis que « les éditions critiques au format imprimé répondent, pour leur part, à des modèles d'évaluation et de validation instaurés dès le XIX^e siècle [...]. Les éditions critiques numériques sont pour le moment peu nombreuses à faire l'objet de telles recensions » [54]. Les méthodes classiques d'évaluation ne s'appliquent pas et les éditeurs scientifiques se heurtent à une difficulté de reconnaissance de leur travail [55]. En France, des groupes de travail pour la mise en place de standards émergent, tels que l'ancien consortium Cahier [56], ou encore le consortium ARIANE-HN [57]. À l'international, on peut citer les *Guidelines for Editors of Scholarly Editions* [58] de la *Modern Language Association* qui proposent une grille pour guider les éditeurs sur les informations importantes à fournir sur le corpus et son établissement, ainsi que sur sa mise en ligne et la question de la préservation des données [59].

22 Comme le soulignait Frédéric Duval, la force du numérique réside dans sa modularité : « une édition numérique peut proposer des parcours plus ou moins ouverts, de la consultation libre d'un dépôt d'archives à des parcours commandés par des approches et intérêts divers [60] ». Cette modularité s'étend également au fait que l'œuvre peut être modifiée à l'infini [61]. Ces éditions pourront être maintenues au cours du temps par une succession d'experts [62]. Cependant, cette modularité engendre de nouvelles problématiques. Alors que les éditions scientifiques papier produisent un nouvel objet en guise de mise à jour, l'édition numérique est modifiée, allant parfois jusqu'à remplacer la version précédente, soulevant dès lors la question du *versioning*. Cette problématique a déjà été abordée par la TEI, qui permet la création d'un journal des versions dans l'élément <versioningDesc>, énumérant les dates et les responsabilités des changements, sans toutefois permettre de remonter à une version antérieure. Pour remédier à ce problème, E. Pierrazzo plaide en faveur d'une gestion du *versioning* au sein des logiciels d'édition numérique, assurant ainsi l'accès aux versions antérieures [63]. Cette gestion du *versioning* peut également être réalisée par l'archivage des fichiers sources.

23 Enfin, si les éditions numériques n'entraînent pas de frais d'impression, les coûts d'hébergement sur un serveur et de maintenance ne doivent pas être sous-estimés. L'utilisation de systèmes informatiques rend ces objets vulnérables à l'obsolescence. Là où un manuscrit se conserve des centaines d'années, la conservation des objets numériques est plus complexe et son obsolescence liée à différents facteurs, tels que l'obsolescence « fonctionnelle » (fichiers endommagés), l'obsolescence de l'entité qui assurait la mise en ligne, ou encore l'obsolescence technologique. Face à ce constat, il nous semble essentiel d'assurer, avant tout, la pérennité des données de l'édition dans des fichiers au format le plus simple et standard possible, comme le XML TEI, dans des dépôts pérennes, et de leur associer des DOI pour faciliter leur accès et leur citabilité (Nakala [64], Zenodo, etc.).

24 Ainsi, l'édition numérique est encore en construction et cherche les critères de sa scientificité. La nouvelle économie dans laquelle elle s'inscrit n'est plus uniquement une économie du texte, mais une économie des données qui demande d'anticiper leur curation et la maîtrise des principes FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) [65].

Vers un changement d'échelle

25 Avec la montée en puissance de l'intelligence artificielle et l'automatisation des chaînes d'acquisition textuelle, une transformation significative se dessine en termes de taille des corpus. Alors que l'édition se limitait autrefois à une œuvre ou à une section cohérente d'un ensemble plus vaste, l'idée d'éditer des œuvres sérielles ou complètes gagne du terrain. Même la très établie TEI s'est adaptée et cherche à fournir un encodage de base comprenant des informations essentielles pour les corpus destinés à une lecture à distance [66]. Cette capacité à accumuler des données textuelles pourrait faciliter la conciliation entre des approches du texte en tant que document et du texte en tant que monument, en permettant d'acquérir plus de texte plus vite et en facilitant, par exemple, une comparaison automatisée des témoins manuscrits, notamment par le biais d'une constitution semi-automatique du stemma. Cependant, cette nouvelle approche du texte comme données risque d'accentuer

le fossé entre les philologues traditionnels et les humanistes numériques :

> As I have argued elsewhere, it is difficult to treat texts responsibly as "data" when much of our data set is inaccurate, whether because of faulty editing or because of the lack of digitization of certain types of texts, particularly those by what we might think of as non-canonical authors. The tensions between such approaches threaten to create splits between digital editing and digital humanities reminiscent of the textual studies wars of the second half of the twentieth century. [67]

26 C'est pourquoi il est important que les spécialistes du texte s'investissent dans ces problématiques afin de contribuer à l'élaboration d'outils numériques (modèles de reconnaissance automatique d'écriture, reconnaissance d'entités nommées, assistance à la collation, alignement de traduction, annotations linguistiques, etc.). L'objectif est de guider la communauté scientifique, qu'elle soit traditionnelle ou spécialisée dans les humanités numériques, pour rendre accessibles un nombre accru de textes, peut-être moins canoniques, tout en respectant les normes scientifiques établies par la communauté, sans laisser des géants commerciaux s'emparer d'une forme de numérisation plus rentable que scientifique des fonds patrimoniaux [68].

Acquisition automatique de texte

27 Grâce à des outils comme *Transkribus* [69] et *eScriptorium* [70], l'acquisition automatique de texte (ATR) permet de travailler sur des documents historiques complexes. Avec des corpus d'entraînement de qualité, les modèles de reconnaissance d'écriture manuscrite (HTR – Handwritten Text Recognition) peuvent atteindre des taux de précision de 92 % à 98 %, voire 99 %, faisant de l'acquisition automatique de texte une tâche résolue du point de vue informatique [71]. Actuellement, l'enjeu principal réside dans la production de données d'entraînement de qualité pour développer des modèles répondant aux critères scientifiques du monde de la recherche.

28 Pour maximiser la qualité des transcriptions, il est impératif d'utiliser des données d'entraînement de première qualité, en établissant des normes de transcription. Ces normes incluent la distinction ou non des allographes, le traitement des abréviations et la préservation des ligatures, mais aussi le choix d'un set de caractères de préférences unicode pour représenter les sources, choix qui doivent être dictés par des critères scientifiques, la nature des sources et les objectifs scientifiques [72]. Le consortium pour la reconnaissance d'écriture manuscrite des matériaux anciens (CREMMA) a élaboré des réflexions approfondies sur la manière de transcrire les vérités de terrain [73]. Deux documents ont émergé de ces efforts : un guide dédié aux documents médiévaux [74] et un ensemble de recommandations pour les documents modernes [75].

29 D'importants jeux de données ont également été produits grâce à la collaboration entre différents projets tels que CREMMA, FoNDUE (FOrmes Numérisées et Détection Unifiée des Écritures) [76], Gallic(orpor)a [77], HTRomance [78] et HTRogène [79]. Ces corpus, englobant une diversité de sources du Moyen Âge aux imprimés du XX^e^ siècle, respectent les normes de transcription précitées, avec quelques ajustements en fonction de la spécificité des sources traitées. Ces corpus représentent une ressource précieuse pour l'entraînement de modèles d'ATR et ont permis l'entraînement de modèles génériques [80] :

- *Gallicorpora+*, pour les imprimés français du XVI^e^ au XIX^e^ siècle [81] ;
- *CATMuS Medieval*, modèle multilingue pour les manuscrits médiévaux compris entre le X^e^ et le XV^e^ siècle [82] ;
- *HTR-United - Manu McFrench* pour les écritures manuscrites modernes françaises [83].

30 Lors de l'utilisation d'un modèle HTR, il est crucial de prendre en compte la technologie sous-jacente, la compatibilité d'environnement et les scores de précision. Ces scores sont dérivés de la capacité du modèle à prédire une transcription sur un échantillon des données d'entraînement réservées à cette fin. La composition du corpus d'entraînement doit être soigneusement considérée pour interpréter correctement le score. Plus le corpus est homogène, plus les scores peuvent être élevés, même avec un corpus d'entraînement restreint, car la précision est évaluée sur un échantillon très similaire au corpus

d'entraînement. À l'inverse, plus un corpus est hétérogène, plus les scores élevés deviennent difficiles à atteindre, mais cela augmente les capacités du modèle à fonctionner sur un corpus inconnu. Les modèles génériques, tels que ceux mentionnés précédemment, conçus pour être utilisés sur de nouveaux corpus, peuvent également être affinés pour répondre à des besoins spécifiques avec un investissement relativement faible en données d'entraînement [84]. Ainsi, le défi actuel consiste à élaborer des modèles génériques capables de traiter des sources les plus variées possibles, objectif que le projet CATMuS [85] tente de relever en développant des modèles multilingues couvrant une vaste période historique.

Les chaînes d'acquisition textuelle

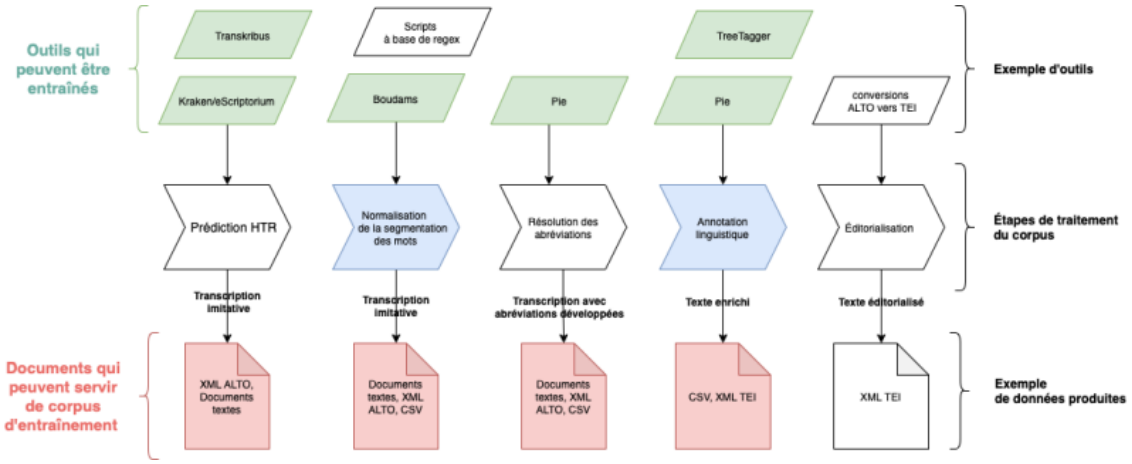
31 L'acquisition automatique de texte à partir de sources numérisées a profondément transformé le rôle de l'éditeur dans la production du texte numérique. La capacité d'acquérir des textes à grande échelle, y compris ceux d'origine manuscrite, a conduit les projets de recherche à aborder des corpus de plus en plus vastes. En outre, l'assise actuelle du standard XML TEI facilite l'émergence de protocoles de publication en ligne [86], notamment avec l'apparition de *TEIPublisher* [87] qui s'appuie sur la modélisation des données en TEI. Enfin, la démocratisation d'outils comme *eScriptorium* ou *Trankribus* a également favorisé la mise en place de protocoles pour transformer les XML ALTO de l'ATR en XML TEI en vue de la mise en ligne des corpus [88]. Le projet *DiScholEd* illustre cette dynamique en visant à établir un protocole facilitant la production de textes et d'éditions en ligne à partir de leur numérisation [89]. Ainsi, l'édition en ligne de la *Correspondance de Constance de Salm* [90], dont le texte a été acquis automatiquement, propose un système de navigation, une édition critique et une édition diplomatique, ainsi qu'un travail sur les entités nommées, avec un accès facilité aux index et à des cartes. La flexibilité de cette approche et son adaptabilité à la source permettent de proposer des protocoles communs tout en évitant une standardisation excessive de l'interface qui ne serait pas en accord avec l'interprétation proposée par l'éditeur [91].

32 Les chaînes d'acquisition textuelle, intégrant une étape d'ATR ou non, peuvent être agrémentées d'une série de normalisations et

d'enrichissements générés automatiquement [92], comprenant le développement des abréviations, la normalisation de la segmentation des mots, l'annotation linguistique et des entités nommées, l'ajout de métadonnées et proposer un balisage automatique vers la TEI en s'appuyant sur les informations de mise en page produites lors de l'ATR [93] (voir figure 2) pour optimiser la hiérarchisation, l'enrichissement et la fouille du texte.

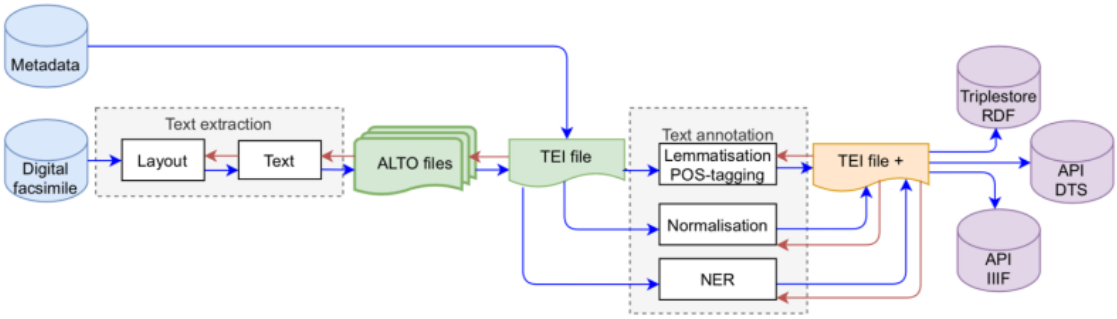
33 Le projet *Gallic(opor)a* [94] illustre ces ambitions, bien que seuls les objectifs liés à l'ATR et à la pré-éditorialisation en TEI aient pu être atteints dans le cadre du financement initial.

Figure 2 : Exemple de protocole, issu de Ariane Pinche, « Guide de transcription pour les manuscrits du X^e^ au XV^e^ siècle », 2022



34 Des initiatives plus récentes, telles que la thèse de Matthias Gille Levenson [95], explorent des chaînes éditoriales semi-automatiques incluant une assistance à la collation. Son logiciel, *teiCollator* [96], décompose le processus en plusieurs étapes, de la transcription du témoin à la réalisation d'une édition critique consultable en PDF via LaTeX.

***Figure 3 : Chaîne de traitement des données dans le projet Gallic(orpor)a,* issue d'Ariane Pinche, Kelly Christensen et Simon Gabay, « Between automatic and manual encoding Towards a generic TEI model for historical prints and manuscripts », *TEI Conference and Members&apos ; Meeting*, Newcastle, 2022.**



35 Ainsi, l'évolution rapide des techniques d'ATR et des chaînes éditoriales numériques redéfinit le paysage de l'édition, plaçant l'éditeur au cœur d'une démarche de traitement de vastes corpus. L'acquisition automatisée permet désormais d'explorer des textes, qu'ils soient originalement imprimés ou manuscrits, à une échelle inédite, qui appelle une évolution des pratiques éditoriales. Les chaînes éditoriales, de plus en plus sophistiquées, englobent des métadonnées variées et des outils automatisés, créant ainsi un écosystème éditorial plus complet. En outre, l'éditeur, désormais confronté à des corpus de grande envergure, est invité à collaborer avec des experts en intelligence artificielle, traitement automatique du langage, et autres domaines connexes. Cette évolution, loin de reléguer la philologie traditionnelle au second plan, offre l'opportunité de construire des méthodes éditoriales solides, tout en explorant des approches de lecture adaptées à divers profils d'utilisateurs. Ainsi, en embrassant ces transformations, l'éditeur peut non seulement relever les défis de l'ère numérique, mais également contribuer activement à l'élaboration d'une édition numérique accessible, riche et en constante évolution.

36 Les évolutions marquantes et les débats philologiques ont façonné l'édition numérique depuis les années quatre-vingt jusqu'à nos jours [97]. Tout au long de cette analyse, nous avons souligné la pertinence de l'édition numérique dans les débats philologiques. Notre examen a mis en lumière les modifications induites par les

innovations techniques, transformant la perception du texte et redéfinissant l'objet éditorial. Cependant, ces avancées n'ont pas été sans susciter de nouvelles difficultés, perturbant des méthodes séculaires et demandant encore aujourd'hui de réfléchir à l'évaluation, la pérennisation, l'accessibilité et la citabilité des éditions numériques. Enfin, nous espérons avoir montré comment le passage au format numérique, suivi de l'émergence des chaînes éditoriales automatisées, représente aujourd'hui une opportunité inédite. Cette évolution permet la constitution de corpus d'une ampleur sans précédent et facilite l'accès aux textes, répondant ainsi aux besoins d'un public toujours plus vaste. Toutefois, cette transition souligne également la nécessité de continuer à adapter nos pratiques et réflexions aux défis posés par ces nouvelles méthodologies, afin de garantir la qualité des éditions numériques dans le paysage académique.

NOTES

1 Notre propos ne traitera pas de l'écriture numérique ni des circuits d'édition commerciale. Édition scientifique désigne, ici, les éditions critiques qui s'intéressent à l'histoire des différents témoins du texte, les éditions génétiques qui s'intéressent aux différentes strates du texte au cours de son processus de création, mais également les éditions qui donnent un contexte à l'œuvre : matérialité du support, illustrations, réception, etc.

2 Jerome McGann, *Radiant textuality: literature after the World Wide Web*, New York, 2001, p. 55.

3 Hans Walter Gabler, « 6. Theorizing the Digital Scholarly Edition », dans *Id.*, *Text Genetics in Literary Modernism and Other Essays*, Cambridge, Open Book Publishers, 2019, p. 121-141.

4 Sotera Fornaro, « Karl Lachmann et sa méthode », *Revue germanique internationale*, 2011, p. 125-138. Paolo Trovato et Michael D. Reeve, *Everything you always wanted to know about Lachmann's method: a non-standard handbook of genealogical textual criticism in the age of post-structuralism, cladistics, and copy-text*, Limena, Libreriauniversitaria.it edizioni, 2014.

5 Richard J. Tarrant, « Classical Latin litterature », dans David C. Greetham (dir.), *Scholarly editing : a guide to research*, New York, The Modern Language Association of America, 1995, p. 96.

6 Richard J. Tarrant, *op. cit.*

7 Bernard Cerquiglini, *Éloge de la variante : histoire critique de la philologie*, Paris, Seuil, 1989.

8 Elena Pierazzo, « 3. Modelling Digital Scholarly Editing: From Plato to Heraclitus », dans Matthew James Driscoll, (dir.), *Digital Scholarly Editing: Theories and Practices*, Cambridge, Open Book Publishers, 2017, p. 41-58.

9 Peter Shillingsburg, « Author, Texts, and Polemics of Textual Criticism », dans *Id.*, *Devils and angels: Textual editing and literary theory*, Charlottesville, 1991, p. 26.

10 John Bryant, « Witness and Access: The Uses of the Fluid Text », *Textual Cultures*, vol. 2 1, 2007, p. 16-42.

11 « La critique génétique envisage l'œuvre littéraire comme processus d'écriture ; […] en conséquence, son objet n'est pas le texte édité et clos, mais la diversité et la multiplicité des traces de ce processus, qui, comme des témoins muets, nous sont livrées dans les fonds d'archives littéraires, notamment des auteurs modernes. » : Bénédicte Vauthier, « Éditer des états textuels variants », *Genesis. Manuscrits – Recherche – Invention*, n° 44, mai 2017, p. 39-55., texte traduit de l'allemand d'Almuth Grésillon, « "Critique génétique" : Probeartikel », Beihefte zu Editio, n° 36, 2013, p. 197.

12 Jerome McGann, *op. cit.*, p. 56.

13 Dans les éditions des œuvres de Tertullien ou l'édition des commentaires de Servius, l'apparat peut occuper jusqu'au trois quarts de la page, rendant la lecture du texte fastidieuse. Voir Tertullien, *Le manteau*, éd. Marie Turcan, Paris, France, Le Cerf, 2007. Paul Wessner (éd.), *Donat, Aeli Donati quod fertur Commentum Terenti : Accedunt Eugraphi commentum et Scholia Bembina*, Leipzig, B. G. Teubner, 1905.

14 « Pour des raisons économiques, le papier est contraint de se limiter à un seul état textuel (*single text editions, one text editions*), qu'il soit ou non reconstruit. Grâce au "multifenêtrage" et à l'hypertexte, le numérique s'est imposé comme le médium approprié et indispensable à l'application des "nouvelles" théories textuelles. » : Frédéric Duval, « Pour des éditions numériques critiques. L'exemple des textes français », *Médiévales*, n° 73,

2017, Le *texte à l'épreuve du numérique*, dossier thématique sous la direction d'Anne Rochebouet, p. 13-30.

15 Pierre Mounier, « Manifeste des *Digital Humanities* », *Journal des anthropologues*, n° 122-123, 2010, p. 447-452.

16 Hans Walter Gabler, *op. cit.*

17 « *It pushes traditional scholarly models of editing and textuality beyond* » : Dino Buzzetti et Jerome McGann, « Critical Editing in a Digital Horizon », dans Lou Burnard, John M. Unsworth et Katherine O'Brien O'Keeffe (dir.), *Electronic textual editing*, New York, Modern Language Association of America, 2006.

18 Patrick Sahle, « 2. What is a Scholarly Digital Edition? », dans Matthew James Driscoll et Elena Pierazzo (dir.) *Digital Scholarly Editing: Theories and Practices*, Cambridge, Open Book Publishers, 2017, p. 19-39.

19 Jonas Carlquist, « Medieval Manuscripts, Hypertext and Reading. Visions of Digital Editions », *Literary and Linguistic Computing*, n° 19 / 1, avril 2004, p. 105-118. G. Franzini, S. Mahony et M. Terras, « A Catalogue of Digital Editions », dans *Matthew James Driscoll et Elena Pierazzo* (dir.), *op.cit.*, p. 161-182.

20 Peter Robinson, « Towards a Theory of Digital Editions », *Variants. The Journal of the European Society for Textual Scholarship*, n°10, 2013, p. 123.

21 « Éditions en ligne de l'École des chartes (ELEC) », consulté le 3 janvier 2024, (http://elec.enc.sorbonne.fr/).

22 Mats Dahlström, « How Reproductive is a Scholarly Edition? », *Literary and Linguistic Computing*, n°19/1, avril 2004, p. 17-33.

23 Peter Robinson et Elizabeth Solopova, « Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue », *The Canterbury Tales Project. Occasional Papers*, juillet 1993. Edition en ligne consultée le 3 janvier 2024 (https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=572872612f33aaae5717c128dc2d73567e539e76)

24 Peter Robinson, « What is a critical digital edition? », *Variants: The Journal of the European Society for Textual Scholarship*, n°1, janvier 2002, p. 43-62.

25 Elena Pierazzo, *Digital scholarly editing: theories, models and methods*, Londres, Routledge, 2015.

26 Ariane Pinche, Bruno Bureau et Christian Nicolas, « Hyperdonat, digital edition project », *TEI Conference and Members' Meeting 2016*,

septembre 2016, Vienne (Autriche) ⟨hal– 01413479⟩. Le projet est resté au stade expérimental.

27 Richard Trachsler et Lino Leonardi, « L'édition critique des romans en prose : le cas de Guiron le Courtois », *Manuel de la philologie de l'édition*, Berlin/Boston, De Gruyter, 2015, p. 44-80.

28 <https://obvil.sorbonne-universite.fr/bibliotheque>.

29 Voir 2.1 U*n standard* : XML *TEI.*

30 <https://github.com/OBVIL>.

31 Frédéric Glorieux, « Dramagraphie (sur le pamphlet 6 de Franco Moretti) », *J'attends des résultats*, 2016, en ligne et consulté le 8 octobre 2024 (https://doi.org/10.58079/tod3).

32 Gregory Crane, « The Perseus Digital Library and the future of libraries », I*nternational Journal on Digital Libraries*, n° 24, juin 2023, p. 117-128, en ligne consulté le18 décembre 2023 (https://www.perseus.tufts.edu/hopper/ ; https://doi.org/10.1007/s00799-022-00333-2)

33 *Uniform Resource Name*, standard informatique qui permet d'identifier une ressource indépendamment de sa localisation et de son accessibilité par internet, ce qui permet à cet identifiant d'être pérenne.

34 Christiane Marchello-Nizia, Alexey Lavrentiev et Céline Guillot-Barbance, « 6. Édition électronique de la Queste del saint Graal », *Manuel de la philologie de l'édition*, Berlin/Boston, De Gruyter, 2015, p. 155-176 ; Christiane Marchello-Nizia et Alexei Lavrentiev (éd.), *Queste del saint Graal*, Lyon, ENS éditions, 2019. Publié en ligne par la Base de français médiéval, http://catalog.bfm-corpus.org/qgraal_cm, Dernière révision le 2018-11-30.

35 Céline Guillot, Serge Heiden et Alexei Lavrentiev, « Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique », *Diachroniques. Revue de Linguistique. française diachronique*, n° 7, 2018, p. 168.

36 Soit un équivalent de ce que nous appelons la nature grammaticale.

37 Michael Robertson, « The Walt Whitman Archive », *Journal of American History*, vol. 99/3, décembre 2012, p. 1019-1020. Article consulté en ligne le 18 décembre 2023 (https://whitmanarchive.org/about/index.html).

38 Jerome McGann, « Dante Gabriel Rossetti and the Betrayal of Truth », *Victorian Poetry*, vol. 26/4, 1988, p. 339-361. Article consulté en ligne le 18 décembre 2023 (http://www.rossettiarchive.org/index.html).

39 Voir le cas du tableau et du poème : « The Girlhood of Mary Virgin », produits sensiblement à la même époque (1847–1848), en ligne consulté le 25 mars 2024 (http://www.rossettiarchive.org/docs/s40.rap.html).

40 On peut également citer le projet *Blake archive* qui donne accès aux illustrations de ces œuvres « illuminées » pour rendre au mieux la richesse des œuvres de l'auteur (https://www.blakearchive.org).

41 « Shelley-Godwin Archive », site consulté le 20 décembre 2023 (http://shelleygodwinarchive.org/about/). Le projet est le résultat d'un partenariat entre la New York Public Library et le Maryland Institute for Technology in the Humanities, en coopération avec la Bodleian Library d'Oxford, la Huntington Library, la British Library, la Houghton Library et le Victoria and Albert Museum.

42 Exemple de page issu du projet : « Shelley, M. "Frankenstein, MS. Abinger C. 56 », dans The Shelley-Godwin Archive, MS. Abinger c. 56, 2v. Retrieved from http://shelleygodwinarchive.org/sc/oxford/ms_abinger/c56/#/p8.

43 « Édition des manuscrits de Madame Bovary de Flaubert | Transcriptions | Classement génétique », en ligne et consulté le 20 décembre 2023 (https://www.bovary.fr/).

44 Jerome McGann, *op. cit.*, p. 25.

45 Anne Mangen, « Hypertext fiction reading: haptics and immersion », *Journal of Research in Reading*, vol. 31/4, 2008, p. 404-419.

46 Nancy M. Ide et Michael Sperberg-McQueen, « The TEI: History, Goals, and Future », *Computers and the Humanities*, vol. 29/1, 1995, p. 515.

47 Greta Franzini, Simon Mahony et Mélissa Terras, « A Catalogue of Digital Editions », dans Matthew James Driscoll et Elena Pierazzo (dir.), *Digital Scholarly Editing : Theories and Practices*, Cambridge, Open Book Publishers, 2017, p. 161-182.

48 Daniel Apollon et Claire Bélisle, « Le destin de l'appareil critique dans l'édition numérique scientifique », dans Daniel Apollon, Philippe Régnier et Claire Bélisle (dir.), *L'édition critique à l'ère du numérique*, Paris, L'Harmattan, 2017, p. 105.

49 Franco Moretti, *Distant reading*, Londres, Verso, 2013. Le *distant reading* propose d'appréhender le texte de manière quantitative avec des analyses de réseaux, des analyses lexicométrique ou stylométrique.

50 Ioana Galleron et Fatiha Idmhand, « De l'interopérabilité à la réutilisabilité des éditions électroniques », *Humanités numériques*, n° 1, 2020, mis en ligne le 1er janvier 2020, consulté le 10 octobre 2024 (http://journals.openedition.org/revuehn/350 ; DOI : https://doi.org/10.4000/revuehn.350).

51 Identification des entités nommées, annotations morphosyntaxiques.

52 Ariane Pinche, « Exploitations et valorisations des données numériques connexes à l'édition », dans Robert Alessi, Marcello Vitali-Rosati (dir.), *Les éditions critiques numériques : entre tradition et changement de paradigme*, Montréal, Presses de l'Université de Montréal, 2023, p. 133-153.

53 Paul A. Broyles, « Digital Editions and Version Numbering », *Digital Humanities Quarterly*, vol.14/2, juin 2020, article en ligne consulté le 11 octobre 2024 (https://hcommons.org/deposits/objects/hc:38104/datastreams/CONTENT/content) : « Digital editions, as I understand them, are simultaneously scholarly publications and data sources ».

54 Joana Casenave, « La fonction de l'éditeur-auteur dans les éditions critiques numériques », *Humanités numériques*, n° 6, 2022, mis en ligne le 1er décembre 2022, consulté le 14 octobre 2024 (http://journals.openedition.org/revuehn/3241).

55 Peter Robinson, « Electronic editions for everyone », *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, Openbooks Publisher, 2010, p. 145-163.

56 « Consortium Cahier – Corpus d'auteurs pour les humanités : informatisation, édition, recherche » en ligne, consulté le 15 octobre 2024 (https://cahier.hypotheses.org/).

57 « Ariane », document en ligne consulté le 22 décembre 2023 (https://cst-ariane.huma-num.fr/).

58 « Guidelines for Editors of Scholarly Editions | Modern Language Association », document en ligne consulté le 8 décembre 2023 (https://web.archive.org/web/20190803225011/https://www.mla.org/Re
Reports-and-Other-Documents/Publishing-and-Scholarship/Reports-from-the-MLA-Committee-on-Scholarly-Editions/Guidelines-for-Editors-of-Scholarly-Editions).

59 Philippe Régnier, « Les enjeux de l'édition critique numérique », dans Daniel Apollon, Philippe Régnier, Claire Bélisle (dir.), *L'édition critique à l'ère du numérique*, Paris, L'Harmattan, 2017, p. 79-100.

60 Frédéric Duval, *op. cit.*

61 Jerome McGann, *op. cit.*

62 Peter Robinson, « Where We Are with Electronic Scholarly Editions, and Where We Want to Be », *Jahrbuch für Computerphilologie*, vol. 5/5, 2003, p. 126-146 : « *These editions will not be made or maintained by one person or by one group, but by a community of scholars and readers working together: they will be the work of many and the property of all.* »

63 Elena Pierazzo, *op. cit.*, p. 185-186.

64 Nakala est un entrepôt de données de recherche pour les sciences humaines et sociales et offre des services sur plusieurs étapes du cycle de vie des données de recherche en SHS : sur leur préservation, leur publication et leur réutilisation. Le service est maintenu par Huma-Num.

65 <https://www.go-fair.org/fair-principles/>.

66 Lou Burnard, Christof Schöch et Carolin Odebrecht, « In search of comity: TEI for distant reading », *Journal of the Text Encoding Initiative* [en ligne], n°14, April 2021 – March 2023, mis en ligne le 8 juillet 2021, consulté le 15 octobre 2024 (http://journals.openedition.org/jtei/3500).

67 Amy E. Earhart, « The Digital Edition and the Digital Humanities », *Textual Cultures*, vol. 7 / 1, 2012, p. 18-28.

68 Robert Darnton, A*pologie du livre : demain, aujourd'hui, hier*, trad. Jean-François Sené, Paris, Gallimard, 2010, p. 100-109.

69 Philip Kahle, Sebastian Colutto, Günter Hackl, *et ali*, « Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents », *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 04, 2017, p. 19-24. En ligne (https://ieeexplore.ieee.org/abstract/document/8270160)

70 Peter A. Stokes, Benjamin Kiessling, Daniel Stökl Ben Ezra, *et al.*, « The eScriptorium VRE for Manuscript Cultures », *Classics@ Journal*, vol.18, 2021, en ligne (https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/).

71 « *From a computer science point of view, the recognition of handwriting seems to be a resolved task. The latest recognition engines allow for the*

successful recognition of specifically trained hands producing a text as reusable data. », Tobias Hodel, David Schoch, Christa Schneider, *et ali*, « General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example », *Journal of Open Humanities Data*, juillet 2021, p. 13, en ligne (https://boris.unibe.ch/157474/3/46-597-1-PB.pdf).

72 En effet, comme le remarquait déjà Peter Robinson le passage au format numérique demande un effort de traduction en caractère dédié. Dans la lignée de ces réflexions, on peut également cité le projet TypoRef : Rémi Jimenes, « TypoReF : étudier les matériels d'imprimerie de la Renaissance française », *Bibliothèques Humanistes*, en ligne depuis le 24 mai 2023 (https://bvh.hypotheses.org/9654).

73 Le projet CREMMA est un projet partenaire de l'équipe Scripta (EPHE-PSL) à l'origine du développement de *eScriptorium* et de *Kraken* (Benjamin Kiessling, « Kraken – an Universal Text Recognizer for the Humanities », Abstract of paper 0673 presented at the Digital Humanities Conference 2019 (DH2019), Utrecht, the Netherlands 9-12 July, 2019, en ligne (https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/Z9G2EX)).

74 Ce guide est le fruit d'un séminaire de recherche qui s'est déroulé à l'École nationale des chartes en 2021-2022 : Ariane Pinche, « Guide de transcription pour les manuscrits du X^{e} au XV^{e} siècle », 2022, en ligne (⟨hal-03697382⟩).

75 Alix Chagué, Thibault Clérice et CREMMA, « Règles générales de transcription pour les corpus CREMMA » article en ligne depuis le 23 septembre 2022 (https://gist.github.com/alix-tz/6f89444521bf1cab0522da520f7e4ff4.)

76 Simon Gabay, Pierre Kuenzli, Jean-Luc Flacone et Christophe Charpilloz, « FoNDUE : Documentation, University of Geneva », en ligne depuis 2022 (https://github.com/fonDUE-HTR/Documentation).

77 Benoît Sagot, Laurent Romary, Rachel Badwen, *et ali*, « Gallic(orpor)a: extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue », en ligne depuis 2022 et consulté le 16 décembre 2023 (https://gallicorpora.github.io/).

78 « The HTRomance project », en ligne et consulté le 10 décembre 2023 (https://htromance-project.github.io/).

79 « HTRogène », en ligne, consulté le 10 décembre 2023 (https://projet.bibliss ima.fr/fr/appels-projets/projets-retenus/htrogene).

80 Tous les modèles cités ci-dessous ont été entraînés avec Kraken.

81 Ariane Pinche et Simon Gabay, « Segmentation and HTR Model », *Gallicorpora*, 2023, en ligne, consulté le 16 décembre 2023 (https://github.com/Gallicorpora/ Segmentation-and-HTR-Models).

82 Ariane Pinche, Thibault Clérice, Alix Chagué, *et ali*, « CATMuS Medieval », *Zenodo*, en ligne depuis novembre 2023 (https://zenodo.org/records/12743230).

83 Alix Chagué et Thibault Clérice, « HTR-United - Manu McFrench V1 (Manuscripts of Modern and Contemporaneous French) », *Zenodo*, en ligne depuis juin 2022.

84 Environ cinq images contenant une trentaine de lignes suffisent, pour un coût de calcul relativement modéré, ce qui permet d'entraîner le modèle sur une machine personnelle sans nécessiter de carte graphique ou de serveurs dédiés, voir Ariane Pinche, « Generic HTR Models for Medieval Manuscripts. The CREMMALab Project », *Journal of Data Mining & Digital Humanities*, en ligne depuis le 16 octobre 2023 (https://jdmdh.episciences.org/11592).

85 Ariane Pinche, Thibault Clérice, Alix Chagué, *et ali*, « CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts: A generalized set of guidelines and models for Latin scripts from Middle Ages (8th – 16th century) », *DH2024*, ADHO, Aug 2024, Washington DC, en ligne ⟨hal-04346939⟩.

86 Anne Baillot et Julie Giovacchini, « TEI Models for the Publication of Social Sciences and Humanities Journals: Opportunities, Challenges, and First Steps Toward a Standardized Workflow », *Journal of the Text Encoding Initiative*, en ligne depuis mars 2023 (https://journals.openedition.org/jtei/3419).

87 Amit Kumar, Susan Schreibman, Stewart Arneil, *et ali*, « <teiPublisher>: A Repository Management System for TEI Documents », *Literary and Linguistic Computing*, vol. 20/1, mars 2005, p. 117-132.

88 Hugo Scheithauer, Alix Chagué et Laurent Romary, « From eScriptorium to TEI Publisher », *Brace your digital scholary edition!*, en ligne depuis novembre 2021 (hal-03538115).

89 La chaîne a été mise en place par Floriane Chiffoleau, sous la supervision d'Anne Baillot et de Laurent Romary, avec l'aide d'Alix Chagué et

de Manon Ovide.

90 Exemple : « Letter of Constance de Salm to the King of Prussia Friedrich Wilhelm III. (Dyck, June 24th 1824) ». Encadrement scientifique et technique du projet Anne Baillot Mareike König Floriane Chiffoleau. Réalisation de la chaîne de traitement par Sébastien Biay, <https://discholed.huma-num.fr/exist/apps/discholed/index_cds.html?collection=cds%2Fcorpus>.

91 Edward Vanhoutte, « 5. Defining Electronic Editions: A Historical and Functional Perspective », *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, Cambridge, Open Book Publishers, 2013, p. 119-144.

92 Des contrôles qualité sont toutefois vivement conseillés.

93 Le projet *SegmOnto* propose un vocabulaire contrôlé pour décrire ces « zones », favorisant la création de jeux de données compatibles, la collaboration entre différents projets, et la mise en place de scripts de transformation, Simon Gabay, Ariane Pinche, Kelly Christensen, *et ali*, « SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles », 2023, en ligne ⟨hal– 04343404⟩.

94 Ariane Pinche, Kelly Christensen et Simon Gabay, « Between automatic and manual encoding Towards a generic TEI model for historical prints and manuscripts », *TEI 2022 conference: Text as data*, en ligne depuis 2022 ⟨hal– 03780302⟩. Kelly Christensen, Ariane Pinche et Simon Gabay, « Gallic(orpor)a: Traitement des sources textuelles en diachronie longue de Gallica », DataLab de la BnF, Paris, 2022, en ligne ⟨hal– 03716534⟩.

95 Matthias Gille Levenson, *Le Regimiento de los prínçipes et sa glose : étude et édition numérique de la partie sur le gouvernement de la cité en temps de guerre (III, 3)*, thèse de doctorat, École normale supérieure de Lyon, 2023.

96 Matthias Gille Levenson, « TeiCollator : une chaîne de traitement ecdotique semi-automatisée », XXX^{e} *Congrès International de Linguistique et Philologie Romane*, 2022, en ligne (hal– 03715059).

97 Cet exposé s'est efforcé de dresser un vaste panorama des possibilités offertes par les éditions numériques, en présentant une gamme variée de projets allant des pionniers de l'édition numérique à des projets actuels. Afin d'illustrer les cas pratiques de la troisième partie, nous avons mis en avant des initiatives avec lesquelles nous avons collaboré ou échangé, et dont les fichiers XML TEI, ainsi que les scripts et la documentation, sont librement accessibles.

RÉSUMÉS

Français

La philologie est une discipline complexe et variée. L'introduction de cet exposé aborde rapidement les diverses pratiques de l'édition savante, mettant en évidence les tensions méthodologiques entre les approches centrées sur le texte et sa transmission et celles davantage orientées vers le document et sa matérialité. À travers un panorama des parcours de lecture proposés par les pionniers de l'édition numérique tels que *Perseus*, *La Queste del saint Graal* ou encore *The Rossetti Archive*, nous examinons les liens qu'ils entretiennent avec les débats de la philologie traditionnelle, ainsi que les réponses que les éditions numériques ont cherché à apporter, de la simple mise à disposition des textes à la création de portails mettant en avant le processus créatif des auteurs. Dans un second temps, nous exposons les critères scientifiques, les standards tels que XML TEI, et les usages des éditions numériques. Cette partie met en lumière les nouveaux enjeux auxquels l'édition numérique doit faire face, notamment dans la transition du texte vers des données pérennes, lisibles et exploitables par les ordinateurs, ouvrant ainsi le texte à des usages inconnus des éditions traditionnelles. Enfin, nous abordons l'impact des innovations technologiques sur les systèmes de production textuelle. L'utilisation de l'intelligence artificielle entraîne des changements méthodologiques dans le domaine de la philologie, notamment en termes de taille de corpus, en facilitant son acquisition et son enrichissement. Ce changement amène également de nouvelles pratiques de traitement du texte avec la mise en place de chaînes d'acquisition plus ou moins automatisées de la numérisation des sources à la mise à disposition d'un texte numérique.

English

Philology is a complex and diverse discipline. This introduction of this article briefly explores various scholarly editing practices, highlighting methodological tensions between approaches centered on the text and its transmission, and those more focused on the document and its materiality. Through a survey of reading pathways offered by digital editing pioneers such as Perseus, La Queste del saint Graal, and The Rossetti Archive, we examine the connections they have with traditional philological debates, as well as the responses digital editions have sought to provide: from simply making texts available to creating portals that showcase authors' creative processes. Next, we delve into the scientific criteria, standards such as XML TEI, and the uses of digital editions. This section sheds light on the new challenges digital editing faces, especially in transitioning text into data that is durable, readable, and usable by computers, thereby opening up the text to uses unknown to traditional editions. Finally, we discuss the impact of technological innovations on textual production systems. The use of artificial intelligence brings about methodological changes in the field of

philology, particularly in terms of corpus size, facilitating its acquisition and enrichment. This shift then introduces new text processing practices with the implementation of more or less automated acquisition pipelines, from digitizing sources to providing digital text.

INDEX

Mots-clés
édition, édition numérique, XML TEI, chaîne éditoriale, reconnaissance automatique de texte

Keywords
edition, digital édition, XML TEI, editorial pipeline, automatic text recognition

AUTEUR

Ariane Pinche
CNRS, CIHAM UMR 5648
IDREF : https://www.idref.fr/253121914
ORCID : http://orcid.org/0000-0002-7843-5050
HAL : https://cv.archives-ouvertes.fr/ariane-pinche
BNF : https://data.bnf.fr/fr/18166519