Texte numérique

Jean-Baptiste Camps

Texte numérique

Derrière l'apparent oxymore - des chiffres et des lettres… (Genêt 2003) - se cache une réalité aujourd'hui omniprésente, à laquelle n'échappe pas l'historien, et dont les origines (lointaines) remontent au moins au code élaboré par Émile Baudot, utilisant le principe du chiffre binaire (*binary digit* ou *bit*) qui forme l'unité de base de nos encodages de caractères actuels tels qu'Unicode, voire aux métiers à tisser du lyonnais Basile Bouchon (1725), dont les cartes perforées seront reprises pour sa machine mécanographique par Hollerith en 1887, puis par IBM.

Ce sont ces cartes perforées qui serviront aux premières initiatives de numérisation de sources anciennes. On retient généralement parmi ces projets pionniers celui de l'*Index Thomisticus* de dom Roberto Busa. Naissant dans l'esprit de Busa à la fin des années 1940, ce projet, titanesque pour l'époque, de lemmatisation et indexation complète de l'œuvre de Thomas d'Aquin, nécessitera pour être réalisé un partenariat avec IBM et l'ouverture d'une école de formation pour opératrices à Gallarate, et n'atteindra sa complétion qu'en 1980 (Busa 2004). Moins resté dans les mémoires, un autre projet pionnier concerne l'indexation des manuscrits de la Mer Morte (Tasman 1957). En dépit des contraintes techniques encore très lourdes, la décennie 1960 voit un véritable essor de l'utilisation de l'ordinateur pour le traitement des sources. L'essentiel des projets relève alors, soit du tri et de l'extraction d'informations à partir de la saisie de fiches dérivées des sources (cf. *NUMÉRIQUE), soit comme le note Froger (1968), de l'*index verborum* et de la lexicographie. L'élaboration automatique de concordances et d'index de mots occupe ainsi une place centrale dans les développements des années 1950 et 1960, et même des décennies suivantes (Burton 1981a-c; 1982), allant de pair avec la mise en place de partenariats avec IBM ou la création de centres dédiés, à l'instar du *Trésor de la langue française* (*NUMÉRIQUE).

Les limites techniques et la difficulté de représentation des textes, avec des systèmes alors essentiellement limités aux caractères non accentués de l'alphabet latin, sans distinction de casse (Hockey 2004), engendre assez tôt une réflexion sur le sujet, par exemple pour les langues sémitiques (Weil 1964-1965). Bientôt, le besoin se fait jour d'un encodage des textes dépassant le simple contenu brut des séquences de caractères, pour inclure des informations de mise en forme, ou, mieux encore, des éléments sémantiques et de structure. Motivé d'abord par le besoin de générer des concordances et identifier des portions de texte (Hockey 2004), cette recherche donna naissance à différentes solutions. COCOA en est une des plus fameuses (Russel 1967; cf. Hockey, 2004), dont l'un des apports est la capacité à définir des structures adaptées à un type particulier de document. Un classique du genre, COCOA sera perfectionné par l'*Oxford Concordance Program* (OCP) dans les années 1980 et le programme TACT (*Textual Analysis Computing Tools*) dans les années 1990 (Hockey 2004). En Allemagne, c'est à Tübingen et sous l'impulsion de Wilhelm Ott qu'est développé TÜSTEP (*Tübinger System von Textverarbeitungs-Programmen*; EADH, s.d.).

Cette réflexion implique également les milieux industriels: en 1969, pionnier des langages à

balise, naît le *Generalized Markup Language*, ou GML, développé par Goldfarb, Mosher et Lorie (Goldfarb, Mosher et Peterson 1970; Goldfarb 1996), qui sera repris par IBM comme *IBM's Document Composition Facility GML*. De GML descendent SGML (*Standard Generalized Markup Language*), devenu norme ISO en 1986 (ISO 8879 1986), puis XML (*eXtensible Markup Language*; W3C 1998).

La question d'éviter le travail en doublon, pour favoriser la cumulabilité que permet le numérique, et d'assurer la pérennité des ressources produites a occupé assez tôt les esprits, dès les décennies pionnières (cf. Delatte, 1965), comme en témoigne aussi la création de l'*Oxford Text Archive* (OTA) en 1976 (Hockey 2004). Cette préoccupation s'est traduite par la volonté de faciliter, d'un point de vue conceptuel et technique, l'échange et le partage des fichiers, qui est l'une des sources, sinon la source principale, de la conception progressive de standards. C'est ainsi qu'en 1987, à l'initiative de Nancy Ide et de l'*Association for Computing in the Humanities*, rejointes bientôt par les autres sociétés savantes du champ (ALLC et ACL), se tient une réunion au Vassar College, à Poughkeepsie, avec pour but de définir des principes pour l'établissement d'un standard de représentation et d'échange des textes (Hockey 2004). Ces « Principes de Poughkeepsie » (créations de recommandations définissant un standard d'échange, modulaire autour d'une base commune, avec sa syntaxe et son métalangage, définies collectivement par la communauté) sont à la base de la *Text Encoding Initiative*, dont les différentes propositions se succèdent, de la proposition 1 (TEI P1) en 1990 à la P5 en 2007 (TEI Consortium, 2020). D'une richesse qui confine parfois à l'excès et d'une évolutivité qui en fait un produit en perpétuelle transformation, la TEI est et demeure le cadre conceptuel et technique incontesté de la représentation savante des données textuelles. Sa modularité permet une variété d'applications, des éditions facsimilaires, génétiques ou critiques aux dictionnaires, en passant par les corpus linguistiques, les textes de performance, les notices de manuscrits, etc. Conçues par une communauté de chercheurs, les recommandations de la TEI sont en outre orientées vers les besoins de l'édition savante, même si l'importance de la communauté anglo-saxonne au sein du consortium, de même que de certains courants philologiques concomittant du développement des *Guidelines* a pu favoriser les modèles documentaires aux éditions critiques, qui se font attendre (Duval 2017).

En effet, dès les années 1980, Bernard Cerquiglini voit dans le numérique un moyen de dépasser la fixité et l'unicité, qu'il juge factice, du texte imprimé d'une édition critique. En démultipliant les états consultables du texte – éditions diplomatiques des différents témoins, éditions anciennes, conjointes à des éditions d'obédience conservatrice ou reconstructionniste –, en laissant le lecteur libre de son parcours et en y ajoutant analyses et statistiques sur les textes, le médium numérique permettrait ainsi de retrouver la fluidité et la variance du texte médiéval, qui serait sa caractéristique essentielle (Cerquiglini 1983). Son *Éloge de la variante*, republié sous forme d'ouvrage quelques années plus tard (Cerquiglini 1989), annonce la naissance d'une *New Philology* revendiquée comme telle (*Speculum* 1990), qui donne le primat aux éditions documentaires d'une source donnée, voire à une quête de l'inclusion de plus en plus d'informations sur celle-ci, qui se poursuit

jusqu'à aujourd'hui et jusqu'à plus soif – *où s'arrêter ?* se demande par exemple Elena Pierazzo (2011). La situation est pourtant moins univoque qu'il n'y paraît, et certains ne perdent pas de vue que l'accès au texte original est ce qui justifie la production d'éditions individuelles des versions scribales (Duggan 1995). La même année que l'ouvrage de Bernard Cerquiglini sort une proposition qui va transformer durablement nos sociétés, et dont l'auteur, Tim Berners-Lee, est désormais universellement connu (Berners-Lee 1989). Ainsi, dans les décennies suivantes, c'est le *World Wide Web* qui va déchaîner l'intérêt et augmenter encore, si besoin était, la focale sur l'édition électronique des textes au sein du mouvement plus large des sciences humaines computationnelles (Hockey 2004).

Dans ces entreprises, le choix technique du langage XML n'est pas anodin, et correspond à une modélisation de ce que sont les textes comme « hiérarchies ordonnées d'objet de contenu » (en anglais, *Ordered Hierarchies of Content Objects*, OHCO); autrement dit, un texte a un sens (un début et une fin) tout en ayant également une structure hiérarchique dans laquelle s'emboîtent des items de contenu, dont la nature dépend de la perspective d'analyse choisie (Renear, Mylons et Durand 1996). Ainsi, les livres contiennent des chapitres, constitués de paragraphes; les cahiers contiennent des feuillets, constitués de pages; un cartulaire contient des copies de chartes, dont la teneur peut se constituer d'un protocole, d'un texte et d'un eschatocole, décomposables en un certain nombre de clauses, dont les éléments eux-mêmes peuvent encore être décomposés, etc.

Ainsi, en dépit de l'aspect extérieurement mécanique que pourrait sembler prendre la numérisation, il n'existe pas de transformation univoque d'un document matériel, une source, à une séquence numérique lisible par l'ordinateur. Encore faut-il d'emblée distinguer, comme le propose Patrick Sahle, « édition numérisée », qui reproduit à l'écran le paradigme de l'édition imprimée, de l'édition numérique de plein droit, qu'il serait impossible d'imprimer sans supprimer de l'information ou des fonctionnalités (Sahle 2017). Car en effet le paradigme numérique, par l'explicitation et la systématisation qu'il nécessite, et son emphase sur l'annotation sémantique, fait voir plus nettement que jamais la dimension interprétative et sélective de tout acte éditorial. Car transcrire, c'est faire des choix: parmi tous les faits concernant sa source, l'éditeur en opère une sélection dont il présente les résultats sous la forme d'une édition (Sperberg-McQueen 2009; cf. aussi Pierazzo, 2011 et Camps, 2016). En ce sens, une transcription peut-être vue comme une restitution partielle, une traduction d'un système à un autre (Robinson et Solopova 1993) ou une description d'une source (Stutzmann 2011), une métadonnée.

Dès lors que l'on admet que l'édition d'une source est un acte sélectif et interprétatif, il en découle naturellement que les choix sont faits en fonction d'une perspective d'analyse donnée et pour des buts de recherche spécifiques (comme dans le modèle OHCO). L'édition électronique et la constitution de corpus a par exemple amené, pour les textes français médiévaux, à remettre en cause les règles dites de Meyer-Roques: issues d'un contexte et répondant à un questionnaire linguistique daté, elles occultent en effet de nombreux faits (segmentation, ponctuation ancienne, allographes et signes abréviatifs, notamment) qui

intéressent aujourd'hui les linguistes de l'écrit (Duval 2012; Camps 2016). Là aussi, le numérique permet de jouer sur la granularité des éléments envisagés (de l'œuvre complète au signe individuel), comme sur la quantité et la nature des informations que l'on enregistre (des identifications de personnes aux annotations linguistiques ou paléographiques), conjoignant ainsi au texte un entrepôt de données pouvant servir à l'interrogation comme à l'analyse computationnelle (*COMPUTATION).

Là où l'emphase était, dans les années 1990 et 2000, sur le numérique comme moyen de diffusion d'éditions pouvant avoir été réalisées par des moyens très traditionnels, les développements des méthodes computationnelles incitent désormais à intégrer mieux intelligence humaine et artificielle à toutes les étapes du travail sur les sources (Andrews 2012). On peut dès lors ambitionner de parvenir à une philologie centrée sur les données ou intensive en données, selon le nouveau paradigme énoncé par Jim Gray (2009), c'est-à-dire à une approche qui revisite et fasse monter en puissance les modes de production comme ceux d'analyse des données textuelles, qui restent néanmoins absolument centrales : de nouvelles données permettent ainsi de nouveaux questionnements, et l'élaboration de nouveaux savoirs scientifiques sur les textes, et, tant que les données sont rendues librement disponibles et réutilisables de manière transparente et documentée (cf. Wilkinson et alii 2016), elles permettent aussi la réalisation concrète des idéaux de reproductibilité et de cumulabilité des entreprises de recherche (Camps 2018).

[Jean-Baptiste Camps.]

Bibliographie

En pied de notice

Andrews (Tara), « The Third Way: philology and critical edition for a digital age », *Variants: the Journal of the European Society for Textual Scholarship*, 10 (2012), http://boris.unibe.ch/43071/.

Burton (D. M.), « Automated Concordances and Word Indexes », *Computers and the Humanities*, 15 (1981), p. 1-14, 83-100, 139-154, et 16 (1982), p. 195-218.

Camps (Jean-Baptiste), « Où va la philologie numérique? », *Fabula-LHT*, (20) 2018, http://www.fabula.org/lht/20/camps.html.

Digital Scholarly Editing: Theories and Practices, éd. Matthew James Driscoll et Elena Pierazzo, Cambridge, 2017, p. 19–39 (Digital Humanities Series), http://books.openedition.org/obp/3397.

Hockey (Susan), « The history of humanities computing», *A Companion to Digital Humanities…*, 2004, p. 3-19.

TEI Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, originally

edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALLC-ACL Text Encoding Initiative, now entirely revised and expanded under the supervision of the Technical Council of the TEI Consortium, version 4.0.0 (2020), rév. ccd19b0ba.

En bibliographie générale

Andrews (Tara), « The Third Way: philology and critical edition for a digital age », *Variants: the Journal of the European Society for Textual Scholarship*, 10 (2012), http://boris.unibe.ch/43071/.

Berners-Lee (Tim), « Information Management : A Proposal », *CERN*, mars 1989, https://www.w3.org/History/1989/proposal.html.

Burton (D. M.), « Automated Concordances and Word Indexes: The Fifties », *Computers and the Humanities*, 15 (1981), p. 1-14.

–, « Automated Concordances and Word Indexes: The Early Sixties and the Early Centers », *Computers and the Humanities*, 15 (1981), p. 83-100.

–, « Automated Concordances and Word Indexes: The Process, the Programs, and the Products », *Computers and the Humanities*, 15 (1981), p. 139-54.

–, « Automated Concordances and Word Indexes: Machine Decisions and Editorial Revisions », *Computers and the Humanities*, 16 (1982), p. 195-218.

Busa (Roberto A.), «Foreword: Perspectives on the digital humanities», *A companion to digital humanities*, , 2004, http://www.digitalhumanities.org/companion/.

Camps (Jean-Baptiste), *La 'Chanson d'Otinel': édition complète du corpus manuscrit et prolégomènes à l'édition critique*, thèse de doctorat, dir. Dominique Boutet, Paris-Sorbonne, 2016, https://halshs.archives-ouvertes.fr/tel-01664932.

Camps (Jean-Baptiste), « Où va la philologie numérique? », *Fabula-LHT*, (20) 2018, http://www.fabula.org/lht/20/camps.html.

Cerquiglini (Bernard), « Éloge de la variante », *Langages*, 69 (1983), p. 25–35, doi: 10.3406/lgge.1983.1140.

Cerquiglini (Bernard), *Éloge de la variante : histoire critique de la philologie*, Paris, 1989 (Des Travaux, 8).

Delatte (L.), «En guise d'éditorial» et «Liste des membres de l'Organisation», *Revue Informatique et Statistique dans les Sciences humaines*, 1 (1965), p. 1-14 et 15-41.

Duggan (Hoyt N.), « The Electronic Piers Plowman Archive and SEENET », *The Electric Scriptorium, Université de Calgary, 12 nov. 1995*, 1995, http://xml.coverpages.org/duggan-piers1.html.

Duval (Frédéric), « Transcrire le français médieval : de l' `instruction' de Paul Meyer à la description linguistique contemporaine », *Bibliothèque de l'École des chartes*, 170 (2012), p. 321-342.

Duval (Frédéric), « Pour des éditions numériques critiques », *Médiévales*, 73 (2017), p. 13–29, doi:10.4000/medievales.8165.

EADH, « Wilhelm Ott », s.d., http://eadh.org/wilhelm-ott, consulté le 14 mars 2020.

Froger (Jacques), *La Critique des textes et son automatisation*, Paris, 1968 (Initiation Aux Nouveautés de La Science).

Genêt (Jean-Philippe), « Des chiffres et des lettres: quelques pistes pour l'historien », *Histoire et Mesure*, 18 (2003), p. 2-7

Goldfarb (Charles F.), *The Roots of SGML: A Personal Recollection*, 1996, http://www.sgmlsource.com/history/roots.htm.

Goldfarb (Charles F.), Mosher (Edward J.), and Peterson (Theodore I.), « An Online System for Integrated Text Processing », *Proceedings of the American Society for Information Science*, 1970, vol. 7, p. 147-150, http://www.sgmlsource.com/history/jasis.htm.

Gray (Jim), « Jim Gray on eScience: A transformed scientific method», dans *The fourth paradigm: Data-intensive scientific discovery*, T. Hey, S. Tansley, K. Tolle (éd.), Washington, 2009.

Hockey (Susan), «The history of humanities computing», A companion to digital humanities…, 2004, p. 3-19.

ISO 8879:1986, *Information processing — Text and office systems — Standard Generalized Markup Language (SGML)*, Technical Committee: ISO/IEC JTC 1/SC 34, Document description and processing languages, ICS : 35.240.30, IT applications in information, documentation and publishing.

Pierazzo (Elena), « A rationale of digital documentary editions », *Literary and Linguistic Computing*, 26-4 (2011), p. 463-477.

Renear (Allen H.), Mylonas (Elli) et Durand (David), « Refining our notion of what text really is: The problem of overlapping hierarchies », dans Research dans *Research in Humanities Computing 4: Selected Papers from the 1992 ALLC/ACH Conference*, éd. N. Ide et S. Hockey, Oxford, 1996.

Robinson (Peter) et Solopova (Elizabeth), « Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue », *The Canterbury Tales Project Occasional Papers*, 1 (1993), p.19-52.

Sahle (Patrick), « 2. What Is a Scholarly Digital Edition ? », dans *Digital Scholarly Editing:*

Theories and Practices, éd. Matthew James Driscoll et Elena Pierazzo, Cambridge, 2017, p. 19–39 (Digital Humanities Series), http://books.openedition.org/obp/3397.

Speculum, 65-1 (1990), numéro spécial « The New Philology ».

Sperberg-McQueen (C. M.), « How to teach your edition how to swim », *Literary and Linguistic Computing*, 24-1 (2009), p.27-39, doi: 10.1093/llc/fqn034.

Stutzmann (Dominique), « Paléographie statistique pour décrire, identifier, dater… Normaliser pour coopérer et aller plus loin? », dans *Kodikologie und Paläographie im digitalen Zeitalter 2 / Codicology and Palaeography in the Digital Age 2*, éd. Franz Fischer, Christiane Fritz et Georg Vogeler, Norderstedt, 2011, p. 247-277 (Schriften des Instituts für Dokumentologie und Editorik, 3), https://halshs.archives-ouvertes.fr/halshs-00596970/.

Tasman (P.J.), « Literary Data Processing », *IBM Journal of Research and Development*, 1 (1957), https://dl.acm.org/doi/10.1147/rd.13.0249.

TEI Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, originally edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALLC-ACL Text Encoding Initiative, now entirely revised and expanded under the supervision of the Technical Council of the TEI Consortium, version 4.0.0 (2020), rév. ccd19b0ba.

Weil (Gérard Emmanuel), « Méthodologie de la codification des textes sémitiques servant aux recherches de linguistique quantitative sur ordinateur », *Bulletin d'information de l'Institut de Recherche et d'Histoire des Textes*, 13 (1964-1965), p. 115-133, doi: 10.3406/rht.1966.1033.

Wilkinson (Mark D.), *et alii*, « The FAIR Guiding Principles for scientific data management and stewardship », *Scientific Data*, 3 (2016), 160018. doi:10.1038/sdata.2016.18.

World Wide Web Consortium (W3C), *Extensible Markup Language (XML) 1.0*, W3C Recommendation 10-February-1998, 1998, https://www.w3.org/TR/1998/REC-xml-19980210.