Contours du document numérique connecté

Evelyne Broudoux

Contours du document numérique connecté

Outlines of a connected digital document

Evelyne Broudoux

Laboratoire Dicen, Conservatoire national des arts et métiers, Paris
evelyne.broudoux@cnam.fr

Mots-clés. Document numérique, annotations, commentaires, données liées, humanités numériques

Keywords. Digital document, annotations, comments, linked open data, digital humanities

1 Introduction

1.1 Le document boîte à outils ?

Si l'on considère l'évolution de l'objet qu'est le document numérisé ou nativement numérique ces quinze dernières années, plusieurs points s'offrent à nous quant à son observabilité. Car au fur et à mesure de son développement, le document numérique est devenu une véritable boîte à outils ; porteur de métadonnées, il offre des instruments d'observation à ceux qui veulent l'étudier. De « boîte noire » technique, il s'ouvre à la médiation avec des fonctionnalités de communication.

Dans un contexte interdisciplinaire de redéfinition du document numérique, Pédauque l'avait théorisé sous trois angles :

- forme : approche structurelle,
- signe : approche sémiotique,
- médiation : approche communicationnelle.

Si aujourd'hui, nous devions réfléchir aux tendances prises depuis Pédauque, il nous faudrait constater qu'existe actuellement une convergence entre l'approche structurelle et l'approche communicationnelle.

Nous partirons donc de la forme de l'objet lui-même et non pas de sa structure. Comment nous apparaît aujourd'hui un document numériquement natif, tel qu'on peut le trouver sur le web, pour observer ce qui en émerge. Deux variables alors se distinguent : ses contours et sa connectivité.

Le focus est porté sur l'aspect communicationnel du document, celui d'être un espace inscriptible matérialisant les médiations entre auteurs, éditeurs et lecteurs. Nous nous appuierons principalement sur l'exemple du document scientifique.

1.2 Les contours du document numérique

Les contours du document numérique construisent une forme. Celle que réalisent ses limites internes. Or, ce qui vient à l'esprit en premier c'est que le document numérique n'a justement plus de limites : ses liens hypertextes entraînent

le lecteur dans un espace extra-documentaire, il existe donc des limites externes au document numérique que nous sommes en mesure de matérialiser.

De plus ses marges sont susceptibles de varier. Car si les marges continuent de clôturer visuellement le document, en réalité, celles-ci sont devenues inscriptibles. Une clôture interne au document existe qui peut être appropriée par les lecteurs. Mais les limites internes au document sont aussi celles des sous-parties qui le composent.

L'inscription étant la trace d'actions, nous nous intéresserons à celles spécifique de la construction de connaissances.

Nous commencerons par examiner :

- le document publié, connecté sur la toile,
- le document-processus, le protodocument collaboratif,
- le document support d'écrilecture.

2 Le document connecté

2.1 Les limites externes aux documents

L'exemple de l'article scientifique connecté à ses sources

Dans le cadre de la lecture d'articles scientifiques se produisant au moment de la recherche d'informations - liée par exemple à la constitution d'états de l'art - les articles au format html dit « enhanced », c'est-à-dire amélioré ou même augmenté, offrent un certain nombre de possibilités. La bibliothèque scientifique en ligne Wiley permet par exemple à partir d'un article :

- d'afficher dans les marges les références citées et d'un clic, consulter dans un nouvel onglet du navigateur, l'article cité en référence lorsqu'il est librement accessible ;
- d'examiner les figures de manière séparée, les télécharger sous forme de diapo « powerpoint » ;
- de solliciter une demande de réutilisation du contenu pour respecter un éventuel copyright ;
- d'ouvrir dans le navigateur le document au format pdf et le télécharger.

L'affichage du pdf dans le navigateur intègre les fonctionnalités de « ReadCube », le service Google de LGRB[1]. Ce qui autorise :

- l'importation du document dans sa bibliothèque,
- la consultation de références proposées par le moteur de recommandations, en rapport avec l'article affiché,
- l'inclusion de la barre d'outils « ReadCube » avec ses quatre services :
    - l'information générale du document,
    - le surlignage et l'annotation de portions de textes,
    - l'accès aux références listées et reliées à Google Scholar,
    - la diffusion via les médias sociaux (Twitter, Facebook, Google+, LinkedIn).

Cette augmentation du document le replace dans le contexte des articles liés par ses citations et de ceux qui lui sont virtuellement liés, fournis par les moteurs de recommandations.

[1] Logiciel de gestion de références bibliographiques

2.2 Les limites internes aux documents

Mais le concept de « publication augmentée » est susceptible de transformer l'unité fondamentale du document. Pour certaines disciplines scientifiques, le document qui fournit des résultats n'est qu'une partie d'un tout représenté par les jeux de données ayant servi aux calculs, les visions analogiques des graphiques traduisant les informations quantifiées, les sources et les ressources (vidéo, audio, etc.). Connecté aux corpus de documents apportant les preuves du travail, l'article scientifique divulgue des résultats qui doivent pouvoir être interrogés, les expériences reproduites. En ce sens, le document enrichi par ses connexions est distribué.

Le document « container »

Conceptualisé au milieu des années 1990, le « document compound » dont la conception est protégée par brevet[2] est représentatif de la fragmentation progressive de l'unité d'information qu'est le document. Le document-compound est constitué d'une partie statique – le texte – et d'une partie dynamique – un ensemble de formules mathématiques que l'on appelle au choix pour effectuer des calculs.

Il s'agit d'un cas particulier de « document composite », bien connu en GED, qui est le résultat de l'assemblage dynamique de fragments documentaires hétérogènes. Ce qui nécessite une gestion sous forme de workflow du travail collaboratif.

L'exemple d'un « document-dossier » au service de l'écriture collaborative

Le « document-dossier » est une appellation adoptée par l'ANR C2M (Chaînes éditoriales Collaboratives Multimédia) de l'Université de Compiègne pour nommer un répertoire partagé sous Scenari[3] qui permet l'écriture à plusieurs (éditeur, co-auteurs et contributeurs) de grains d'informations. L'objectif étant de faciliter la ré-éditorialisation de contenus (ex : édition collaborative d'un manuel scolaire). L'écriture se déroule à plusieurs en mode wiki et l'historique consiste à pouvoir visualiser l'état antérieur d'un fragment et tout ses liens. D'un point de vue communicationnel, les utilisateurs sont avertis des modifications en cours et sont libres de conserver les versions des unités d'informations qui leur conviennent.

Ce « document-dossier » a ceci de particulier qu'il représente un état stable d'un travail collectif prêt à être rééditorialisé (Crozat, 2012). Il s'agit d'un document « container » qui peut servir aussi à l'archivage avec des contenus figés et les métadonnées nécessaires à leur exploitation.

3 Le document numérique comme espace d'écrilecture

3.1 L'écrilecture

L'annotation vue en tant que processus est un acte d'écriture qui intervient au moment de la lecture. Cet enrichissement de la lecture qui repose sur des actes scripturaux est ancienne et date de la recopie des textes à la main. Avec Gérald Kembellec, pour le projet « Ecrilecture augmentée »[4], nous avons resitué l'écrilecture dans l'évolution des pratiques historiques de construction de l'érudition en convergence avec la « lettrure » observée dès le moyen âge qui, comme l'a

[2] https://www.google.com/patents/US5630126

[3] http://scenari-platform.org/projects/scenari/fr/pres/co/

[4] http://www.hesam.eu/blog/2015/09/07/appel-communications-ecrilecture-augmentee-dans-les-communautes-scientifiques-humanites-numeriques-et-construction-des-savoirs/

rappelé Emmanuel Souchier (2012), désignait une seule et même activité mêlant la lecture et le commentaire.

Le néologisme écrilecture a permis de rendre compte de pratiques littéraires créatrices avec l'ordinateur. Sous le terme (ecrileitura), Pedro Barbosa dès 1992 a décrit ce phénomène de délégation au lecteur de la constitution de textes à lire, l'auteur se situant en amont dans un texte-programme générant de multiples variations. Alain Vuillemin l'avait repris pour caractériser ce nouveau comportement du lecteur entraîné dans des manipulations créatrices face à l'écran· « L'acte d'écrilecture, d'écriture et de lecture interactives, est alors conçu comme une action périphérique, faite par l'utilisateur d'un ordinateur autour d'un fragment de texte de référence » (Vuillemin, 1990, p. 103). Si cette délégation de la constitution du sens d'un texte au lecteur est classique en littérature (la théorie de la réception de Wolfgang Iser, le lecteur interprète d'Anne Jorro), le concept d'écrilecture va encore plus loin, puisqu'il part du principe que le travail d'écriture intérieure pendant la lecture peut être externalisé. Le premier « système d'annotation dynamique » a été ainsi conceptualisé dès 1990 pour les lecteurs de la BNF dans un programme de numérisation : « Il sera possible de constituer un corpus de texte à partir des collections, de l'organiser en y introduisant des signets ou des balises, puis d'y associer des annotations et des commentaires à propos de fragments qui auront été sélectionnés au préalable » (Vuillemin, 1990, p. 103). Malheureusement, ce projet a fait long feu et s'il a été question d'une « seconde génération » de postes de lecture, force est de constater que ceux-ci ont été bridés, empêchant au contraire tout acte d'écrilecture, avant de disparaître totalement.

Si aujourd'hui, les processus d'écrilecture sont soutenus par de multiples fonctionnalités logicielles, leur autonomie réalisée par les raisonnements computationnels sur la sémantique pourraient transformer les processus cognitifs liés à la lecture.

3.2 L'interopérabilité et l'annotation sémantique

Nous rappelons ici brièvement la distinction entre documents, ressources, métadonnées et annotations.

Avec le web sémantique, le document est une ressource comme une autre qui peut être unifiée pour les documents identifiables : que ce soit une adresse unique URI ou un identifiant de type URN : ISBN, ISNI, etc. La ressource devient « citable » ou « adressable » même hors de tout contexte numérique.

Les métadonnées sont des ensembles structurés de données descriptives qui renseignent les ressources comme les notices le faisaient avec les documents. Associées aux documents numériques dans un processus d'indexation, connectées aux référentiels, elles vont permettre de créer des liaisons à la demande et de rendre visible les ressources dans les requêtes des moteurs et de créer des collections thématiques.

Nous reprenons la distinction opérée par (Garlatti, Prié, 2006) entre métadonnées et annotations : une métadonnée est attachée à une ressource identifiée en tant que telle (sa description est normalisée et on peut mettre en place des inférences) alors que l'annotation est « plus située au sein de cette ressource et écrite au cours d'un processus d'annotation-lecture ».

La notion d'annotation sémantique des documents a été définie par les acteurs informaticiens du web sémantique comme « l'accrochage d'un élément de l'ontologie à un fragment de document, qui implique de reconnaître (ou de déduire) la présence de l'élément ontologique dans la forme de surface du document ». Il

s'agit d'une tâche automatisable ou semi-automatisée qui relève de la fouille de textes plutôt qu'un acte d'écrilecture.

L'annotation sémantique peut aussi être comprise comme une manipulation technique indispensable à la liaison des données. Un outil comme Pundit[5] propose d'annoter sémantiquement des textes et donc de faire des liaisons, en éditant directement les triplets des données liées.

Un exemple typique de ce que l'on peut faire avec des données sémantiquement liées est Isidore, un projet Adonis-HumaNum et CCSD-CNRS : ce méta-moteur de recherche scientifique spécialisé en SHS créé par Laurent Capelli, Jean-Luc Minel et Stéphane Pouyllau moissonne depuis 2011 les données OAI-PMH, RDFa, RSS, de Revues.org, Persée, HalSHS, theses.fr, pour ne citer que les premiers. Isidore est précisément une plateforme de collecte, d'enrichissement et de diffusion de documents et de données en libre accès, qui participe à la construction du web de données francophones en SHS – et depuis 2014 en anglais et espagnol – et contribue à valoriser la recherche dans nombre de domaines.

Ce service est devenu progressivement indispensable au montage de projets d'humanités digitales qui ont été créés grâce au potentiel de connectivité d'Isidore qui vient d'intégrer - pour compléter son offre couteau-suisse en humanités digitales – un volet d'analyse de données.

Les collections contextualisées

Selon les objectifs, l'article scientifique peut être appelé dans les différents contextes créés par les axes théoriques disciplinaires : construction d'un état de l'art, preuve étayant un article de vulgarisation, réfutation dans le cadre d'une controverse, etc. Les résultats de la recherche d'information sur les moteurs scientifiques forment à eux seuls des collections virtuelles de documents, des corpus.

C'est dans la constitution de collections et la valorisation des archives que les projets d'humanités numériques prennent leur essor.

Un exemple de constitution de collections est la plateforme scientifique Criminocorpus[6] spécialisée en histoire de la justice, des crimes et des peines et développée depuis 2008. Elle se présente en octobre 2015 sous la forme d'un site-portail et comprend :

- un musée contenant 31 expositions et des visites de lieux de justice,
- une bibliothèque de 79 011 pages,
- deux éditions de corpus juridiques contenant 4 385 articles et 7 442 versions d'articles,
- 12 chronologies contenant 754 événements,
- des statistiques concernant 21 708 données numériques,
- une bibliographie de 68 828 références,
- une revue,
- un blog d'actualités,

et 101 collaborateurs.

Cet inventaire à la Prévert ne doit pas faire écran au travail de reéditorialisation des ressources réalisé pour leur mise à disposition au public.

En partenariat avec les Archives nationales de France, le CNRS et le ministère de la Justice été créé le *Clamor* en 2015, le premier centre d'humanités numériques dédié à l'histoire de la justice et responsable de la plateforme Criminocorpus.

[5] http://www.thepund.it/

[6] https://criminocorpus.org

3.3 L'annotation comme processus

Mais revenons sur le processus de l'annotation qui consiste à associer à un document des informations complémentaires comme des remarques (explication, commentaire critique, etc.) ou des notes (référence bibliographique, url, etc.) qui peuvent être de différente nature (textuelle, imagière, multimédia). Ces ajouts interviennent à des moments de réflexion se traduisant par le repérage et la sélection de fragments de documents.

Résultat : d'une part, un ancrage au sein du document qui se manifeste sous la forme de soulignés et surlignages, appels de notes, etc. et la production de notes se positionnant au regard de cet ancrage : pour ce qui est de l'imprimé, dans ses marges, au bas de des pages, à la fin des chapitres ou à la fin du document lui-même ; pour ce qui est du web et des documents imitant l'imprimé, ce sera souvent les mêmes dispositions du livre qui seront d'abord reprises ; et pour ce qui est des livres numériques, tout reste à inventer.

Les outils d'annotation

Depuis plus de vingt ans, de nombreux projets informatiques de laboratoire conçoivent des outils d'annotation qui dépassent rarement l'état de prototypes. Sans doute, ne sommes-nous pas encore prêts à les utiliser car il manque une culture générale sur l'annotation et des projets éducatifs capables de l'enseigner. Cependant, les enseignants faisant coïncider les potentialités du numérique avec des approches pédagogiques innovantes se sont rapidement saisis des potentialités de l'écrilecture des documents et travaillent la réécriture de manière renouvelée.

De multiples outils peuvent être utilisés dans des projets éducatifs comme celui permettant des annotations vidéo comme Polemic Tweet[7] de l'IRI.

Un exemple d'enracinement numérique d'une pratique ancienne est celui de la transcription qui consiste à retranscrire les annotations des textes manuscrits en « mode texte » numérique.

La transcription collaborative

Transcrire des images en texte est une occupation courante pour l'édition de manuscrits. Dans les projets de numérisation, les faiblesses de la reconnaissance optique de caractères (OCR) ont rendu indispensable une intervention humaine pour éditer un texte fidèle à l'original et corriger les erreurs d'interprétation des algorithmes d'OCR. C'est pour suppléer à ces carences que le projet Wikisource a été fondé dans l'objectif de retranscrire des livres du domaine public, en s'appuyant sur le modèle éditorial du Wiki et en misant sur la collaboration de personnes enregistrées. Transcribe Bentham est basé sur un Wikisource et vise à transcrire les notes d'un philosophe prolifique (14 172 manuscrits sur 60 000 ont déjà été transcrits au 30 octobre 2015). Le crowdsourcing scientifique trouve ici sa justification comme avec E-Recolnat qui vise à identifier les étiquettes de 350 ans de collections d'histoire naturelle française.

Une panoplie d'outils dédiés a vu le jour dont la majorité repose sur un moteur wiki. Citons Scripto[8], un outil opensource installable dans un CMS, qui aide à la transcription pour les collections. Il est maintenant à la base de nombreux projets dont celui de Do It Yourself History[9] qui vise depuis 2011 à transcrire les collections des Bibliothèques et Archives de l'université de l'Iowa.

---

[7] http://polemictweet.com

[8] http://scripto.org/

[9] http://diyhistory.lib.uiowa.edu/

3.4 La commentarisation

L'appropriation des outils d'annotation ne fait que commencer et c'est dans les projets de lecture savante qu'on observe leur installation durable : par exemple, le programme collaboratif Comment-R du Labex Hastec qui porte sur la pratique du commentaire théologique, philosophique ou scientifique pendant l'Antiquité tardive, le Moyen Age et l'Époque Moderne. Son objectif est de produire à la fois des études particulières et des instruments de travail (catalogues, éditions, bases de données) en combinant « l'étude érudite des textes et la mise en œuvre des nouvelles technologies en humanités numériques ».

Du côté web 2.0, citons le projet éditorial Medium, plateforme de blogging installée avec un module de commentarisation depuis 2012. L'originalité de cette initiative tient à ce qu'elle a réussi a agréger des blogueurs autoritatifs mais aussi des journalistes, des intellectuels et des professionnels qui n'hésitent pas faire connaître des opinions, des recherches ou des initiatives qui n'auraient pas trouvé d'éditeur autrement. Ex : La *Proposition pour une vraie réforme ferroviaire* de Jean-Daniel Guyot fait également connaître une entreprise fondée en 2009[10]. Il serait dommage aussi d'ignorer la plateforme Genius[11] qui regroupe les textes de chansons mais aussi des textes classiques à partir desquels des interprétations sont réalisées sous la forme de commentaires mais aussi des remarques sur ces commentaires qui s'affichent sous forme de listes.

Dans le processus de travail rédactionnel, il parait important de distinguer l'annotation qui concerne l'écrilecture préparatoire à la rédaction, de la commentarisation qui concerne le feedback. Citons le travail de Johanna Daniel[12] qui a soumis à commentaires son mémoire de Master2 de l'Ecole des Chartes, pendant sa rédaction. Ce mémoire intitulé « *Les outils d'annotation et l'édition scientifique de corpus textuels. L'exemple du projet "Les Guides de Paris"* » a été commenté avec CommentPress, un plug-in fonctionnant sous WordPress.

D'autres outils comme Hypothes.is[13] se placent directement dans le navigateur et comme les services de bookmarking proposent un champ de réception de commentaires partageables en mode public. Est ainsi actuellement expérimenté un dispositif d'évaluation ouverte par les pairs pour la revue VertigO avec *Hypothèses*[14].

4 Conclusion

La transformation du document numérique poursuit son cours. L'œuvre de connexion sémantique des liens hypertextes a commencé et concerne des projets d'édition scientifique de grande ampleur qui valorisent des collections, construisent l'appareil critique des œuvres, contribuent à la construction des connaissances, favorisent la compréhension des auteurs.

---

[10] https://medium.com/@jdguyot/propositions-pour-une-vraie-reforme-ferroviaire-849812ebbd86

[11] http://genius.com/

[12] Retour d'expérience de l'auteur sur son mémoire annoté ici : http://johannadaniel.fr/isidoreganesh/2014/12/ecriture-connectee-experience-redaction-memoire/

[13] https://hypothes.is/

[14] http://vertigo.hypotheses.org/category/commentaire-ouvert

Les trois types de documents évoqués en introduction : connecté, processus collaboratif et support d'écrilecture sont la trace de pratiques érudites persistantes. Les propositions des outils font apparaître des invariants :

- le surlignage, la commentarisation,
- l'activité d'écrilecture à partir du navigateur web ou à travers une application dédiée.

Les humanités numériques se construisent sur la convergence du document comme forme et du document comme médiation.

L'examen de l'évolution de l'outillage de lecture d'articles scientifiques en ligne, des outils d'annotation et de commentarisation prouve qu'ils s'inscrivent dans la sémantisation du web. Leurs usages suivent et seront à différencier suivant les disciplines, les contextes et les objectifs.

5 Bibliographie

Berra, A. (2015). « Philologia ». Carnet de recherche. URL : http://philologia.hypotheses.org/

Broudoux, E. (2012). « Vers l'objet documentaire (re)contextualisé ». 9e congrès des enseignants documentalistes de l'Education nationale (Fadben). Cnam, Paris. Nathan, 2012. <sic_00715868>

Crozat, S. (2012). « Chaînes éditoriales et rééditorialisation de contenus numériques » in *Le document numérique à l'heure du web de données* (éd. Calderan L, Laurent L., Lowinger H., Millet J.). ADBS, pp. 179-220. <hal-00740268>

Daniel, J. (2014). Les outils d'annotation et l'édition scientifique de corpus textuels. L'exemple du projet « Les guides de Paris ». Mémoire de Master2 de l'Ecole des Chartes. URL : http://fr.slideshare.net/Peccadille/les-outils-dannotation-et-ldition-scientifique-de-corpus-textuels-mmoire

Iser, W. (1985). *L'acte de lecture : théorie de l'effet esthétique*, P. Mardaga, Bruxelles.

Jahjah, M. (2015). « Marginalia ». Carnet de recherche. URL : https://marginalia.hypotheses.org/

Jorro, A. (2000). *Le lecteur interprète*. PUF, collection Education et formation, Paris.

Pédauque, R. (2006). *Le document à la lumière du numérique*. C&F Editions.

Prié, Y. Garlatti, S. (2004). « Annotations et métadonnées dans le Web sémantique », in Revue I3 Information-Interaction - Intelligence, Numéro Hors-série Web sémantique, 24 pp.

Soubrié, Thierry (2001), « Enseigner la lecture intime du texte littéraire grâce à l'édition hypertextuelle. », Document numérique 1/2001, Vol. 5, p. 181-208. URL : www.cairn.info/revue-document-numerique-2001-1-page-181.htm.

Souchier, E., (2012), « La lettrure à l'écran. Lire & écrire au regard des médias informatisés ». *Communication & langages*, n°174, p. 85-108

Vuillemin, A., (1999) « La lecture interactive et l'écrilecture », *Littérature, informatique, lecture*, Vuillemin A., Lenoble M., Limoges, Presses Universitaires de Limoges, p. 101-110.