L'édition électronique de sources historiques

De nouvelles approches de l'indexation ?

Gautier Poupeau

Édition électronique

URL : <http://journals.openedition.org/ccrh/3023>

DOI : 10.4000/ccrh.3023

ISSN : 1760-7906

Éditeur

Centre de recherches historiques - EHESS

Édition imprimée

Date de publication : 10 janvier 2005

ISSN : 0990-9141

Référence électronique

Gautier Poupeau, « L'édition électronique de sources historiques », *Les Cahiers du Centre de Recherches Historiques* [En ligne], 35 | 2005, mis en ligne le 24 mai 2011, consulté le 01 mai 2019. URL : <http://journals.openedition.org/ccrh/3023> ; DOI : 10.4000/ccrh.3023

Ce document a été généré automatiquement le 1 mai 2019.

Article L.111-1 du Code de la propriété intellectuelle.

L'édition électronique de sources historiques

De nouvelles approches de l'indexation ?

Gautier Poupeau

- 1 Dans le cadre de cette journée d'études consacrée à l'indexation du récit, il m'a semblé intéressant de présenter l'édition électronique de sources et ses conséquences sur les pratiques d'indexation. En effet, éditer et indexer une source constituent deux étapes étroitement liées dans le processus de mise à disposition des textes, l'indexation représentant une partie non négligeable du travail de l'éditeur scientifique.
- 2 Conçue à l'origine comme la mise à disposition d'un ouvrage sur le support numérique, l'édition électronique de sources est apparue il y a une quinzaine d'années avec la mise sur le marché de cédéroms rassemblant des corpus de sources. À cet égard, les plus connus des médiévistes sont le Cetedoc édité par Brepols et la *Patrologie latine* par Chadwick-Healey. Dans la plupart des cas, ces projets consistaient à mettre en place des bases de données textuelles interrogeables, sans réelle intention d'édition scientifique, et constituent de ce point de vue des rééditions. L'invention du Web au début des années 1990 a modifié l'intérêt pour l'édition électronique. En effet, alors que le cédérom présente les mêmes défauts que le papier pour la diffusion et la gestion des stocks, le Web permet de s'affranchir de ces problèmes et donne une nouvelle visibilité aux résultats de la recherche. Dans un premier temps, de la même façon que les cédéroms, l'édition électronique sur le Web s'est limitée à la mise à disposition de documents issus de la recherche sans réelle valeur ajoutée pour le chercheur, si ce n'est la disponibilité des textes. Cette phase correspond à une période d'adaptation et de compréhension de ce nouveau média. Avec la multiplication des informations présentes sur le Web et l'expérience acquise, nous assistons aujourd'hui à une réflexion sur « l'art d'éditer » en ligne. Les enjeux ne se limitent plus aujourd'hui à la simple publication d'un ouvrage en ligne, mais à mieux publier en ligne, améliorant le traitement et l'exploitation des informations mises en lignes sur différents sites. Ainsi, un effort particulier est fait pour l'adoption de langages communs, ou standards, et la recherche et l'échange

d'informations, induisant une réflexion sur la structuration des textes. La pratique de l'indexation mise de côté par la recherche en texte intégral retrouve alors une nouvelle jeunesse. Vu le peu de recul que nous avons sur ces questions, je pense ne pouvoir vous donner que des éléments de réflexion. Mais, avant cela, je vais m'efforcer de vous présenter rapidement, en évitant autant que faire se peut le jargon technique propre à cette discipline, l'édition électronique de sources, les raisons du choix de ce support, les apports, les conséquences et les problèmes posés.

L'édition électronique

- 3 Dans le cadre de cette présentation, je voudrais vous faire partager l'expérience menée depuis maintenant trois ans à l'École nationale des chartes et les étapes de la réflexion qui l'ont accompagnée. Cette expérience est partie d'un constat initial qui, soyons honnêtes, n'avait rien de scientifique, mais plutôt d'ordre économique et logistique. Depuis dix ans, l'École des chartes a décidé de dynamiser ses publications et la mise à disposition des résultats de ses recherches. Or, une de ses spécialités est, depuis sa création, au début du XIX^e siècle, l'étude et l'édition des sources historiques. Cette politique a donc naturellement abouti à la publication de plusieurs éditions de sources dans sa collection « Mémoires et documents de l'École des chartes », comme l'édition des chartiers de l'abbaye prémontrée de Saint-Yved-de-Braine et de l'Abbaye-aux-Bois, mais aussi d'ouvrages méthodologiques, comme la série des *Conseils pour l'édition de textes* parue aux éditions du CTHS.
- 4 Or, il n'est un secret pour personne que l'édition universitaire en sciences humaines et sociales doit supporter, comme tous les autres types d'édition, les coûts d'impression des ouvrages, de gestion des stocks et de diffusion. De plus, elle connaît une crise importante qui se traduit par une diminution de plus en plus importante des tirages d'ouvrages, ce qui entraîne une relative confidentialité des parutions. Enfin, les entreprises d'édition de sources sont souvent des projets longs et fastidieux et leur publication sur papier doit attendre l'établissement définitif du texte. Toutes ces raisons ont fait de l'édition de sources historiques un type de publication de plus en plus difficile à éditer.

L'édition électronique : une solution aux problèmes de l'édition de sources historiques

- 5 Ces constats et la mise en place du site Web de l'École des hautes études en sciences sociales en juin 2001 ont alors été le point de départ de la présente réflexion. En effet, l'édition électronique présente plusieurs avantages intéressants.
- 6 Tout d'abord, les coûts de l'édition électronique sont différents de ceux du support papier. En dehors des coûts inhérents à la préparation de l'ouvrage qui sont en gros identiques, la principale différence tient à la rationalisation des dépenses. Alors que dans le cadre de l'édition électronique l'investissement est important pour l'achat de matériel et les développements informatiques, les coûts de production y sont réduits, puisqu'il n'existe pas de frais d'impression, de diffusion ou de gestion des stocks pour chaque ouvrage comme le support papier. En outre, la mise en place de chaînes éditoriales permet la réutilisation des développements informatiques ultérieurs. Enfin, les coûts ne sont pas proportionnels au nombre de pages mis en ligne, ce qui permet de proposer en

plus du texte édité des documents annexes qui donnent un éclairage différent. Il faut toutefois modérer cette remarque avec le coût éventuel des droits de reproductions d'images en ligne, souvent exponentiel.

- 7 L'édition électronique offre également une certaine souplesse éditoriale. Au contraire d'une édition sur le support-papier qui est figée au moment de sa publication, l'édition électronique peut être mise à jour à n'importe quel moment. Une erreur peut être corrigée facilement sans répercussion économique. De plus, de vastes entreprises de publication, comme le sont les éditions de sources, peuvent avec le support électronique être mises à disposition au fur et à mesure de leur réalisation. Ainsi, le *Cartulaire blanc de l'abbaye de Saint-Denis* qui proposera à terme la transcription intégrale des 2600 actes contenus dans ce document est mis en ligne au fur et à mesure, au rythme d'un chapitre par an. Avec le seul support « papier », il aurait fallu attendre de disposer d'un corpus plus significatif, voire de sa réalisation complète, pour en publier l'édition.
- 8 Enfin, le Web offre une disponibilité accrue des résultats de la recherche et donc une meilleure visibilité potentielle des ouvrages. Ainsi, le chercheur peut disposer à n'importe quel moment depuis n'importe quel ordinateur relié au Web d'une masse critique d'informations auxquelles il est souvent difficile d'accéder dans toutes les bibliothèques. L'audience potentielle du Web est mondiale et sans rapport avec la distribution d'un ouvrage papier. Ainsi, le site du *Cartulaire blanc de l'abbaye de Saint-Denis* est consulté en moyenne par 80 visiteurs par jour. Il est inutile de poursuivre la démonstration : vous avez sans doute déjà tous fait cette expérience sur le Web. Malgré tout, il existe un revers à cet attrait apparent : le chercheur doit pouvoir retrouver l'information souhaitée au milieu des huit milliards de pages Web référencés par un moteur de recherche comme Google. Pour mener à bien ce référencement, il faut alors comprendre les mécanismes du Web.

Expériences et premiers constats

- 9 À partir de cette réflexion, une première expérience a été menée en 2002. Elle a abouti à la mise en ligne de la première version du *Cartulaire blanc de l'abbaye de Saint-Denis* en janvier 2003. Elle nous a permis de mieux appréhender ce nouveau média et de dresser une première série de constats, même si elle consistait en une adaptation simple et fidèle d'un type de publication issu du papier.
- 10 Ainsi, l'accès à l'information a été la première spécificité de ce support qui nous est apparue. Dans un ouvrage papier, le support est déterminé par la structure physique du livre, tandis qu'en édition électronique, il dépend de la structure logique déterminée par l'éditeur. Le Web est fondé sur la technologie de l'hypertexte, dont le principe est de relier des pages-écrans entre elles par l'intermédiaire de liens. Cette technologie permet de multiplier les points d'accès directs à l'information. Cela ouvre de nouvelles possibilités dans l'appréhension d'une source, car elle permet la mise en place de parcours de lecture différents en fonction de l'intérêt de chaque utilisateur. Cela pose aussi la question du statut de la source, puisque chaque lecteur « déconstruit » la source originale et la reconstruit en fonction de son parcours au sein du texte édité.
- 11 Outre ce premier constat qui s'est fait presque naturellement au moment de la mise en place de notre édition, les apports du support numérique n'étaient pas évidents avec cette première expérience, et il nous est rapidement apparu que nous avions sous-exploité l'outil. Évidemment, et je ne sous-estime pas ces apports, nous résolvions un

certain nombre de nos problèmes économiques et logistiques et nous proposons un module de recherche assez simple, sans lemmatisation, mais qui permettait tout de même d'effectuer une recherche en texte intégral. Il semblait alors que l'édition électronique constituait surtout pour les chercheurs une évolution non négligeable dans la mise à disposition des données de la recherche. Or, jusque là, l'outil informatique avait représenté une révolution dans le traitement des données au moment de la recherche. Il nous semble donc important de centrer aujourd'hui une partie de notre réflexion sur le rassemblement de ces deux facettes de l'outil informatique dans deux directions : l'analyse lexicographique et les apports de l'indexation.

- 12 Cette préoccupation en rejoint deux autres qui sont en fait des préalables : la possibilité d'interroger efficacement des corpus hétérogènes de sources présents sur différents sites - ce qu'on appelle l'interopérabilité - et la conservation à long terme des textes édités. Ce travail nous occupe depuis un an. Pour le mener à bien, nous nous sommes appuyés sur des recherches menées dans ce secteur depuis une quinzaine d'années et sur les travaux de l'organisme de normalisation des langages utilisés sur le Web, le W3C. Le principe fondamental en est la séparation des mises en forme et des contenus. Nous avons donc d'un côté la structuration logique du texte mise en lumière au moyen de balises. Ces balises permettent de caractériser des portions au sein d'un texte, ce que les informaticiens nomment par un abus de langage la sémantisation du texte. C'est le langage XML (eXtensible Markup Language). La présence d'une balise indique que la portion de texte a une caractéristique et le nom de la balise indique de quel type de caractéristique il s'agit. De l'autre côté, nous avons la mise en forme du texte, c'est-à-dire les interfaces de navigation. Dans ce contexte, il est important que tous les sites utilisent une même grammaire de balisage qui détermine le nom des balises et leurs règles d'utilisation. Il est alors possible de faire des requêtes sur les balises dans plusieurs sites, c'est-à-dire de faire des recherches dans plusieurs textes grâce aux différents types de caractéristiques, même si les interfaces de navigation propres à chaque site sont différentes.
- 13 Ces grammaires sont appelées DTD (définition de type de document) ou schémas. Or, depuis 15 ans, il en existe une qui correspond assez bien à nos besoins : la TEI (Text Encoding Initiative). L'avantage de cette DTD est qu'elle est générique ; on peut ainsi décrire et structurer de la même façon les informations de types de sources complètement différents. Outre l'interopérabilité, ce principe de séparation du contenu et de la mise en forme garantit aussi la conservation, puisque les données ne sont pas spécifiques à une application et à un outil informatique qui deviendra rapidement obsolète. De plus, le fait d'utiliser un standard ouvert et libre, comme la TEI, permet aussi de garantir à tous l'accès aux règles de balisage, donc à la structure logique et à la compréhension du texte. Malgré tout, utiliser un standard n'est pas une garantie éternelle, car personne ne peut prédire les changements dans ce domaine, mais, si une communauté suffisamment importante utilise le même standard, nous pouvons espérer que des outils seront développés pour faire évoluer automatiquement nos textes.
- 14 Cette seconde phase de travail a abouti à la mise en ligne de la deuxième version du *Cartulaire blanc de l'abbaye de Saint-Denis* en octobre 2004 qui va être suivie bientôt par la sortie d'autres éditions. Une fois notre chaîne éditoriale mise en place, même si elle reste

encore perfectible, nous pourrions nous lancer dans une nouvelle étape composée de trois parties :

- La création de nouveaux types de publication complètement adaptés au Web et à ses particularités ;
- la possibilité de proposer aux lecteurs différentes vues de la source éditée : une édition de travail, une édition « imitative », la numérisation et l'édition proprement dite ;
- le traitement de l'information dont l'indexation est une partie intégrante.

- 15 C'est évidemment à cette partie que je vais maintenant m'intéresser, mais vous aurez compris que nous ne sommes qu'au début de notre réflexion.

L'indexation

- 16 L'indexation dans le cadre d'une édition scientifique de sources permet de simplifier l'accès à l'information grâce au repérage de noms propres et/ou de notions présentes dans le texte. Comme le rappelle Olivier Guyotjeannin dans le volume des Conseils pour l'édition de textes consacré aux actes et documents, la forme et l'importance de l'index dépendent souvent des règles de la collection qui accueille l'édition et des buts de l'éditeur scientifique. Ce travail est souvent long, fastidieux et peu valorisé, pourtant il s'avère essentiel et son absence est considérée comme une grave erreur scientifique. La composition des index présente donc souvent des disparités, mais nous pouvons essayer d'en brosser le portrait idéal.
- 17 Dans le cas d'un index nominum, il comporte :
- - le nom du lieu ou de la personne sous une forme lemmatisée ;
 - - l'ensemble des formes de ce nom rencontrées dans le texte et si besoin remis au nominatif ;
 - - des renseignements complémentaires sur le lieu ou la personne indexés.
- 18 De plus, les différentes entrées d'index sont souvent présentées hiérarchiquement, ce qui induit des relations entre elles (exemple). D'après cette description, il est indéniable que l'index constitue une première couche interprétative. Cela est d'autant plus vrai dans le cas d'un *index rerum* dont l'établissement est entièrement conditionné par la problématique de l'éditeur scientifique et à ce titre ne peut prétendre à l'exhaustivité. Mais, l'exploitation de tels index se révèle le plus souvent fastidieuse. Leur présentation sur le papier rend difficile un accès rapide à une information précise. Cette complexité se remarque assez bien dans le cas où un lecteur ne connaît qu'une forme particulière du nom recherché, puisqu'il doit passer en revue toutes les entrées possibles ou utiliser les renvois indiqués par l'éditeur scientifique.
- 19 Depuis une dizaine d'années, la mise à disposition de corpus de sources sur le support numérique a été multipliée, d'abord sous forme de cédéroms et aujourd'hui sur le Web. Comme je l'ai déjà évoqué, ce support offre la possibilité d'effectuer des recherches en texte intégral, par la recherche d'un terme ou d'une expression précise au sein du texte. Cela représente la plus importante innovation du support numérique, à tel point que les formulaires de recherche sont devenus le point d'accès principal, voire unique, aux textes sur certains cédéroms. Les possibilités d'interrogation et d'exploitation des sources ont été augmentées, en permettant d'effectuer plus rapidement des traitements sur un plus grand nombre de textes. Or, si nous effectuons un tour d'horizons des différents titres parus, un constat s'impose : l'absence quasi-systématique d'index. Au moins trois raisons peuvent expliquer cette absence. Le temps et l'investissement qu'auraient demandé leur

mise en place constituent à mon avis la première raison. En effet, les promoteurs de tels projets ont préféré mettre l'accent sur des outils de recherche performants et le nombre de textes, ce qui correspond, me semble-t-il, à une demande de la communauté des chercheurs. De plus, cette absence pouvait sembler être comblée par la présence de puissantes fonctions de recherche. En effet, alors que l'accès à l'information est accéléré par les moteurs de recherche, quel intérêt auraient représenté des index ? Cette concurrence constitue donc la deuxième raison : les index semblaient caduques, à moins de considérer la lemmatisation comme une forme d'indexation. Enfin, comme je l'ai déjà signalé, la plupart de ces corpus ne sont pas conçus comme des éditions scientifiques, mais plutôt comme des bases de données textuelles composées de sources parfois déjà éditées. La présence d'index n'a pas paru se justifier dans ce contexte.

- 20 Pourtant, comme le montre le thème de cette journée d'étude, les pratiques d'indexation réapparaissent. Plusieurs facteurs peuvent expliquer ce nouvel intérêt. Comme je l'ai déjà signalé, nous sommes actuellement dans une phase d'appropriation de ce nouveau média qui se caractérise par l'adaptation de types de publication existants sur le support papier. Dans ce cadre, nous sommes passés de la mise à disposition de bases de données textuelles à l'élaboration d'éditions scientifiques électroniques. À cet égard, l'expérience menée à l'École des chartes est significative. Or, l'indexation est une partie intégrante du travail de l'éditeur scientifique. La question du support ne rentre pas en ligne de compte, mais ce sont plutôt les conventions ou les habitudes qui conditionnent la mise en place des index.
- 21 De plus, on se rend rapidement compte que la recherche en texte intégral possède certaines limites, surtout si aucune lemmatisation n'a été prévue. Ainsi, dans le cadre des *index nominum*, deux intérêts sont à noter par rapport à une recherche en texte intégral :
- proposer des renseignements complémentaires utiles à l'identification ;
 - offrir une lemmatisation des noms propres qui présentent souvent les variations de graphie les plus importantes.
- 22 Par ailleurs, ces outils se fondent sur la comparaison de base du mot recherché et du texte. Il n'est pas possible d'effectuer une recherche sur le sens ou les synonymes, par exemple. Or, les *index rerum* peuvent répondre, même imparfaitement, à ce besoin.
- 23 Évidemment, de ce point de vue, le support numérique n'apporte rien de nouveau aux index. L'innovation se situe plutôt au niveau du statut de l'index à l'intérieur de l'édition, de son utilisation et de son exploitation. Ainsi, grâce à la technologie hypertextuelle, l'index passe d'un outil de repérage souvent difficile à utiliser d'un point de vue pratique à un outil de navigation à l'intérieur de la source. D'un simple clic, le lecteur atteint directement l'endroit indexé par l'entrée et surtout, constitue facilement un parcours de lecture en suivant les différents points du texte indexés pour une entrée. L'accès à l'information est plus rapide que sur le support papier et on assiste à un changement d'ordre cognitif qu'il est encore difficile de mesurer, mais qui est indéniable.
- 24 L'exploitation de l'index est aussi facilitée, puisque l'outil informatique rend possible la recherche sur l'entrée lemmatisée, sur toutes les formes rencontrées dans le texte et sur les renseignements complémentaires. Sur le support papier, cette possibilité de trouver une entrée d'index à partir d'une forme présente dans le texte est permise par le système des renvois à l'intérieur de l'index, dont l'établissement va être pris en charge automatiquement par l'outil informatique. De la même façon, l'index inverse, pratiquement inexistant sur le support papier, peut là aussi être géré automatiquement.

- 25 Enfin, une technologie Web émergente peut permettre de formaliser et d'exploiter les relations induits par la hiérarchie de l'index. C'est le concept de Web sémantique. Il est fondé sur la mise en place de ce que les informaticiens appellent des ontologies, c'est-à-dire des thesaurus. L'idée est ensuite de lier le thesaurus à des documents ou des parties de documents et ainsi indiquer le sens ou la notion abordée. La mise en place *ex nihilo* de thesaurus constitue la partie la plus longue et la plus fastidieuse de cette technologie. Or, la mise en place d'un index peut simplifier le travail, puisque on va construire le thesaurus à partir des mots indexés. Ainsi, l'éditeur scientifique dresse son index, il fabrique donc une liste de noms propres et de notions qui sont déjà reliés à un texte. Le travail consiste ensuite à relier les différentes entrées de l'index entre elles, en indiquant le type d'association. Automatiquement est ensuite généré un graphique représentant les différentes entrées reliées entre elles avec les renvois vers les documents. Dans le cadre des *index nominum*, on peut constituer des réseaux sociaux très intéressants dans le cadre de l'étude d'une région et dans le cadre des *index rerum*, une carte sémantique de l'information contenue dans le document.
-

AUTEUR

GAUTIER POUPEAU

École nationale des chartes