De la chaîne éditoriale à la plateforme de recherche : structurer, enrichir et diffuser de vastes collections numérisées de publications scientifiques

Nathalie Fargier
nathalie.fargier@persee.fr
ENS de Lyon, France

Brief Summary

Persée développe une plateforme de numérisation, d'enrichissement, de diffusion et d'archivage de publications francophones en SHS, de la 1ère parution à la période la plus récente (revues scientifiques, actes, livres). Résultat d'un travail collégial entre chercheurs, éditeurs, bibliothécaires et ingénieurs, cette plateforme diffuse près de 650 000 documents scientifiques en open access. Elle constitue une expérience originale et concrète pour explorer les enjeux croisés de la numérisation patrimoniale, de l'édition électronique et de l'analyse sémantique. La communication vise à décrire le passage de documents imprimés (représentation arborescente) à une publication numérique puis à un ensemble de données structurées (représentation réticulaire sous forme de graphe) et, à analyser la recontextualisation opérée et les transformations à l'œuvre. La bibliothèque numérique, l'entrepôt OAI et le triplestore Persée sont des points d'accès complémentaires et offrent des potentialités radicalement nouvelles de fouille de texte et de données, d'analyse, de mise en relation et de réutilisation.

Full abstract

Persée est une plateforme de production, d'enrichissement, de diffusion et d'archivage de collections de publications francophones en sciences humaines et sociales, de la 1ère parution à la période la plus récente (revues scientifiques en 1er lieu mais aussi actes de colloques et livres). Résultat d'un travail collégial entre chercheurs, éditeurs, bibliothécaires et ingénieurs, cette plateforme diffuse actuellement plus de 650 000 documents scientifiques en texte intégral et en *open access* et, elle poursuit son enrichissement.

Une décennie après l'ouverture de persee.fr, alors que la numérisation du patrimoine documentaire est massive et que de nouveaux modes de publication ont émergé, cette plateforme constitue une expérience originale et concrète permettant d'explorer les enjeux des collections numériques dans un monde connecté et les interactions entre numérisation patrimoniale et édition électronique. Dans le cadre de cette présentation, il est proposé de décrire les conditions du passage du monde de l'imprimé à celui du numérique, de volumes papier à un ensemble de documents structurés et reliés entre eux et, d'analyser la « r*econtextualisation* » qui a été ainsi opérée et les transformations structurelles qui sont à l'oeuvre.

Méthode de fragmentation et chaîne opérationnelle de traitement

Confrontés à des objets qui disposent d'une matérialité évidente (la revue comme un ensemble de documents papier) et d'une pertinence intellectuelle (la revue en tant qu'objet éditorial), nous avons assuré une modélisation des données pour mettre en évidence les structures des documents et leurs dépendances. Ces structures sont représentées dans un modèle qui contrôle la validité des documents. La chaîne éditoriale Persée réunit un ensemble de méthodes et d'outils permettant de constituer des corpus numériques en XML, de manière intégrée et largement automatisée, en vue de leur indexation fine, leur diffusion et leur archivage. Les métadonnées sont encodées selon les schémas Dublin Core, MarcXML et MODS. Le texte intégral issu de la ROC (reconnaissance optique de caractères) est disponible selon le schéma TEI et enfin, l'ensemble des données est décrit et organisé au sein d'un container XML au format METS. La publication des documents s'opère par des algorithmes de transformation qui s'appuient sur le modèle pour publier des documents dans des formats standards.

De l'arborescence au réseau : l'évolution de la représentation logique des documents

Les publications scientifiques imprimées sont organisées selon une logique arborescente avec, pour les revues, par exemple, un agencement de type : titre/année/tome/volume/numéro/article. Cette représentation abstraite fait l'objet d'une retranscription numérique jusqu'à un niveau plus précis qui est celui du plan des articles, de la liste des illustrations, des résu-

més et, des annexes, etc. Considéré isolément et intrinsèquement, l'article devient le nœud d'un nouveau maillage en se fondant sur l'exploitation des citations, des noms d'auteurs et des informations en son sein même. Nous mettons en œuvre un référencement croisé (cite / cited by) et un alignement sur des référentiels d'autorités Auteurs et des sources comme DBpedia. Une analyse plus fine des articles permet d'identifier des entités nommées et d'établir des liens avec des thesaurus disciplinaires. Loin de tout traitement massif, la méthodologie retenue combine des algorithmes de recherche ciblée soumis à validation humaine afin de garantir pertinence et qualité. Les objets numériques ainsi créés se distinguent fondamentalement du matériau papier de base et ils offrent des potentialités majeures en termes de recherche, de Text and Data Mining, d'analyse et de mise en relation. Ainsi, le numérique permet-il de référencer parallèlement la collection appréhendée comme un objet intellectuel à part entière, des segments constitutifs (article/communication/chapitre) et des données, de multiplier les points d'accès à un ensemble structuré, contextualisé et historicisé.

Ouverture de l'accès, ouverture du code source et ouverture des données

Dès l'origine, la voie de l'*Open Access* a été retenue pour la diffusion sans aucune restriction des métadonnées et des documents en texte intégral, l'ouverture et le partage étant considérés comme des instruments essentiels de visibilité et de circulation de la production scientifique dans un environnement où la langue française n'est plus en situation d'hégémonie. Selon une suite logique, les développements informatiques ont été opérés dans un esprit *Open Source* et les données Persée sont intégrées au web de données sous la forme de triplet RDF. L'objectif poursuivi est double : favoriser la réutilisation des contenus dans d'autres contextes que ceux qui ont vu leur création et l'accès de tous au patrimoine scientifique au-delà des cercles académiques.

Bibliography

Babeu, A. (2011). « Rome Wasn't Digitized in a Day »: *Building a Cyberinfrastructure for Digital Classicists*. London : CLIR Publication. 307p

Bachimont, B. (2007). L'indexation multimédia : description et recherche automatique. Paris : Hermès. *Nouvelles tendances applicatives : de l'indexation à l'éditorialisation.* P15-29

Pédauque, R. T. (2006) *Le document à la lumière du numérique. Forme, texte, medium : comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité.* Caen : C&F Editions. 218p

Salaün J M. (2007) « La redocumentarisation, un défi pour les sciences de l'information ». *Études de communication,* Num. 30, p13-23.

Vitali Rosati, M. (2016). What is editorialization? Sens public. Mars 2016 http://www.sens-public.org/article1059.html