

Persée et le web de données

L'UMS Persée, développe le portail www.persee.fr, une bibliothèque patrimoniale de publications scientifiques principalement en sciences humaines et sociales. Le portail diffuse aujourd'hui plus de 600.000 documents scientifiques en libre accès et en texte intégral.

L'équipe Persée est depuis longtemps convaincue de l'intérêt de lier ses données à d'autres sources d'informations. Cela se traduit par l'attribution de DOI aux documents, par un travail conséquent sur l'identification de citations entre les documents, par l'alignement de ses auteurs avec les référentiels nationaux ABES¹ et BNF², mais également avec des sources extérieures au monde des bibliothèques : wikipédia, RePEc, etc.

Persée a fait le choix de ne pas utiliser de procédures permettant de rechercher massivement des correspondances entre ses documents et des vocabulaires et/ou référentiels. Au lieu de cela, plusieurs méthodologies ont été mises en œuvre :

- des algorithmes de recherche ciblés permettent de dégager des liens candidats qui sont soumis à une validation humaine,
- des outils permettant d'interroger des référentiels « pivot » lors de la production des contenus de Persée,
- des outils exploitant le web de données et permettant de construire, à partir de ces liens validés, de nouveaux liens.

Cette démarche permet d'atteindre un haut niveau de pertinence dans les liens qui sont associés à chaque ressource diffusée par Persée.

Aujourd'hui, Persée franchit un nouveau cap en établissant des liens fondés sur les contenus des documents : deux projets aboutiront, fin 2016, à la mise en ligne de collections intégrant de nouveaux types de liens sémantiques :

- Le projet ATHAR, en collaboration avec le laboratoire inVisu de l'INHA, établit des liens entre des publications anciennes et un thésaurus³ d'identification des monuments du Caire ;
- Plusieurs collections en sciences de la vie et de la terre pour lesquelles des liens ont été établis entre les documents de Persée et la classification taxonomique exposée par GBIF⁴ (Global Biodiversity Information Facility).

Pour chacune des ressources indexées, le portail Persée proposera une page composite permettant notamment de :

- lister les contenus présents sur Persée (documents, iconographie, etc.),
- fournir des informations complémentaires issues des sites tiers,
- fournir des liens permettant à l'utilisateur de rebondir vers d'autres sources d'information.

Au-delà de cette exploitation sur le portail web, l'équipe Persée mettra à disposition de la communauté (fin 2016) un triple store permettant, dans une logique d'ouverture et de partage, d'exposer toute la richesse des données produites et/ou collectées.

¹ <http://www.idref.fr>

² <http://www.data.bnf.fr>

³ <http://cairogazetteer.fr/invisu/>

⁴ <http://www.gbif.org/>

Auteur :

Viviane Boulétreau,

Viviane.Bouletreau@persee.fr (@Vivi_lyon)

UMS Persée - Université de Lyon, ENS de Lyon, CNRS

Responsable du pôle Informatique et développement de l'UMS.

Ingénieur de recherche, je travaille depuis 1999, dans le domaine de la diffusion des résultats de la recherche : CyberThèses/CyberDoc, archives ouvertes et Persée. Quelques mots clés résument mon activité et mon engagement : open access, open source, open data, SGML→XML→RDF.