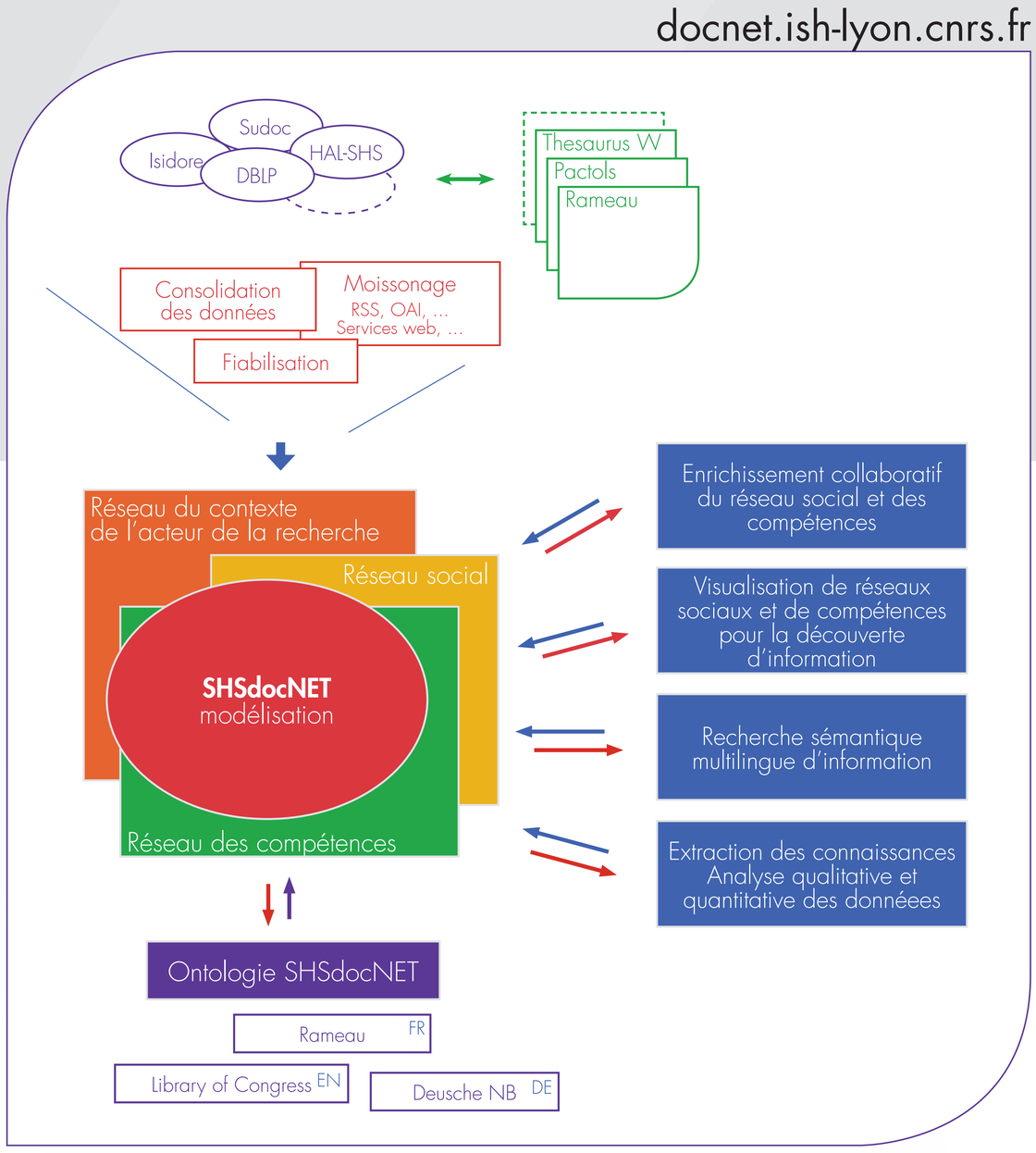
SHSdocNET

Technologies du web sémantique

docnet.ish-lyon.cnrs.fr



Résumé

SHSdocNET est un portail de cartographie et de valorisation des compétences en sciences humaines et sociales en région Rhône-Alpes-Auvergne. Les domaines d'expertise sont issus de productions scientifiques accessibles sur le web de données (plus de 80 000 publications).

Autour de ce portail, l'Institut des Sciences de l'Homme (ISH) a adapté ou développé des outils et méthodes informatiques et statistiques pour proposer aux acteurs de la recherche des analyses spécifiques quantitatives et qualitatives.

Les traitements et analyses des données du portail permettent notamment d'identifier les thématiques et les dynamiques des collaborations inter-laboratoires ou d'étudier plus finement la pluridisciplinarité telle qu'elle s'exerce au sein des équipes de recherche.

Mots-clés : web sémantique, ontologies multilingues, data mining, extraction de contenus, résolution d'entités, enrichissement sémantique, analyse latente, réseaux d'auteurs, visualisation.

Plateforme

Le portail SHSdocNET met en œuvre des technologies du web social et sémantique :

– **moissonnage** de données à partir de sources d'information et de documentation disponibles sur le web de données telles que HAL-SHS, ISIDORE, SUDOC, DBLP, etc. (OAI-PMH, RSS, API Rest) ;

– **indexation** des ressources collectées (moteur basé sur Apache Lucene) ;

– enrichissement par **analyse latente**, sémantique (LSA) ou par distribution (LDA), permettant d'établir des relations entre un ensemble de documents et les termes qu'ils contiennent, en construisant des « concepts » liés aux documents et aux termes ;

– **recherche sémantique** multilingue de compétences proches par des techniques s'appuyant sur les **ontologies** généralistes comme Rameau, Library of Congress, Deutsche Nationalbibliothek, et celle spécialisée en cours d'élaboration à l'ISH ;

– **fiabilisation** des informations s'appuyant sur des outils et des méthodes de fouille de données.

Exploitation et visualisation

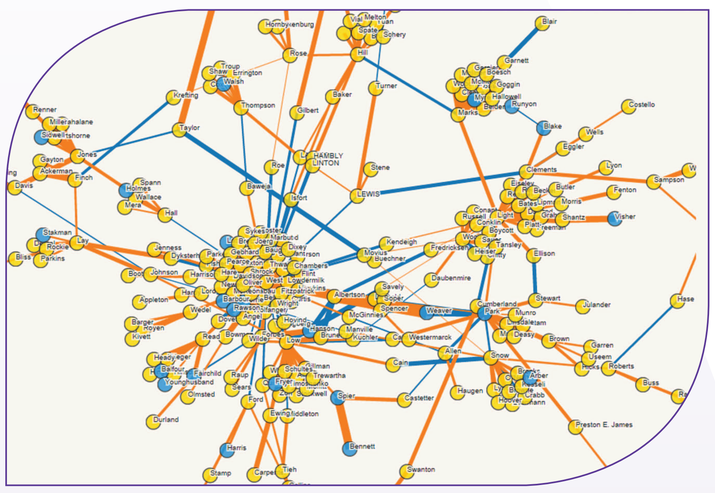
Au sein de PANELS et PAGES, les deux plateformes technologiques pour les sciences humaines et sociales coordonnées par l'Institut des Sciences de l'Homme.

Capture intelligente de réseaux d'auteurs

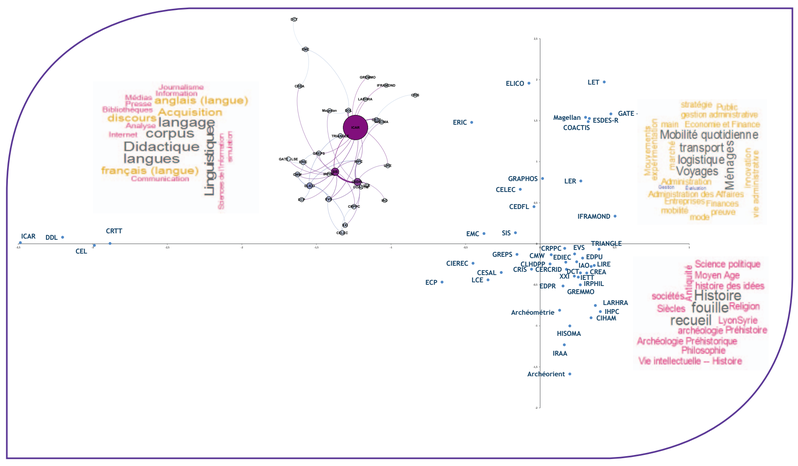
Rôle des communautés de recherche dans la politique environementale. Synthèse de réseaux d'auteurs hétérogènes à travers des publications liées à la crise écologique des années 1930 provoquée par le « Dust Bowl » aux USA.

Méthodologie : construction d'un référentiel fiable d'auteurs, détection automatique de la structure logique des documents et d'entités nommées, extraction de citations d'auteurs explicites et implicites dans les publications, construction des réseaux de collaboration (co-auteurs) et des réseaux de citations entre auteurs.

Langages, outils : Java, ParsCit, SolR, js, NoSQL.
Méthodes : modélisation de graphe, Force Atlas, clustering (lingo, STC).



Visualisation et interactions



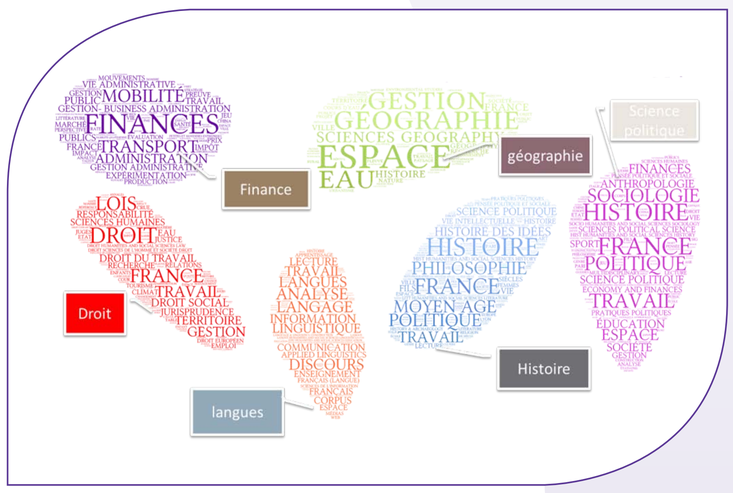
Visualisation des proximités et interaction des laboratoires à partir du corpus (titres, résumés et mots-clés). Les proximités sont quantifiables et basées sur la projection respectives des termes et des laboratoires sur le plan factoriel exprimant le maximum d'information (méthode factorielle AFC symétrique).

Les nuages de mots expriment les contributions de chaque terme à l'inertie totale de chaque axe. Le graphe exprime les collaborations entre les laboratoires en utilisant la modélisation de graphes (force-based algorithm), nous avons appliqué l'algorithme Force Atlas 2 qui a une complexité de O(N*log(N)).

Le travail permet d'avoir une approximation sémantique entre les thématiques sur lesquels les laboratoires et les visualiser grâce à une analyse de positionnement sur le plan factoriel.

Langages, outils : R, Gephi.
Méthodes : analyse factorielle des correspondances, modélisation de graphe.

Détection automatique de disciplines



Analyse de l'interdisciplinarité dans les SHS en confrontant des articles scientifiques contenus dans les bases de données bibliographiques sans domaine scientifique renseigné.

Mise en concurrence des méthodes issues du *machine learning* et de la statistique (svm, knn, tree, bagging, Random forests, boosting, logistic, linear discriminant analysis).

Comparaison des distances de similarités (Cosine, Pearson, Jaccard, Levenshtein) appliquées aux disciplines.

Classification d'articles scientifiques et mise en œuvre d'une approche pour la prédiction des disciplines des articles « SIM-DISP », qui se base sur les similarités entre les disciplines pour minimiser l'erreur de classification.

Langages, outils : R, WEKA, Java.
Méthodes : machine learning, prédiction.

INSTITUT DES SCIENCES DE L'HOMME

Institut des Sciences de l'Homme
14 avenue Berthelot • 69363 Lyon cedex 07 • France • Tél. 33 (0)4 72 72 64 64 • Fax 33 (0)4 72 80 00 08 • www.ish-lyon.cnrs.fr

Coordinateurs du projet SHSdocNET
Djamel A. Zighed • Tél. 33 (0)4 72 72 65 70 • abdelkader.zighed@ish-lyon.cnrs.fr
Sonia Guérin-Hamdi • Tél. 33 (0)4 72 72 79 32 • sonia.guerin-hamdi@ish-lyon.cnrs.fr

Développement informatique, traitement et analyse statistique
Sonia Guérin-Hamdi • Sofiane Bouzid • Adrian Tanasescu • Illyyne Saffar

USR 3385 • CNRS / ENS de Lyon / Sciences Po Lyon / Université Lumière Lyon 2 / Université Jean Monnet St-Etienne / Université Jean Moulin Lyon 3

ISH Communication - Illustrations : Sofiane Bouzid, Sonia Guérin-Hamdi (ISH) – JDEV 2015, juin 2015, Bordeaux (INP)