

# EXPÉRIENCES



**DSI de GE Money Bank, Arnaud Fritz a renégocié ses accords avec IBM et introduit de la variabilité dans son contrat d'infogérance.** P. 26

**Frédéric Charles préconise une organisation de la DSI en start up. Objectif : gagner en flexibilité et en réactivité face aux besoins des métiers.** P. 28



## DOCUMENTS

# Le CNRS utilise le web sémantique pour donner de la valeur à ses publications

Isidore, le portail de la branche sciences humaines et sociales de l'établissement public, centralise les publications de recherche, les enrichit sémantiquement et les ouvre sur l'extérieur via le format de description RDF.

**Thèses, blogs, enregistrements sonores, photos, bases de données, informations sur les colloques...** Depuis le lancement du portail Isidore en mars dernier (sur Rechercheisidore.fr), les publications de 331 laboratoires de recherche de la branche sciences humaines et sociales (SHS) du CNRS sont accessibles en un point unifié. Plus de 1,2 million de documents, émanant de plus de 1 000 sources, sont déjà consultables par les équipes internes, les chercheurs extérieurs ou étrangers, les étudiants ou les journalistes...

### Des données consultables par d'autres référentiels

Principale originalité du projet, au-delà des volumes manipulés : les contenus référencés sont structurés conformément aux standards du web sémantique. Plus précisément, ils respectent le modèle de description RDF (Resource Description Framework). Cette syntaxe particulière des métadonnées, qui s'appuie sur des trios sujet-verbe-complément, les rend directement interrogeables par d'autres applications ou référentiels supportant RDF.

Le CNRS n'en est pas à sa première initiative en matière d'agrégation de publications. Il disposait déjà d'un système d'édition scientifique en ligne, Revue.org, et de plusieurs plates-formes



Ingénieur de recherche au CNRS, Stéphane Pouyllau préside au développement d'Isidore.

de dépôt d'archives ouvertes, la principale étant Hal-SHS (Hyper article en ligne-Sciences de l'homme et de la société). Ce sont d'ailleurs ces initiatives, arrivées à maturité en 2007, qui ont encouragé l'organisme à envisager un niveau supérieur de diffusion de l'information. Le déclenchement final du projet a été amorcé par la publication de deux rapports, l'un du CNRS,

l'autre ministériel, tous deux pointant l'importance du partage des données numériques dans la recherche. Lorsqu'on lui confie la mission Isidore, en 2007, le TGE (très grand équipement) Adonis, une structure interne du CNRS-SHS chargée de valoriser le numérique dans la recherche en sciences humaines et sociales, s'oriente d'abord vers un métaportail agrégeant

JIM WALLACE

les sources existantes. Mais il s'avère que la seule fédération est insuffisante. « Nous voulions que le système de référencement apporte une plus-value en qualifiant les contenus des producteurs de données, rapporte Stéphane Pouyllau, ingénieur de recherche au CNRS, qui dirige le développement d'Isidore. Avec les comités scientifiques, nous avons alors élaboré un processus d'enrichissement des documents. Ce qui s'est ensuite traduit par la création de la chaîne de traitement aujourd'hui au cœur d'Isidore. »

Cet enrichissement s'effectue d'abord sur le plan sémantique. Les mots clés identifiés dans les documents sont rapprochés avec des termes présents dans des référentiels scientifiques. Isidore étant relié à des données métier (archéologie, géographie, histoire...) et techniques (géolocalisation, nomenclature des disciplines scientifiques, périodes chronologiques). « Ce traitement sert à qualifier l'information et à l'insérer dans une hiérarchie. Un article contenant "villa gallo-romaine à plan carré" sera rattaché au niveau supérieur, soit "villa gallo-romaine". L'utilisateur pouvant monter ou descendre d'un cran, élargir ou resserrer sa requête », détaille Stéphane Pouyllau.

**Une signature sémantique pour chaque catégorie**

Autre élément névralgique de la chaîne de traitement sémantique : la catégorisation des documents dans une ou plusieurs disciplines scientifiques. Pour réaliser cette opération, Isidore a dû se prêter à un exercice d'apprentissage : il a ingéré la base d'archives scientifiques Hal-SHS, soit plus de 30 000 documents déjà catégorisés manuellement par les chercheurs. A l'issue de cette étape, chaque catégorie a été associée à une signature sémantique, c'est-à-dire à un corpus de mots. Dès lors, la classification d'un document s'effectue, par comparaison de cette signature, d'une part, et par son rapprochement sémantique avec des documents déjà catégorisés, d'autre part. Seulement, les deux traitements affichaient souvent des résultats contradictoires : le moteur d'enrichissement sémantique rapprochait certains articles catégorisés par les chercheurs de contenus appartenant à une discipline différente, quoique le plus souvent connexe.

« Pour chaque discipline scientifique, nous avons exploré avec les chercheurs les publications qui sont à la frontière. Puis nous avons réglé en conséquence le poids de l'enrichissement sémantique ou de la catégorisation manuelle. »

**La qualification de l'information et sa catégorisation sont les plus-values du traitement sémantique**

Reste qu'avant de procéder à ces traitements, Isidore doit déjà s'attaquer au format des contenus présentés par les producteurs de données. Le portail accepte des producteurs trois formats

ensuite en RDF. Principale difficulté de l'opération : faire accepter aux chercheurs les modèles de transformation des documents vers ce format. « Par exemple, à la différence du protocole OAI-PMH, qui ne reconnaît pas la notion de filiation entre documents, RDF permet de réaliser des hiérarchies entre fichiers parents et enfants. Il a donc fallu que les scientifiques amendent notre façon d'organiser l'information au sein d'Isidore. Avec, au final, un meilleur enrichissement. » Lorsque la source est déjà décrite en RDFa, nul besoin de transformation. La récolte d'information est alors simplifiée car le producteur indique au sein des balises HTML les ressources RDF à mois-

CNRS, il place également Isidore dans la lignée des projets open data, un mouvement militant pour la libération et la réutilisation des données publiques. C'est d'ailleurs l'un des plus gros chantiers de ce type en France. En offrant un point d'accès utilisant SPARQL (langage d'interrogation des entrepôts RDF), il permet à des référentiels extérieurs de récupérer tout ou partie des métadonnées sémantiques. Elles deviennent alors réutilisables selon les clauses de la licence Creative Common, qui impose, notamment, la citation des sources et le partage des résultats de cette réutilisation selon les mêmes conditions. Paradoxalement, cette transparence engendre certaines

documentaire sont une infime minorité. Depuis longtemps, les chercheurs ont besoin de diffuser de l'information et de contractualiser des partenariats avec des laboratoires dans le monde entier, et donc de partager leurs données, surtout dans le domaine des sciences humaines et sociales. » Les premières mises en ligne de données de la branche SHS du CNRS remontent ainsi à 1997.

**Cinq fois moins de temps pour une même recherche**

Les gains sont ailleurs. Dans la qualification des sources et les associations sémantiques des contenus, et dans le temps économisé. Pour une même



L'AVIS DU DIRECTEUR

**Marin Dacos,** directeur du Centre pour l'édition électronique ouverte (Cleo)

**Notre laboratoire héberge trois plates-formes : Revue.org, Calenda et Hypotheses.org.**

Nous avons dû procéder à quelques enrichissements documentaires, mais nous répondions en partie aux deux formats (OAI-PMH et RDF) préconisés par Isidore. La mise à niveau des plus petits laboratoires, qui ne disposaient pas de compétences informatiques sur le web sémantique, a été plus problématique. Reste que les barrières d'entrée d'Isidore sont volontairement faibles (les flux RSS sont acceptés), histoire d'inclure un maximum de laboratoires et de les faire progresser.

**Le CNRS a récemment dénombré plus de 2 000 bases de données de recherche.** Apparemment, seules 10 % d'entre elles seraient interopérables avec Isidore. Il faudra globalement dix ans pour les faire toutes évoluer. En attendant, les laboratoires devront s'aligner sur des bonnes pratiques, comme favoriser au maximum le découpage des informations plutôt que leur agrégation, ou encore pousser à l'adoption et au respect de conventions de saisie.

**PUBLICITÉ**

Détails et conditions de l'offre sur sfrbusinesssteam.fr. SFR Business Team, marque du groupe SFR, est à destination des entreprises.

standards : RDF, RSS/Atom (le plus basique) et OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), un protocole utilisé pour le moissonnage de métadonnées des archives ouvertes. Ce dernier offre l'avantage d'être reconnu par de nombreux moteurs de recherche dont se sert la communauté scientifique. L'objectif étant d'aligner ces formats sur le standard RDFa (RDF encapsulé dans une page web). Pour les documents exposés en RSS et OAI-PMH (les plus nombreux), la gestion documentaire applique une moulinette qui extrait des métadonnées (titre, auteur, mots clés), stockées

sonner dans la page (titre, image, paragraphe, notes...). Ce qui évite de récupérer les éléments inutiles.

**Un important projet open data**

Actuellement, le nombre de sources respectant RDF est assez limité. Mais les volumes de documents à traiter sont conséquents. A l'image des collections exposées sur hypotheses.org, la plateforme de blogs scientifiques (20 000 entrées) utilisée par les chercheurs, ou sur le portail Calame, le catalogue en ligne des archives de l'enseignement supérieur. Si le format RDF facilite ainsi le moissonnage de documents au sein du

réserves chez les chercheurs. Comme la crainte que leurs données se noient dans cet immense réservoir. Ou celle, plus sourde, que leurs travaux soient surveillés de trop près par des chercheurs étrangers. Et pas n'importe lesquels : le deuxième pays après la France à consulter Isidore est les Etats-Unis (et ses universités). Notons au passage qu'Isidore n'a pas d'équivalent outre-Atlantique. Difficile de mesurer les gains arithmétiques de ce portail. A-t-il permis aux laboratoires du CNRS de rendre publics plus de travaux et d'informations ? « Pas vraiment. Les producteurs ayant attendu Isidore pour ouvrir leur socle

recherche, il aurait été divisé par cinq, selon une étude interne au CNRS. Jusque-là, les chercheurs et les étudiants devaient en effet réitérer leurs requêtes sur quatre ou cinq moteurs différents, chacun avec son vocabulaire d'interrogation. D'ici à 2012-2013, le TGE Adonis projette de mettre Isidore – sous forme de marque blanche – à la disposition des collectivités territoriales désireuses de proposer des contenus en lien avec leurs activités ou leur région. Il entend également étudier dans quelle mesure son système documentaire est applicable aux sciences dures du CNRS. VINCENT BERDOT

**! À SAVOIR**

**Le projet Isidore**

- **Budget :** 600 000 €.
- **Intervenants :** Atos (maîtrise d'ouvrage), Centre pour la communication scientifique directe (maîtrise d'œuvre), Sword (intégration), Antidot (gestion documentaire, moteur de recherche, catégorisation, classification, web sémantique avec Information Factory), Mondeca (gestion des référentiels avec Topic Manager).