

ISIDORE, de l'interconnexion de données à l'intégration de services.

Yannick Maignien CNRS UPS 2916

Le CNRS a communiqué largement, le 4 avril 2011, jour de la Saint Isidore¹, sur l'inauguration de sa plateforme SHS, ISIDORE.

Cette plateforme donne accès sur le Web à plus d'un million de documents et publications de la recherche française en sciences humaines et sociales, issues de plus de 800 sources différentes qui sont moissonnées, analysées, traitées et enrichies automatiquement. Rien à voir donc entre cette plateforme savante, et Google avec son indexation de l'ensemble du WEB...mais aussi l'incapacité du Leviathan de Mountain view à interconnecter de façon riche des catégories de connaissance (autre que pour son seul dessein de régie publicitaire) !

Cette information, relayée et amplifiée par la communication de la société ANTIDOT a été largement reprise par de nombreux sites d'informations, soulignant l'appartenance de ce projet à la mouvance Web 3.0 du Web de données ou Web sémantique. ISIDORE était déjà livré sur le Web depuis décembre 2010 en version beta. Comme le dit Pierre COL (ANTIDOT) : « A cet égard, le projet ISIDORE, du CNRS TGE Adonis, constitue effectivement le plus grand projet « web des données » / Linked Data / Open Data mené à bien en France à ce jour ».

Ayant eu la responsabilité de conception et de réalisation de cette plateforme en dirigeant le TGE ADONIS, sous l'autorité du Comité de Pilotage présidé par Michel Spiro, de mars 2007 à septembre 2010, je suis particulièrement heureux et fier de cette réalisation collective et de cette annonce inaugurale de la part de la Direction du CNRS.

À juste titre, les communiqués et commentaires relèvent plus précisément l'ampleur des collaborations qui ont présidé à cette réalisation. À cet égard, la maîtrise d'ouvrage du TGE² a bénéficié de l'assistance d'Atos Consulting, pour élaborer le cahier des charges et définir les conditions de maîtrise d'œuvre. Au terme de cette phase, la Direction générale du CNRS décidait en juin 2009 de procéder par l'ouverture d'un marché public plutôt que de confier cette réalisation à l'INIST. L'ambition du cahier des charges de s'orienter résolument vers le Web de données justifiait largement ce choix. Dans le même mouvement, il était décidé de déléguer la maîtrise d'œuvre au CCSD³ et de passer un marché de réalisation, en octobre 2009

¹ J'avais choisi en effet cet acronyme, ISIDORE non seulement pour sa signification : Intégration de Service et Interconnexion de Données de la Recherche et de l'Enseignement supérieur, mais aussi en référence à Isidore de Séville, (v. 560 – 4 avril 636 ap. J.-C.) auteur des *Etymologies* et choisi comme Saint patron des informaticiens et de l'Internet du fait de l'organisation particulière de son oeuvre encyclopédique. Comme le relève Régis Robineau, du site INSULA, ce n'est nullement une coïncidence avec le choix de l'acronyme ISIDORE. Un autre Collègue me faisait remarquer, à l'époque, que l'on pouvait entendre aussi ...*Easy door*.

² Assumée avec maîtrise surtout par Stéphane Pouyllau et Jean-Luc Minel.

³ Où Daniel Charnay et Laurent Capelli ont joué un rôle décisif

grâce au « Plan de relance gouvernemental », contractualisant avec le consortium ANTIDOT, MONDECA, SWORD⁴.

C'est en effet une réalisation importante du CNRS, ayant impliqué de très nombreux acteurs des Sciences humaines et sociales, laboratoires du CNRS, mais aussi de l'Université ou équipes de recherche privées, dès le début des travaux du TGE Adonis, ceci par le biais de deux appels à projets, en 2007 et en 2009. Ces procédures ont permis d'associer de nombreuses compétences et possibilités d'innovation déjà en oeuvre dans environ 70 équipes, soutenues financièrement sur des objectifs très ciblés. C'est à travers ce dialogue fourni, complexe, parfois conflictuel mais toujours fructueux que s'est élaboré et affiné le cahier des charges finalement rédigé en mai juin 2009, en fonction de ces contraintes utilisateurs ou des typologies de donnée concernées (image, texte, enregistrement sonores, manuscrits, données quantitatives, etc.) .

Ce réseau de projets, tissé patiemment pendant 3 ans est essentiel à plus d'un titre, car il a permis largement de conditionner les accords d'accès aux données des laboratoires, mais aussi des revues agrégées par Revues.org, Cairn et Persée, puis moissonnées et indexées par ISIDORE et maintenant accessibles pour les usages de la recherche. De plus, fin 2009, un accord avec la BNF permettait d'associer également les données patrimoniales de GALLICA, si importantes pour compléter ou référencer nombres de données secondaires de la recherche dans les humanités, mais aussi les sciences sociales.

En terme de méthodologie de projet, cette élaboration théorique d'un cahier des charges (dont la portée finale requerrait les technologies du Web sémantique et l'expression des données en RDF) via la construction d'un réseau de travail (au sens propre du *NetWork*) par appels à projets, mériterait de plus amples développements⁵. C'est en tout cas la raison d'être du très court délai de conception et de réalisation et de l'efficacité de cette étape importante de la construction de l'infrastructure numérique du Très Grand Equipement ADONIS. Ajoutons cependant que l'époque, et les technologies, étaient sans doute mûres pour que cette réalisation se fasse aussi efficacement. Rappelons qu'une première initiative du TGE ADONIS, lancée en 2006 avait rapidement échoué, sans doute prématurée⁶

Cela est largement souligné, l'appartenance de ce projet à la mouvance du Web de données, à l'ouverture des données, est ce qui signe son originalité. D'une part, comme le précise Pierre COL d'ANTIDOT, soulignant cet aspect de l'interconnexion de données grâce au format RDF:

« L'Open Data, et plus largement la vague du Linked Data et du « web des données », concerne les États, avec leurs administrations et services publics, ainsi que les collectivités locales et aussi toutes les organisations, y compris les entreprises privées, petites et grandes, qui ont intérêt à partager ouvertement certaines informations (pas toutes évidemment) avec leurs clients, fournisseurs, partenaires, bref avec leur écosystème. (...) Et les technologies du

⁴ Ce marché public était d'un montant d'environ 600 K€, ce qui est un investissement minime au regard de l'ambition réalisée.

⁵ Benoît Habert, alors directeur adjoint du TGE ADONIS, a joué un rôle décisif dans cette étape de conception « collaborative ».

⁶ Cf. Un écho dans la presse, encore en ligne :

http://www.lefigaro.fr/sciences/20061110.FIG000000064_cnrs_le_scandale_d_une_numerisation_ratee.html

web des données, ou web 3.0, en donnant directement accès à des données interconnectées plutôt qu'en ouvrant des API spécifiques à chaque source d'information ou silo de données, apportent un gain considérable en matière d'interopérabilité. (...) Les métadonnées de tous ces documents ont été normalisées et alignées sur des référentiels et thésaurus scientifiques, automatiquement classifiées, articulées entre elles et enrichies et, pour finir, publiées dans un triple store RDF de plusieurs millions de triplets, où elles sont librement interrogeables en SparQL. Une démo de ce qu'il est possible de développer à partir de ce point d'entrée SparQL est disponible ici : <http://www.lespetitescases.net/semweblabs/isidore/>⁷ ».

Le développement d'ISIDORE, à partir des recommandations du W3C sur l'utilisation de RDF pour le Web de données, était explicite dès 2009⁸, non seulement pour ce volet de l'interconnexion des données, mais aussi pour l'intégration des services que cette interconnexion devait permettre à terme⁹ : « La transformation des bases de données scientifiques en Web de données est au coeur des réalisations des infrastructures numériques pour les sciences : plateformes de publication, puis collaboratives Web 2.0 (Blogs, Wiki, etc.) et plateformes de calcul (grilles, traitement, bases de données relationnelles) ; elles doivent être des plateformes d'intégration de services, faisant converger le rôle de nombreux opérateurs différents, et d'autre part une interconnexion de données hétérogènes des laboratoires agrégées à partir de sites distribués ».

Si on réduit ISIDORE à l'interconnexion de données, il peut sembler que ce « moteur de recherche » n'apporte pas grand chose de plus que Google, sauf à travailler sur le seul champ des sources SHS indexées. Mais ce ne serait qu'une apparence trompeuse. Potentiellement, cette interconnexion de données, issues de sites Web divers, hétérogènes, ouvre sur une possibilité de constituer des corpus nouveaux de données pour des problématiques nouvelles, à partir de l'autonomie des acteurs détenteurs ou producteurs de ces données. C'est cette constitution de *Triple Store* RDF nouveaux, librement interrogeables en SparQ, constitués de masse énormes de triplets, qui doit faire émerger de nouveaux services scientifiques. C'est cette problématique d'intégration de services qui sera amenée à se développer à l'avenir.

On le sait, la croissance de Google dans le domaine des services est largement exogène, horizontale, conquérante, à partir du moteur de recherche d'information et vers tous les terrains successifs des usages. Pour autant, les données restent disjointes de ces différents usages (à part pour des *mash up* simples comme le couplage entre géolocalisation de Google map et des données textuelles). Mais une véritable productivité par croisement intégré (par des traitements et calculs) est impossible hors de l'expression en RDF de ces données, hors de l'autonomie des données et de la liberté des producteurs de données à entrer dans ces réseaux

⁷ Les Petites Cases, Site de Gautier Poupeau, qui a joué un rôle majeur dans cette orientation décisive du cahier des charges dans le sens du Web de données. Plus d'informations sur ce projet et sur les outils logiciels pour le réaliser, fournis par ANTIDOT (éditeur de logiciel français très impliqué dans les outils pour le web de données) :
- <http://bit.ly/CasClientISIDORE> (PDF de 4 pages présentant le projet ISIDORE)
- <http://bit.ly/AIF-v1> (PDF de 4 pages présentant la solution Antidot Information Factory)

⁸ Cf « Construire le web de données pour les sciences humaines et sociales » Stéphane Pouyllau, Shadia Kilouchi, http://halshs.archives-ouvertes.fr/sic_00494227/fr/

⁹ Cf. « Les nouvelles frontières numériques des sciences » Yannick Maignien, <http://www.tge-adonis.fr/Les-nouvelles-frontieres>

d'intégration. C'est cette révolution là, à venir, révolution fortement liée à l'aspect hautement collaboratif du travail scientifique, que Google est en train de freiner par son hégémonie centralisatrice.

À notre sens, cette intégration de service peut se développer, dans un contexte d'ouverture aux données publiques (Open data). Ce Cadre doit être favorable pour traiter non seulement les données administratives, mais également l'ouverture des données scientifiques, certes dans une problématique spécifique ; au même titre que la problématique d'ouverture des données culturelles¹⁰

Cette intégration de services devrait alors pouvoir se développer dans, au moins, quatre directions :

- l'intégration de champs culturels et linguistiques hétérogènes. Interconnecter des données d'autres langues suppose de croiser d'autres référentiels, d'autres communautés, d'autres traditions historiques et patrimoniales. La collaboration avec EUROPEANA, projet également avancé en matière de Web de données, devrait permettre de sceller cette internationalisation d'ISIDORE¹¹

- L'intégration de disciplines scientifiques autres que les SHS. L'interdisciplinarité est le champ privilégié du Web de données et des services qu'il peut faire émerger. La biologie est pour l'instant le terrain privilégié de telles interdisciplinarités , mais « restreintes » à des champs d'ontologies pour l'essentiel pré-établies. L'interconnexion de champs hétérogènes, hors référentiels permettant leur congruence, oblige à définir des services nouveaux intégrant les données et les finalisant dans des problématiques inédites, « polydisciplinaires », comme le dit Morin .

- L'intégration des acteurs. Cela suppose que les réseaux sociaux (et les données personnelles qu'ils comportent) soient réellement maîtrisées par les communautés qui les mettent en œuvre, et non par les possesseurs des logiciels d'applications Web 2.0 qui centralisent et utilisent ces données personnelles à des fins marketing. Ceci est un enjeu majeur des données nominatives sur la base desquelles tout un ensemble de pratiques académiques de « notoriété » et de contrôle sont institutionnalisées, véritable frein, pour l'instant, à la libre circulation des données et à l'émulation transparente des acteurs.

- L'intégration des fonctions logiques. Pour l'heure, les syntaxes par lesquelles les données exprimées en RDF permettraient des intégrations de services, se bornent à utiliser des logiques prédicatives. Une orientation serait d'exprimer les triplets RDF des données via des logiques modales. L'exploration d'hypothèses à partir de masses considérables de données pourrait alors passer par la puissance automatisée de calcul utilisant les gisements dispersés sur l'universalité du Web. Cette segmentation de « mondes » potentiels permettra de donner corps à des hypothèses « invisibles », indépendamment de ces connexions de données.

¹⁰ CF. Le rapport de M. Ory-Lavollée et Mme J. Pierre. « Partager notre patrimoine culturel »

¹¹ Le TGE ADONIS a également obtenu de coordonner, à partir d'un Siège parisien, avec des partenaires allemands (Niedersächsische Staats- und Universitätsbibliothek Göttingen) la phase de réalisation du projet européen DARIAH de la feuille de route européenne ESFRI <http://www.dariah.eu/>

Bien sûr, ISIDORE n'en est qu'à ses débuts, et ces orientations (et sans doute d'autres) devront être analysées, développées, discutées, critiquées, pour être graduellement mises en œuvre, et afin de mettre réellement le Web de données au service du Web des sciences, c'est-à-dire au service de communautés élargies de chercheurs.