

Compte rendu synthétique de la journée d'étude sur le Web sémantique

Le 27 juin 2008 à l'auditorium du MNHN à Paris, l'Aula, club des utilisateurs des produits Archimed, a organisé en partenariat avec la société Tosca consultants, une journée d'étude sur le Web sémantique.

En introduction, Marc Maisonneuve a rappelé la croissance exponentielle du Web, si rapide qu'aucune statistique n'en est désormais publiée. Diversification des langues employées, fort développement du Web invisible, absence de descriptif structuré des documents publiés... compliquent singulièrement la tâche des moteurs de recherche. L'enjeu du « Web sémantique » (chacun, au cours de la journée s'est accordé à préférer l'expression « Web de données »¹) est de faciliter l'accès aux informations disponibles sur le Web grâce à l'interopérabilité des métadonnées qui va permettre à ces moteurs de trouver et de mettre en relation des données jusqu'alors confinées dans leurs sites (ou leurs bases de données, puisqu'une grande partie du Web invisible est stocké dans des bases de données).

Raphaël Troncy du *Centre for Mathematics and Computer Science* à Amsterdam a ensuite dressé un panorama complet du Web sémantique, partant des activités du W3C (<http://www.w3.org/2001/sw/>) et de l'évolution historique de l'Internet depuis 1992, date de la première version d'html, pour aboutir à la présentation des travaux qui permettront aux robots de comprendre le contenu des sites, grâce à l'emploi de langages (le RDF ou *Resource Description Framework*) et d'ontologies (le OWL : *Web Ontology Language*) adéquats. L'idée de base qui soutient le Web sémantique est d'exprimer les métadonnées dans un modèle entité relation (qu'illustrent fort bien les FRBR) et d'identifier toutes les entités à l'aide d'URI (Uniform Resource Identifier, soit littéralement « identifiant uniforme de ressource »). La description des métadonnées à l'aide de RDF sous forme de triplets facilite le traitement automatique de l'information et le rapprochement des métadonnées issues de sites web divers et variés. On peut ainsi mettre en place un enrichissement progressif de la connaissance grâce à la propagation des propriétés d'inférence (A a une sœur qui se nomme B, B a un enfant qui se nomme C, donc A est tante de C). Le Web sémantique permet ainsi de construire des graphes RDF (modèle entité relation de métadonnées) de plus en plus gros et de plus en plus connectés : il s'agit bien d'un Web de données dont l'enjeu est l'interopérabilité des métadonnées. Certains projets sont déjà conçus dans la perspective du Web sémantique, citons DBpedia (<http://dbpedia.org/About>) ou Geonames (<http://www.geonames.org/>). Interconnectés, ils permettent d'enrichir de leurs données respectives les résultats d'une recherche.

En conclusion à son intervention, Raphaël Troncy, a montré le projet *MultimediaN N9C Eculture* (<http://e-culture.multimedien.nl/demo/search>) qui présente le résultat de travaux sur l'interopérabilité des vocabulaires employés par plusieurs musées ; ce projet illustre l'importance des URI attachés à chaque élément descriptif des œuvres présentées.

¹Citons Tim Berners-Lee, James Hendler et Ora Lassila qui en mai 2001 écrivent un article fondateur : *The Semantic Web : A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*, où les auteurs précisent " The Semantic Web will enable machines to COMPREHEND semantic documents and data, not human speech and writings".

vermeer, Jan
<http://e-culture.multimedien.nl/ins/gettyulan#500032927>

no image


Links

- permanent link
- annotate

Johannes Vermeer (October 31, 1632 - buried on December 15, 1675) was a Dutch painter, who lived and worked in Delft. He is also sometimes referred to as Vermeer of Delft or Johannes van der Meer. Alongside Rembrandt, Vermeer is the best known painter of the Dutch Golden Age, and his paintings are admired for their transparent colours, careful composition, and brilliant use of light.

| Property | Value |
|-----------------------|---|
| date of birth | 1622; 1632 |
| place of birth | Delft |
| death date | 1670; 1675-12 |
| place of death | Delft |
| preferred event | _file:///export/data1/e-culture.multimedien.nl/demo/RDF/vocabularies/getty/ULAN-painter.rdf#_Description2813 |
| gender | Male |
| preferred nationality | Dutch |
| ULAN preferred parent | Person |
| alternative role | painter; schilder; tekenaar |
| preferred role | artist |
| id | 500032927 |
| labelNonPreferred | Jan Vermeer; Vermeer van Delft, Jan; Vermeer van Delft, Jan Reyniersz; Vermeer van Delft, Johannes; Vermeer, Johannes; Vermeer, Johannes van Delft; van der Meer, Jan |
| labelPreferred | Vermeer, Jan |
| Source | Johannes_Vermeer |
| hasStyle | Baroque |
| Source | Rijksmuseum Amsterdam |
| type | Person (ULAN); Agent |
| sameAs | Vermeer van Delft, Johannes; Vermeer, Johannes |
| in scheme | Getty ULAN |
| name | Vermeer van Delft, Johannes; Vermeer, Johannes |

used as metadata in:

| Property | Subject |
|-----------|--|
| hasArtist | Baroque |
| Creator |  |

<http://e-culture.multimedien.nl/demo/search>

Olivier Walbecq de la société Archimed nous a ensuite présenté la position de sa société par rapport aux moteurs sémantiques et les partenariats qu'elle a engagés en la matière. Quelles sont les réelles nouveautés ? Les systèmes sont-ils vraiment intelligents ? Quelles sont les difficultés, les contraintes dans la mise en place de ces nouveaux outils et comment les intégrer dans l'existant ? Il nous a exposé comment ces moteurs prennent en compte les comportements des utilisateurs pour fournir des réponses pertinentes (quelles ont été les réponses choisies par les utilisateurs présentant les mêmes profils, quels sont les historiques de navigation, etc.).

Lucile Grand de la Direction des archives de France nous a ensuite présenté le projet de guichet unique d'accès à toutes les ressources en ligne du ministère de la culture et de ses établissements publics (soit 35 bases de données) afin de faciliter la consultation de ces ressources par le grand public et les professionnels. A l'arrivée : deux guichets uniques distincts, l'un pour la généalogie, l'autre pour tout le reste. Lucile Grand nous a présenté quelques exemples de recherche sur ces portails et les règles d'indexation qui ont été utilisées notamment pour l'accès aux noms propres de la base Nomina. (<http://nomina.france-genealogie.fr/nomina/>)

Nous avons terminé la matinée avec la présentation de Dominique Stutzmann, Archiviste paléographe, conservateur à la BnF qui travaille au sein du Département de l'information bibliographique et numérique. Il est revenu sur les enjeux que doivent saisir les bibliothécaires : Il convient d'exposer les données structurées dont disposent les bibliothèques et pour cela d'atomiser la connaissance en proposant une description au niveau le plus élémentaire. Cet enrichissement des données nécessite la reconnaissance d'entités nommées (de façon à ce que les moteurs reconnaissent les noms de personne, les lieux, les événements), leur catégorisation et leur

géolocalisation. On comprend que les bibliothèques sont très bien placées pour réaliser ce travail. Les données décrites en RAMEAU, par exemple, comportent déjà une multitude de termes retenus ou rejetés ce qui constitue un référentiel précieux qu'il sera possible d'« aligner » avec d'autres référentiels pour enrichir des corpus : mettre en relation les informations de Gallica avec celle de BDpedia, par exemple ou de Géoportail (<http://www.geoportail.fr/>)

L'après-midi a repris avec la présentation de Yann Nicolas de l'ABES qui, à son échelle, reproduit la diversité de corpus et de catalogues que connaît à une toute autre échelle, le Web. L'enjeu pour l'ABES était donc de permettre une interopérabilité des trois catalogues que sont le Sudoc, Calames (Calames est le catalogue des archives et des manuscrits des bibliothèques universitaires françaises, de grands établissements nationaux et de bibliothèques de recherche) et STAR (Signalement des thèses électroniques, Archivage et Recherche). Les trois défis peuvent être résumés à « interagir, s'exposer, s'enrichir ». Yann Nicolas nous a montré quelle a été la réponse des bibliothèques aux trois âges du web qui sont passées de l'interopérabilité à la mode Z39.50 puis au SRU/SRW pour aspirer aujourd'hui au Web sémantique qui permettra l'interopérabilité de tous les catalogues dans et en dehors des bibliothèques. Les étapes pour y parvenir consistent à modéliser les données contenues dans les catalogues en RDF (nommer les métadonnées de manière standardisée, attribuer un URI à chaque occurrence de ces métadonnées) ; exposer les métadonnées ; les lier à d'autres corpus ; permettre à d'autres de les interroger grâce à SPARQL². Au bout du compte et grâce à l'emploi d'un même formalisme RDF, d'un même vocabulaire et d'un même modèle FRBR, des corpus éloignés pourront être interrogés en une seule fois et de façon transparente pour les utilisateurs... ainsi bientôt les catalogues de l'ABES pourront être enrichis de ceux de la BnF, des Archives de France, etc. Cet objectif pourra être atteint tout en gardant une structure décentralisée au sens où chaque producteur sera libre de gérer ses propres URI, à condition que les alignements soient possibles (mise en correspondance univoque).

Isabelle Westeel de la BM de Lille nous a ensuite montré comment dans son établissement on prépare les métadonnées d'une base de documents numérisés de façon à ce que dans un futur proche elles soient visibles et accessibles pour tout. Elle a rappelé aux participants que les bibliothèques doivent rencontrer les usagers là où ils se trouvent, c'est-à-dire rarement sur les portails des bibliothèques. Pour permettre cette rencontre, Isabelle Westeel a repris les plaidoyers qu'avaient entamés Yann Nicolas et Dominique Stutzmann, à savoir : exposer massivement les données invisibles des bibliothèques et pour cela : trancher la question de la granularité (Faut-il décrire les données au niveau du document, de l'article, de la photo ?) ; travailler sur les questions de nommages des fichiers ; mettre en place des guides de bonnes pratiques.

Elle a rappelé que des travaux rigoureux doivent être dès à maintenant engagés par les bibliothécaires qui ont beaucoup d'atouts pour réussir cette tâche car ils connaissent les enjeux de la normalisation et de l'interopérabilité depuis longtemps ! Elle a mis en garde les bibliothécaires pour leurs futurs cahiers des charges où devront figurer clairement l'impératif d'URL pérennes et univoques.

² Le langage SPARQL définit la syntaxe et la sémantique nécessaire à l'expression de requêtes sur une base de données de type RDF et la forme possible des résultats.

La journée s'est terminée par la présentation de Gautier Poupeau qui a retracé l'histoire des catalogues des bibliothèques qui ont d'abord été accessibles via des interfaces complexes et dans une logique « top down ». Puis est venu l'âge du Web social qui a permis de « mettre la bibliothèque dans le flux » sous la contamination des pratiques issues du site marchand amazon.com. La logique « top down » s'est enrichi d'un flux « bottom up ». Enfin va arriver le web des données où la bibliothèque fera partie intégrante de l'infosphère dans la mesure où les données qu'elle décrit vont pouvoir être « portées » vers d'autres et mises en relation via des *mashup*³ réunissant le tout sur un seul site. L'ouverture progressive du catalogue va nécessiter un changement dans la granularité de l'information qui ne se fera plus au niveau de la notice bibliographique, mais, à l'intérieur de la notice, au niveau de chacun des accès (données contrôlées par un fichier d'autorité).

Enfin, le vendredi soir à 17 heures, les participants étaient encore nombreux pour écouter les quelques questions qui ont porté sur les outils qu'il sera nécessaire de développer pour interroger ces sites enrichis que l'on nous a dit extrêmement difficiles à développer...et sur la faisabilité d'un « super catalogue » qui réunirait toutes les critiques, commentaires, illustrations et fichiers numérisés qui se rapporteraient à un même ouvrage.

Cécile Touitou, le 30 juin 2008

³ Le Grand dictionnaire terminologique de l'Office québécois de la langue française traduit ce terme par application composite et le définit ainsi « Application qui amalgame le contenu provenant de différentes sources ou de différents éditeurs de contenu afin de fournir un nouveau produit ».