Séminaire IST Inria du 30/09/2012 au 05/10/2012 à Carnac (56)

« Le document numérique à l'heure du web de données »

Obsolescence ou persistance du "document"

Intervenant : Jean-Michel Salaün (ENS Lyon)

Même si le mot est ancien, la référence à la notion de document est récente dans l'histoire, sans doute en résonance avec l'organisation de la société industrielle, de sa régulation et de ses valeurs. On lui a parfois préféré dans les années 70 celle d'information qui insistait sur le contenu au détriment du support. Le web au tournant du millénaire s'est appuyé sur un renversement du circuit documentaire, jusqu'à, dans le web des données, un court-circuitage radical. S'agit-il de l'effacement d'une notion périmée au profit d'une autre ou d'un simple décalage ? Le succès du web accompagne des transformations sociales et économiques profondes. Quelles sont alors les conséquences sur nos régimes de vérité, de preuve, de transmission de l'éventuelle obsolescence de la notion de document ?

Histoire(s) de notices

Intervenant : Gautier Poupeau (Antidot)

Choisir, modéliser, décrire, classer, repérer, telles sont les missions des professionnels de l'information qui se sont incarnées et s'incarnent encore la plupart du temps sur l'établissement et la mise à jour de notices. Que ce soit Paul Otlet, Vanevar Bush ou même Suzanne Briet, leur système reposait sur la description d'objets ou de documents sous cette forme synthétique. Alors que les technologies du web sémantique semblent rebattre une nouvelle fois les cartes, nous voudrions interroger la notion de «notice» et en étudier l'évolution dans les trente dernières années afin de la situer par rapport à cette nouvelle étape et, ce faisant, dégager l'apport des professionnels de l'information dans la construction du web de données.

Les technologies du web appliquées aux données structurées

Intervenant : Emmanuelle Bermès (Centre Pompidou)

Il s'agit de dérouler "pas à pas" un exemple concret de transformation puis de mise en ligne d'un jeu de données suivant les principes du Linked Data. Cette démarche sera illustrée par un exemple concret d'application aux données culturelles.

Nous commencerons par les bases techniques du web sémantique, en déroulant les étapes suivantes :

- adopter l'architecture du web et en particulier les URI,
- définir un modèle de données,
- transformer les données en RDF,
- choisir les bons vocabulaires et référentiels,
- interroger les données avec SPARQL.

L'objectif de cette première partie est d'acquérir un niveau suffisant pour comprendre ces différentes technologies et éventuellement, préparer un cahier des charges en vue de la réalisation d'un projet de création de données en RDF.

Relier, réutiliser, partager : l'apport du web de données

Intervenant : Gautier Poupeau (Antidot)
En poursuivant l'exemple entamé précédemment, nous aborderons cette fois la perspective de l'ouverture des données sur le Linked Data :

- présentation des données disponibles dans le "Linked Data Cloud",
- aligner ses données avec d'autres,
- mettre à disposition des interfaces techniques (négociation de contenu, SPARQL endpoint),
- créer un service de type mash-up basé sur ces données.

L'objectif est de comprendre ce qu'est le web de données, et ce qu'il peut apporter dans le contexte des données culturelles et scientifiques.

Les référentiels : typologie et interopérabilité

Intervenant : Antoine Isaac (Europeana, VU University Amsterdam)
Tout d'abord, nous préciserons ce que la notion de référentiel peut recouvrir pour une approche linked data. A savoir, des artefacts de type:

- "metadata element sets" (ontologies formalisées à la OWL),
- "value vocabularies" (thesauri, fichiers d'autorités...),
- autres "datasets" (données sur les "objets du monde").

(cf. http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/)

Pour chacune de ces catégories, nous discuterons les caractéristiques de référentiels typiques. Nous nous intéresserons à leur provenance et à la manière dont ils ont été conçus (top-down vs. bottom-up, choix de modélisation). Nous insisterons également sur ce que la technologie Linked Data (et en particulier le rôle crucial des URIs) permet de changer par rapport aux artefacts qui correspondent à ces catégories, dans des approches plus traditionnelles. Par exemple:

- pour les thesauri et autres "value vocabularies", la transition de plus en plus visible d'une approche orientée termes (et noms) à une approche orientée concepts, voire "entités du monde réel",
- pour les schémas de données, les possibilités de réutilisation, voire d'"édition distribuée".

Si l'interopérabilité au niveau des "datasets" de référence aura pu être abordée dans la matinée, on pourra dans l'après-midi se concentrer sur l'interopérabilité au niveau des "value vocabularies" et des "metadata element sets". Pour ces deux familles on étudiera ce à quoi la notion d'alignement peut faire référence, ainsi que les scenarios de réutilisation par extension et/ou contrainte, en particulier au travers du concept d'"application profile" pour les schémas de données.

Donner du sens à des documents semi-structurés : de la construction d'ontologies à l'annotation sémantique

Intervenant : Nathalie Aussenac-Gilles (IRIT)

- **Partie 1 :** construction et peuplement d'ontologies à partir de textes
  - démarche générale
  - critères de bonne structuration d'une ontologie
  - outils de Traitement Automatique des Langues pour faciliter la construction d'ontologies
  - ouvertures
- **Partie 2 :** "donner du sens" à des contenus : l'annotation sémantique
  - associer des données et des modèles sémantiques
  - démarche générale
  - quel type de ressource pour caractériser "sémantiquement" des contenus/ des données ?
  - où l'on retrouve le TAL / ouvertures

Les biens communs de la connaissance, une utopie pragmatique

Intervenant : Valérie Peugeot (Orange Labs, association Vecam)
L'histoire pré numérique des biens communs de la connaissance.
La percolation du numérique :

- les biens communs comme réponse à l'abondance,
- le réseau infrastructure de production de communs.

Biens communs, marché et puissance publique : complémentarités et porosités.
Les biens communs au risque du politique :

- les nouvelles enclosures,
- les mouvements sociaux du numérique.

Du web de documents au web de données : implications juridiques de l'ouverture des données

Intervenant : Lionel Maurel (Auteur du blog S.I.Lex (BnF)
Le passage d'un web de documents à un web de données a eu des implications juridiques qui deviennent de plus en plus sensibles à mesure que progresse le mouvement d'ouverture des données. Dans le web de documents, les données étaient généralement "encapsulées" dans des documents, objets soumis au droit d'auteur ou intégrées dans des bases de données qui font l'objet d'un droit de propriété intellectuelle particulier. Ces régimes juridiques se sont rapidement avérés trop contraignants pour épouser les exigences d'ouverture qui faisaient jour avec le passage à un web de données. De nouvelles solutions juridiques ont vu le jour qui s'attachent à fluidifier et à encadrer la réutilisation des données. Deux voies coexistent actuellement qui ne reposent pas sur le même fondement juridique. Des modèles de licences libres ont été adaptées pour traiter le cas particulier des données, avec toujours un ancrage dans la propriété intellectuelle, mais en renversant ses principes de fonctionnement. Parallèlement, un droit des données publiques, d'origine européenne, a mis en place un autre cadre, consacrant un véritable droit des citoyens à la réutilisation des données et faisant lui aussi l'objet de déclinaison en licences.

Ces nouveaux cadres juridiques ont permis le développement des initiatives d'Open Data qui présentent à l'heure actuelle une grande variété dans les modèles juridiques déployés. Des différences sensibles existent au niveau local, national et international qu'il est intéressant d'analyser. Certains domaines présentent de fortes spécificités, comme celui de la diffusion des données de la recherche.

Les stratégies juridiques sont également différentes entre le secteur public et le secteur privé. Ces évolutions dessinent un cadre juridique mouvant et complexe pour le web de données, et fragile également, dans la mesure où les data peuvent toujours avoir du mal juridiquement à être considérées indépendamment des documents qui les incorporent.

Chaînes éditoriales et rééditorialisation de contenus numériques : enjeux actuels et prospectifs

Intervenant : Stéphane Crozat (UTC)
Les chaînes éditoriales XML sont orientées vers la création et la gestion de documents structurés. Les documents structurés sont des documents dont la structure est manipulable automatiquement, mais pas le contenu qui reste inintelligible à la machine. Elles vont ainsi au delà des approches traditionnelles de gestion documentaire via les métadonnées, en permettant d'adresser des éléments plus fins que le document.

Cette approche ouvre notamment sur de nouvelles possibilités de rééditorialisation, c'est à dire de réutilisation et de remise en contexte de fragments, via des techniques telles que le polymorphisme, la transclusion, la dérivation,

la paramétrisation... La rééditorialisation telle qu'elle est pratiquée dans les chaînes éditoriales impacte les processus documentaires à la fois au niveau de l'écriture et au niveau de la gestion, dont l'enjeu devient le maintien du sens au sein d'un réseau de fragments vivants.

Parmi les enjeux prospectifs l'on pourra citer :

- Le multimédia : Exemple de pratiques de rééditorialisation multimédia effectives ou en genèse (Ina, médiathèques, conférences, webdocumentaires...)
- Annotation et collaboratif : le lien entre document et activité (illustrations fonctionnelles à travers Scenari4, chaîne éditoriale collaborative issue du projet ANR C2M)
- Documents et données : quelle articulation entre le document et la donnée (après la séparation, les retrouvailles) ; "Data-isation" des documents (comment le document structuré peut être considéré comme une ressource riche pour des pratiques de type web sémantique ou web de données) ; chaînes éditoriales et traitement de données (quelles proximités, quelles différences ?)
- Enjeux sociétaux, organisationnels, économiques des chaînes éditoriales

Du cycle de vie des données au cycle de vie des objets

Intervenant : Alexandre Monnin (Paris 1, IRI, Inria)
Retour sur le cycle éditorial des données : à partir de l'exemple de DBpedia.fr voir comment l'on passe de contributions sur un Wiki tel que Wikipedia à des entités dans une base de connaissances faisant office de référentiel de fait pour le web de données. Quelles en sont les étapes, comment penser cette transformation, comment l'enrichir et la pérenniser, etc.

L'avenir du web au prisme de la ressource

Intervenants : Fabien Gandon (Inria) **et Alexandre Monnin** (Paris 1, IRI, Inria)

- Retour sur l'architecture du web, afin de récapituler un certain nombre de notions énoncées au cours de la semaine
- Éclairage sur la notion ressource en tant qu'elle permet d'unifier différents webs :
    - La ressource à l'épreuve du web de document (qu'appelle-t-on une page ?),
    - La ressource à l'épreuve du web sémantique (retour sur la crise d'identité - toujours en cours ! du web sémantique),
    - La ressource à l'épreuve du web application (un web de calcul ?),
    - La ressource à l'épreuve du web des objets (en quoi s'agit-il toujours du web ? Dissolution de la frontière sur/en dehors du web).