

“Une goutte de sémantique dans un monde de liens”

Une réunion du W2S sur le web sémantique à La Cantine, par Fabien Gandon (Inria)
<http://fabien.info>

Environ 30 personnes, dont plusieurs chercheurs, une personne d'Orange Consulting, plusieurs d'Orange Labs, formateur ESEO, entrepreneurs, chargés de mission, conseillers, thésards voire même quelques curieux. La salle du premier étage déborde sur le palier, c'est un signe indiscutable de succès !

La plus grande avancée du web sémantique est en train d'arriver, grâce à l'implémentation de RDF dans Facebook. Grosse génération de triplets RDF en cours : la croissance des triplets est plus importante que la loi de Moore. BestBuy vient de basculer son catalogue en format RDF. Lorsque des millions d'informations deviennent disponibles d'un coup, le standard prend tout son sens ...

Rappels sur les technologies – tout ceci est déjà assez ancien

Liens d'association (Vannevar Bush, juillet 1945) du mémex entre les informations qui nous importent, pour outiller l'information et structurer l'accès. Afficher un document entraîne l'affiche de documents associés – mémoire associative, mais seulement technologie mécanique disponible

Hypermédia, hypertexte (Ted Nelson, 1965) – une structure de fichier pour l'information complexe, changeante et indéterminée – mais ordinateurs isolés à l'époque

web (Tim Berners-Lee, 1989) – liens utilisant des références – à travers le réseau – premiers développement sur la station Next livrée au CERN – le premier navigateur permettait d'éditer n'importe quelle page, fonction qui disparaît à la sortie du CERN, et qui reviendra avec le wiki, mais plus tard

Dans le document original de Tim Berners-Lee, les liens sont typés, et l'objectif est de construire des URI. Finalement, l'essor se fera sur des liens banalisés, et avec des URL. Le typage des liens paraît complexe, et on ne voit pas bien son utilité.

Guerre du web, puis création du w3c en 1994 – consortium pour faire avancer les standards du web
échelle de lecture du w3C : working draft (travail collaboratif) → last call (document figé pour 3 mois) → candidate recommendation (deux prototypes) → proposed recommendation (Technical Architecture Group, modéré par Tim Berners-Lee) → recommendation (en vue d'usage généralisé par les grands acteurs, vision différente des standards techniques)

Méfiez-vous de ceux qui font référence à des documents qui ne sont pas des recommandations officielles : note, incubator group report, member submission

Fabien participera à un workshop sur RDF2 prévu en juin 2010, pouvant déboucher sur un nouveau groupe de travail w3c si suffisamment d'intérêt est démontré.

Le web sémantique est introduit par Tim Berners-Lee, lors d'une conférence www en 1994, en tant que meta-plan du web, et connexion des plans entre eux. Le typage des liens envisagé s'inspire des graphes conceptuels. <http://www.w3.org/Talks/WWW94Tim>

L'innovation réussit lorsque l'évolution est limitée (une idée) et progressive par rapport à l'existant (incrémentale) Si vous voulez réussir vous aussi, regardez ce qu'ont fait les autres ...

Les limitations des moteurs de recherche actuels – exemple : « book victor hugo » - faux positif (boutique sur le boulevard Victor Hugo) et absence de résultat (page sur la bibliographie de Victor Hugo, où ne figure pas le mot « book »)

Nous identifions et interprétons l'information (panneau « perdu »), les machines, non. Elles simulent.

Vers un web structuré : séparation du CSS du contenu (normé en 1996, utilisé vers 2000)

2008 : XML, puis explosion de spécifications complémentaires – amusant : Graffiti ML

Mais la structure n'est pas la sémantique ! <structure/> == <blabla/>

Nos échanges d'information entre humains sont basés sur des ontologies partagées, qui permettent le raisonnement.

Ontologie (=la science de ce qui est) n'est pas ontologies (= structures de connaissance identifiées derrière les occurrences réelles)

Sémiotique de Pierce : un signe est une entité qui représente une autre entité pour un agent. Trois types de signes : icône (montre la forme de quelque chose), indice (pointe vers quelque chose – « il n'y a pas de fumée sans feu »), symbole (représentation basée sur une convention)

Les ontologies sont des graphes orientés : « est un » ou « partie de » ou « équivalent à » ou ...

Réticence d'usage des ontologies : similaire à celle rencontrées lors de l'introduction des langages à base d'objets en son temps ; et la structure gérée dans l'ontologie doit être cachée, comme une base de donnée SQL n'est pas visible sur une page web.

Eléments d'architecture actuels :

- URI/IRI
- XML
- RDF (le web sémantique commence ici)
- RDF-S (ontologie légère)
- SPARQL (interrogation)
- Ontology : OWL (en trois couches)
- Rule : RIF
- Unifying Logic

RDF est un modèle qui décompose toute connaissance en triplet (sujet, prédicat, objet), équivalent en logique à un prédicat binaire, ou, en théorie des graphes, à un arc orienté.

RDF est aussi une syntaxe XML pour mieux échanger ces graphes.

Le modèle de diffusion du web s'appuyait sur l'extension des liens HTML. C'est le même modèle qui s'applique pour la diffusion de RDF dans différents domaines de connaissance.

Les principes de l'initiative Linking Open Data

1. Publier ses données en RDF (ce principe a été ajouté par Fabien)
2. Utiliser des URI pour nommer les éléments d'information
3. Utiliser des URI HTTP (URL) pour que l'on puisse les suivre (= consulter l'information)
4. Lorsqu'un URI est suivi, donner des informations adaptées (HTML pour un humain, RDF pour une machine) – deux formats différents, servis sur négociation de contenu
5. Inclure dans ces informations des URI vers d'autres données (pour découvrir d'autres liens)

Dbpedia est la version RDFisée de Wikipedia

Problème de traçabilité des triplets, confiance dans les équivalences des URI – exemple : les raies manta sur le site de la BBC, à partir des informations fournies par Wikipedia

Représentation des tags sous forme de triplets RDF : information (label, « tag », URL), acte social

(personne, « has tagged », URL), ...

Introduction des graphes nommés dans les futures propositions de RDF2 pour grouper les triplets par famille (origine des triplets, contexte d'usage, etc...)

Voir <http://sindice.com/> engin de recherche sémantique, qui fournit des sources de triplets, qu'un utilisateur humain peut importer. Consolidation des triplets attachés à un URI.

SPARQL, langage de requête envoyée à un SPARQL endpoint, langage de description des résultats, basé sur RDF, et aussi protocole transactionnel – 1.0 est normalisé, 1.1 est en cours de standardisation au W3C (working draft actuellement) avec un enjeu sur l'update – la recommandation officielle ne permet pas encore l'analyse de chemin sur les graphes sociaux, quoique l'INIRIA ait déjà développé des extensions ad hoc

70% des serveurs web sont derrière des pare-feux – toutes les communautés RDF ne sont pas ouvertes

L'un des enjeux est la décentralisation des déréférencements, pour ne pas répéter l'erreur faite sur DNS avec la centralisation de la gestion des TLD

Live Social Semantics - mélange web sémantique, web social et RFID, lié par SPARQL – badge de conférence émetteur unidirectionnel, avec opt-in, et mise en relation directe entre personnes en train de converser - « les meilleurs prédicats viennent des relations sociales »

Aujourd'hui Tim Berners-Lee promeut le web of data plutôt que le web sémantique. A-t-il introduit le concept de web sémantique trop tôt ?

Passerelles techniques entre le web et la structure :

- RDFa (RDF dans des attributs de HTML – exemple : introduction d'attributs Dublin Core dans HTML)
- GRDDL pour déclarer des profils d'extraction de RDF à partir de XML/HTML

Le web de données : les pages peuvent être soit consultées par un humain, soit traitées par un logiciel (RDFa ou GRDDL)

Voir plugin Tails Export <https://addons.mozilla.org/en-US/firefox/addon/2240/> pour Firefox

Voir activation de profil sur engin de recherche de Yahoo!

Voir javascript IMDB pour préparer l'actualisation de Facebook (« I like... »)

RDFS pour définir les relations, leur hiérarchie et leurs signatures

« Une petite goutte de sémantique peut générer une grande vague du web »

OWL, pour l'arsenal généraliste, avec toute la puissance des opérateurs ensemblistes (transitivité, symétrie, union, disjonction, intersection, ...)

OWL 2.0 avec sucre et sel en plus (compacité d'expression, et absence de propriétés)

OWL et OWL 2.0 sont au statut de recommandation depuis octobre 2009

Bon exemple d'usage de RDF : Creative Commons – petite ontologie (qui réutilise Dublin Core), forte viralité du système, intégration de la gestion des droits sous forme de méta-data dans les pages, intégré dans les processus d'indexation/restitution de Google (qui, pourtant, clâme ne pas faire de web sémantique)

Cambridge Semantics (spin-off MIT) système Anzo de plugin dans Excel pour typer les données saisies avec le schéma d'entreprise (le vocabulaire RDFS de l'entreprise) – application: envoi d'e-mail aux participants à un projet, pour faire de l'intégration de données dans l'environnement bureautique

web 2.0 === brouhaha 2.0

Comment aider à filtrer les activités issues des réseaux sociaux ? En revisitant les outils d'analyse disponible. En réseau social, les personnes incarnent plusieurs profils. En désactivant certains profils, on change les caractéristiques du graphe.

Ipernity.com en RDF fournit une base, transformé en FOAF.

Aaussi appliqué à l'ADEME (communautés thématiques)

Réutilisation du web sémantique en sortie, pour annoter les réseaux avec les caractéristiques identifiées, et crowdsourcing de cette information pour la faire vérifier par des humains.

En gestion de la connaissance, le cercle vertueux commence lorsqu'elle est intégrée à l'activité normale (assistance au job). Ca ne marche pas si c'est fait en plus du reste.

Beaucoup de recherche sur la base des distances ontologiques (ex: recherche de tutoriels sur le sans-fil) – l'enjeu : transformer les ontologies en espaces métriques

Simple Knowledge Organization System (SKOS) – schéma RDFS adapté aux bibliographies et thésaurus

Rule Interchange Format (RIF) – vers l'interopérabilité des règles d'inférence du web sémantique

POWDER, langage d'annotation d'un ensemble de ressources d'un seul coup – certification des assertions – listes, domaines, expressions

services web sémantiques, pour l'annotation sémantique des web services – description en RDF et recherche en SPARQL

Vers l'affordance sémantique

« L'avenir tu n'as pas à le prédire, mais à le permettre » Saint Exupéry

Mythe : On n'a pas besoin de moteur d'inférence pour utiliser une ontologie

Mythe : le web sémantique n'est pas le retour de l'IA par la petite porte

Question de Dominique : quid du web conversationnel ? Problème de contexte hyper-spécialisé et de surcodage des échanges – l'ontologie pour Skyblog n'est pas pour demain

Avenir : ontologie des hash tags de twitter

Avenir : début de thèse avec Alcatel Lucent sur l'application de ces concepts à la présence sur le web social (petits graphes, mais très dynamiques)

Le piège du domaine : le domaine d'application n'est pas forcément celui de l'ontologie

evidence based reasoning

« Je n'ai pas vu une ontologie universelle » (exemple : machine à café + téléphone en Corée)

Les ontologies sont à la fois des briques et des objets vivants (comment construire sur des bases instables ?)

Reuse !

VoCamp = camp pour hacker des vocabulaires, avec des experts du domaine et des usagers qui ont des besoins

« a lightweight ontology allows us to do lightweight reasoning »

le mot important, dans « web sémantique », c'est « web »

problèmes : scale, security, semiotics, mobile, hypermnnesia (le cerveau oublie, pas l'ordinateur)

Comment intégrer le web physique (objets connectés au réseau) ?

Il n'y a pas que la pile sémantique du web, mais tout plein d'autres (accessibilité, ...)

Fabien : « Celui qui controlera les méta données contrôlera les informations et services à toutes les échelles » La diversité des métadonnées est une condition nécessaire pour éviter la future main-mise sur le monde ! Attention donc

Fabien : « Pour gérer une diversité, rien de tel que d'utiliser une autre diversité »

Semantic media wiki : accessible uniquement aux geeks, à cause de la complexité d'expression des triplets dans le contenu