

Web Sémantique

Master 2

David Genest

Université d'Angers

Année universitaire 2009-2010

Chapitre I

Introduction

- 1 Un bref historique
- 2 Pourquoi le web est mal adapté à certains usages ?
- 3 Comment faire ?

- Le *web sémantique* est une *extension* du web qui facilite l'*automatisation* du traitement des connaissances disponibles.
- C'est une extension du web classique (HTML, HTTP, etc. ne sont pas remis en cause).
- Les connaissances ne sont pas représentées dans une langue naturelle mais formalisées à l'aide de langages pouvant être interprétés par des machines.
 - Agrégation de connaissances venant de plusieurs sources, comparaisons
 - Publication de données sémantiques sous différentes formes
 - Génération de nouvelles connaissances (inférence)
 - etc.
- Pas uniquement utile pour le web-internet.

L'initiative *web sémantique* est soutenue par le W3C.

Chapitre I

Introduction

- 1 Un bref historique
- 2 Pourquoi le web est mal adapté à certains usages ?
- 3 Comment faire ?

Historique

- 1989 : *Tim Berners-Lee* (CERN, Genève) commence le développement d'un système hypertexte.
- 1990 : Premières définitions pour HTTP, HTML, URL.
- 1992 : Premier annuaire de sites web. 26 sites.
- 1994 : Netscape Navigator 1.0, Fondation du W3C.
- 1995 : Microsoft ne croit pas au web, puis change d'avis.
- 1998 : Plus de 2 millions de sites. Création de Google.
- 2000 : XHTML 1.0.
- 2006 : 100 millions de sites.
- 2007 : Web 2.0.
- 2009 : Plus de 240 millions de sites.

Historique

Ce qui a changé...

- Nombre de sites
Nécessité d'avoir des outils pour rechercher des informations.
- Types des informations disponibles
Texte seulement, images, données issues de bases, documents techniques, etc
Certains types de données se prêtent mal à la recherche type moteur de recherche : Comparaison de documents, recherches « type SGBD », raisonnement sur les connaissances disponibles.
- Organisation des informations côté serveur
Web statique (HTML pur) \Rightarrow Web dynamique (CGI, PHP, SGBD, JSP, Java, Services web). Le web ne contient pas uniquement des pages mais aussi des « services ».

Chapitre I

Introduction

- 1 Un bref historique
- 2 Pourquoi le web est mal adapté à certains usages ?
- 3 Comment faire ?

Problèmes avec les langages du web : Données

Les données sont « cachées » dans le code HTML.

Exemple

Horaires de trains, horaires d'avion → documents HTML avec tables

Comment croiser les deux documents pour un trajet train puis avion ? Les documents HTML ne peuvent être utilisés (sauf ad-hoc) car les documents HTML sont une *présentation* des données.

Pourtant, à la base, les données sont souvent stockées de façon structurée : dans un SGBD.

Mais le schéma de la base des trains est sans doute très différent de celui de la base des avions.

Il faudrait une représentation « commune », utilisant un langage standard pour pouvoir croiser les données (automatiquement).

Problèmes avec les langages du web : Informations

Les informations sont « cachées » dans le code HTML qui contient l'expression dans une langue naturelle des informations.

... ou dans des images, des fichiers sonores, des vidéos, etc.

On peut utiliser des moteurs de recherche (sur le texte), mais pour des raisons de performance (et de taille du web), ces moteurs ne font aucun traitement sophistiqué (TALN) sur les textes → recherche de mots.

... ce qui est très différent de recherche d'informations.

Il faudrait représenter (indexer) les informations dans un langage standard, en utilisant un vocabulaire standard (contrôlé) (ou mieux une ontologie) → Comparaison de documents, possibilité de raisonnement pour résoudre une requête, prise en compte de documents multimédias, réponses formées de plusieurs documents ou de parties de documents.

Problèmes avec les langages du web : Services

Exemple

Achat de billets de trains, validateur html, web mail, etc.

Le service rendu est « caché » dans du code HTML.

Comment connaître ce que propose un service ? Comment utiliser conjointement plusieurs services ?

Mêmes solutions.

Meta-données : « données sur les données », association de données (exploitables par ordinateur) à ... tout ce qui peut être accessible sur le web (ou pas).

Chapitre I

Introduction

- 1 Un bref historique
- 2 Pourquoi le web est mal adapté à certains usages ?
- 3 Comment faire ?

Comment faire ?

... sur le web classique

Séparer la présentation du contenu...

- SGBD + Présentation (PHP, ...) : le SGBD n'est pas visible.
- HTML + CSS : mise en page « à part », mais toujours pas de description (utilisable par une machine) de ce que « contient le document ».
- XHTML : Évite le fouillis d'HTML, mais il s'agit toujours de documents.
- XML + XSLT → (X)HTML : Mieux, mais le XML n'est pas toujours visible... en plus, XML n'est pas un langage (mais un métalangage) : comment comparer deux documents XML écrits avec des DTD différentes ?

Comment faire ?

Solutions apportées par le web sémantique

- Utilisation d'un langage commun (RDF) pour exprimer des informations sur des ressources.
- Chaque ressource (document, personne, objet, etc.) est identifiée par un identificateur (URI).
- Expression d'assertions simples sous la forme de triplets (sujet, prédicat, objet). Le sujet est une ressource (URI), l'objet est une ressource ou un littéral, le prédicat est une relation entre les deux.

Exemple : « *La vie, l'univers et le reste* » a pour « auteur » « *Douglas Adams* ».

Mais il ne s'agit pas de trois chaînes de caractères :
identification (désambiguisation) du titre, de la relation, de l'auteur.