Introduction au Web Sémantique

Bernard ESPINASSE

Aix-Marseille Université
LIS UMR CNRS 7020

Septembre 2019

- **Bref historique**
- **Faiblesses du Web actuel**
- **Définition du Web sémantique**
- **Architecture du Web Sémantique**
- **Quelques applications du Web Sémantique**

Références

- **Livres, articles et rapports :**
  - O. Corby and F. Gandon and C. Faron-Zucker, Le Web sémantique : comment lier les données et les schémas sur le web ? Dunod, 2012.
  - G. Antoniou, Van Harmelen F., A Semantic Web Primer, The MIT Press Cambridge, Massachusetts London, England, 1999.
  - John Hebeler and Matthew Fisher and Ryan Blace and Andrew Perez-Lopez and Mike Dean, Semantic Web Programming, Wiley, 2009.
  - B. Menon, Comprendre les standards du web de données, https://www.cairn.inforevue-i2d-information-donnees-et-documents-2016-2-page-32.htm
  - …
- **Cours :**
  - Cours de M. Gagnon, Ecole Polytechnique de Montréal, 2007.
  - Cours de S. Staab, ISWeb – Lecture Semantic Web, Univ. Koblenz-Landau.
  - Cours de J.-B. Hook, Université Paris Sud, 2013.
  - Cours de C.A. Caron, Université de Lille 3, 2014.
  - Cours de D. Genest, Université d'Angers, 2009.
  - Cours de O. Papini, Aix-Marseille Université, 2015.
  - …

Plan

- **Bref historique**
  - Le Web … son histoire
  - Grandes générations du Web
- **Faiblesses du Web actuel**
  - Le Web aujourd'hui
  - Problèmes avec le Web actuel
  - Solutions aux problèmes du Web actuel
- **Définition du Web sémantique**
  - Du Web actuel au Web Sémantique
  - Web sémantique et Web des données et données liées
  - Problématiques du Web Sémantique
- **Architecture du Web Sémantique**
  - Briques représentation
  - Briques requêtes
  - Briques raisonnement
  - Briques confiance
- **Quelques applications du Web Sémantique**
  - Web Sémantique et commerce électronique
  - Web Sémantique et gestion des connaissances

1. Bref historique

- **Le Web … son histoire**
- **Grandes générations du Web**

Le Web …

- **Le WEB** = Système hypertexte public fonctionnant sur Internet
- Permet de consulter des **pages** (page Web) mises en ligne dans des **sites**, grâce à un **navigateur**
- **Hyperliens entre les pages** : donnent la métaphore de la toile d'araignée
- A l'origine du Web on a :
  - **HTML 1.0** - HyperText Markup Language (1992)
  - La notion d'**URL** - Uniform Resource Locator (1994)

⇨ Naissance du Web 1.0 (1990)

Bref historique du Web

- **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.
- **2004** : **Firefox** 1.0.
- **2005** : **Plus de 60 millions de sites**
- **2019 : Plus d'un milliard de sites !!!!**

Grandes générations du Web

- **Web 1.0** (1990)
  - Web 1.0 : **web statique**, a évolué vers un *Web dynamique* (pages générées par des outils, à partir de BD par exemple).
  - Plus tard : langages de script, navigateur Web plus riches (Netscape + javascript, 1996).
- **Web 2.0** (Web actuel)
  - **Interaction entre les utilisateurs** : *web social, web participatif*
  - Evolutions technologiques permettant **d'utiliser tout type d'application via un navigateur** (AJAX, Rich Internet Application)
  - **Commerce électronique**.
- **Web 3.0** – Web Sémantique ou Web des Données (en développement)
  - Technologies pour rendre le **contenu des ressources** du Web **accessible et utilisable** par les **programmes** et **agents logiciels**,
  - Grâce à des **métadonnées formelles** formalisées dans des **langages développés par le W3C**.

2. Faiblesses du Web actuel

- **Le Web aujourd'hui**
- **Problèmes avec le Web actuel**
- **Solutions aux problèmes du Web actuel**

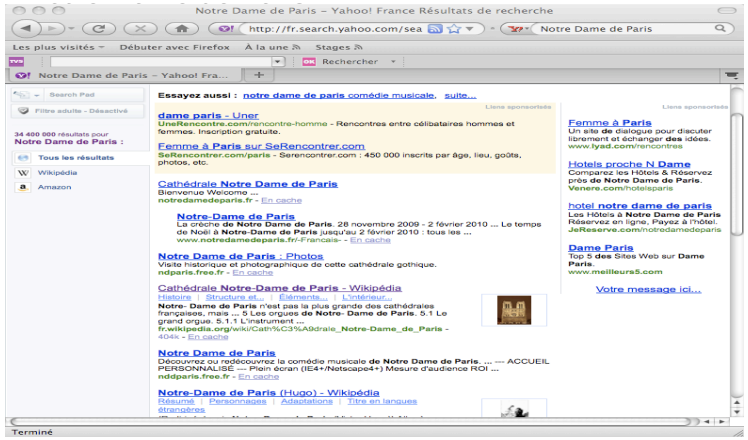
Le Web aujourd'hui

- Information sur le Web essentiellement prévue pour être affichée (écran, imprimante) et lue par des humains : pages Web
- Seuls les humains peuvent interpréter le contenu des pages Web
- L'interaction entre un utilisateur et le Web passe surtout par un moteur de recherche.
- Ces moteurs font essentiellement une recherche syntaxique de mots-clés
- Usage de « wrappers » pour une extraction automatique à partir de pages Web (par exemple pour indexer les documents selon les mots)

⇨ Le Web actuel est essentiellement **syntaxique** : contenu quasi inaccessible aux traitements machines

Exemple du Web actuel

- Recherche : "Notre Dame de Paris"



- C'est l'utilisateur humain qui seul interprète les résultats, c.a.d. leur attribue une sémantique

Problèmes avec le Web actuel (1)

Séparer la PRESENTATION 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 :
  - XML n'est pas toujours visible
  - XML n'est pas un langage (mais un métalangage) : comment comparer 2 documents XML écrits avec des DTD différentes ?

⇨ **Comment faire ?**

Problèmes avec le Web actuel (2)

- **Le Web actuel entrave la recherche, l'extraction, la maintenance et la génération d'INFORMATION**
  - La majorité des **données sur le Web** est sous une **forme qui ne permet pas de l'utiliser** à **grande échelle**.
  - **Pas de système global de publication de données** permettant aux **machines** et aux **humains** de les **traiter** :

    Exemple : Évènements sportifs, météo, guides TV, guides cinéma, ... sont présentés par de **nombreux sites Web**, mais presque tous au **format HTML** (structure logique + présentation)
- **Actuellement, pas d'accès réel au CONTENU des documents**
  - Contenu et Information pas accessible ni interprétable par des machines
  - Pas possible de composer dynamiquement des documents cohérents et adaptés aux utilisateurs

Problèmes avec le Web actuel (3)

- **INFORMATIONS « cachées » dans le code HTML :**
  - code HTML contient l'expression dans une **langue naturelle** (LN) des **informations**, images, fichiers sonores, vidéos, …
  - **moteurs de recherche** (sur le texte), pour des raisons de performance (et taille du Web) ne **font aucun traitement sophistiqué** (TALN) des textes :

    → **recherche de mots,** ce qui est **très différent** d'une **recherche d'informations**
- **SERVICES « cachés » dans le code HTML :**
  - Comment **connaître** ce que **propose** un **service** ?
  - Comment **utiliser conjointement** plusieurs **services** ?

    Ex: Achat de billets de trains, validateur html, Web mail, …

Problèmes avec le Web actuel (4)

Exemple : un service consistant à organiser un voyage…

- Ex: **Horaires de trains** & **Horaires d'avion** = 2 documents HTML avec tables :
  - **Comment croiser les 2 documents** pour un trajet train puis en 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** (par ex. 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** (**automatiquement**) les données.

Problèmes avec le Web actuel (5)

- **Faiblesses des moteurs de recherche par mots-clés**
  - Faible précision
  - Résultats : seulement pages Web
  - Seul l'humaine peut interpréter et combiner les résultats
  - Résultats des recherches pas lisibles par d'autres logiciels
- **Signification du CONTENU du Web pas accessible aux machines**
  - Les machines ne font pas de traitement de la langue naturelle (TALN) et ne peuvent pas ainsi lever les ambiguïtés
  - Les machines manquent de **sémantique**

Solutions aux problèmes du Web actuel

Usage de **Meta-données,** pour annoter les données du Web dans un **langage standard**, en utilisant un **vocabulaire standard** (contrôlé) permettant aux machines :

- des **comparaisons de documents**
- des **raisonnements** pour résoudre une requête
- la prise en compte de **documents multimédias**
- des **réponses formées** de **plusieurs documents** ou de parties de documents
- **communiquer** entre elles, **coopérer** dans une recherche d'information, échanger leurs résultats et les fusionner

⇨ **Web des données, Web sémantique**

3. Définition du Web Sémantique

- **Définition du Web Sémantique**
- **Du Web actuel au Web Sémantique**
- **Web sémantique et Web des données**
- **Web des données et données liées**
- **Problématiques du Web Sémantique**

Définition du Web Sémantique

« The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation »

« Le Web sémantique est une **extension du Web actuel** dans lequel **l'information est munie d'une signification** bien définie permettant aux **ordinateurs** et aux **personnes** de mieux **travailler** en **coopération** »

Tim Berners-Lee, James Hendler, Ora Lassila
The Semantic Web, Scientific american, May 2001
http://www.scienfificamerican.com

Du Web actuel au Web Sémantique (1)

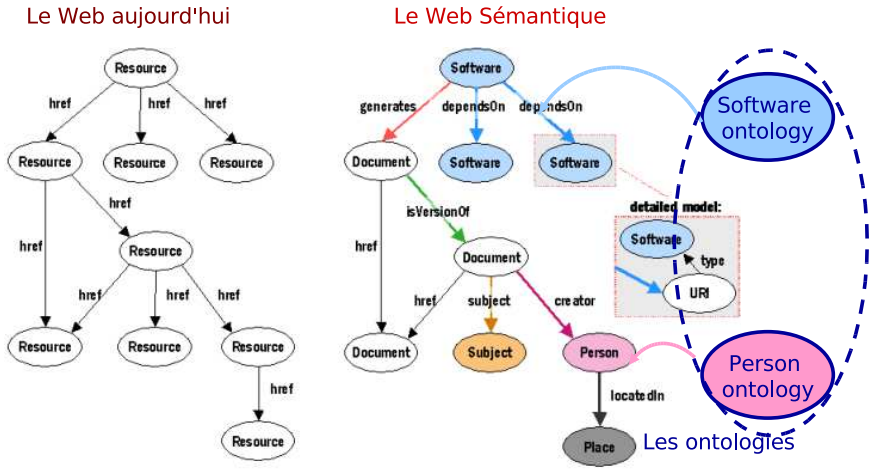
- **Du Web actuel :**
  - Pas de structure explicite globale
  - Liens non exploitables sémantiquement
  - Travail limité sur les informations
  - …
- **Au Web sémantique :**
  - Connaissances formalisées
  - Lien sémantique entre informations
  - Annotations plus riches
  - Standard à base d'XML

- Ensemble de documents
- Recherche par mots clés
- Utilisable par l'humain

→

- Ensemble de connaissances (XML et RDF(S)) (Descripteurs sémantiques)
- Recherche par concepts
- Exploitable par des machines

Du Web actuel au Web Sémantique (2)



Source : W3C Semantic Web Activity, Koivunen and Miller, 2001

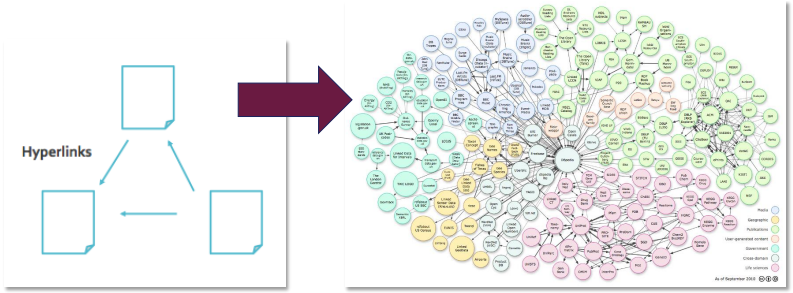
Web sémantique et Web des données

- *« Le **Web de données** est le Web des données qui peuvent être traitées directement ou indirectement par des machines pour aider leurs utilisateurs à créer de nouvelles connaissances » Tim Berners Lee*
- Le **Web de données** consiste à lier et structurer l'information pour accéder simplement à la connaissance qu'elle contient déjà
  Source : « W3C Semantic Web Activity » , W3C, 7 novembre 2011
- Objectif du **Web des données** :
  - Mettre à disposition des données en utilisant des techniques **standardisées** qui garantissent **l'interopérabilité**
  - **Relier** les données elles-mêmes (linked data)
  - Rendre ces données **interprétables** par les **machines**
  - « Permettre aux données d'être **partagées** et **réutilisées** au-delà des limites applicatives, organisationnelles ou communautaires » (W3C)

⇨ le **Web Sémantique** met en œuvre le **Web de données**

Web des données et données liées (1)

- Du Web des documents au Web des **données liées** (linked data) :



Données ouvertes (open data) : données peuvant être publiées et rendues publiques sous une licence ouverte sans les lier à d'autres sources

≠

Données liées (linked data) : données peuvant être liées aux URIs d'autres sources de données, en utilisant des standards ouverts tels que RDF, sans être disponibles publiquement sous une licence

Web des données et données liées (2)

- *« Les **données liées (linked data)** sont un ensemble de principes de conception pour le partage de données lisibles par les machines sur le Web pour une utilisation par les administrations publiques, les entreprises et les citoyens. » Source : EC ISA Case Study: How Linked Data is transforming eGovernment*
- Les **4 principes de conception des données liées** (Tim Berners Lee):
  1. Utiliser des URI (identificateurs de ressources uniformes) pour les noms des choses
  2. Utiliser des URIs http, de sorte que les humains puissent consulter ces adresses
  3. Fournir sur les URI (ressources) des informations utiles, en utilisant les standards du Web Sémantique (RDF *, SPARQL)
  4. Inclure des liens vers d'autres URIs afin qu'ils puissent découvrir plus de choses

Problématiques du Web Sémantique

- **Liens avec des problématiques existantes :**
  - Représentation des connaissances
  - Traitement et compréhension du langage naturel
  - Déduction automatique ...
- **Domaines concernés :**
  - Sciences cognitives
  - intelligence artificielle
  - linguistique, logique …
- **Spécificités** :
  - Ressources du Web, structuration de données existantes sur le Web
- **Toujours les mêmes difficultés :**
  - Ambiguïté
  - Données temporelles
  - Incertitudes
  - Confiance ...
- *Exemple :*
  - *Notre Dame de Paris est aussi un roman de Victor Hugo*
  - *La construction de Notre Dame de Paris a commencé en 1163*
  - *…*

Web Actuel Vs Web Sémantique

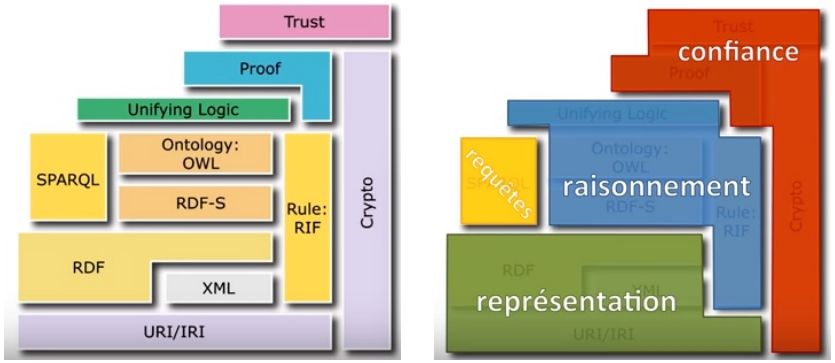
Web actuel	Web Sémantique (WS)
▪ Ensemble de **documents**	▪ Ensemble **d'information / connaissances**
▪ Basé essentiellement sur **HTML**	▪ Basé essentiellement sur **XML et RDF(S), OWL**
▪ Recherche par **mots-clés**	▪ Recherche par **concepts** (**ontologie**)
▪ Utilisable par **l'être humain**	▪ Utilisable par la **machine**

4. Architecture du Web Sémantique

- **Briques représentation**
- **Briques requêtes**
- **Briques raisonnement**
- **Briques confiance**

La Semantic Web Stack

- La **Semantic Web Stack** (appelé aussi Layer Cake) est une illustration représentant **l'architecture du Web sémantique** spécifiant des **briques fonctionnelles**
- Une brique s'appuie sur les briques de dessous



- Ces briques peuvent être regrouper en grandes fonctions : représentation, requêtes, raisonnement, et confiance.

Briques représentation : URL, URI, IRI

- Objectif général : **nommer les ressources**
- **URI** : Uniform Resource Identifier (août 98) / **IRI** (International Resource Identifier) permet d'identifier une ressource (physique ou abstraite) sur le Web
- Les **URL** : URI qui donnent le moyen d'accéder à la ressource. Ex : http://www.wikipedia.org/
- **URN** : URI qui permettent d'identifier une ressource par son nom dans un espace de noms. Ex : urn:isbn:0-395-36341-1
- Par la suite, on ne s'intéresse qu'aux URI references, c.a.d aux URI telles qu'on dispose d'un algorithme pour construire une URI absolue.
- en **XML**, les espaces de noms permettent d'utiliser des vocabulaires XML différents, dans un même document.

Briques représentation : XML

Objectif général : **représenter les ressources pour les machines**

- **XML** : eXtensible Markup Language, recommandation XML 1.0 en février 98)
- **XML** : séparation fond/forme
- Méta-langage, qui permet de définir des langages de documents.
- Nombreux dialectes : MathML, XSLT, XACML, SVG, XHTML, …
- Document bien formé : le parenthèsage (balises ouvrantes-fermantes) est correct
- Document valide : conforme à un schéma, défini par une DTD, un XML-schema, un schéma RelaxNG
- Outils de manipulations, basés sur Xpath : XSLT, XQuery
- Internationalisation : Unicode

Briques représentation : RDF (1)

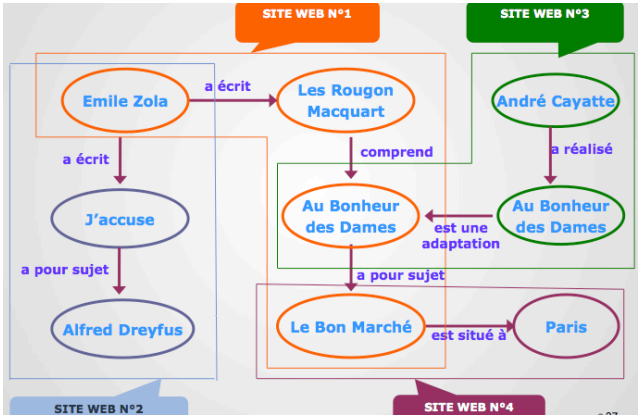
Objectif général : **exprimer les ressources et les relier**

- **RDF = la base du Web sémantique**
- **RDF** = cadre pour décrire des données sur les ressources du Web (Resource Description Framework)
- Représentation des ressources par Triplets (**S**ujet, **P**rédicat (propriété), **V**aleur) :
  - Sujet : une ressource qui peut être identifiée par un URI
  - Prédicat : une spécification réutilisée et identifiée par URI de la propriété
  - Objet : une ressource ou constante à laquelle le Sujet est lié
- Triplets sérialisés de différentes façons (XML, Turtle, …)
- Permet de constituer des graphe RDF, des bases de données RDF (Triples-stores)

Briques représentation : RDF (2)

Source : Transition bibliographique : Web sémantique et web des données, Sensibilisation à l'évolution des catalogues

- Triplets RDF interconnectés :



Briques requêtes : SPARQL

Objectif général : faire des **requêtes** sur ressources exprimées en **RDF (et RDF-S)**

- **SPARQL = Simple Protocol and RDF Query Language :**
  - Un langage de requête pour RDF
  - Un protocole : spécification pour émettre et envoyer des requêtes SPARQL (services Web) vers des serveurs dédiés et en recevoir les résultats
  - Un format XML pour l'affichage des résultats obtenus (requêtes de type SELECT et ASK)
- Permet de réaliser des requêtes fines et précises
- Permet aussi de réaliser des opérations : ajout, modification, suppression, tris, ... de données RDF
- Inspiré de SQL pour la syntaxe et les fonctionnalités

Briques raisonnement : Ontologies (1)

Objectif général : **formaliser des connaissances** sur un **domaine spécifique** pour leur utilisation par les **machines**

- Le petit Larousse :

  Ontologie = du grec : *ontos* être, et *logos* science, étude de l'être en tant qu'être …

  c'est-à-dire « l'étude des propriétés générales de ce qui existe ».

- Par extension, en **informatique** :

  Définition [Gruber 1993]: "An ontology is a formal, explicit specification of a shared conceptualization of a domain of interest".

  - L'ontologie est un ensemble structuré de termes et concepts, relation entre concepts, **représentant le sens** (les connaissances) d'un domaine d'informations particulier
  - L'ontologie doit permettre aux machines de **raisonner** sur ces connaissances formalisées du domaine

Briques raisonnement : Ontologies (2)

Rôle des ontologies dans le Web Sémantique :

- **Définir** de manière **déclarative** un **vocabulaire commun** résultat d'un **consensus social** dans un **domaine donné** :
  - Chaque élément de vocabulaire possède une interprétation unique partagée par tous les membres du domaine
- Décrire la **sémantique** des **termes** et leurs **relations** :
  - L'**interprétation** de chaque terme est unique er résulte d'une **sémantique formelle**.
  - L'ensemble des **termes** et leurs **relations** fournissent **un cadre interprétatif** dépourvu d'ambiguïté pour **chaque terme**.
- Fournir des **mécanismes d'inférence** qui respectent la **sémantique** formelle.

Briques raisonnement: langage d'ontologies

Langages de description d'ontologie du W3C :

RDF-Schema :

- Permet de décrire un **vocabulaire RDF** spécifique à un domaine (*RDF vocabulary description language)*
- Fournit une **sémantique** à ce vocabulaire en décrivant les propriétés et les classes des ressources RDF
- Utilisé pour ***formaliser des ontologies légères*** en permettant un *raisonnement limité* sur ces ontologies

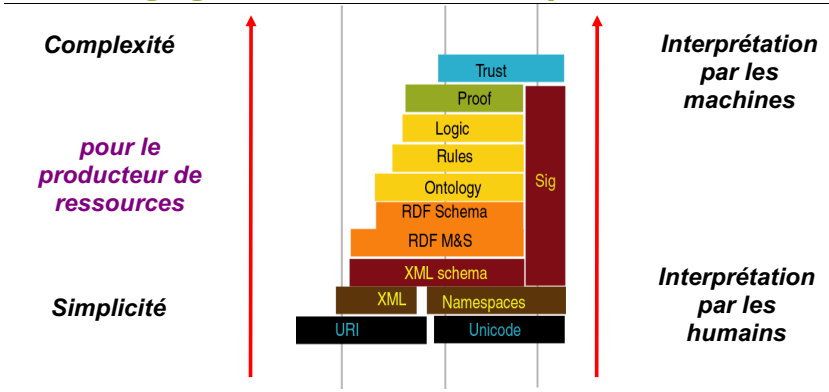
OWL : Web Ontology Language (issu de DAML+OIL)

- Langage plus puissant de formalisation d'ontologies : relations entre classes, contraintes de cardinalité propriété de typage plus riches, …
- Utilisé pour ***formaliser des ontologies lourdes*** en permettant un *raisonnement puissant* sur ces ontologies en s'appuyant sur les logiques de description (LD)

Briques confiance: Proof, Trust, Crypto

- Ces briques (Proof, Trust et Crypto) sont **en cours de développement**
- Elles ne correspondent pas encore à des technologies standardisées
- **Objectifs :**
  - Définir un **langage universel** pour une logique monotone, qui permet de décrire des relations **plus riches qu'en OWL**
  - Définir un **système d'inférence à base de règles** permettant de dériver des assertions à partir des assertions connues (décrites en RDF, OWL, ...)
  - Usage de la **cryptographie** pour assurer que les assertions utilisées et le système d'inférence sont fiables (on peut faire confiance à la fois au système et aux données) : *signature numériques, recommandation, …*

Les langages du W3C et complexité

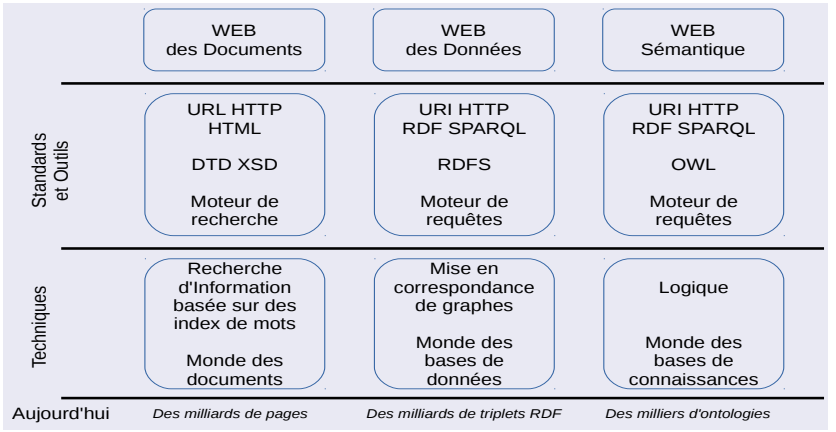


Quel niveau de complexité est nécessaire ?

- *complexité **algorithmique** des mécanismes d'inférences*
- *complexité **technique** pour les constructeurs d'outils*
- *complexité **conceptuelle** pour l'utilisateur moyen*

Synthèse …

- Web des documents, Web des données et Web sémantique :



Source : Inspiré de la présentation de MC Rousset : http://www.college-de-france.fr/media/serge-abiteboul/UPL5540483766962034581_Rousset20120530.pdf

5. Quelques applications du Web Sémantique

- **Web sémantique et commerce électronique**
- **Web Sémantique et gestion des connaissances**

Web Sémantique et commerce électronique B2C (1)

Source : O. Papini

- **Un scénario typique B2C (Business to Customer):**
  - les *clients* visitent des sites de *magasins en ligne*, étudient leurs *offres* sélectionnent et *commandent* des produits
  - activité importante dans l'industrie
  - idéalement, les clients devraient visiter beaucoup de sites de magasins MAIS cela prend trop de temps !
  - les "**shopbots**" (robots logiciels) : le font à leur place
- **Limitation des "shopbots" :**
  - reposent sur des "*wrappers*"(conteneurs à contenu caché) nécessitant une programmation importante
  - extraient de l'information sur la base d'une *analyste textuelle*
  - doivent être *reprogrammés* lorsqu'un magasin change ses outils

Web Sémantique et commerce électronique B2C (3)

Source : O. Papini

- **Le Web sémantique adapté commerce électronique B2C :**
  - *agents logiciels de collecte* **récupèrent l'information** sur le produit et les termes du service :

    prix, information sur le produit, livraison, politique de confidentialité seront interprétés et comparés aux besoins de l'utilisateur
  - informations sur la ***réputation*** des magasins
  - *agents logiciels acheteurs* sophistiqués pourront conduire des ***négociations***

Web Sémantique et commerce électronique B2B (1)

Source : O. Papini

- **Le Web sémantique : impact sur le commerce électronique B2B (Business to Business) :**
  - la plus **grande promesse économique**
  - actuellement repose la plupart du temps sur des **EDI** (Electronic Data Interchange)
    - technologie seulement comprise par des experts
    - difficultés de programmation, de maintenance, error-prone programmation séparée pour chaque communication B2B
  - le Web semble être une parfaite infrastructure **mais B2B mal géré par les standards web**

Web Sémantique et commerce électronique B2B (2)

- **Le Web sémantique adapté commerce électronique B2B :**
  - **Enregistrement des partenariats** sans charges indirectes
  - **différences entre terminologies** résolues par l'utilisation de modèles de domaine abstrait standards
  - **échange de données** par l'utilisation de service de translation
  - **enchères, négociations, ébauche de contrats** automatiquement (ou semi-automatiquement) réalisés par des agents logiciels

Web Sémantique et gestion des connaissances (1)

Inspiration : O. Papini

- **Le Web sémantique : impact sur gestion des connaissances**
  - la gestion des connaissances concerne : l'acquisition, l'accès, le maintien des connaissances dans une organisation
  - activité importante dans l'industrie
  - importance pour des organisation internationales dispersées géographiquement
  - la plupart des informations disponibles sont faiblement structurées (textes, sons, images, · · ·)

Web Sémantique et gestion des connaissances (2)

- **Limitation des technologies actuelles de gestion des connaissances :**
  - **Recherche d'information :**
    - moteurs de recherche à base de mots-clés …
  - **Extraction d'information :**
    - intervention humaine nécessaire pour naviguer, chercher, interpréter, combiner
  - **Maintenance de l'information :**
    - incohérences de terminologie, information dépassée
  - **Visualisation de l'information :**
    - impossible de définir des vues sur la connaissance Web

Web Sémantique et gestion des connaissances (2)

- **Le Web sémantique adapté à la gestion des connaissances :**
  - les **connaissances** sont **organisées** en **espaces conceptuels** selon leur signification
  - **outils automatiques** pour la **maintenance** et la **découverte** de **connaissances**
  - **réponse** à des **questions sémantiques**
  - **réponse** à des **questions sur plusieurs documents**
  - possibilité de définir **qui peut voir certaines parties de l'information**