INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Cours Web Sémantique- 4IF

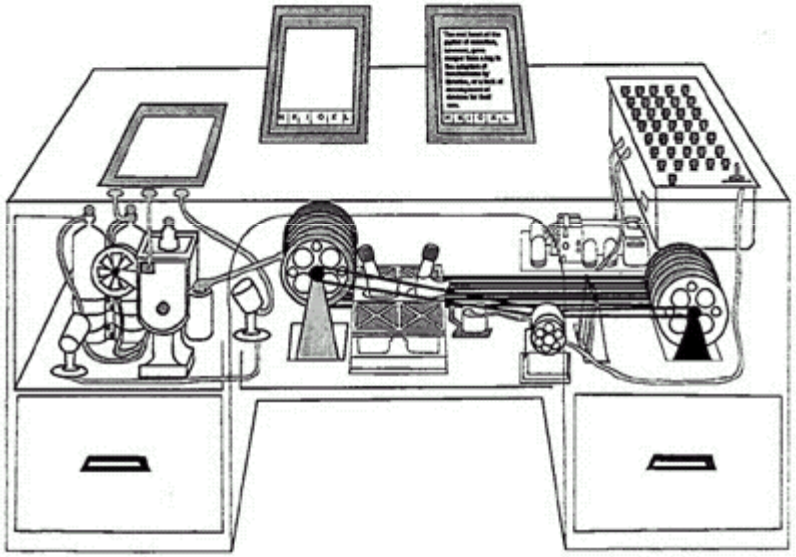
Partie 1 : Introduction au web sémantique

Sylvie Calabretto/Mehdi Kaytoue

Plan du cours

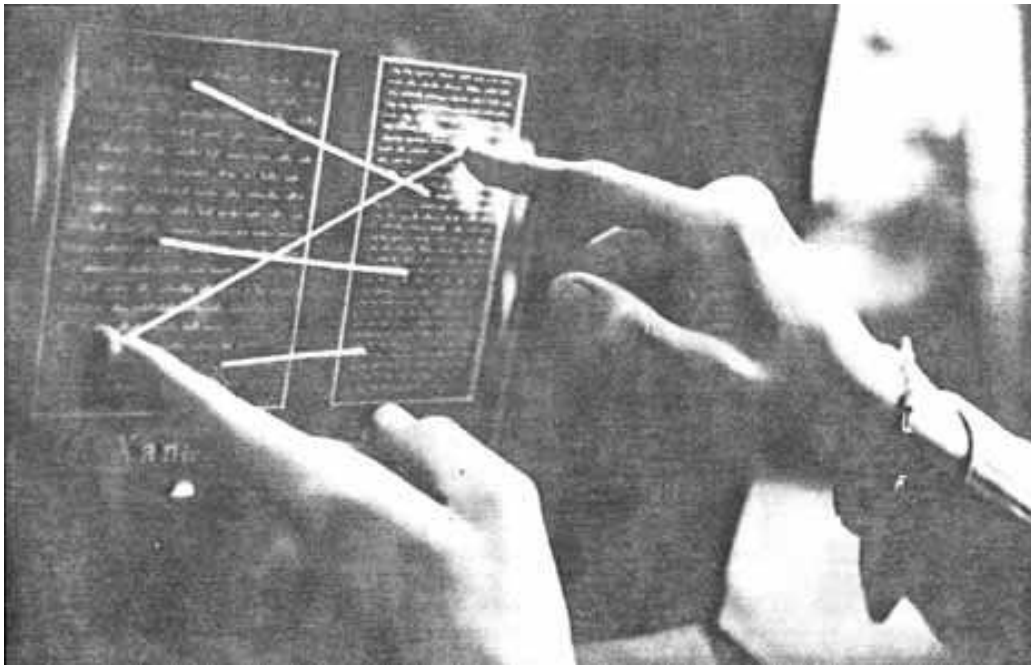
- Introduction au web sémantique
- Décrire les données du Web : RDF
- Représenter les connaissances : ontologies, RDF-S et OWL
- Interroger le Web : SPARQL
- Raisonner : inférences et RIF

La préhistoire du web :
Memex (Vannevar Bush, 1945)



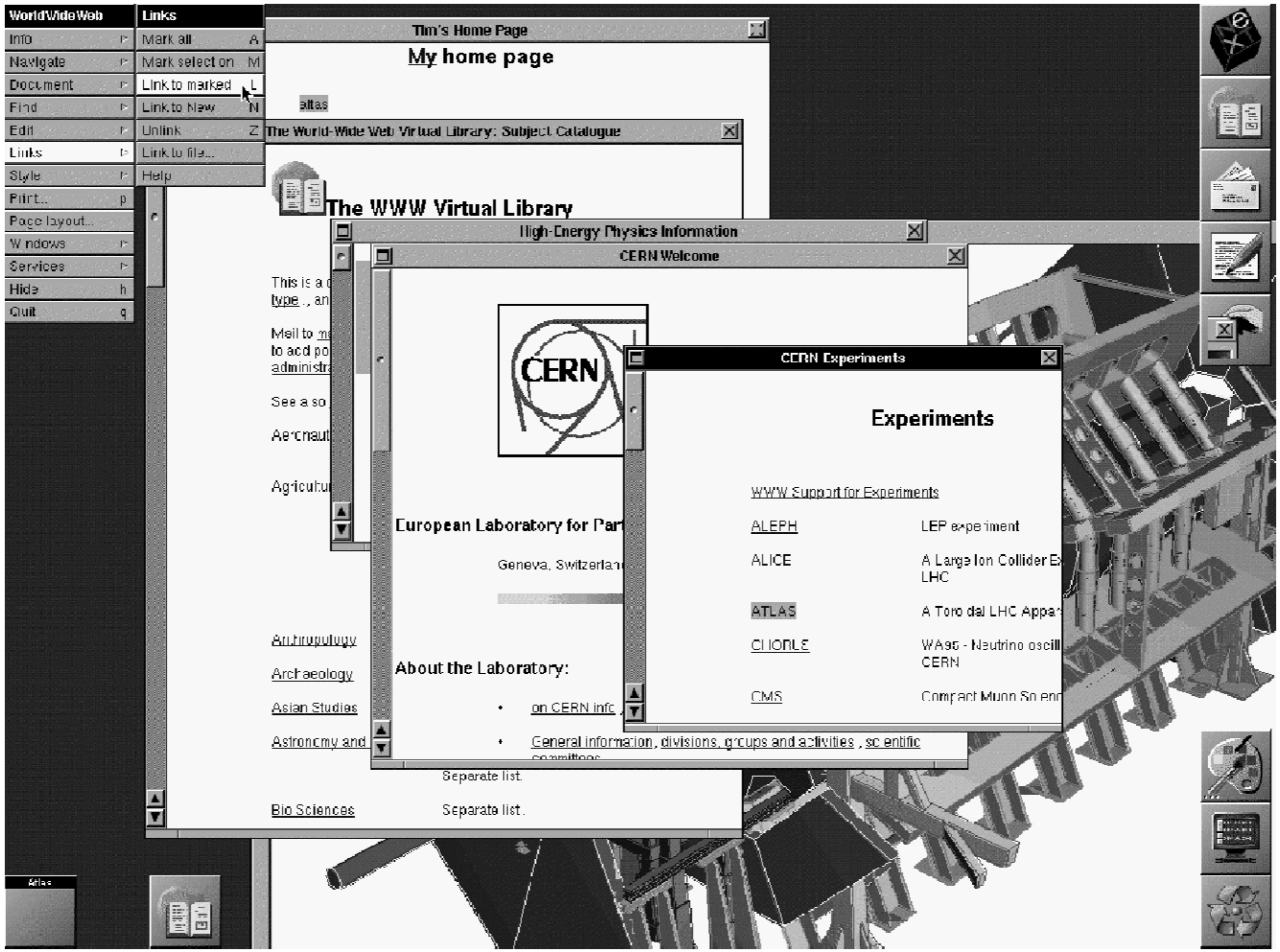
- Défi scientifique : améliorer les moyens d'accès aux connaissances
- Système de gestion d'un réseau complexe de notes (publications, …)
- Machine bureau imaginaire Memex (pour Memory extender)
- Qui n'utilisait pas d'outil informatique
- Prémices des systèmes hypertextes

Xanadu (Ted Nelson, 1960)

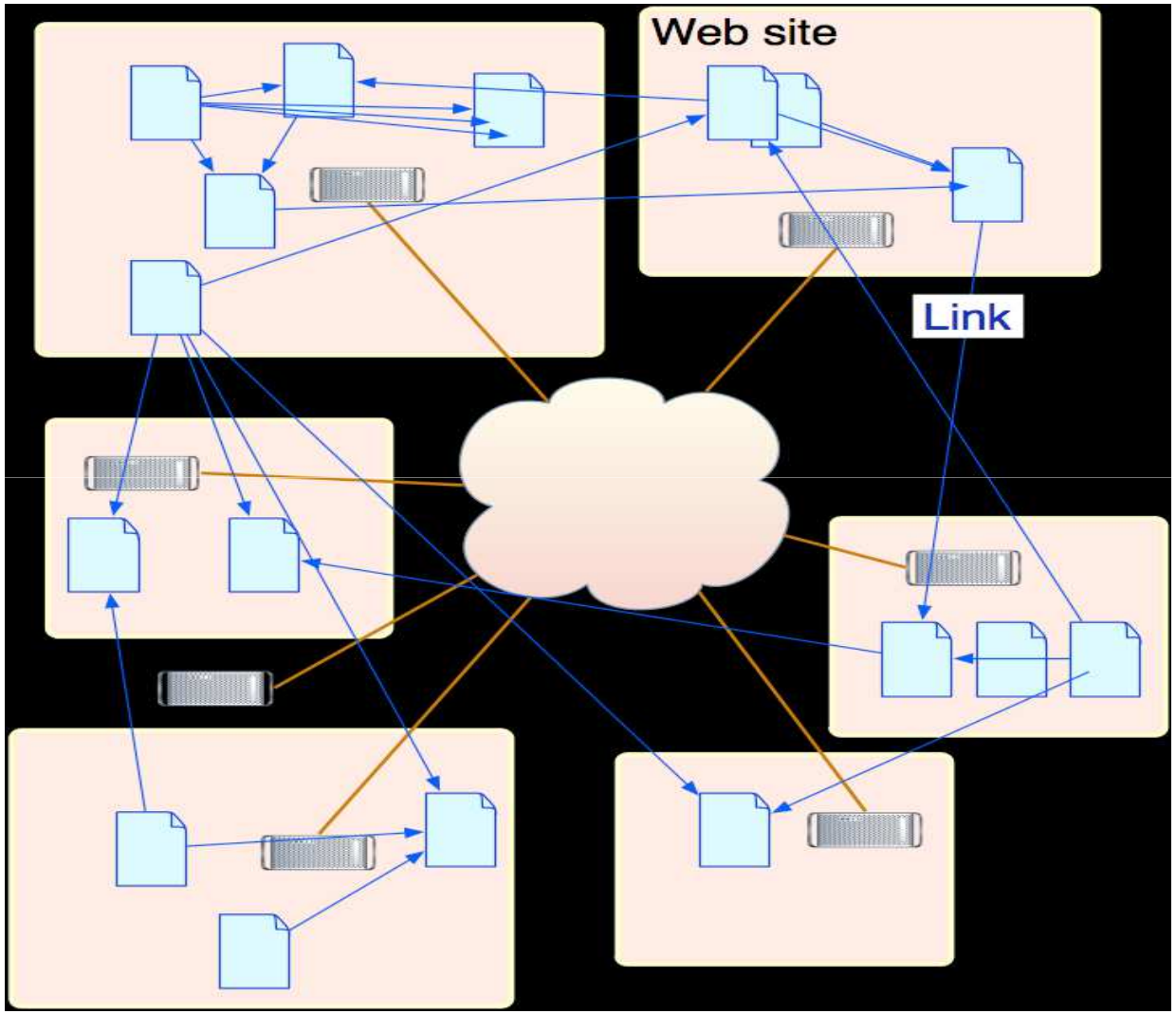


- **Projet de système d'information permettant le partage instantané et universel de données informatiques**
- **Invention du concept d'hypertexte**
- **Invention de la souris**

World Wide Web (Tim Berners-Lee, 1989)



World Wide Web



Le web des documents

- **Le web = Système d'information hypermédia sur INTERNET**
- **Conçu en 1989 par Tim Berners-Lee (au CERN) pour permettre aux chercheurs-visiteurs du CERN d'échanger des informations scientifiques (articles, rapports) après leur séjour**
- **Fondé sur l'extension du concept « hyperdocument » aux réseaux internationaux (hyperdocument réparti)**
- **Une amélioration de l'existant (et non une révolution) :**
  - permet un accès plus convivial à des serveurs existants (WAIS, GOPHER, FTP, … existaient avant mais exigeaient des « clients » spécifiques)

Le web des documents

- **Le web est un système d'information hypermédia réparti qui exploite l'Internet en s'appuyant sur 3 mécanismes :**
  - **un *schéma uniforme de nommage* des ressources : les URI (*Universal Resource Identifier)*,**
  - **des *protocoles* pour accéder aux ressources nommées sur le Web (comme HTTP),**
  - **les *hyperdocuments textuels*, qui consistent à inclure des hyperliens dans un texte.**

URI

- Les URI permettent de nommer des ressources sur le Web. Elle sont de la forme
- **protocole://serveur[:port][/chemin][/nom de document][#position]**

  Exemples :

  http://www.w3.org

  mailto:mkaytoue@insa-lyon.fr

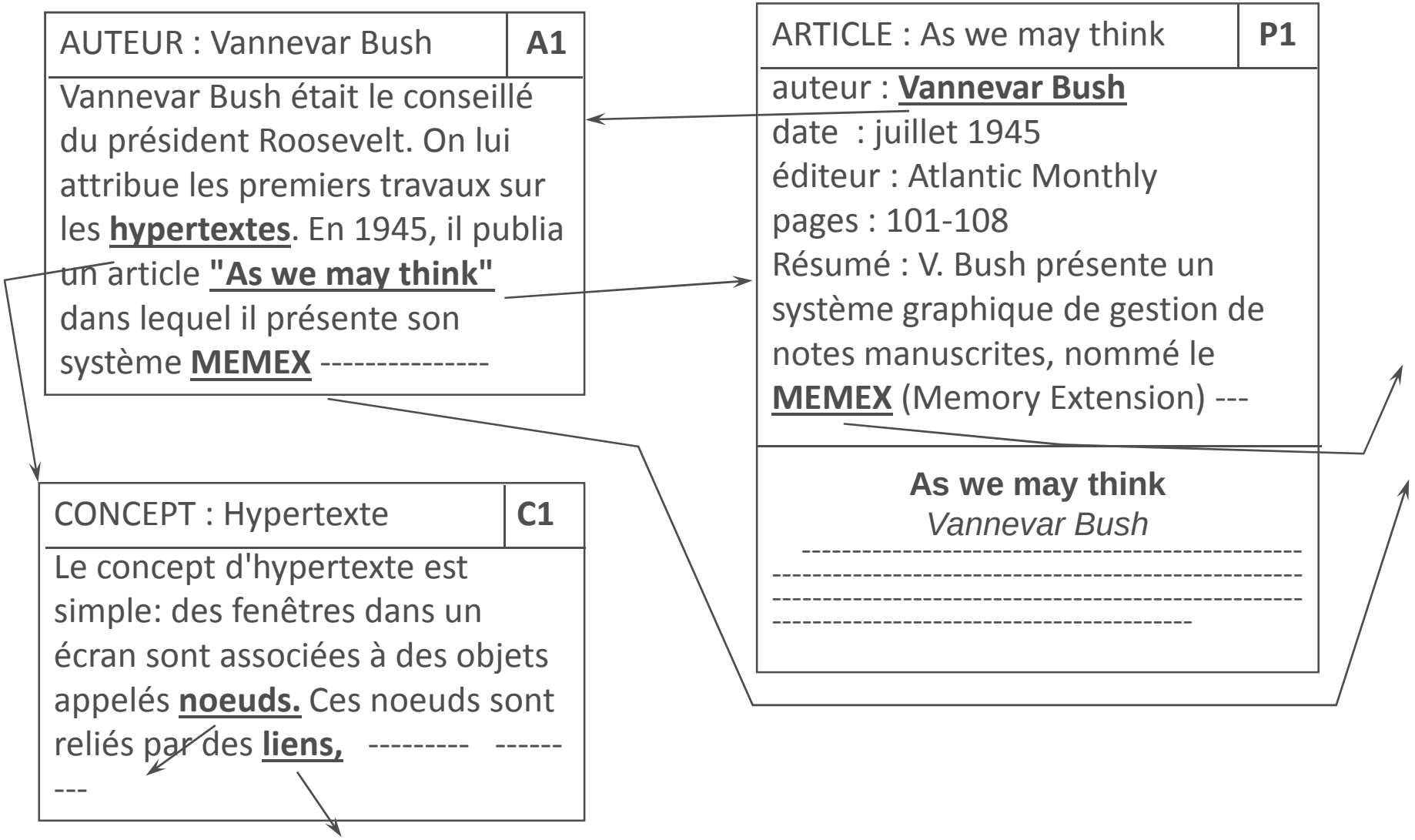
  ftp://ftp.insa-lyon.fr
- Les URIs qui identifient des ressources d'information sont appelées **URL** (*Uniform Resource Locator*).
- Les **IRI** sont des URI internationalisés

Définition des hyperdocuments

- **Il faut distinguer système hypertexte (logiciel) des hyperdocuments textuels (données gérées)**
- **Un hyperdocument est un ensemble structuré de nœuds et de liens**
- **Les nœuds sont associés à des contenus**
  - Textuels (avec ou sans formule)
  - Graphiques (géométriques ou photographiques)
  - Sonores ou vidéo

  } Multimédias
- **Les liens (typés) entre les structures définissent la structure**
  - Hypergraphe / Graphe / Arbre

Exemple d'hyperdocument



HTML

- **HyperText Markup Language**
- **Modèle de représentation d'hyperdocuments**
- **Utilisé par les serveurs et les clients web**
- **Défini selon le standard SGML (c'est une DTD)**

  Définit à la fois la structure logique d'un nœud, sa structure physique et sa présentation

- **En constante évolution :**

  •HTML-1 (1989) : du texte, quelques styles, des liens hypertextes

  •HTML-2 (1994) : HTML-1 + des images + des formulaires interactifs

  •HTML-3 (1996) : HTML-2 + des graphiques vectoriels, du son, des applets

  •HTML-4 (1998) : HTML-3 + vidéo, outils pour INTRANET, …

La "Document Type Definition" HTML

Définit la structure "logico-physique" d'un "document" WEB

Structure simple avec un nombre réduit de types d'éléments SGML

- Noeud : **<HTML>** *contenu du noeud* **</HTML>**
- entête : **<HEAD>** *contenu entête* **</HEAD>**
- titre : **<TITLE>** *titre* **</TITLE>**
- paragraphe : **<P>** *texte du paragraphe* **</P>**
- ancre : **<A** *URL> texte-ancre* **</A>**
- etc.

Les **formats** des images, des polices, des enrichissements, ... sont prédéfinis (JPEG, GIF, ...)

Exemple HTML (G. Antoniu, F. van Harmelen)

```
<h1>Centre de kinésithérapie Agilitas  </h1>
Bienvenue à la page d'accueil du Centre de kinésithérapie
   Agilitas. Ressentez-vous de la douleur? Avez-vous eu un
   accident? Notre personnel
Lise Davanport,
Josiane Bouville (notre charmante secrétaire) et
Etienne Matthieu vont prendre soin de vous.
<h2>Horaire des consultations</h2>
Lun 11.00 - 19.00<br>
Mar 11.00 - 19.00<br>
Mer 15.00 - 19.00<br>
Jeu 11.00 - 19.00<br>
Ven 11.00 - 15.00<p>
Veuillez noter que nous n'avons pas de consultations les
   semaines de
<a href=". . .">State Of Origin</a> games.
```

Document HTML

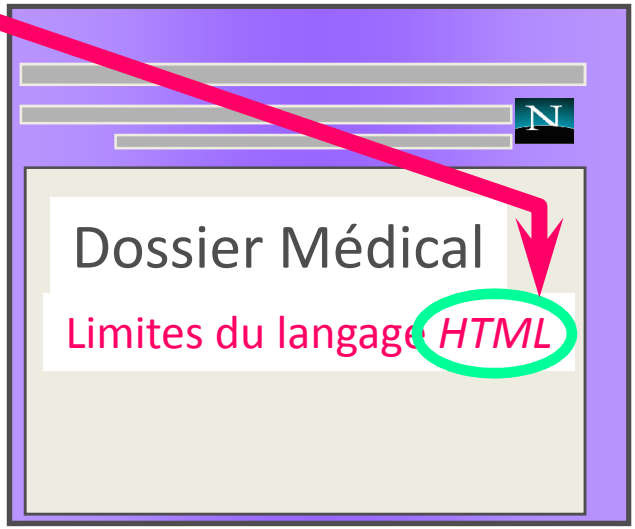
```
<HEAD>
 <TITLE>Rapport PFE</TITLE>
</HEAD>
<BODY>
 <H1>Dossier Médical</H1>
 <P> Limites du langage <I>HTML </I> </P>
</BODY>
```

Problèmes

Le Modèle logique confond la présentation

- Pb de recherche d'Information pertinente
- Aspect statique des Tags sur le Client

Document visualisé par un Browser



- *Besoin d'une sémantique*

  → **<Langage> HTML </Langage>**

- *Dissocier la présentation*

  → **Tag Langage → Style : *Italique***

La maturité du web

- **Séparer forme et contenu**

  La nécessité de traiter et échanger les données du web font apparaître le besoin de séparer les données des traitements et de la présentation dans le navigateur

- **Web = Grande Base de Données de Documents Structurés**

- **Dates Clés**

  1996 : première définition des feuilles de style

  1998 : définition de XML 1.0

Web structuré : la famille XML

- **1998 : XML 1.0 (Extensible Markup Language)**
  - Séparation entre contenu et présentation
  - Format textuel d'échange de données structurées
  - Standard pour définir des langages balisés

  1998-2006 : langages de schémas DTD, XML Schéma, RELAX NG, langages XHTML, SVG, MathML, …, langages de manipulation Xpath, Xlink, Xpointer, XSL, XSLT, XQuery

Exemple XML (G. Antoniu, F. van Harmelen)

```
<société>

   <traitementProposé>Kinésithérapie</traitementPro
   posé>
   <nomSociété>Centre de Kinésithérapie
   Agilitas</nomSociété>
   <personnel>
        <kiné>Lise Davanport</kiné>
        <kiné>Etienne Matthieu</kiné>
        <secrétary>Josiane Bouville</secrétaire>
   </personnel>
</société>
```

DTD, XML Schéma, XPath

- **DTD (Document Type Definition)**

  Une DTD définit les balises autorisées, leurs attributs et leur enchaînement

- **2004 : XML Schéma**

  Contraintes sur structure et contenu / Notion de type et héritage

- **1999 : Xpath 1.0 (XML Path Language)**

  Description des chemins dans un document XML

- **2001: XLink (XML Linking Language)**

  Généralisation XML du concept de lien HTML

- **2003 : Xpointer (XML Pointer Language)**

  Extension des URL pour pointer sur des éléments d'un document XML

- **2006 : Xquery (XML Query Language)**

  Langage de requête sur les structures XML, inspiré de SQL / s'appuie sur les systèmes d'adressage Xpath, Xlink, XPointer

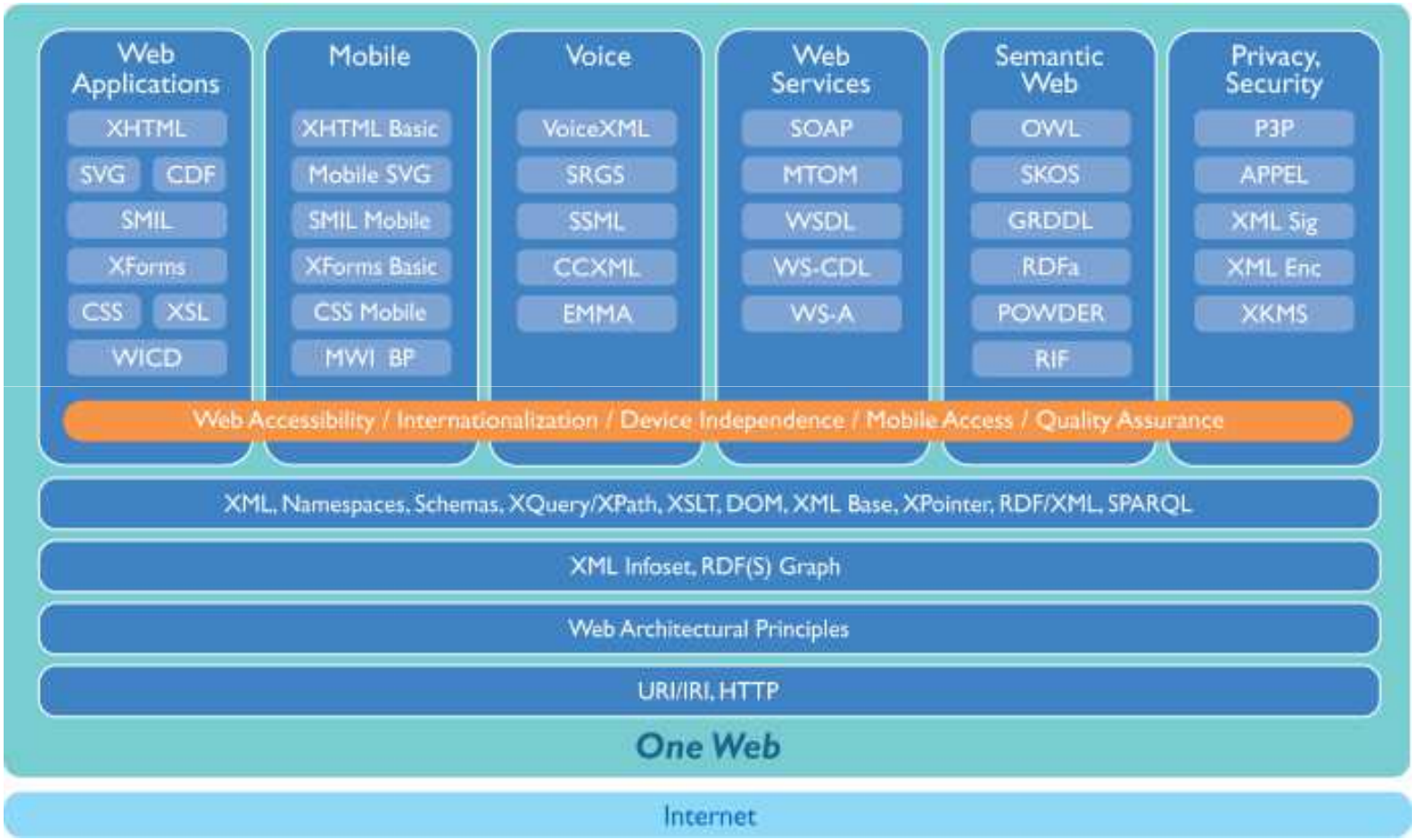
XHTML

- **XHTML™ 1.0 : Extensible HyperText Markup Language**
  - Reformulation de HTML 4 en XML 1.0 (bien formé)
  - Construit au dessus d'XML : bénéficie des outils XML (parser, valider, transformer, etc) et mécanismes de modularisation et extension (composer avec d'autres langages)
- **XHTML™ 2.0 : W3C Working Draft 26 july 2006**
  - Ne cherche pas la compatibilité ascendante
  - Générique, moins de présentation, plus de structure, accessibilité et utilisabilité, moins de scripts, indépendance au terminal

W3C : les architectes du web

- **W3C créé en 1994**
  - Membres fondateurs : MIT, INRIA, Université Keio
  - Organisé en groupes de travail
  - Forum ouvert et neutre (compagnies et organisations)
  - Futur du web et de ses standards
  - Conception et décision par consensus
- **Devient organisme de normalisation**
  - Standard = Recommendations
- **Statuts des standards du W3C**
  - Working Draft / Last Call / Candidate Recommendation / Proposed Recommendation / Recommendation
- **Site du W3C : www.w3.org**

Vision unifiée des activités du W3C



Quelques dates clés

- **1991 : Premier serveur Web hors Europe : Standford Linear Accelerator Center**

  Linus Torvalds : un système UNIX pour PC

- **1992 : Navigateur allégé Mosaïc CERN gratuit**

  50 serveurs Web sur INTERNET

- **1993 : Mosaïc graphique, souris, UNIX/X, MacOS, Windows/DOS**

  341 634 sites Web

- **1994 : Web = deuxième service après FTP et avant Telnet**

  Yahoo! (Yet Another Hierarchical Officious Oracle) / Nescape (remplace Mosaïc)

- **1995 : Web = premier service sur INTERNET**

  JAVA, JAVAScript / Netscape 2.0, Windows, Mac et UNIX avec applets

- **1996 : Digital lance Alta Vista**

  Internet Explorer 3.0 et guerre des navigateurs

- **1998 : Mozilla / 1999 : Google / 2008 : Chrome / 2009 : Bing**

Du web des documents au web sémantique

- **La masse des informations stockés sur le web augmente à vitesse exponentielle**
  - •Chaque seconde, 29 000 Go d'informations sont publiés dans le monde
  - •De 2013 à 2020, la masse de données de l'univers digital va doubler tous les deux ans !
  - •En 2012, plus de 580 millions de sites web
  - •Chaque seconde ce sont près de 100 000 recherches sur le moteur de recherche Google
- **Le problème du web n'est plus d'augmenter la taille des « autoroutes de l'information » mais de concevoir et réaliser des systèmes permettant de filtrer les informations et de les délivrer de façon « intelligente »**

Le 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 co-operation.“**

  [Berners-Lee et al., 2001]

Un web sémantique : les machines pourront accéder aux documents par leur contenu.
Intelligence Artificielle : les machines pourront proposer des services Intelligents aux humains.

Le web sémantique

Le web aujourd'hui

- Ensemble de documents
- Basé essentiellement sur HTML
- Recherche par mots-clés
- Utilisable par l'humain

Le web sémantique

- Ensemble de connaissances
- Basé sur XML, RDF(S), OWL
- Recherche par concept
- Utilisable par la machine

Recherche d'Information sur le Web

La situation

- Le mode d'interaction de l'utilisateur avec le WEB passe prioritairement par un moteur de RI,
- L'interrogation et la recherche sont faites de façon syntaxique
- L'utilisateur humain interprète les résultats, i.e. leur attribue une sémantique, et reformule sa requête au besoin

Exemple de requête

- "Hugo"
- liste ordonnée de documents du Web contenant la chaîne de caractères Hugo sur des critères syntaxiques
- L'utilisateur affine sa requête selon qu'il cherche un titre de roman de Victor Hugo, la date de naissance de Victor Hugo, ou encore un calecon Hugo Boss.

Que faut-il ajouter au web actuel ?

- Il faudrait que les programmes (les services) puissent **interpréter les données** : ce document correspond à un hôtel, un hôtel est un mode d'hébergement, dans un hôtel …

- Un prérequis est de **représenter les connaissances liées aux données** pour faire des inférences : cette page représente un hôtel, un hôtel est un mode d'hébergement donc cette page représente un mode d'hébergement

Les chalenges du Web sémantique

- **Représenter les connaissances dans un monde ouvert (tout type de connaissance), à l'échelle du Web et avec une très large variété de protocoles, de langues, de culture, … pour pouvoir manipuler les connaissances liées au contenu du Web.**

Le web sémantique

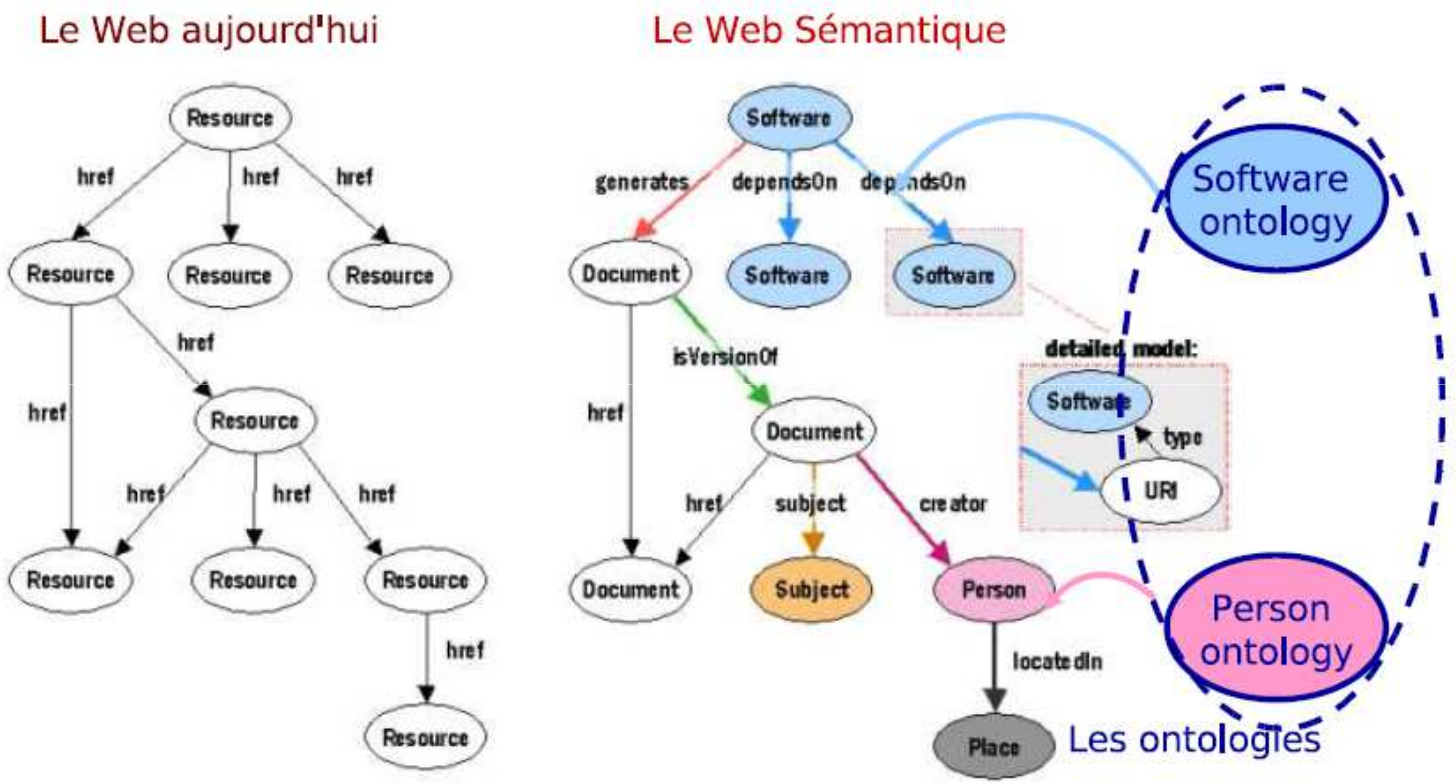


Figure: source : W3C Semantic Web Activity, Koivunen and Miller, 2001

Ontologies

- **Spécification formelle et explicite d'une conceptualisation**
- **Structurées en termes de concepts et de relations entre concepts**
- **Formalisation partagée sur un domaine**
- **Utiles pour l'organisation, la navigation sur les sites web, les recherches sur le web, la récupération de l'information interprétée**

Ontologies

hémopéritoine : *« épanchement hématique localisé au niveau du péritoine »*

Hiérarchie de concepts

ETAT_PATHOLOGIQUE
- ETAT_PATHOLOGIQUE_LOCAL
  - **LESION**
    - adénopathie
    - …
    - épanchement
      - épanchement gazeux
      - épanchement liquidien
        - épanchement de pus
        - **épanchement hématique**

ANATOMIE
- …
- ANA_TISSU_ENVEL
  - capsule
  - duremère
  - mésentère
  - peau
  - …
  - **péritoine**

Hiérarchie de relation

LOCALISATION
- à_côté_de
- à_l'extérieur_de
- au_dessus_de
- …
- **au_niveau_de**

Liens

LESION (**LOCALISATION**) **ANATOMIE**

Concept défini

épanchement hématique **(au_niveau_de)** **péritoine**

Ontologies

fracture à la base du crâne

Hiérarchie de concepts

LESION
épanchement
fracture

ANATOMIE_OBJET
OS
crâne

SITUATION
angle
base
bord

Hiérarchie de relations

LOCALISATION
à_côté_de
à_l'extérieur_de
au_dessus_de
…
au_niv_de
…

OBJET
…
DE
…

Liens

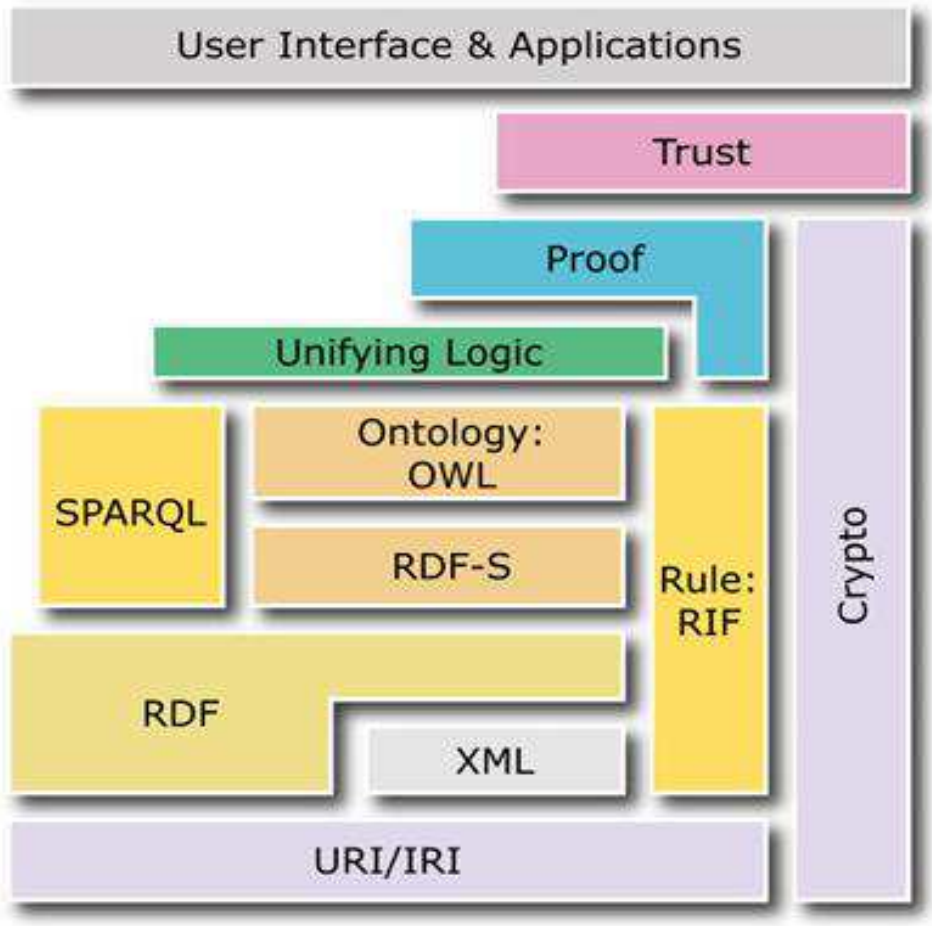
LESION (LOCALISATION) SITUATION

SITUATION (OBJET) ANATOMIE_OBJET

Concept défini

fracture (au_niveau_de) base (DE) crâne

Le web sémantique : une approche par couches



Les quatre principaux standards du web sémantique

- **RDF (1999) : modèle de triplets pour décrire et connecter des ressources anonymes ou identifiées par un URI**

  (sujet, prédicat, objet) ***/*** **graphe orienté étiqueté**
- **SPARQL (2006) : langage de requête sur les graphes RDF**
- **RDFS (2004) : langage de description de vocabulaires**

  subClassOf, subPropertyOf, range, domain
- **OWL (2004) : Extension de RDFS muni d'une sémantique formelle**

RDF : Resource Description Framework

- **RDF**
  - •Annotation sémantique des ressources (web)

    Ex : doc.html a pour auteur Mehdi et parle du web
  - •Assertion de liens entre ressources (donner du sens, c'est lier des informations ou connaissances)
  - •Plus riche qu'une annotation syntaxique par mot-clé
- **Triplet RDF <sujet, prédicat, objet>**
  - •Décrit un sujet (ressource identifiée par une URI)
  - •Associe au sujet un prédicat (qui dénote une relation, une propriété du sujet identifiée par une URI)
  - •Donne une valeur à une propriété = l'objet. Une valeur est soit une ressource identifiée par une URI, soit une valeur primitive (un littéral)
  - •Les triplets RDF forment des graphes

Triplets RDF

• Exemples :

<http://www.u-picardie.fr/~furst/ , dc:creator , "Fürst">

<http://www.u-picardie.fr/~furst/ , dc:contributor , http://w3c.org>

<http://www.u-picardie.fr/~furst/ , #a_destination_de , #etudiants_info>

<http://w3c.org , #has_for_member , "CERN">

• Les URL introduites par # sont définies localement

• Des prédicats sont définis dans des vocabulaires existants :

– dc (Dublin Core) : schéma de métadonnées pour les documents (*dublincore.org)*

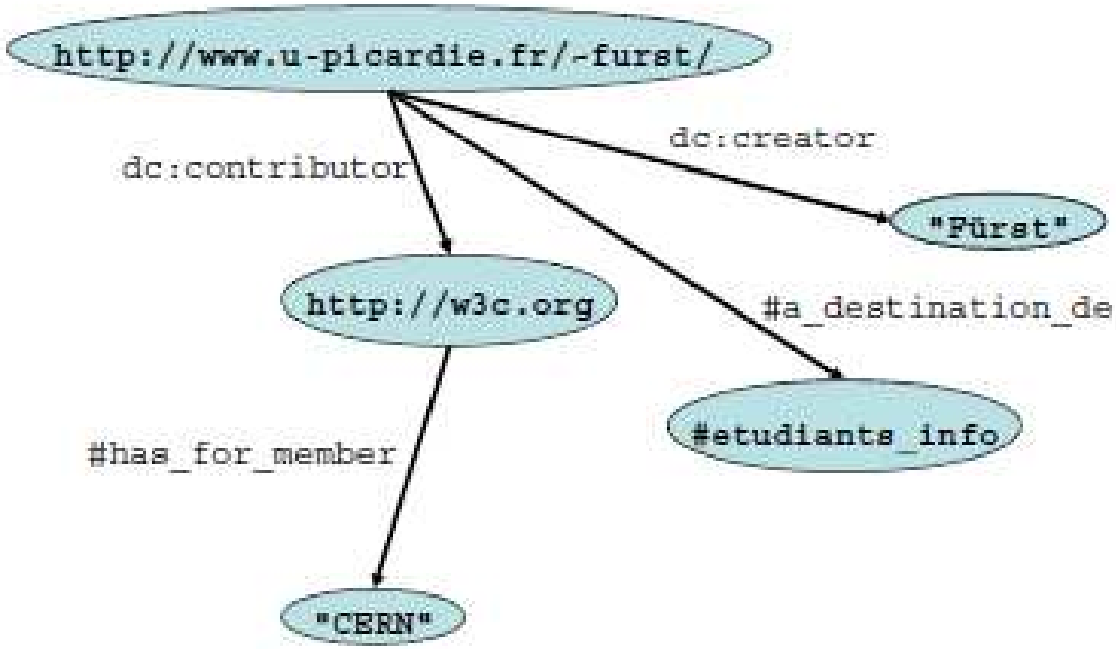
– Il existe d'autres vocabulaires (ex. foaf Friend of a Friend, *www.foaf-project.org)*

– rdf définit aussi ses propres prédicats

Triplets RDF

Les triplets RDF forment des graphes d'entités

```
<http://www.u-picardie.fr/~furst/ , dc:creator , "Fürst">
<http://www.u-picardie.fr/~furst/ , dc:contributor , http://w3c.org>
<http://www.u-picardie.fr/~furst/ , #a_destination_de , #etudiants_info>
<http://w3c.org , #has_for_member , "CERN">
```



SPARQL langage de requêtes *"à la SQL"* pour interroger des graphes RDF

- **SPARQL Protocol And RDF Query Language**
- **SPARQL : 3 outils importants :**
  1. Un langage de requête sur des graphes RDF permettant de spécifier le type de données recherchées
  2. Un protocole pour soumettre une requête à un serveur distant et recevoir les résultats
  3. Un format (JSON, XML, CSV, etc.) pour représenter les résultats d'une requête

Exemple de requête SPARQL

- **Donner les ?t et ?p tels que ?p est le dessinateur de ?t**

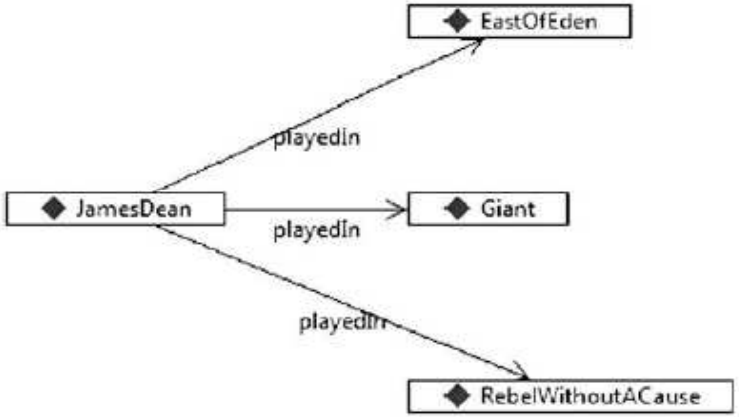
```
PREFIX bd: <http://www.collection.com/bd#>

SELECT * WHERE {
{ ?t bd:dessinateur ?p }
}
```

Résultats (sérialisés en XML)

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="t"/>
    <variable name="p"/>
  </head>
  <results>
    <result>
      <binding name="t"><uri>http://www.collection.com/bd/serie/Laufeust
de Troy/Thanos l incongru</uri></binding>
      <binding name="p"><uri>http://www.collection.com/personne/Tarquin</
uri></binding>
    </result>
    <result>
      <binding name="t"><uri>http://www.collection.com/bd/Laufeust de
Troy/L ivoire du Magohamoth</uri></binding>
      <binding name="p"><uri>http://www.collection.com/personne/Tarquin</
uri></binding>
    </result>
  </results>
</sparql>
```

Autre exemple de requête SPARQL

```
Ask: SELECT ?what WHERE {:JamesDean :playedIn ?what}
Answer: :Giant, :EastOfEden, :RebelWithoutaCause.
```



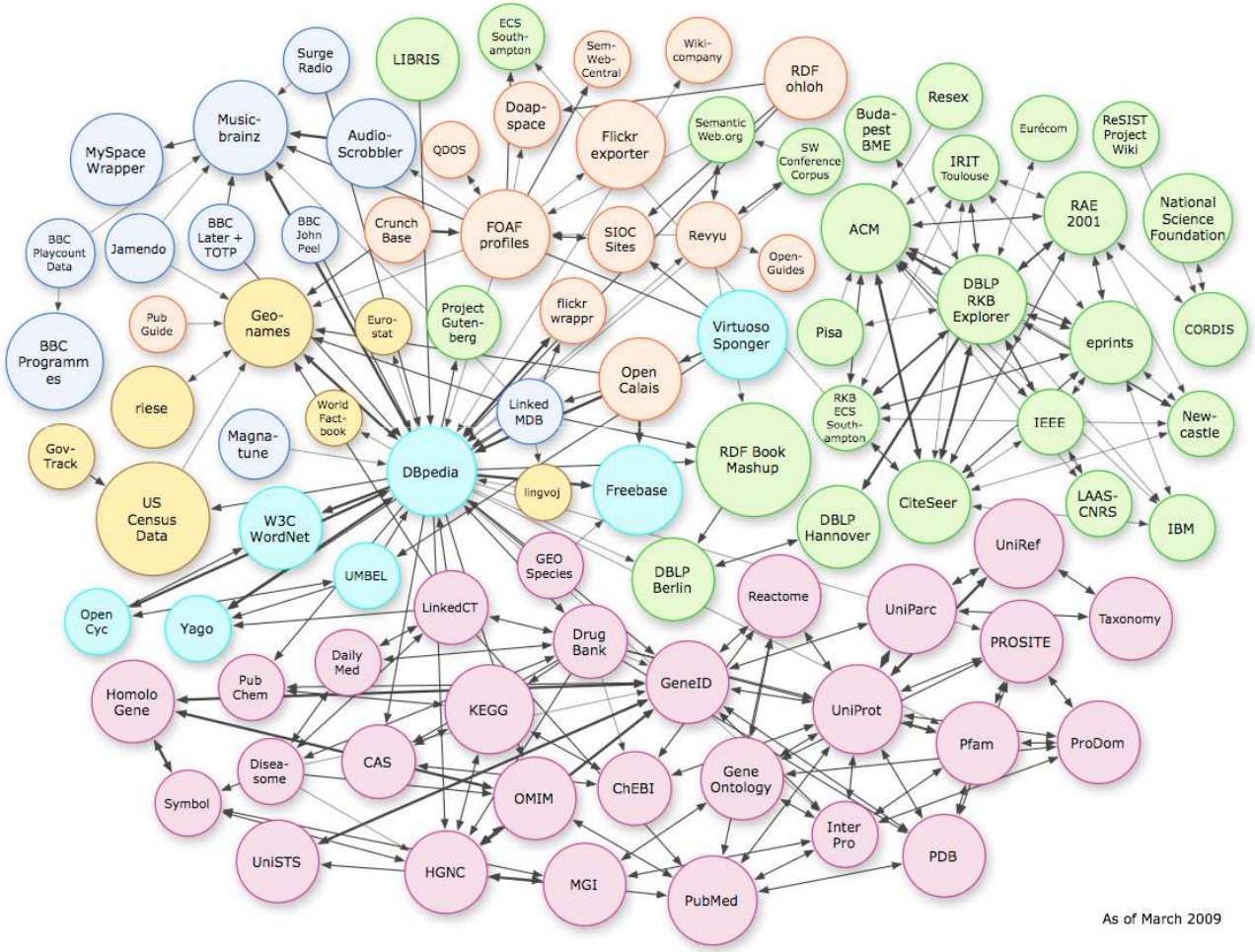
Graphe données

Graphe requête

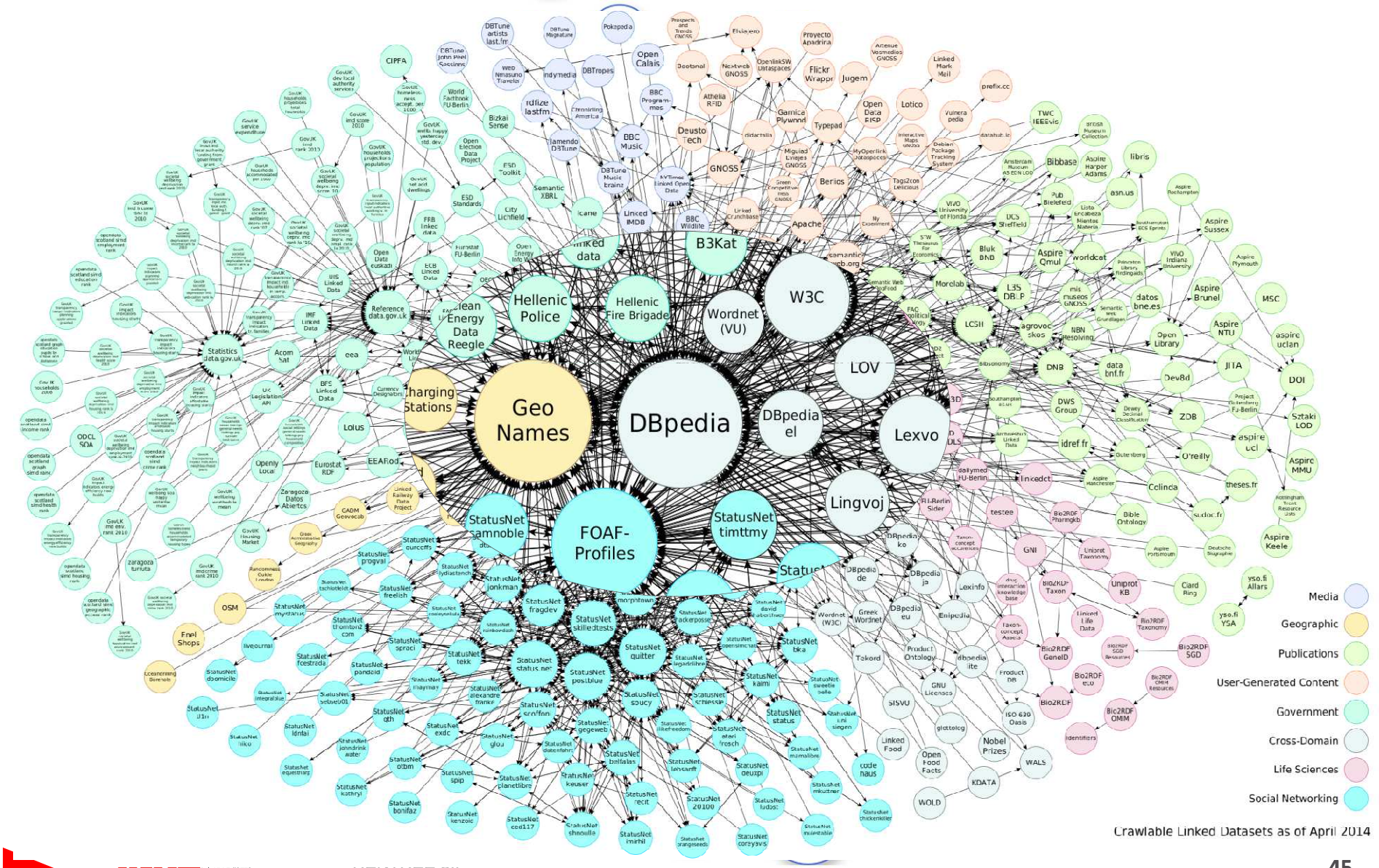
Le Web de Données

- **Première phase de déploiement massif du Web sémantique reposant sur RDF et SPARQL**
  - RDF : modèle de Données
  - SPARQL : langage d'interrogation
- **Linked Open Data (LOD) : mise en ligne de données libres en utilisant URI, HTTP, RDF et SPARQL**

Linked Open Data

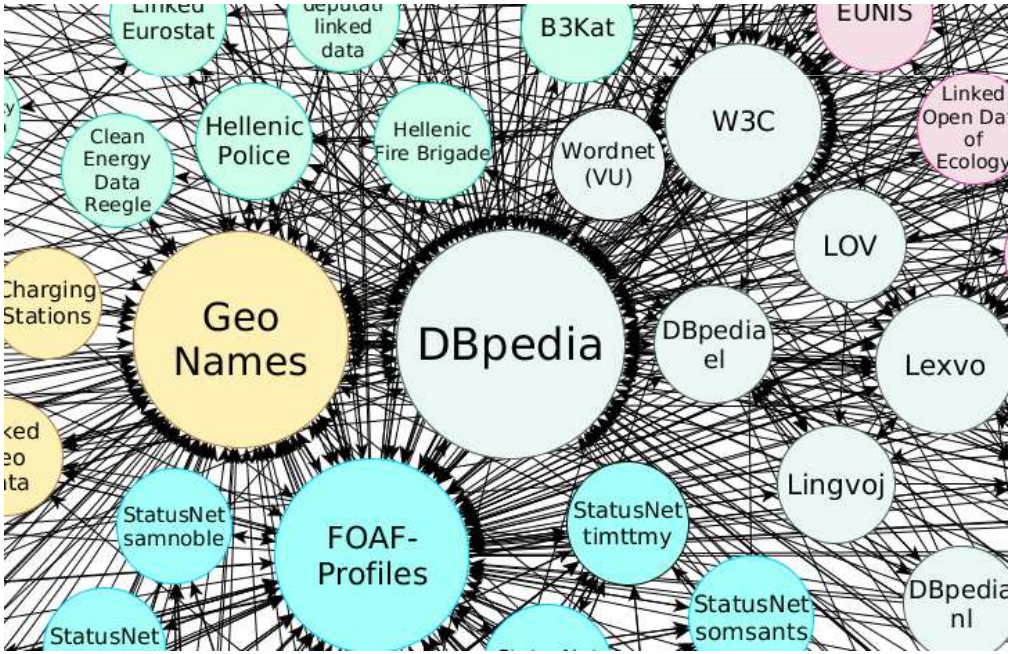


L’évolution du Linked Open Data (LOD)



DBpedia

- **DBpedia contient des informations structurées issues de Wikipedia.**
- **Des informations importantes disponibles sur Wikipedia sont transformées en triplets RDF.**



Langages de représentation des ontologies

- **RDF Schema ou RDFS :**

  RDF : modèle de données pour les objets et leurs relations

  RDF Schema : langage de description du vocabulaire

  Décrit les propriétés et les classes des ressources RDF

- **OWL (Ontology Web Language) :**

  Langage plus riche

  Relations entre classes

  Contraintes de cardinalité

  Propriétés de typage plus riches

RDF Schema

- **Déclaration et description :**
  - Des types de ressources manipulées (classes)

  Exemple : les livres, les films, les personnes

  - Des types de relations entre ces ressources (propriétés)

  Exemple : « a pour auteur », « a pour acteur », « a pour titre »

- **Définition du vocabulaire utilisé dans les graphes RDF**

RDF Schema : ontologies légères

- **Nommer et définir un vocabulaire conceptuel consensuel et faire des inférences élémentaires**

  Nommer les classes de ressources existantes

  Nommer les relations qui existent entre ces classes

  et donner leur signature

  Liens hiérarchiques entre classes **et entre propriétés**

  Donner un URI aux concepts qui vous sont importants

- **Squelette taxonomique d'une ontologie**

OWL: Ontologies lourdes

- **OWL sur une restriction de RDF/S**

  OWL Lite / DL / Full

  Logiques de description

  Vérification, classification, identification

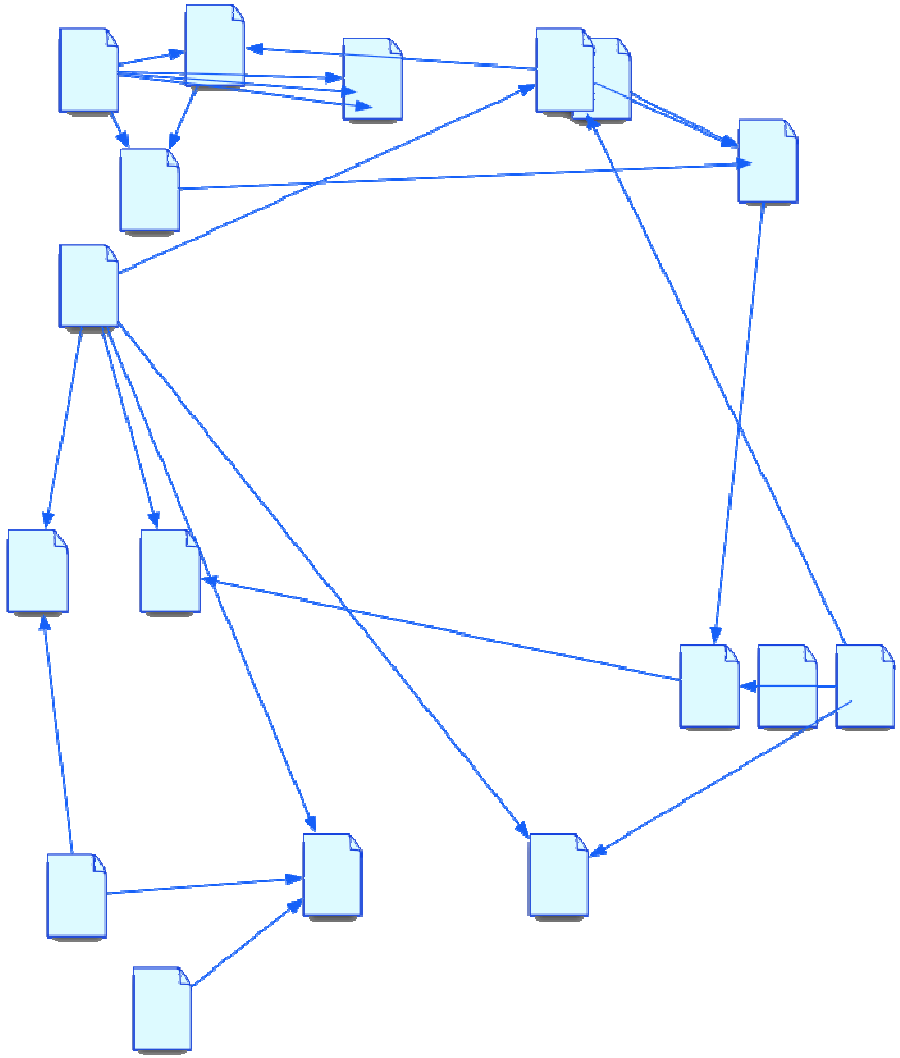
- **Définition de classes (énumération, union, intersection, complément, disjonction, restriction valeur et cardinalité des propriétés)**
- **Caractérisation des propriétés (symétrique, transitive, fonctionnelle, inversement fonctionnelle, inverse)**
- **Gestion des équivalences, versions, documenter**

Rule Interchange Format : RIF

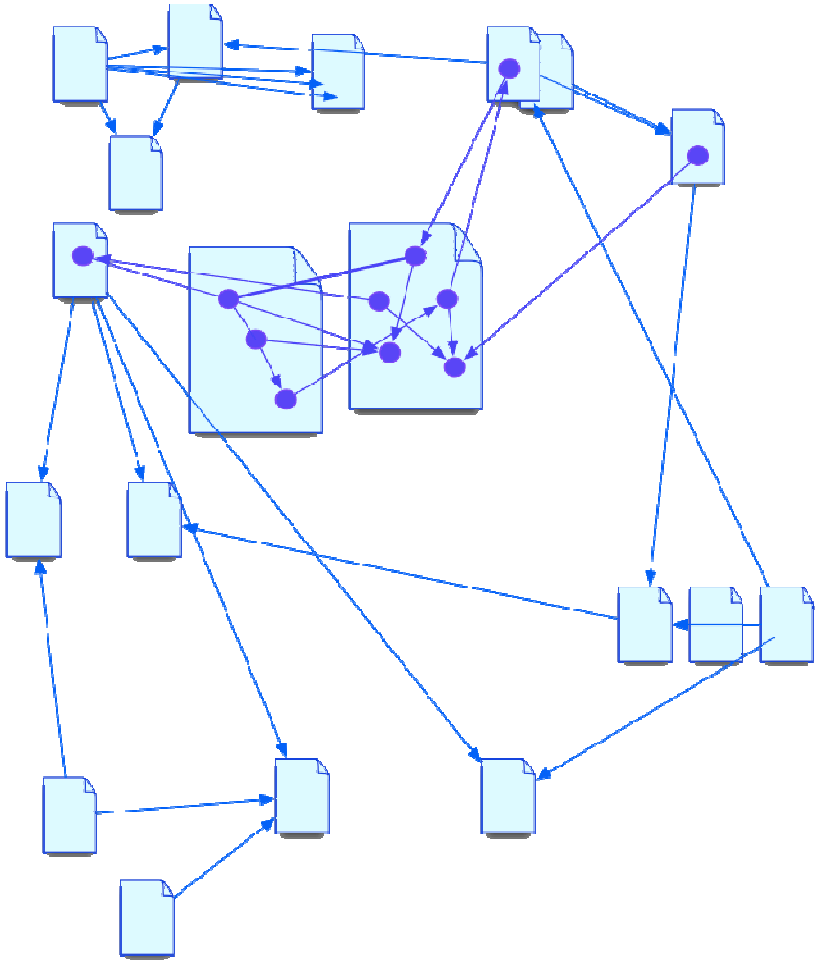
- **RIF est dédié à l'échange de règles d'inférence ou de règles de production sur le web sémantique**
- **Famille de langages : RIF Core, RIF BLD et RIF PRD**
- **RIF Core : échange de connaissances de déduction sous la forme de règles logiques proches des clauses de Horn**

  SI une hypothèse est vraie ALORS une conclusion peut être déduite ou ajoutée
- **Exemple : SI une personne a écrit une thèse sur le séquençage du génome ALORS cette personne peut être rangée dans la classe des docteurs et on peut ajouter la génomique parmi ses centres d'intérêt**

Le web de documents : les liens cachent les ordinateurs



Le web sémantique : les entités cachent les documents

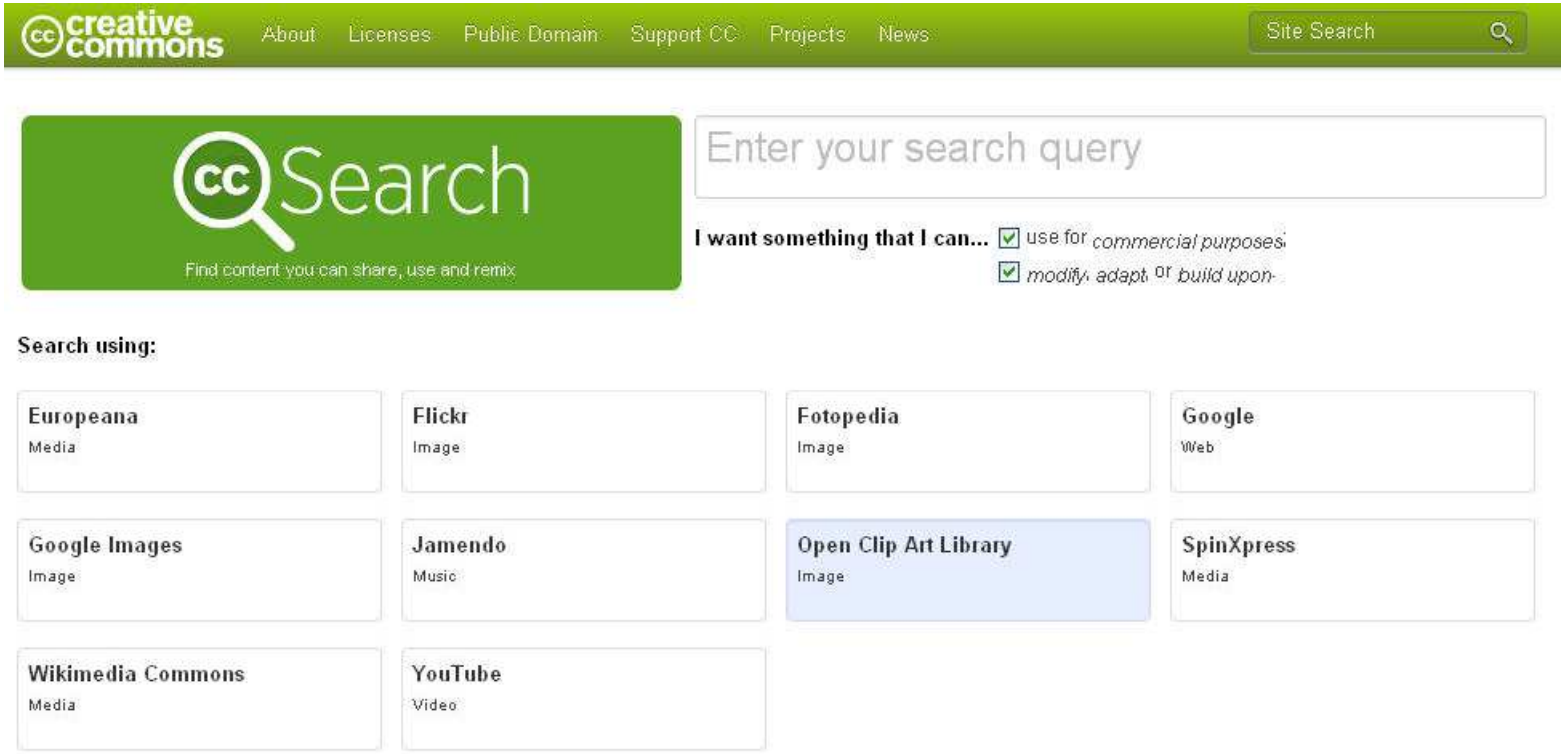


Des exemples d'application

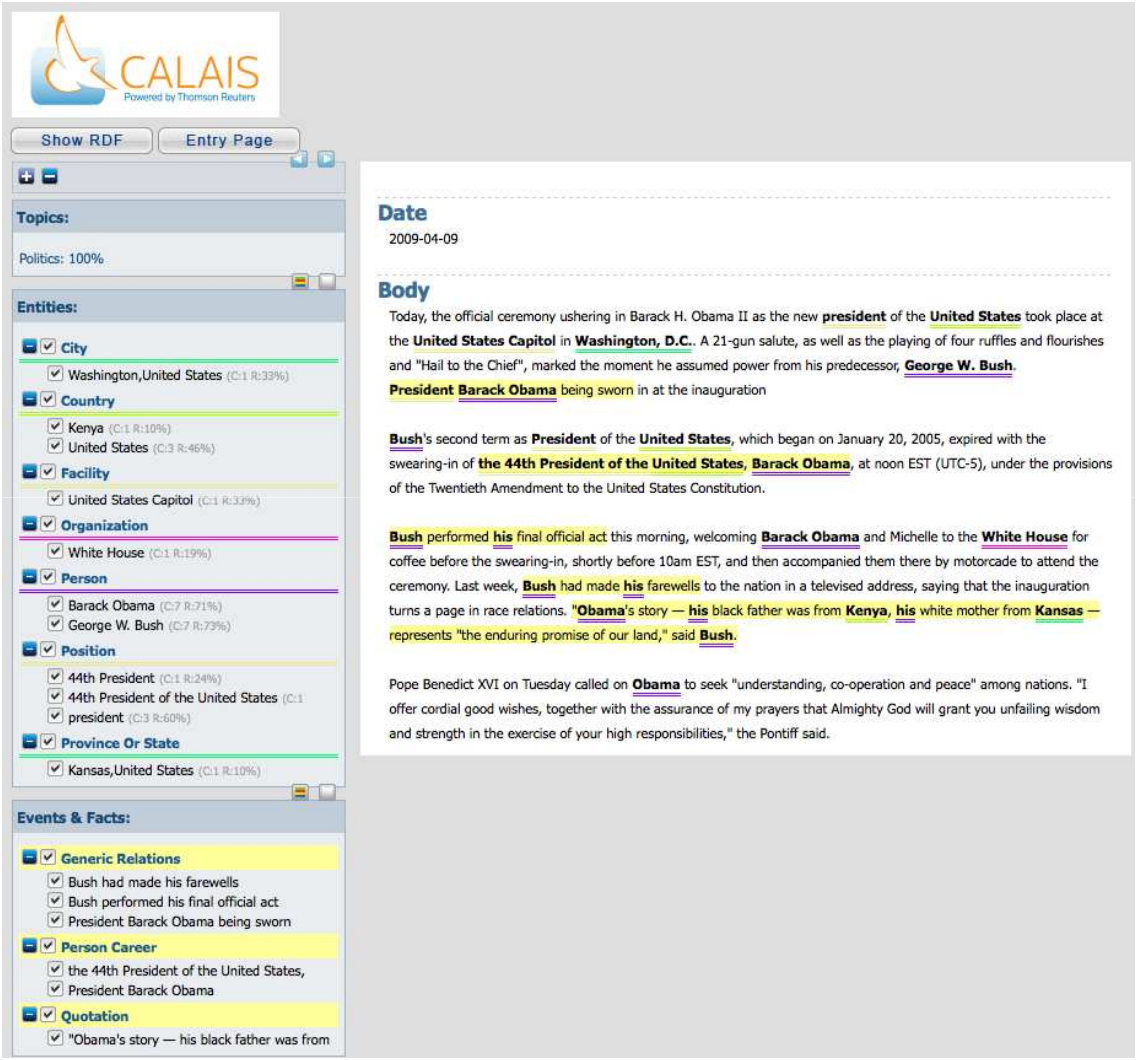
- SearchMonkey de Yahoo! (2008)
- Creative Commons (2008)
- OpenCalais (2008)
- Google Knowledge Graph (2013)
- EnSEn : Enhanced Search Engine (2014)

SearchMonkey de Yahoo!

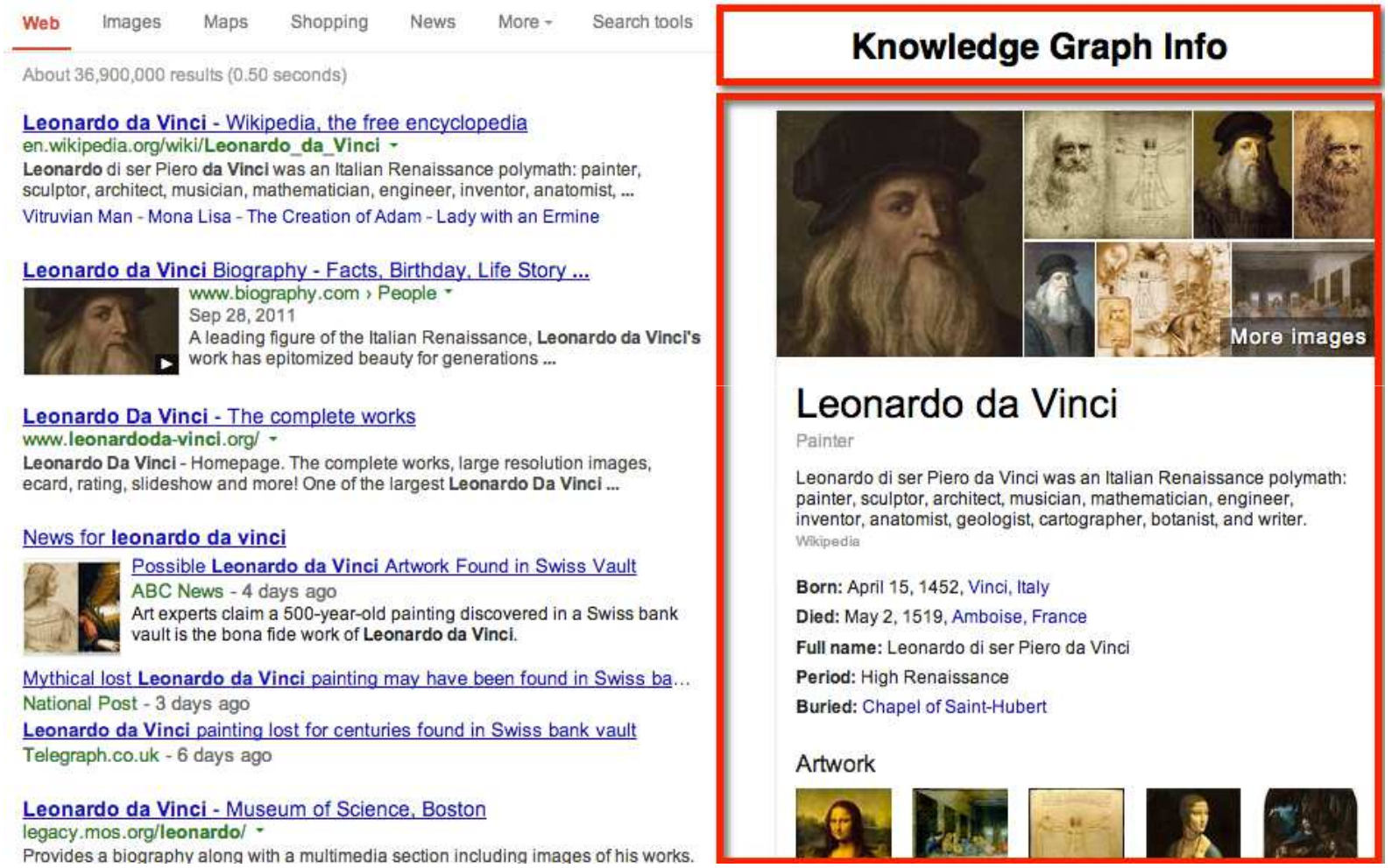
Creative Commons



OpenCalais


CALAIS
Powered by Thomson Reuters
Show RDF
Entry Page
Topics:
Politics: 100%
Entities:
City
Washington,United States (C:1 R:33%)
Country
Kenya (C:1 R:10%)
United States (C:3 R:46%)
Facility
United States Capitol (C:1 R:33%)
Organization
White House (C:1 R:19%)
Person
Barack Obama (C:7 R:71%)
George W. Bush (C:7 R:73%)
Position
44th President (C:1 R:24%)
44th President of the United States (C:1
president (C:3 R:60%)
Province Or State
Kansas,United States (C:1 R:10%)
Events & Facts:
Generic Relations
Bush had made his farewells
Bush performed his final official act
President Barack Obama being sworn
Person Career
the 44th President of the United States,
President Barack Obama
Quotation
"Obama's story — his black father was from
Date
2009-04-09
Body
Today, the official ceremony ushering in Barack H. Obama II as the new president of the United States took place at the United States Capitol in Washington, D.C.. A 21-gun salute, as well as the playing of four ruffles and flourishes and "Hail to the Chief", marked the moment he assumed power from his predecessor, George W. Bush.
President Barack Obama being sworn in at the inauguration
Bush's second term as President of the United States, which began on January 20, 2005, expired with the swearing-in of the 44th President of the United States, Barack Obama, at noon EST (UTC-5), under the provisions of the Twentieth Amendment to the United States Constitution.
Bush performed his final official act this morning, welcoming Barack Obama and Michelle to the White House for coffee before the swearing-in, shortly before 10am EST, and then accompanied them there by motorcade to attend the ceremony. Last week, Bush had made his farewells to the nation in a televised address, saying that the inauguration turns a page in race relations. "Obama's story — his black father was from Kenya, his white mother from Kansas — represents "the enduring promise of our land," said Bush.
Pope Benedict XVI on Tuesday called on Obama to seek "understanding, co-operation and peace" among nations. "I offer cordial good wishes, together with the assurance of my prayers that Almighty God will grant you unfailing wisdom and strength in the exercise of your high responsibilities," the Pontiff said.

Google Knowledge Graph



ENsEN : Enhanced Search Engine

Outils/Implémentations disponibles

- Outils : http://www.w3.org/2001/sw/wiki/Tools
- Langages de développement : C, C++, Java, PHP, Javascript, Python, Perl, C#, Prolog, …
- Plus de 40 Triple Stores : Jena, TripleStoreJS, …
- Plus de 50 outils de développement : Protégé, SWObjects, …
- Beaucoup de livres : http://esw.w3.org/topic/SwBooks

Bibliographie

- GANDON Fabien et al. *Le web sémantique : comment lier les données et les schémas sur le Web.* Paris, Dunod, 2012.
- ALLEMANG Dean et HENDLER James. *Semantic Web for the Working Ontologist. Effective Modeling in RDFS and OWL. Morgan Kaufmann, 2011.*
- *Site web du W3C : http://www.w3.org*