Institut
Mines-Télécom

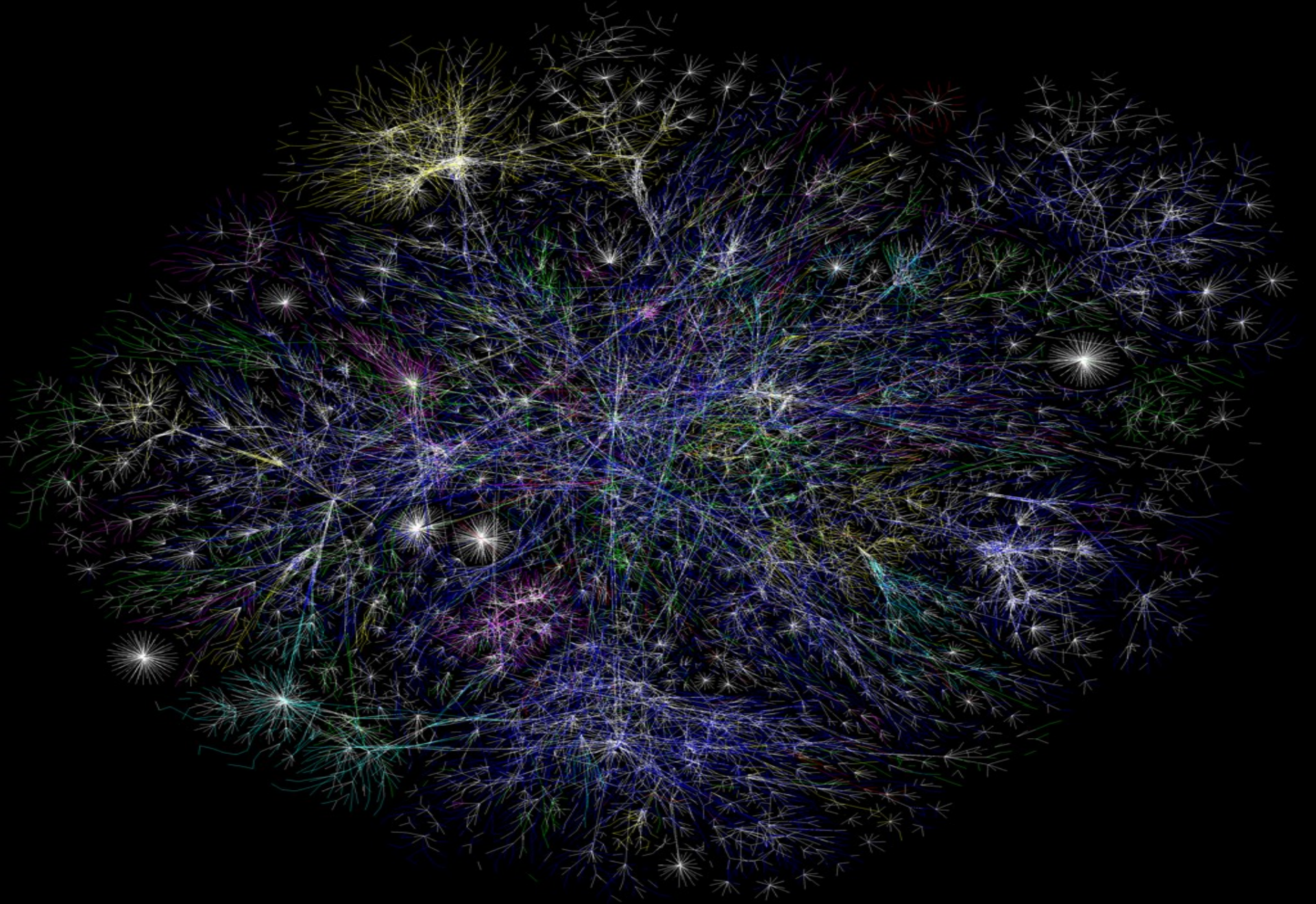
Collecter des informations sur le Web

Robots, Crawler, scraper, spider,...

Télécom ParisTech

Jean-Claude Moissinac – Mai 2019

Avec des éléments de Fabian Suchanek, Pierre Sennellart, Cyril Concolato



http://www.opte.org/

Plan

- **Introduction**
- **Sources de données**
- **Principes du crawling**
- **Analyse de pages**
  - Contenu
  - Liens
  - Données structurées
- **Services**
- **Web sémantique**
- **Outils**
- **Conclusion**

Plan

- **Introduction**
- Sources de données
- Principes du crawling
- Analyse de pages
  - Contenu
  - Liens
  - Données structurées
- Web sémantique
- Outils
- Conclusion

Objectif

- **Des milliards de pages**
- **Des entrepôts de données**
- **Des services de données**
- **Voir comment obtenir des données du Web**
  - Extraction de données de pages Web
  - Obtention de données disponibles
    - OpenData
    - APIs/services
    - Web sémantique

Plan

- Introduction
- **Sources de données**
- Principes du crawling
- Analyse de pages
  - Contenu
  - Liens
  - Données structurées
- Web sémantique
- Outils
- Conclusion

Types de sources de données sur le Web

Web classique ***approfondi dans la section crawler***

- **Pages Web HTML**
- **Pages dynamiques**
  - Dont négociation de contenu
- **Autres types de contenus**
  - Pdf, Gif, Txt…

- **Pages Web sémantique** ***abordé plus loin***
  - *Pages web avec des indications formalisées sur la nature des contenus*

- **APIs, Web Services**
- **OpenData**

Web Services: définition

- **Un Web Service est**
  - Un logiciel
  - Qui expose des fonctions via un protocole de communication (sur le Web, en général HTTP)
  - À l'aide de méthodes d'exploitation standardisées qui en systématisent l'utilisation
    - Indépendance des langages et des systèmes

- **Cela permet**
  - De rendre accessible des services
  - De les distribuer
  - De composer des services évolués à partir de services élémentaires
  - De bénéficier d'une infrastructure réseau bien établie

Exemples

REST altitude

- http://api.geonames.org/astergdem?lat=45.64&lng=1.85&username=demo
- https://elevation-api.io/api/elevation?points=(45.64,1.85),(62.52417,10.02487)

REST POI voisins

http://api.geonames.org/findNearbyPlaceName?lat=45.64&lng=1.85&username=demo&style=full

http://api.geonames.org/findNearbyPlaceNameJSON?formatted=true&lat=45.64&lng=1.85&username=demo&style=full

REST: Representational state transfer

- **Ni un protocole, ni un format**
- **Un style de mise en œuvre de système distribué**
  - De ce fait, on peut s'inspirer du modèle sans en respecter tous les principes
  - Proposition initiale: thèse de Roy Fielding

- **Principes de base:**
  - Il suffit de connaître l'URI d'un service et son mode d'emploi pour y accéder
  - HTTP fournit toutes les fonctions nécessaires
    - GET, PUT, POST, DELETE
    - Les verbes de HTTP utilisés comme commandes d'actions sur le serveur
  - Fonctionnement sans état
    - Si on est puriste!

Avantages de REST

- **Simplicité de mise en œuvre**
  - En tout cas pour des développeurs habitués au développement de sites Web dynamiques
- **Avantages liés à l'absence d'état**
  - Moindre charge du serveur => meilleure capacité de réponse
  - Facilité de mise au point
  - Facilité de répartition de la charge
- **Très bonne intégration dans l'univers HTTP**
- **L'association URL/ressource facilite l'exploitation de caches**

APIs Web

- On parle souvent d'API Web pour les services accessibles sur Internet

- ProgrammableWeb

https://www.programmableweb.com/category/all/apis

- exemple de répertoire d'API
- des centaines d'API recensées: cartographie, réseaux sociaux, traduction…

Plan

- Introduction
- Sources de données
- **Principes du crawling**
- Analyse de pages
  - Contenu
  - Liens
  - Données structurées
- Web sémantique
- Outils
- Conclusion

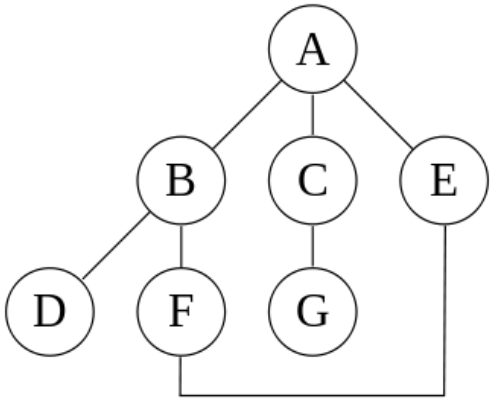
Crawler

- **Parcours automatisé du Web**
- **Extraction de données de pages Web**
- **Actions principales**
  - Choisir des pages à parcourir
  - Parcourir les pages
  - Exploiter les pages obtenues
- **Choix principaux**
  - Choix des pages où commencer
  - Choix des URLs à suivre dans ces pages
  - Méthodes de recherche de données dans la page

Crawler: rôle

- **Crawlers, (Web) spiders, (Web) robots**
  - Outils qui parcourent automatiquement des pages Web
- **Buts**
  - Sauvegarde/Archives
  - Analyse de site
  - Consultation offline
  - Extraction de données (veille, fusions…)
    - Par analyse au vol
    - Par analyse offline
- **Fonctions contraintes**
  - Limiter aux pages importantes le périmètre du contenu parcouru
  - Eviter les pièges à robot
  - Faire face à la variabilité des contenus

Crawler: Parcours



Voir plus loin 'analyse de page' pour les sources de liens

- **Parcours en profondeur** *(ABDFECG)*
  - URL1->URL2->URL3->URL4…
  - On peut se perdre et ne jamais parcourir d'autres parties du web
- **Parcours en largeur** *(ABCEDFG)*
  - URL1->URL1.1
    - -> URL 1.2
    - -> URL 1.3
    - …
- **Combinaison des deux en mettant une borne sur chaque parcours en profondeur**
- **Il peut y avoir des boucles**

Crawler: Procédé général

- **Sélectionner un ensemble d'URLs à traiter**
- **Récupérer une page en suivant une URL**
  - Maintenir un index des pages visitées
- **Analyser la page**
  - Par exemple avec l'API DOM
- **Sauver le contenu important pour le projet courant**
- **Extraire les URLs de la page et en choisir certaines**
- **Ajouter ces URLs dans la liste des URLs à traiter**
- **… et boucler là-dessus**
  - Soit tant qu'il y a des URLs à traiter
  - Soit que vous ayez récupéré assez de données
  - Soit que vous ayez trouvé l'information que vous cherchez
  - Soit que trop de temps est écoulé
  - …

Crawler: Limiter le parcours

- **Taille du Web: le Web est infini!**
  - Pièges à robots, boucles
  - Pages dynamiquement créées
- **Garder un focus sur des pages importantes**
  - Dans un contexte donné
    - Par domaines DNS
    - Par sujets
    - …

Limiter le parcours (1)

- **Focus sur des pages importantes**
  - Dans un contexte donné
  - [Abiteboul et al., 2003]
- **Focus sur une liste de domaines DNS**
  - filtrage simple des URLs
- **Focus sur un sujet**
  - techniques de crawling ciblé [Chakrabarti et al., 1999, Diligenti et al., 2000]
  - basé sur la classification de page Web et évaluation/prédiction de l'intérêt d'un lien

Limiter le parcours (2)

- **A l'intérieur d'un domaine**
  - Limiter la profondeur de visite
  - Limiter le nombre pages visitées
- **Limiter à une liste de noms de domaine**
  - Ex: le même que l'URL de départ
- **Limiter en définissant une condition d'arrêt**
  - Ex: termes trouvés dans la page
  - Ex: critère de filtrage d'URL
- **Limiter à des types de contenus**
  - Ex: arrêter d'explorer une branche quand on trouve un PDF ou une image

Crawler: Bonnes pratiques

- **Eviter le DOS (Denial Of Service)**
  - Attendre de 100ms à plusieurs secondes avant de solliciter à nouveau un domaine déjà sollicité
  - Ex: WikiCFP->délai 5s; DBPedia-> délai >10ms
- **Respecter les exclusions**
  - Fichier robots.txt
    - Fichier à la racine d'un serveur qui indique les pages qu'un robot peut parcourir [Koster, 1994]
      - User-agent: *
      - Allow: /tupeuxyaller
      - Dissallow: /nyvaspas
  - Exclusion par meta dans une page

    <meta name=« ROBOTS » content=« NOINDEX, NOFOLLOW »>
  - Exclusion sur un lien

    <a href=« mapagesecrete.html » rel=« nofollow »>…

Crawler: Traitement parallèle

- **Délais des réponses réseau**
  - => attente des réponses et ‘callback’
- **File d’attente par domaine**
  - Et réglages associés: délai, parseur, filtres
- **Traitements parallèles des requêtes**
  - Programmation multi-thread
  - Entrées/sorties asynchrones
- **Utilisation de l’option keep-alive (ou HTTP/2) pour diminuer la charge des connexions**
- **Distribution**
  - Map-reduce

Crawler: contraintes (1)

- **Eviter de revisiter des pages déjà visitées**
  - Si elles n'ont pas été modifiées
  - Peuvent être accédées par des URLs différentes
  - Peuvent conduire à des boucles dans le parcours
- **Prévoir une fréquence de mise à jour**
- **Faire face à la variabilité des contenus Web**
  - Types de resources Ex type MIME
  - Versions des normes (HTML…) et non respect des normes (TAGSOUP)
- **Définir des méthodes d'extraction d'information**
- **Tenir compte des limites placées par les serveurs**

Crawler: Contraintes (2)

- **Identifier les pages mises à jour**
- **Une méthode courante:**
  - le hachage d'URL
    - = calcul d'une valeur numérique représentative de l'URL
    - Quand on trouve deux URL associées à la même valeur numérique, on peut approfondir la comparaison
  - Le hachage de contenu
    - Même principe, mais sur le contenu de la page
    - Pour détecter une page où on arrive par plusieurs URLs
- **Difficulté: pages presque identiques**
  - Ex: à la l'heure près pour une page qui affiche l'heure

Note: des fonctions de hachage sont disponibles dans la plupart des langages

Crawler: éviter les pages visitées estampille temporelle

- **HTTP Timestamping: 2 mécanismes, potentiellement utilisés avec chaque requête**
  - **entity tags**
    - **identificateur unique du document; change si le document change**
    - **Peut être utilisé comme sélecteur dans la requête (If-Match)**
  - **modification dates**
    - **Peut être utilisé comme sélecteur dans la requête (If-Modified-Since)**

```
If-Modified-Since: Wed, 15 Oct 2008 19:40:06 GMT
```

```
ETag: "497bef-1fcb-47f20645"
Last-Modified: Tue, 01 Apr 2008 09:54:13
```

Souvent fournis pour les contenus statiques

Rarement fournis pour les contenus dynamiques

Crawler: HTTP Cache-Proxy

- Deux autres indications de ‘fraicheur’ d’un contenu, pour les caches et les proxies:

```
Cache-Control: max-age=60, private  Expires: Tue, 01 Apr 2008 13:25:55 GMT
```

- max-age: délai maximum en secondes où un document est garanti rester à jour
- Expires: date à laquelle un document sera considéré comme dépassé
- Souvent fourni...
- ... Mais avec 0 ou un délai d’expriration très court.
- ⇒ information de faible portée

- Données meta et autres dans le contenu des pages

Crawler: autres informations temporelles

- **Des fichiers autres que HTML peuvent avoir une information de version et/ou de date**
  - PDF, documents Open Office, etc.: des metadonnées contiennent des informations de date de création et de dernière modification.
  - RSS feeds: contiennent des estampilles temporelles fiables
  - Images, Sons: EXIF metadata (ou similaire). Pas toujours exploitable
  - Sitemaps

Crawler: content type negociation

- **Navigateur et serveur se comportent différemment en fonction du type de contenu**
  - Mise en page d'un contenu HTML
  - Affichage brut d'une page de texte
  - Affichage d'une image
- **Le client peut indiquer les types qu'il préfère**
- **MIME est le standard de déclaration de type de contenu**
  - Exemple: image/jpeg, text/plain, text/html, application/xhtml+xml, application/pdf
- **Les documents texte et HTML doivent aussi être accompagnés d'une indication sur le jeu de caractère utilisé**

Exemple

```
HTTP/1.1 200 OK
Content-Type: text/html;
charset=UTF-8
```

Client et serveur: identification

- Les clients Web clients et les serveurs peuvent s'identifier avec une chaîne de caractère (protocole HTTP)
- Utile pour servir des contenus différents à différents navigateurs, pour détecter des robots…
- … mais n'importe quel client peut s'annoncer comme étant un autre

Exemple

User-Agent: Mozilla/5.0 (X11; U; Linux x86_64; fr; rv:1.9.0.3) Gecko/2008092510 Ubuntu/8.04 (hardy) Firefox/3.0.3

Server: Apache/2.0.59 (Unix) mod_ssl/2.0.59 OpenSSL/0.9.8e

Crawler: Doublons

- **Détecter les doublons ou les presque doublons**
  - Pour éviter l'indexation multiple d'un contenu
- **Cas trivial: même ressource issue de même URL**
  - Détecté avec version canonique de l'URL
- **Détection de contenu strictement identique**
  - Comparaison par hachage
- **Contenus presque identiques**
  - Date, Conseil du jour, Publicité…
  - Personnalisation Ex: nom du visiteur connecté
  - => plus compliqué à détecter
  - => on peut essayer de détecter et d'éliminer une partie du contenu variable

Crawler: Doublons stricts et hachage

- **Détection de contenu strictement identique**
  - Comparaison par hachage
- **Une fonction de hachage est une fonction mathématique transformant un objet numérique (nombres, chaîne de caractère, binaire,…) en un nombre pseudo-aléatoire de taille fixe**
- **Par exemple, pour une chaîne**
  - $\sum s[i]*31^{n-i-1} \bmod 32$

Crawler: pages très proches

- **Distance d'édition (edit distance) Ex** Levenshtein
  - Compter le nombre de modifications élémentaires – ajout, suppression, échange- pour passer d'une chaîne de caractère à l'autre
  - Ne passe pas à l'échelle sur un très grand nombre de documents où il faudrait comparer toutes les paires possibles
- **Shingles**
  - Principe: 2 documents sont similaires si ils partagent un grand nombre de k-grams (suite d'éléments de longueur k)
  - Exemple: `I like to watch the sun set with my friend.`
  - `My friend and I like to watch the sun set.`
  - `S ={i like, like to, to watch, watch the, the sun, sun set, with my, my friend}  T ={set with, with my, friend and, and i}`

Institut
Mines-Télécom

Analyse de page Web

HTML (HyperText Markup Language) [W3C, 1999]

- Normalisé par le W3C (World Wide Web Consortium) formé d'industriels (Microsoft, Google, Apple. . .) et d'institutions academiques (ERCIM, MIT, IMT, etc.)
- Format ouvert: exploitation, traitement par de nombreux softwares et hardwares
- Fichier text avec tags (ex: …)
- décrit la structure et le contenu du document
- recommande d'éviter les données de présentation (ce role est dévolu au CSS)
- Pas de description de comportement dynamique (ce role est dévolu au Javascript ou aux traitements côté serveur)

The HTML language

- HTML est un langage alternant texte et tags (<blabla> or </blabla>)
  - Les tags permettent de structurer des parties d'un document, et sont notamment utilisés par les navigateurs pour guider la mise en page du document
  - Un document est structur en deux parties principales: l'entête <head> ... </head>)

    Et le corps <body> ... </body>)

Tags

- Syntaxe: (tag ouvrant et fermant)

```
<tag attributs>contenu</tag>
```

ou (élements sans contenu)

```
<tag attributs />
```

- Pour certains tags, le contenu peut être composé de texte et d'autres tags
- Les attributs représentent les paramétres associés avec l'élément, sous forme de liste nom=« valeur » séparés par des espaces

The different versions of HTML

- HTML 4.01 (1999) strict and transitional

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
```

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
```

- XHTML 1.0 (2000) strict and transitional
- XHTML 1.1 and XHTML 2.0
- HTML5: standard actuel, continuellement mis à jour

```
<!DOCTYPE html>
```

Tag soup

- **De nombreux documents HTML du Web datentd'avnt HTML 4.01**
- **En pratique:**
  - de nombreuses pages Web ne respectent pas trictement un des standards
  - Les navigateurs ne respectent pas trictement un des standards
- **⇒ tag soup!**
- **Lorsqu'on exploite des pages Web, il est nécessaire d'appliquer des heuristiques pour interpréter les pages**

Structure d'un document HTML

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <!-- Header of the document-->
  </head>
  <body>
    <!-- Body of the document -->
  </body>
</html>
```

- <!DOCTYPE ...> spécifie quelle version de HTML est utilisée

Header

- Le header est delimité par les tags
  <head> . . . </head> .
- Le header contient des meta-informations

  tels que: title, encoding, fichiers associés, etc.

  Le modèle de représentation des caractères, normalement spécifié au tout début du header:

  <meta http-equiv="Content-Type" content="text/html; charset=utf-8">

  Utf-8 très utilisé (compatible avec l'historique ascii)

  Le titre (seul item obligatoire dans le header). Information affichée dans la barre de titre des navigateurs

  <title>Mon super site web</title>

body

- <body>... </body> tags delimitant le corps du document.
- structuré en sections, paragraphes, listes, etc.
- 6 tags descrivent les sections, par ordre décroissant d'importance order of importance:
  - <h1>Title of the page</h1>
  - <h2>Title of a main section</h2>
  - <h3>Title of a subsubsection</h3>

  ...
- ... tags délimitent des paragraphes de texte.
- tag délimitent des blocs sans signification précise

Sources de liens dans une page HTML

- **Hyperliens**
  -
- **Media**
  - <img src=« … »>
  - <embed src =« … »>
  - <object data=« … »>
- **Frames**
  - <frame src=« … »>
  - <iframe src=« … »>
- **Scripts. Exemples:**
  - Différentes formes d'appels AJAX
  - Window.open(« … »)
- **Sitemaps (cf.sitemaps.org)**

Liens

- Ce qui différencie les pages Web (pages hypertext) de simples documents: les liens!
- Introduits avec <a> ...
- Un lien peut envoyer vers:
  - Un document sur un autre serveur
  - Un autre document sur le même serveur
  - Une autre partie du même document

```
<a href="http://www.cnrs.fr/">
  <img src="images/cnrs.gif" alt="CNRS">
</a>
<a href="#marqueur1">Notes</a>
<a href="bio/indexbioinfo.html">Bioinformatics</a>
```

Analyse de page

- **Du très empirique…**
  - Ex: Récupérer le contenu du 1er paragraphe du 3eme div
  - Pb: fragile dans le temps si la page évolue
- **Au très sophistiqué**
  - Identification de régularités sur des séries de page et définition automatisée de règles d'analyse
  - Analyses linguistiques ou sémantiques
- **Cas particuliers**
  - Recherche de suites de mots à partir d'un répertoire de référence
  - Recherche d'un 'motif' à partir d'expressions régulières (regular expression)

Analyse d'une page (exemple)

- http://bibliothequenumerique.tv5monde.com/livre/

105 liens

<a> 50
<img> 29
<script> 26

Tous n'ont pas besoin d'être suivis

Outils

Outils pour capter des pages

- **Outils tout prêt**
  - WinHtTrack (php)
  - Selenium (automatisation du Web)
  - Bixo (s'appuie sur Hadoop-Map Reduce)
  - Heritrix
  - Apache Nutch
- **Développement**
  - Scrapy (python), moteur de crawl
  - Crawler4j (java)

Extracteur de site, crawler

- **Exemple: WinHTTrack**

- **Aspire des pages reliées à une ou plusieurs pages de départ données**
- **Obtient une vision statique du site**
  - État des pages générées à un instant donné

- **Usage**
  - Sauvegarde
  - Analyse de site
  - Consultation offline
  - Extraction de données par analyse offline

Exemple PHP (principe): récupérer une page

```
<?php

$ch = curl_init("http://www.example.com/page1");
$fp = fopen("example_homepage.txt", "w");

curl_setopt($ch, CURLOPT_FILE, $fp);
curl_setopt($ch, CURLOPT_HEADER, 0);

curl_exec($ch);
curl_close($ch);
fclose($fp);
?>
```

Outils pour analyser les pages

- **DOM API**
  - Tous langages
  - Les pages doivent être ‘bien formées’…
  - … sinon utiliser HTML Tidy

- **XSLT**
- **Beautiful Soup (python)**
- **Boilerpipe (java)**
- **Apache Tika (Java)**
- **Jtidy (java)**
- **Readability**
  - Ciblé uniquement sur le texte des pages

Exemple PHP (principe): trouver les <a>

```
<?php
$file = "test.html";
$doc = new DOMDocument();
$doc->loadHTMLFile($file);

$elements = $doc->getElementsByTagName('a');

if (!is_null($elements)) {
  foreach ($elements as $element) {
     // ici trouver l'attribut href et l'ajouter dans une liste…
  }
}
?>
```

Trouver des informations dans les pages

- **Analyser (parser) la page pour y trouver des motifs**
  - Par exemple avec des expressions régulières
- **Parcourir la page avec l'API DOM**
- **Transformer la page avec XSLT**

- **Trouver des informations structurées dans les pages qui en ont**
  - Json-ld, RDFa, microformat
  - *Abordé plus loin*

Institut
Mines-Télécom

Web sémantique
et
Web des données

Idées du Web Sémantique

- **Rendre les données du Web exploitables**
  - Par les humains
  - Par des machines
  - *(De préférence par les deux)*

- **Pour cela, il faut**
  - Marquer/typer des données dans les pages du web
  - Définir une méthode pour publier des données sur le Web

- **Résultats attendus**
  - Faire traiter des données par des machines
  - Tisser des liens entre des données dispersées

Web Sémantique

- **Définir une infrastructure qui permet aux machines d'opérer sur les données en les 'comprenant'**
- **C'est-à-dire:**
  - Permettre à des machines d'opérer sur les données d'autres machines
  - Assurer l'interopérabilité
  - Permettre aux données de se décrire elles-mêmes
  - Permettre aux machines de raisonner sur les données
  - Permettre aux machines de fournir des réponses à des requêtes 'sémantiques'
- **Méthode: s'appuyer sur le WWW pour rendre les données disponibles d'une façon standard, notamment dans les pages web**

Marquage sémantique

https://developers.google.com/search/docs/guides/intro-structured-data

- RDFa
- Microdata
- Json-ld

Schema.org

- Event, Organization, Person, Product, Review, AggregateRating, Offer
- **schema.org**

Représentation des connaissances

Granules de connaissances

- Les triplets RDF
- (sujet)(prédicat)(objet)
- Sujet: l'entité sur laquelle porte la connaissance
- Prédicat: l'affirmation qu'on fait sur le sujet; une propriété applicable au sujet
- Objet: valeur qu'on associe au prédicat (valeur de la propriété

L'ensemble constitue une connaissance sur le sujet

Resource Description Framework

RDF

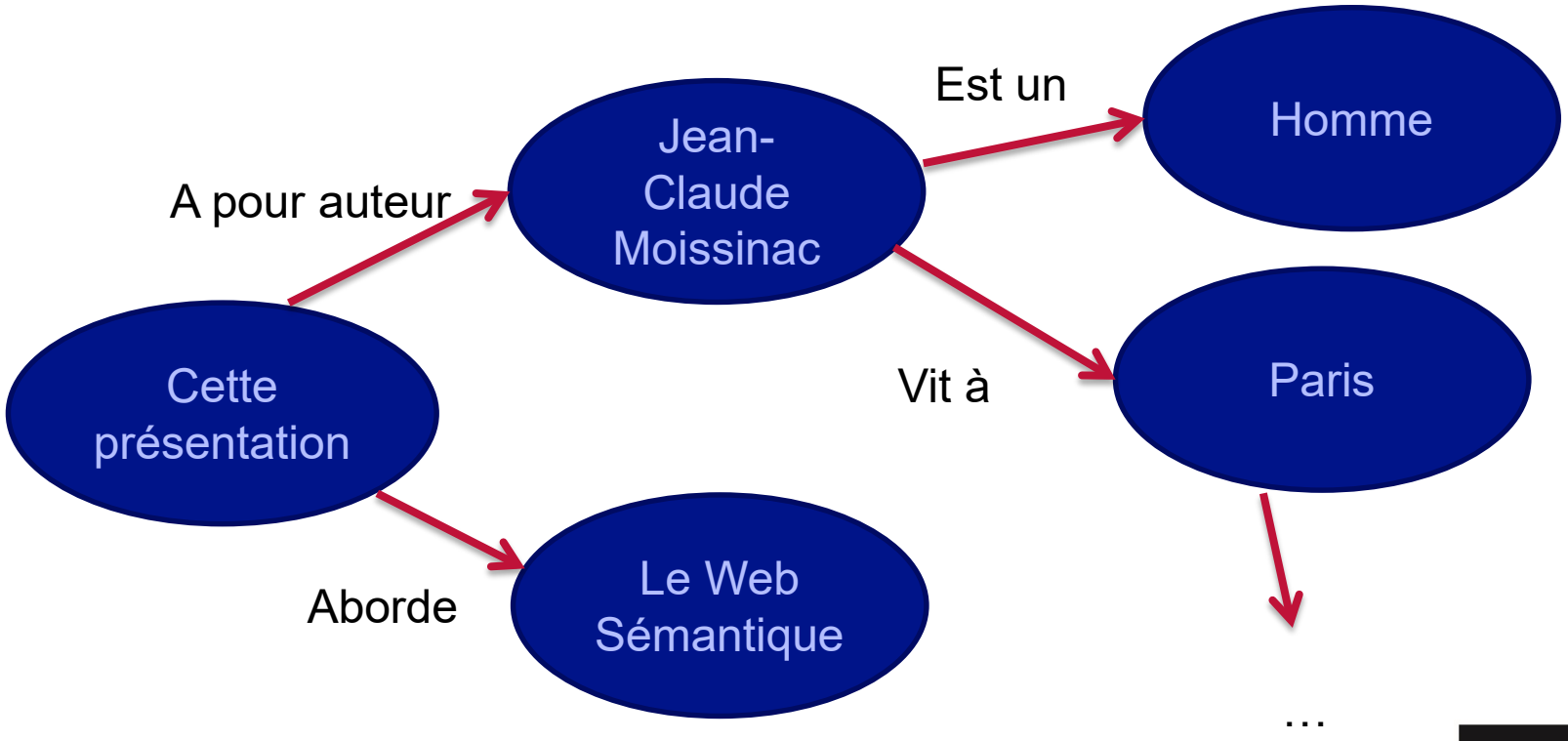
- **Resource**
  - Pages, images, vidéo, données…
  - Accessibles par une URI (ex: http://monsite.fr/...)
- **Description**
  - Propriétés et relations de la ressource
- **Framework**
  - Modèle (simple), langage, syntaxes pour ces descriptions

RDF, le modèle

- Décrire tout ce qu'on peut par des triplets
- (sujet, prédicat, objet)
- Cette présentation a pour auteur Jean-Claude Moissinac et aborde le Web Sémantique
- (cette présentation, a pour auteur, Jean-Claude Moissinac)
- (cette présentation, aborde, le Web Sémantique)

RDF définit des graphes

- **Un ensemble de triplets RDF peut être vu comme un graphe orienté et étiqueté**



Utilisation d'URIs

- Les URIs sont uniques par construction
- Si deux entités (machines, personnes…) utilisent des URIs différentes, il se peut qu'elles traitent de la même chose
- Si deux entités (machines, personnes…) utilisent une même URI, il est sûr qu'elles traitent de la même chose

Construction d'URIs – modèle des URLs

- **Modèle des URLs**
- **<protocole>:<domaine>/<chemin d'identification>**

- **En pratique, pour les données liées:**
- **http://monsupersite.com/data/geo/Paris**

- **Protocole http**
- **Domaine possédé par un propriétaire de nom**
- **Chemin désignant de façon unique un concept**

(d'autres modèles d'URIs existent)

Construction d'URIs – modèle des URLs (2)

- **Le chemin désignant de façon unique un concept**
  - Peut être totalement abstrait
  - Peut ne pas amener à une page web

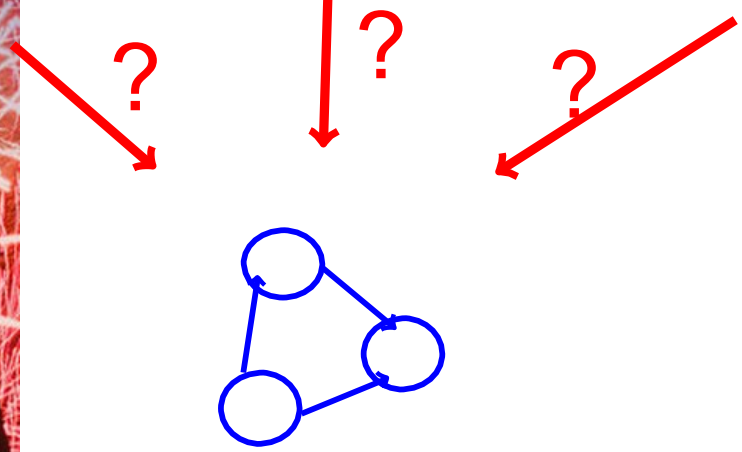
- http://www.geonames.org/2988507/
- http://fr.dbpedia.org/resource/Paris
- http://dbpedia.org/resource/Paris
- http://yago-knowledge.org/resource/Paris

- **Mais les recommandations (Linked Data)**
  - Indiquent notamment comment ramener à une page web (demo DBPedia)

RDF dans les pages Web

Exploiter du RDF dans des pages Web?

Basic Specifications	
Resolution:	8.00 Megapixels
Sensor size:	1/2.5"
Lens:	5.00x zoom (35-175mm eq.)
Viewfinder:	LCD
ISO:	80-3200
Shutter:	2-1/1000
Max Aperture:	3.5
Dimensions:	3.6 x 2.3 x 0.9 in. (92 x 59 x 22 mm)
Weight:	6.1 oz (172 g) includes batteries
MSRP:	$400
Availability:	03/2007



Homepage

Gerhard Weikum

Max-Planck-Institut für Informatik
Department 5: Databases and Information Systems
Building E1.4, Room 402
Campus E1.4
66123 Saarbrücken
Germany

Email: weikum@mpi-inf.mpg.de
Phone: +49 681 9325 500
Fax: +49 681 9325 599

RDFa

RDFa est une syntaxe pour annoter des pages HTML avec du RDF
https://rdfa.info/

** Jean Mois
**

**Chercheur en dessin animé 1957-
**

Roubaix, Nord

** **

RDFa Lite

Définir le vocabulaire

Localement, tous les termes associés à un noeud HTML vont venir du vocabulaire défini dans 'vocab'.

Jean Mois

Chercheur en dessin animé 1957-

Roubaix, Nord

Définir le sujet

Toutes les propriétés associées au nœud HTML ont pour sujet l'entité désignée dans 'resource'.

```
<div vocab="http://schema.org/"
resource="http://moissinac.wp.mines-telecom.fr/">
Jean Mois<br>
Chercheur en dessin animé 1957-<br>
Roubaix, Nord
</div>
```

Définir un type

Le type du sujet est donné par ’typeOf’.

<div vocab=“http://schema.org/“
resource=“http://moi... “ typeOf=“Person>

Jean Mois

Chercheur en dessin animé 1957-

Roubaix, Nord

Triplet

<http://moissinac...> rdf:type <http://schema.org/Person> .

Définir un fait avec une valeur

Un tag avec 'property' défini un fait sur le sujet courant; la valeur associée est celle du atg

```
<div vocab="http://schema.org/"
resource="http://moi..." typeOf="Person">
<span property="name">Jean Mois</span><br>
Chercheur en dessin aimé 1957-<br>
Roubaix, Nord
</div>
```

Triplet

<http://moi...> <http://schema.org/name> "Jean Mois" .

Solutions alternatives à RDFa

Standards similaires à RDFa:

- Microdata
- Json-ld
  - Désormais recommandé par Google

Demo

- **http://www.w3.org/2012/pyRdfa/Validator.html**
- **https://rdfa.info/play**
  - http://givingsense.eu/foaf/moissinacRdfa.htm
- **https://developers.google.com/structured-data/testing-tool/**

Marquage utilisé par les moteurs de recherche

Demo

Aller sur http://www.trustedreviews.com/lenovo-p2-review

Copier ce lien dans https://search.google.com/structured-data/testing-tool

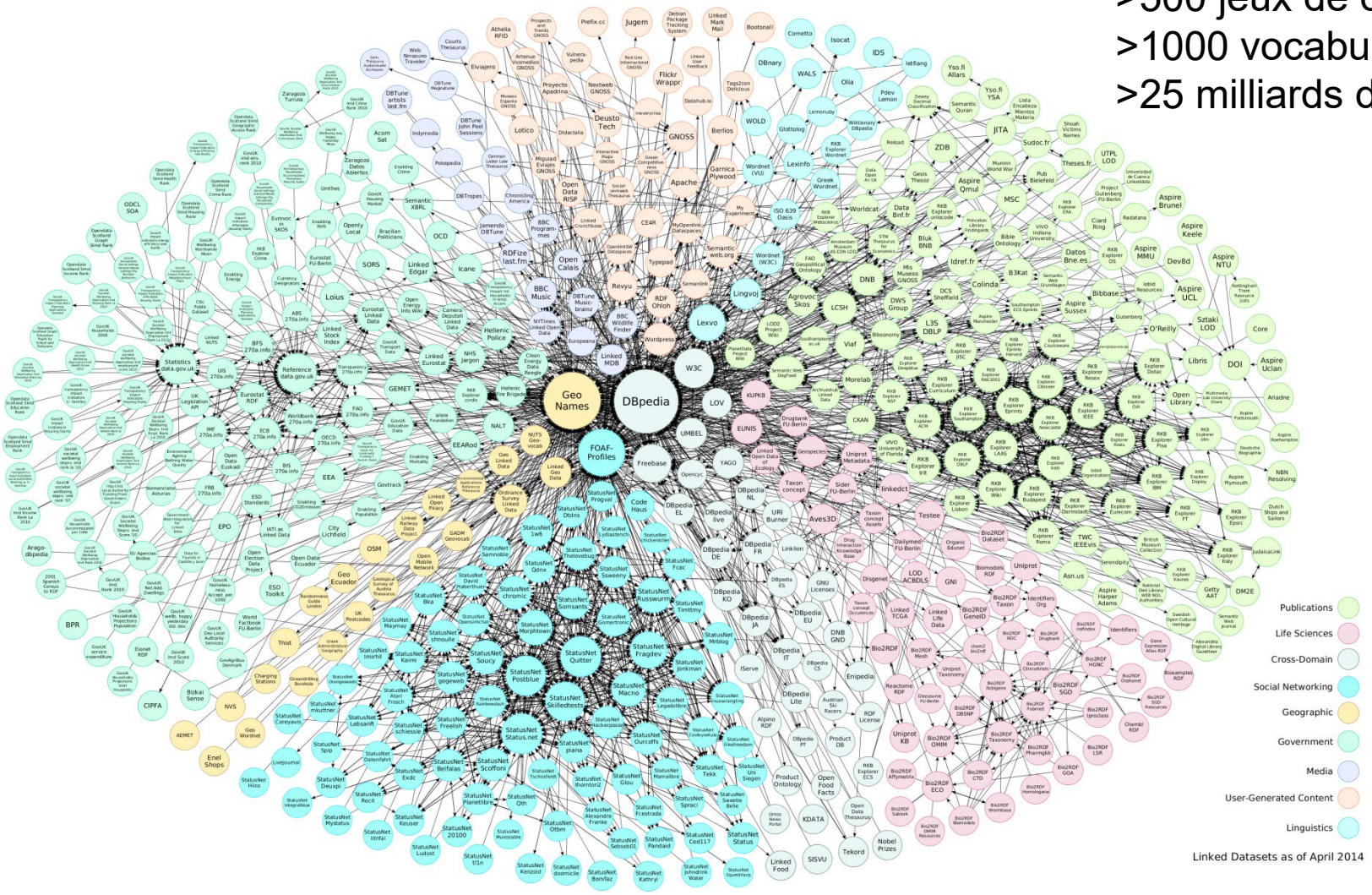
JSON-LD: exemple

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Organization",
  "url": "http://www.your-company-site.com",
  "contactPoint": [{
    "@type": "ContactPoint",
    "telephone": "+1-401-555-1212",
    "contactType": "customer service"
  }]
}
</script>
```

Web des données

Linked Open data

>500 jeux de données
>1000 vocabulaires
>25 milliards de triplets



TELECOM ParisTech

Bases de connaissances extraites du Web

- DBPedia
- Yago
- Wikidata

DBPedia

- Initiative pour tirer une représentation sémantique du contenu de Wikipedia
- Défini des gabarits d'extraction de faits (triplets sujet, prédicat, objet) à partir de portions de page de DBPedia
- L'extraction est automatique sur la base de ces gabarits

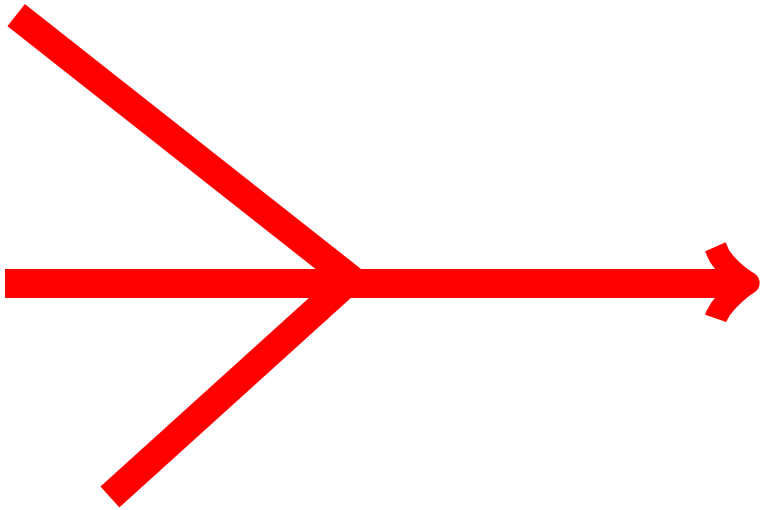
- Demo

YAGO

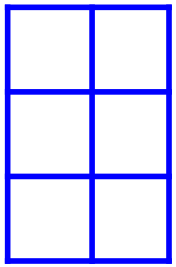
Le projet YAGO project extrait des informations de Wikipedia et d'autres sources.

WordNet

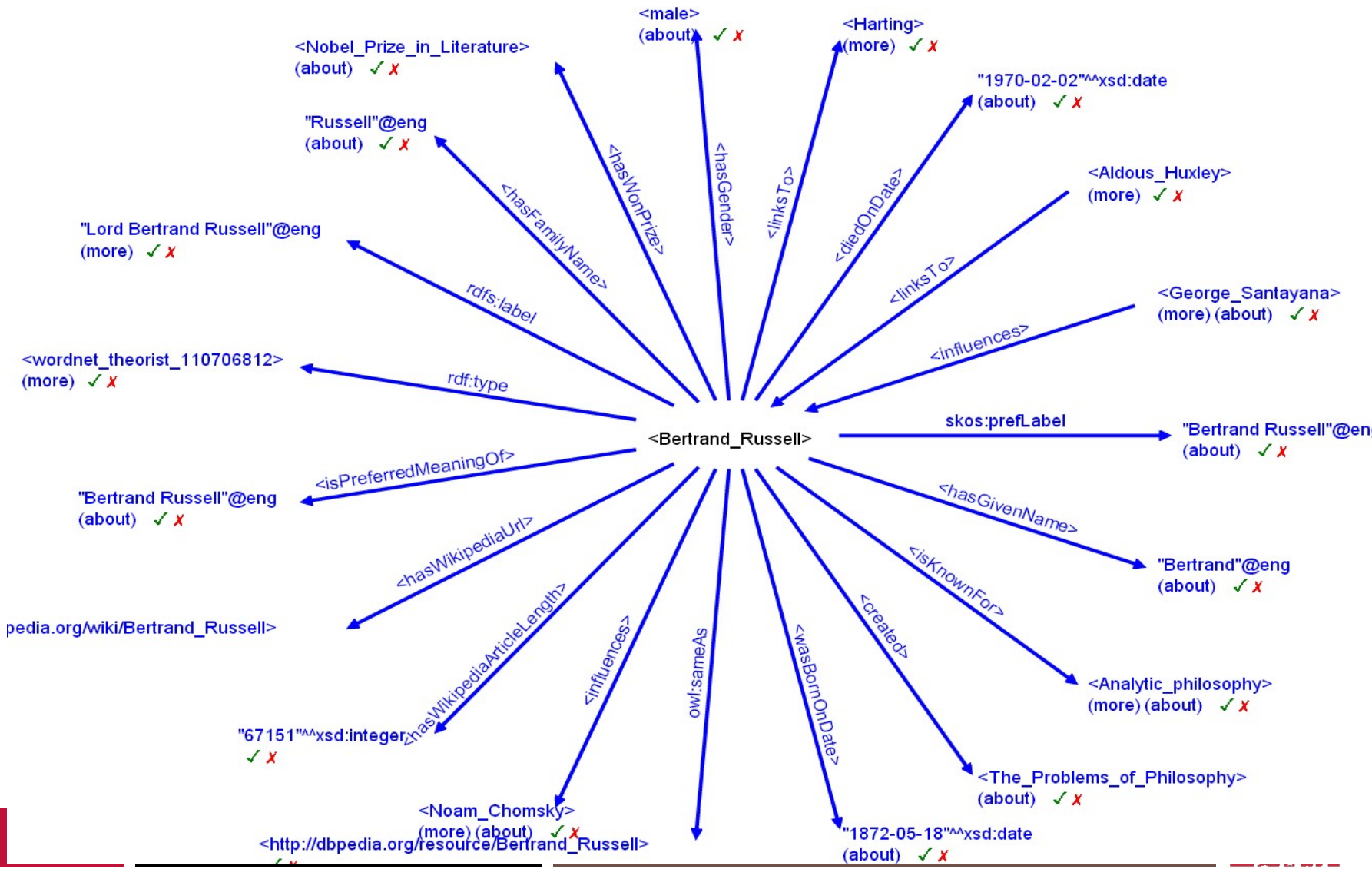
GeoNames



yago
select knowledge



YAGO


<Bertrand_Russell>
<male> (about) ✓ ✗
<Harting> (more) ✓ ✗
"1970-02-02"^^xsd:date (about) ✓ ✗
<Aldous_Huxley> (more) ✓ ✗
<George_Santayana> (more) (about) ✓ ✗
"Bertrand Russell"@eng (about) ✓ ✗
"Bertrand"@eng (about) ✓ ✗
<Analytic_philosophy> (more) (about) ✓ ✗
<The_Problems_of_Philosophy> (about) ✓ ✗
"1872-05-18"^^xsd:date (about) ✓ ✗
<Noam_Chomsky> (more) (about) ✓ ✗
<http://dbpedia.org/resource/Bertrand_Russell>
"67151"^^xsd:integer ✓ ✗
pedia.org/wiki/Bertrand_Russell>
"Bertrand Russell"@eng (about) ✓ ✗
<wordnet_theorist_110706812> (more) ✓ ✗
"Lord Bertrand Russell"@eng (more) ✓ ✗
"Russell"@eng (about) ✓ ✗
<Nobel_Prize_in_Literature> (about) ✓ ✗
skos:prefLabel
<hasGivenName>
<isKnownFor>
<created>
<wasBornOnDate>
owl:sameAs
<influences>
<hasWikipediaArticleLength>
<hasWikipediaUrl>
<isPreferredMeaningOf>
rdf:type
rdfs:label
<hasFamilyName>
<hasWonPrize>
<hasGender>
<linksTo>
<diedOnDate>
<linksTo>
<influences>

Vocabulaires généraux

- **Rdf**
  - Rdf:type
- **Rdfs**
  - *rdfs:subClassOf, rdfs:property, rdfs:domain, rdfs:range*
- **Dublin Core**
  - xmlns:dc=http://purl.org/dc/elements/1.1/
  - **dc:title … description de documents**
- **(Dolce)**
- **Geo84**
  - Geo:lat, geo:lon
- **Foaf**
  - **foaf:Person -> foaf:name**
- **…**

Autres vocabulaires

- Schema.org (for Web content)

http://schema.org

- Creative Commons (types of licences)

http://creativecommons.org/ns#

- Facebook Open Graph (for Web content)

http://ogp.me/

Exemple: Dublin Core

Dublin Core pour décrire des documents.

dc:creator, dc:title, dc:format, dc:MediaType, dc:language...

see this KB

dc:creator

“Elvis: An auto-biography”

dc:title

dc:description “All about my life&wife”

dc:issued “1980”

Exemple FOAF

- **Vocabulaire pour décrire une personne et ses relations avec d'autres; en format RDF/XML**
- **<rdf:RDF …>**
  - <foaf:Person> <foaf:name>Jimmy Wales</foaf:name> <foaf:title>Mr.</foaf:title> <foaf:givenName>Jimmy</foaf:givenName>
  - <foaf:familyName>Wales</foaf:familyName>
  - <foaf:mbox rdf:resource="mailto:jwales@bomis.com"/>
  - <foaf:homepage rdf:resource="http://www.jimmywales.com/"/> <foaf:nick>Jimbo</foaf:nick> <foaf:depiction rdf:resource="http://www.jimmywales.com/aus_img_small.jpg"/>
  - <foaf:interest> <rdf:Description rdf:about="http://www.wikimedia.org" rdfs:label="Wikipedia"/> </foaf:interest>
  - <foaf:publications rdf:resource="http://www.jimmywales.com/pubs/publications.rdf"/> …
  - <foaf:knows>
    - <foaf:Person> <foaf:name>Angela Beesley</foaf:name></foaf:Person>
  - </foaf:knows>
  - <foaf:knows>
    - <foaf:Person rdf:about="http://jimmycricket.com/me"> <foaf:name>Jimmy Criket</foaf:name> </foaf:Person>

    </foaf:knows>
  - </foaf:Person>

</rdf:RDF>

Trouver un vocabulaire

- Lov
- http://lov.okfn.org/dataset/lov/

- Demo

Des outils pour trouver des vocabulaires et ensembles de données

- http://datahub.io/
- http://lov.okfn.org/dataset/lov/
- http://prefix.cc/
- http://data.gouv.fr

Des jeux de données de référence

Généralement associés à un vocabulaire, éventuellement défini par une ontologie

- **Dbpedia**
- **Geonames**
- **Bnf**
- **Europeana**
- **DBLP**
- **BBC**
- **British Museum**
- **Library of Congress**
- **Fondation Getty**
- **…**

Open Data

Rapports avec l'OpenData

- **Open Data**
  - Mouvement international qui tend à rendre disponibles publiquement les données produites sur fonds publics
  - S'étend à une tendance à rendre des données utilisables publiquement sur le Web

- **Open Data n'implique pas Web Sémantique et Données Liées (Linked Data), mais le permet**
  - Ex: publications de données au format CSV

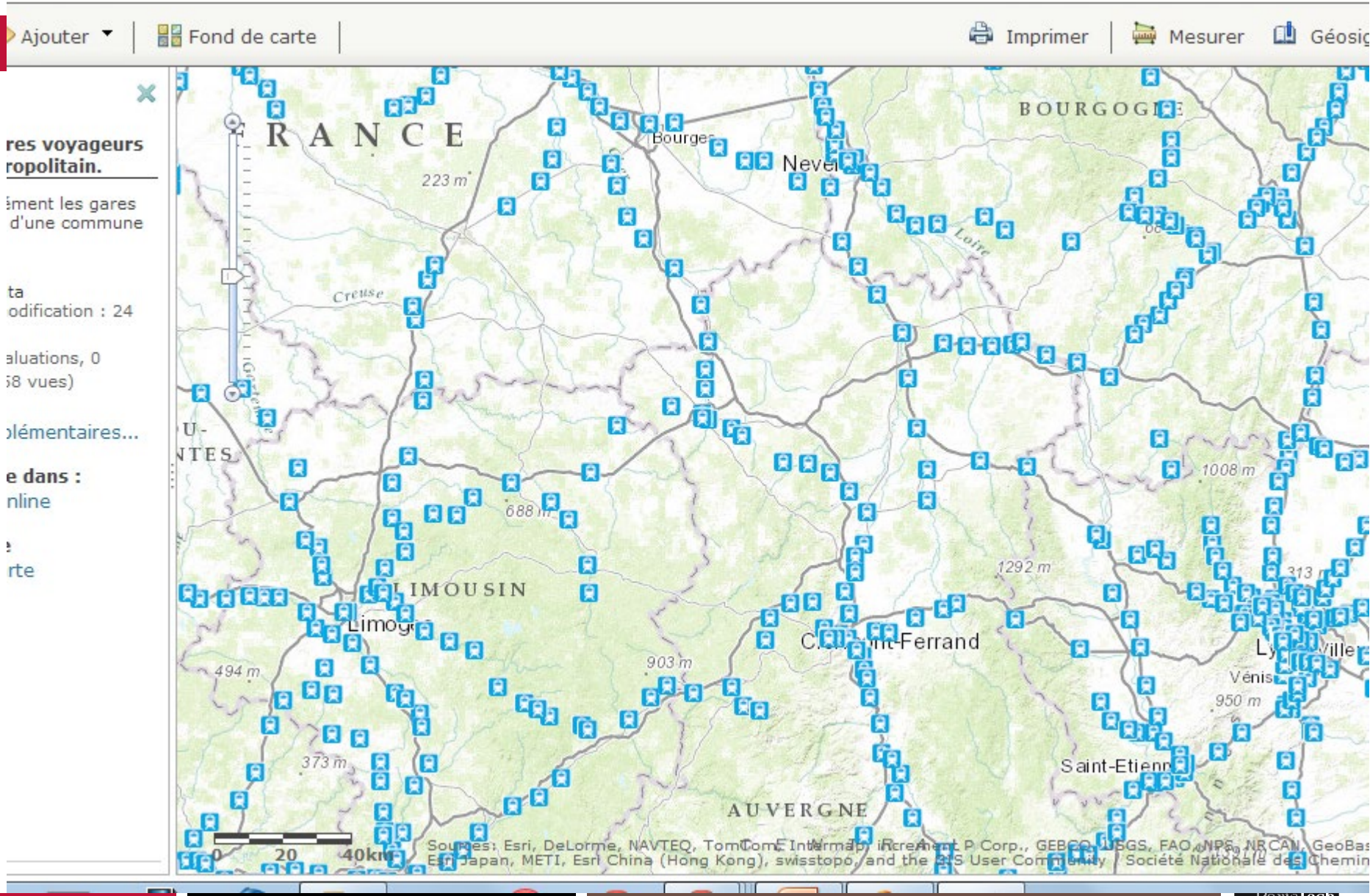
Linked Open Data Project

- US census data
- BBC music database
- Gene ontologies
- DBpedia general knowledge, + YAGO, + Cyc etc.
- UK government data
- geographical data in abundance
- national library catalogs (USA, Germany etc.)
- publications (DBLP)
- ...and many more

Data.gouv.fr

- Presque toutes les réutilisations mises en avant sont des cartes
- Exemples de données
- Exemples de cartes: cf OpenGeoData.fr

Ensemble des gares voyageurs du territoire métropolitain.

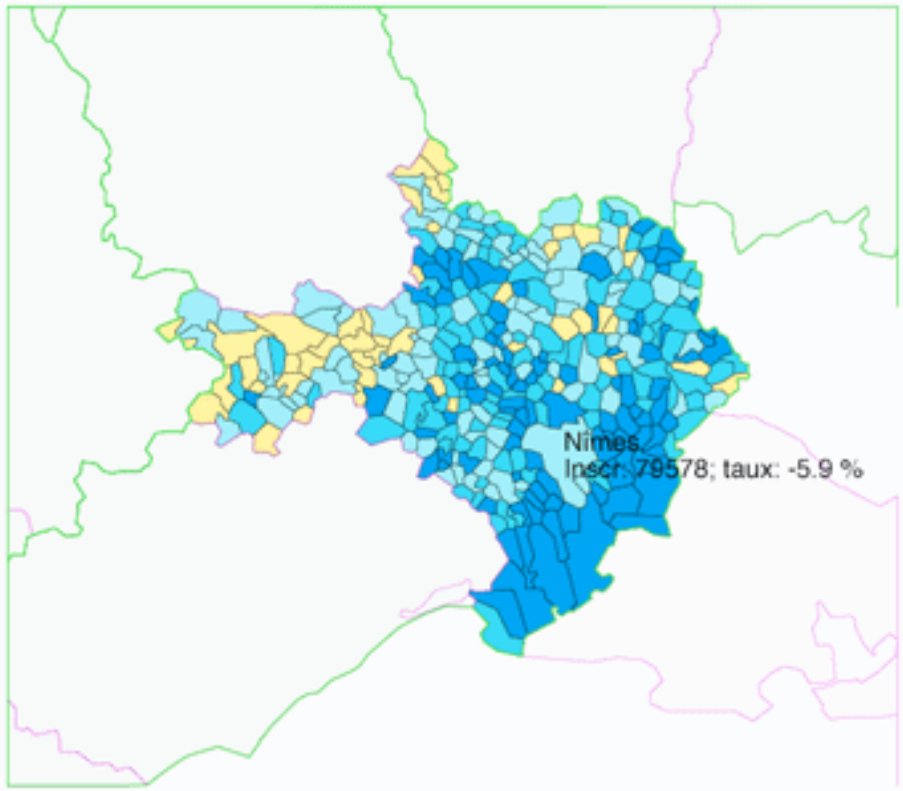


Source: http://www.opengeodata.fr/

France 2002 : premier tour des élections présidentielles

Si votre souris survole une commune, son nom et ses résultats sont affichés. Zooms et déplacements possibles! (avantages du SVG).

Retour à la documentation



Gard

Variables décrites en %:
+ Chirac
- Le Pen

Éric Guichard
ENSSIB et ENS

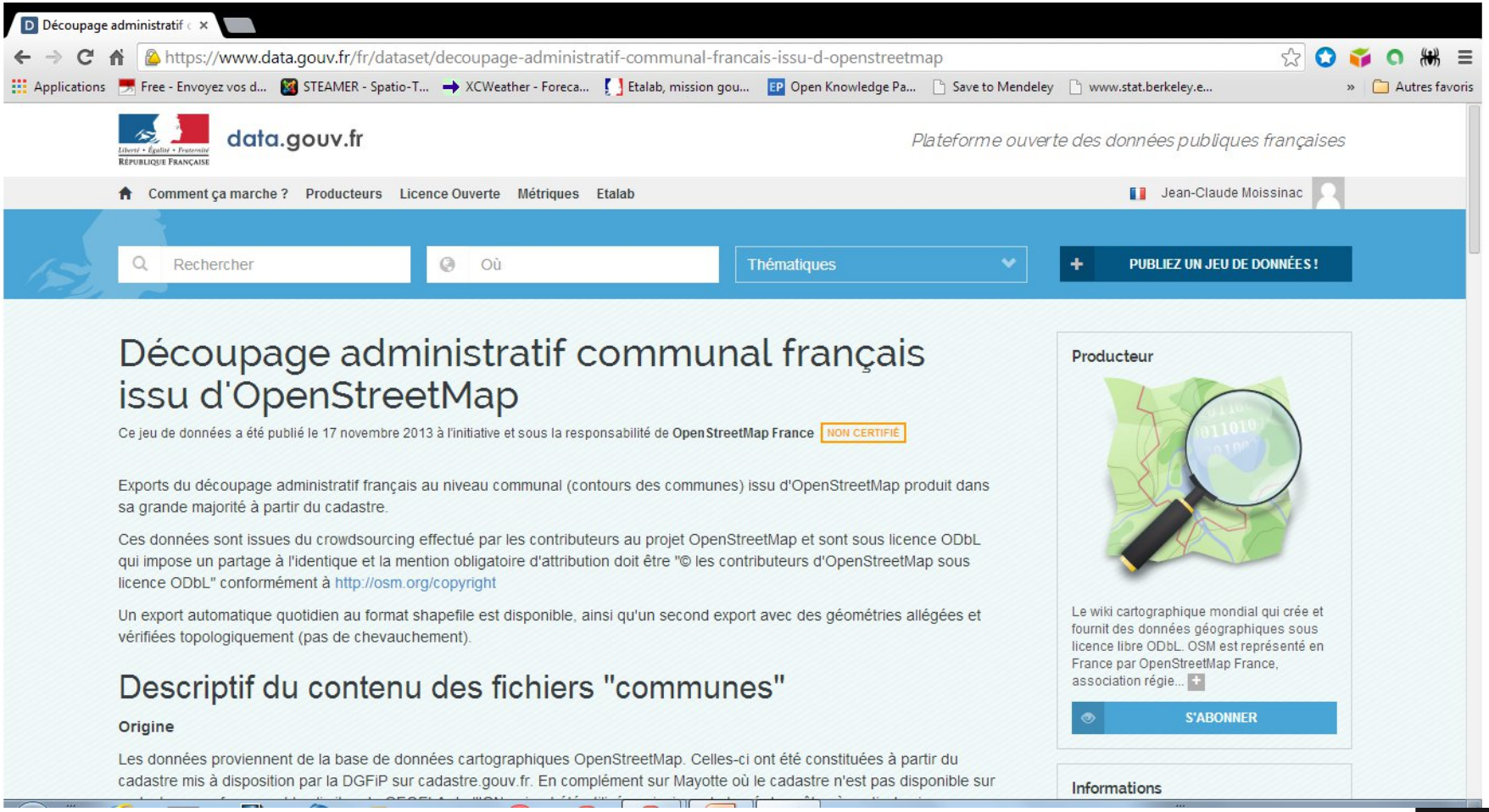
-12.8 -7.5 -0.3 19.6

-33.8 -12.7 -7.4 0.0

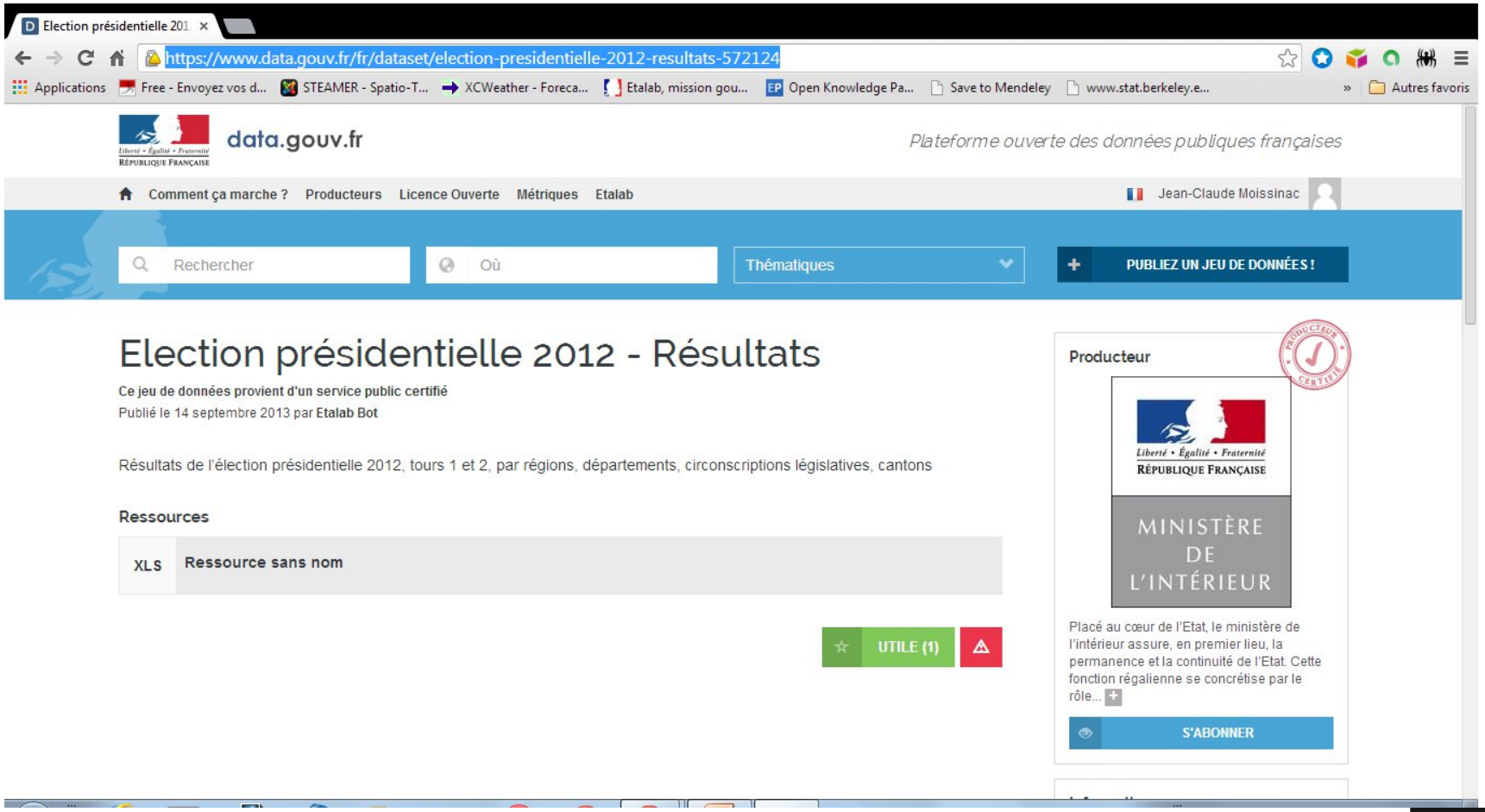
Méthode: Discrétisation uniforme

Auteur: Éric Guichard. Logiciel Érathostène, écrit en Perl (2000 pour la version ps, 2004 pour la version svg).
Sources: Ministère de l'intérieur, CD-Atlas de France (revu et largement corrigé), Éric Guichard.
Remerciements : J. Beigbeder, A. Danzart, J.-C. Moissinac, C. Potier, H. Théry.

Les contours des communes



Les résultats



Conclusion

Quelques points importants

- **Grandes variétés de protocoles, langages, technologies utilisées sur le Web**
- **Crawler, c'est parcourir un graphe**
- **Construire un crawler est une tâche non triviale d'ingénierie**
  - en particulier si on veut crawler à grande échelle

Licence de droits d'usage

Contexte public } avec modifications

Par le téléchargement ou la consultation de ce document, l'utilisateur accepte la licence d'utilisation qui y est attachée, telle que détaillée dans les dispositions suivantes, et s'engage à la respecter intégralement.

La licence confère à l'utilisateur un droit d'usage sur le document consulté ou téléchargé, totalement ou en partie, dans les conditions définies ci-après et à l'exclusion expresse de toute utilisation commerciale.
Le droit d'usage défini par la licence autorise un usage à destination de tout public qui comprend :
–le droit de reproduire tout ou partie du document sur support informatique ou papier,
–le droit de diffuser tout ou partie du document au public sur support papier ou informatique, y compris par la mise à la disposition du public sur un réseau numérique,
–le droit de modifier la forme ou la présentation du document,
–le droit d'intégrer tout ou partie du document dans un document composite et de le diffuser dans ce nouveau document, à condition que :
– L'auteur soit informé.

Les mentions relatives à la source du document et/ou à son auteur doivent être conservées dans leur intégralité. Le droit d'usage défini par la licence est

personnel et non exclusif.

Tout autre usage que ceux prévus par la licence est soumis à autorisation préalable et expresse de l'auteur : sitepedago@telecom-paristech.fr