The Semantic Web and Related Challenges

Antoine Amarilli

École normale supérieure, Département d’informatique
The Web

The Web is the largest public source of information that can be accessed by programs.
The Web is decentralized:

- Information and websites are not trustworthy.
- Distributed, globally fault-tolerant but information can disappear.
- You have to crawl it (no dumps!).
The Web is unstructured:

- Standards but they are disobeyed.
- Lots of natural language text.
- Only hints of structure from the markup (e.g., tables).
How does the Web compare to relational databases as a way to store information?

- Relational databases are *structured*.
- They support *expressive queries*.
- You have to choose a *schema* beforehand, and stick to it.
- They are not designed to be *integrated* with one another.

⇒ Could we have the best of both worlds? What could we use it for?
The Semantic Web (example)

The Semantic Web: RDF

- Identify items by a **URI** (which may be a **URL**).
- **Triples** between three URIs: subject, predicate, object.
- **Federated**: the URIs can be managed by independent organizations.
- **Literal** values (with language and datatype annotations).
- Several **representations**: bulk text or XML, relational databases, triple stores, SPARQL endpoints, implicit graph representation.
- Structure is **optional**: nothing, a simple class taxonomy, or full-fledged constraint languages.

→ What do we want to do with it?
Get answers, not results

Google search for "author of hamlet"

About 22,300,000 results (0.33 seconds)

William Shakespeare
Hamlet, Author

Google search for "date of birth of shakespeare"

About 8,670,000 results (0.31 seconds)

1564
William Shakespeare, Date of birth
Get answers, not results

→ No support for complex queries!
You probably thought Wolfram Alpha was better?
Another example

Say I want to find out about the works which have won both the Hugo award and the Nebula award. Fortunately, someone materialized the view for me:

[List of joint winners of the Hugo and Nebula awards - Wikipedia, the ...](en.wikipedia.org/.../List_of_joint_winners_of_the_Hugo_and_Nebula...)

This is a list of the works that have won both the Hugo Award and the Nebula Award, awarded annually to works of science fiction literature. The Hugo Awards ...

What if I want to restrict to the works written by a female author? Why can’t I write something like:

Google

```
```

If we were dealing with a database, it would make no sense to materialize explicitly all “useful” views by hand!
Other examples

- **Aggregate** relevant information, don’t centralize it (e.g., reviews).
- **Visualize** heterogeneous information (e.g., on a map).
- Today: popular services approximate this with home-grown, domain-specific, incompatible **APIs**.
Many independent information sources.

Links between these sources.
→ Linked Data Cloud.
École normale supérieure - Paris
www.ens.fr/ - Translate this page
Établissement public d'enseignement supérieur et de recherche pour les études prédoctorales et doctorales en sciences, lettres, sciences humaines et (...)
Concours Lettres - Entrer à l'ENS - Concours Sciences - Débouchés et carrières

English - École normale supérieure - Paris
www.ens.fr/?lang=en
Formations prédoctorales - Étudiants internationaux - Contacts and Maps - DMA

ENS.A.D.
www.ensad.fr/ - Translate this page
Sous l'égide de la Fondation Paris Sciences Lettres, SACRe est le fruit d'une collaboration entre l'École normale supérieure et les Écoles d'art et de ...
Score: 23 / 30 - 13 Google reviews

École Normale Supérieure - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/École_Normale_Supérieure
The École normale supérieure (French pronunciation: [eukɔl nɔʁmal syperyœʁ]; also known as Normale sup', Normale, and ENS) is a French grande école ...
Overview - Influence abroad - Free online content - Alumni and faculty

ENS de Lyon
www.ens-lyon.eu/ - Translate this page
ENS de Lyon : offre de formation, concours d'entrée, recherche, international, diffusion
Sig.ma - Live views on Web(s) of Data

Sig.ma does on the fly, interactive information visualization with bits coming from up to hundreds of sources at the same time. Sig.ma pages have permalinks and can be embedded in web pages.

Search on out live Sig.ma installation:

- venus

Use:
- Sindice
- OKKAM
- YBoss
- Lod Sparql Endpoint
- Your own data

Examples: Tim Berners Lee, Barack Obama, Michael Jackson

Use it online or Download Sig.ma EE
Where does the data come from?

1. Web pages with semantic markup.
2. Existing databases published on the Web.
3. Structured content extracted from Web pages.

→ The last option is the most interesting one!
Add semantic markup to Web pages

Here is an example from schema.org (Google, Bing, Yahoo!).

```html
<div itemscope itemtype="http://schema.org/Person">
  <span itemprop="name">Jane Doe</span>
  <img src="janedoe.jpg" itemprop="image" />

  <span itemprop="jobTitle">Professor</span>
  <div itemprop="address" itemscope itemtype="http://schema.org/PostalAddress">
    <span itemprop="streetAddress">
      20341 Whitworth Institute
      405 N. Whitworth
    </span>
    <span itemprop="addressLocality">Seattle</span>,
    <span itemprop="addressRegion">WA</span>
    <span itemprop="postalCode">98052</span>
  </div>
  <span itemprop="telephone">(425) 123-4567</span>
  <a href="mailto:janedoe@xyz.edu" itemprop="email">
    jane-doe@xyz.edu
  </a>
</div>
```

People are reluctant to use that, though.
Publish existing databases

- We can express relational databases in RDF.
- However, we have to align them with other data sources.
- We must find out instances which already exist in the other data sources.
- We must reuse the predicates used by the other data sources.
- Several challenges:
  - Literal ambiguity: “Titanic”
  - Variants: “Charles Brackett” vs “Charles William Brackett”
  - Complex datatypes: “1952-03-11” vs “1952”
  - Structure: “birthDate” vs “event” and “date”
Ontology alignment examples

imdb:tt0046435 → imdb:label → 'Titanic'
dbp:Titanic_(1953_film) → foaf:name → 'Titanic'
dbp:Charles_Brackett → foaf:name → 'Charles William Brackett'
imdb:p138992 → imdb:producerOf → imdb:tt0046435


dbp:Douglas_Adams → dbp:birthDate → '1952-03-11'

bio:date → '1952'
rdf:type → bio:Birth
Information extraction: Wikipedia

Wikipedia, a centralized island in the decentralized Web.

- No need to crawl: use **dumps**.
- No **copyright** problems.
- Essentially **factual** information.
- More **trustworthy**.
- Hints of **structure**: categories, infoboxes, consistent conventions, etc.


Hearst patterns in natural language text

The players had undergone, it seems, a “transference of emotion,” Dr. Pepping and his colleagues wrote. Emotions such as happiness and confidence are known to be contagious, with one person’s excitement sparking rolling biochemical reactions in onlookers’ brains.

Instance

philadelphia_76ers is a sports team
joe_mcdonald is an athlete
alison_wearing is a monarch
h_tel_emory_conference_center_hotel is a hotel
paul_johansson is a celebrity
avenal is a city located in the state or province california

Information extraction: the Deep Web

A lot of Web information is contained in result pages produced from structured back-ends and hidden behind forms.

Recherche

Veuillez entrer un ou plusieurs mots-clé. Cela peut être :

- le nom, et/ou le prénom d'une personne ;
- la fonction d'une personne ;
- un numéro de téléphone, de fax ;
- une adresse électronique ;
- un département, un service, un laboratoire.

Mots-clé :isabelle  
ok

[Diadem Project, University of Oxford.  
http://diadem.cs.ox.ac.uk/]
Information extraction: the Deep Web

- ANDRE Isabelle Charge de communication
- AUJARD Isabelle
- BARBOSA Isabelle Technicien
- BELLANGER Isabelle Developpeur
- BORG Isabelle Assistante
- BRUNET Isabelle Technicien
- CHARNAVEL Isabelle Doctorant
- CHORT Isabelle Allocataire de recherche
- CREPY Isabelle Bibliothecaire adjoint
- DAJOZ Isabelle Enseignant-chercheur
- DAUTRICHE Isabelle Doctorant
- DE VENDEUVRE Isabelle Directeur des relations internationales
- DELAIS Isabelle Secrétaire
- DERIS Isabelle Pilotage et controle de gestion
- DUHA Isabelle Professeur des Universites
- GOUARNE Isabelle Post-Doctorant
- HAVELANGE Isabelle Ingenieur de recherche
- JOUANNEAU Isabelle Coordinatrice
- KALINOWSKI Isabelle Chercheur
- LAVALEIX Isabelle Responsable de la gestion financiere
- LELIEVRE Isabelle Secrétaire
- LIN Isabelle Étudiant
- MISTRAL Isabelle
- MOTTAR Isabelle Doctorant ENS Doctorant
- PANTIN Isabelle Directeur du departement LILA
- PERRAS Claire-Isabelle
- PIMOUGUET-PEDARROS Isabelle Enseignant-chercheur
- PORTE Isabelle Responsable logistique des sites de Jourdan et Montrouge et du pole administratif
- VERITE Isabelle Ingenieur de recherche
Information extraction: the Deep Web

Isabelle DELAIS

Fonctions :
- Secrétaire

Affectations :
- Département d'informatique • UMR 8548 Laboratoire d'informatique de l'ENS (LIENS)

Adresse électronique : isabelle.deilais@ens.fr

Adresses postales :

<table>
<thead>
<tr>
<th>Adresse</th>
<th>Étage/Bureau</th>
<th>Téléphone</th>
<th>Fax</th>
</tr>
</thead>
<tbody>
<tr>
<td>45, rue d'Ulm 75230 Paris cedex 05</td>
<td>Aile Rataud RDC, bureau B10</td>
<td>20 45</td>
<td>20 75</td>
</tr>
</tbody>
</table>
Information extraction: the Deep Web

All these result pages will have a similar DOM structure.

The variable (red) parts are what we’re interested in.
Entity disambiguation

Disambiguation (disambiguation)

This disambiguation page lists articles associated with the same title. If an internal link led you here, you may wish to change the link to point directly to the intended article.

Word-sense disambiguation is the process of identifying the sense of a word in a sentence.

- Wikipedia provides a mapping between names and entities.
- Use several assumptions:
  - Prominence of entities: Paris, France vs Paris, Texas.
  - Context similarity: Venus the planet vs Venus the goddess.
  - Coherence between the assignments: Mars and Venus.
  https://d5gate.ag5.mpi-sb.mpg.de/webaida/
Web sources with conflicting facts.

Use several assumptions:
- A trustworthy source provides many correct facts.
- A correct fact is provided by many trustworthy sources.

Probabilistic model, fixpoint computation.


B. Zhao, B. I. P. Rubinstein, J. Gemmell, J. Han. *A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration*. VLDB’12.
The two previous steps lead to **uncertainty**.

Different ways to **model** this uncertainty:

**Probabilistic databases**

<table>
<thead>
<tr>
<th>container</th>
<th>containee</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>aquarium</td>
<td>45</td>
<td>0.97</td>
</tr>
<tr>
<td>salle_r</td>
<td>45</td>
<td>0.95</td>
</tr>
<tr>
<td>mc203</td>
<td>45</td>
<td>0.02</td>
</tr>
</tbody>
</table>

**Probabilistic XML**

- Interesting **problems**: optimal representation, expressiveness, running time, etc.
Access patterns

- Consider the usual relational algebra.
- Predicates can only be used through access patterns.
- We need to answer a query.
- Is a certain access relevant to the query?
- “Long-term relevance of dependent accesses for conjunctive queries is NEXPTIME-complete.”
- Related to query containment (which is linked in turn to finite model theory, tree automata, etc.).

“Find two pupils with two classes in common.”

\textbf{IsIn}(pupil, class)

<table>
<thead>
<tr>
<th>Pupil:</th>
<th>fechant</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>pupil</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>fechant</td>
<td>physique3</td>
</tr>
<tr>
<td>fechant</td>
<td>physique42</td>
</tr>
</tbody>
</table>

\textbf{IsIn}(pupil, class)

<table>
<thead>
<tr>
<th>Class:</th>
</tr>
</thead>
<tbody>
<tr>
<td>algoprog</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pupil</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>bourgeat</td>
<td>algoprog</td>
</tr>
<tr>
<td>delpeuch</td>
<td>algoprog</td>
</tr>
<tr>
<td>forest</td>
<td>algoprog</td>
</tr>
</tbody>
</table>
An interdisciplinary field!

- **Relational databases** as a useful model.
- **Finite model theory** which is the math behind relational databases.
- **Natural language processing** for information extraction.
- **Artificial intelligence** and links with computer reasoning.
- **Information theory** and minimum description length.
- **Formal languages** and tree automata.
- **Logic** for constraint languages.
- **Web technologies** to actually implement things and benchmark them.
To find out more...

Thanks! 

Thanks for your attention! 
Questions welcome.

Image credits
Frame 2: http://www.threebrackets.com/web_develop.htm
Frame 3: http://openp2p.com/pub/a/p2p/2001/12/14/topologies_one.html
Frames 8, 9, 11, 14: http://google.com/
Frame 10: http://wolframalpha.com/
Frame 15: http://sig.ma/
Frame 18: http://schema.org/
Frames 23 to 25: http://annuaireweb.ens.fr/
Frames 27 and 28: http://en.wikipedia.org/
Frame 31: https://xkcd.com/755/