The Semantic Web and Related Challenges

Antoine Amarilli

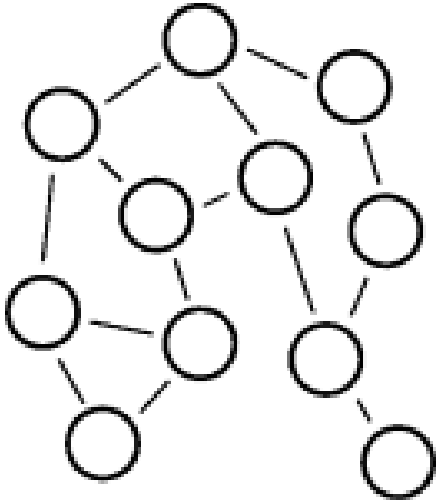
École normale supérieure, Département d'informatique

The Web

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

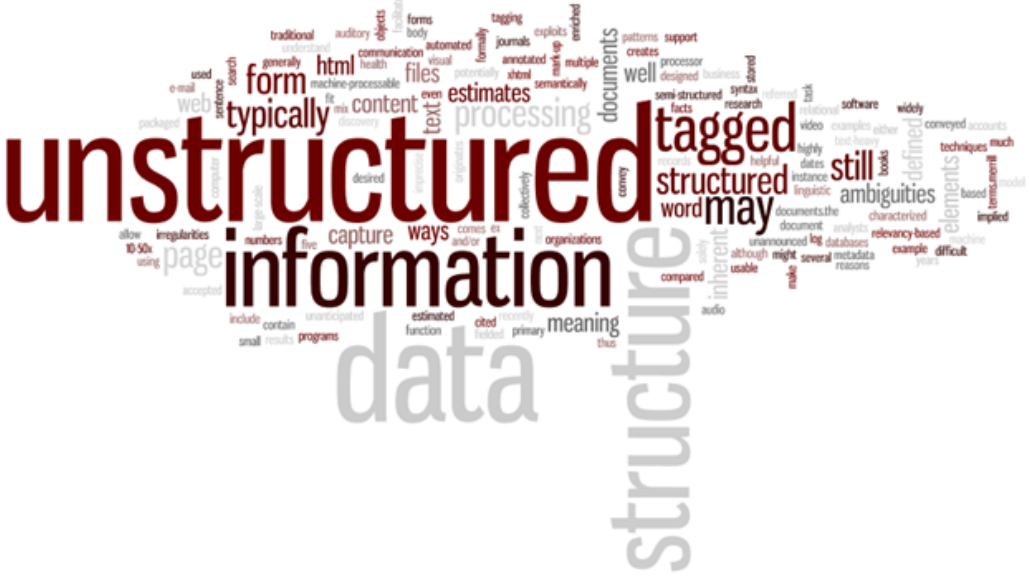
Decentralized

- 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!).



Unstructured

- The Web is unstructured:
  - Standards but they are disobeyed.
  - Lots of natural language text.
  - Only hints of structure from the markup (e.g., tables).



Relational databases

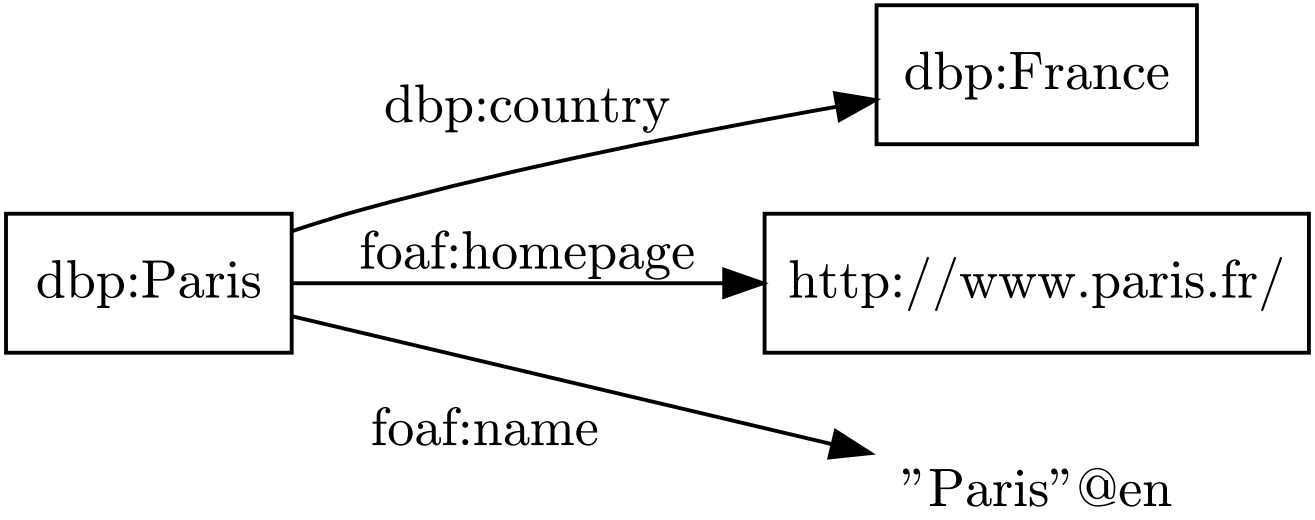
id	name	evo
1	Bulbausaur	2
2	Ivysaur	3
3	Venusaur	NULL
4	Charmander	5
5	Charmaleon	6
6	Charizard	NULL
7	Squirtle	8
8	Wartortle	9
9	Blastoise	NULL
10	Caterpie	11
11	Metapod	12
12	Butterfree	NULL
13	Weedle	14
14	Kakuna	15

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)



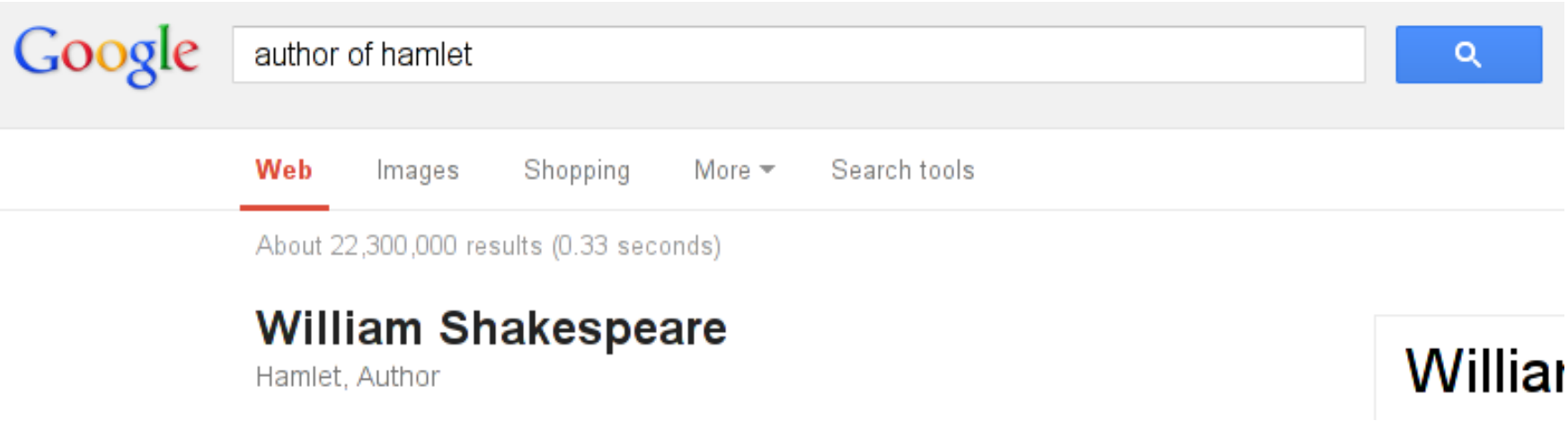

```
<dbp:Paris> <dbp:country> <dbp:France> .
<dbp:Paris> <foaf:homepage> <http://www.paris.fr/> .
<dbp:Paris> <foaf:name> "Paris"@en .
```

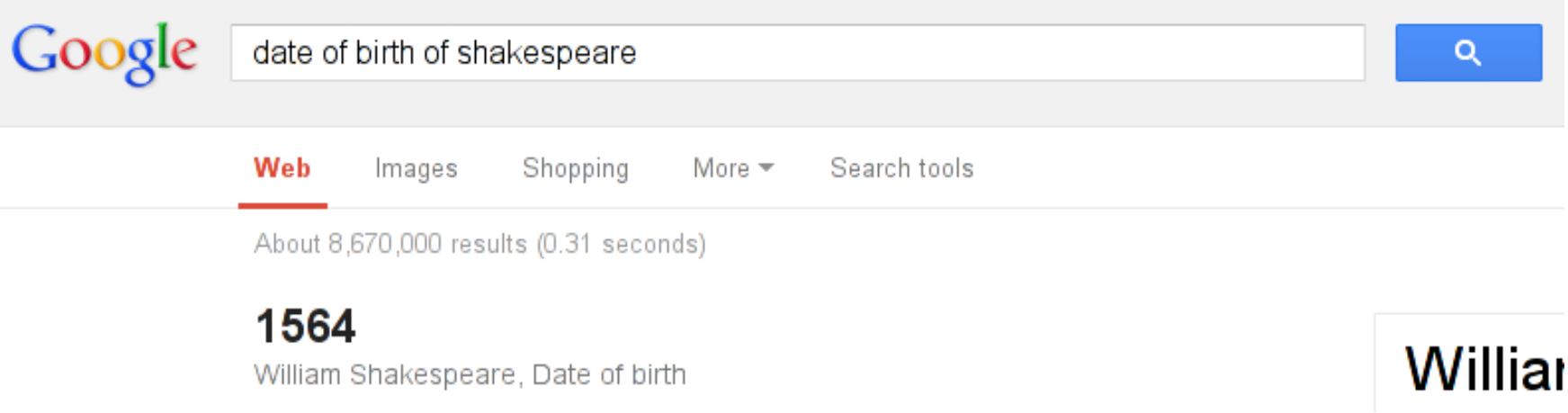
The Semantic Web: RDF

- 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.

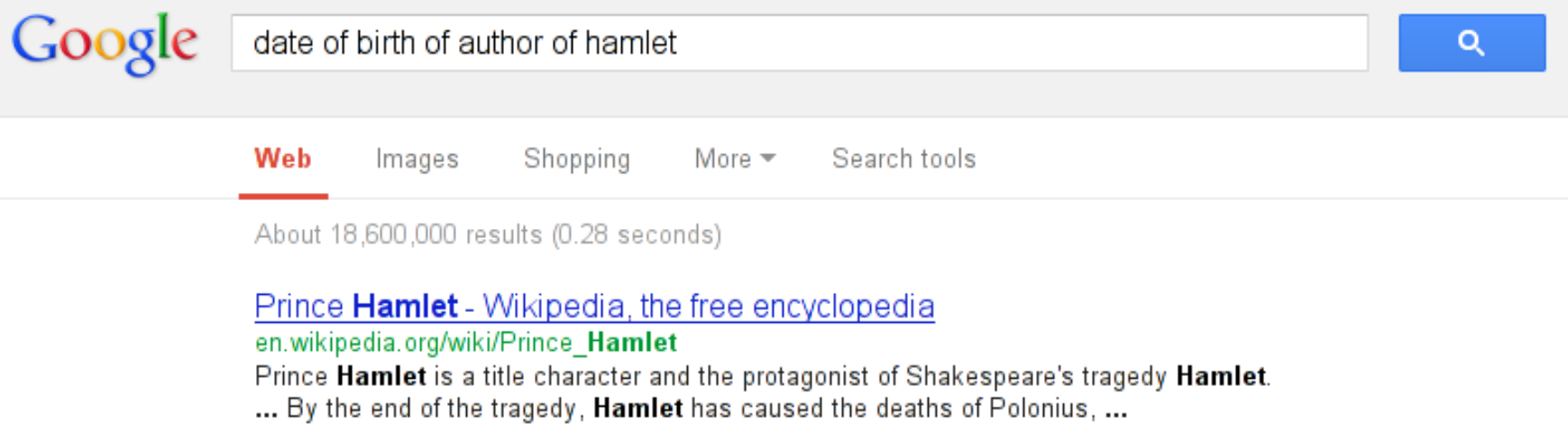
→ What do we want to do with it?

Get answers, not results



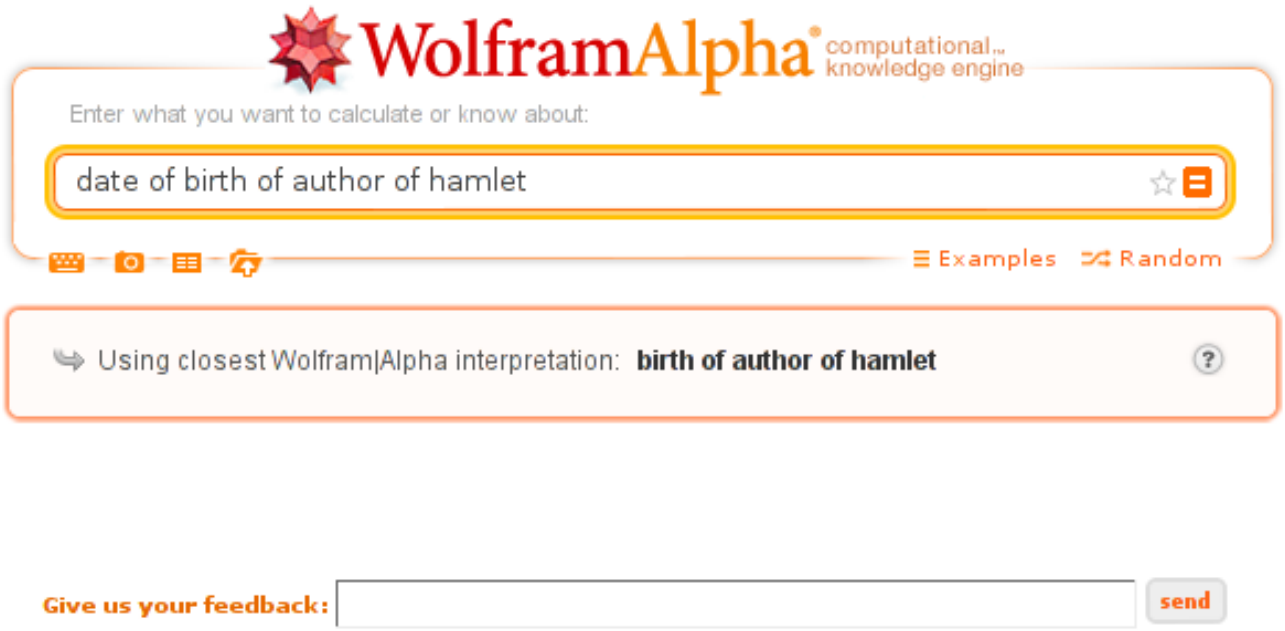


Get answers, not results



→ 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_Nebul**...
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:

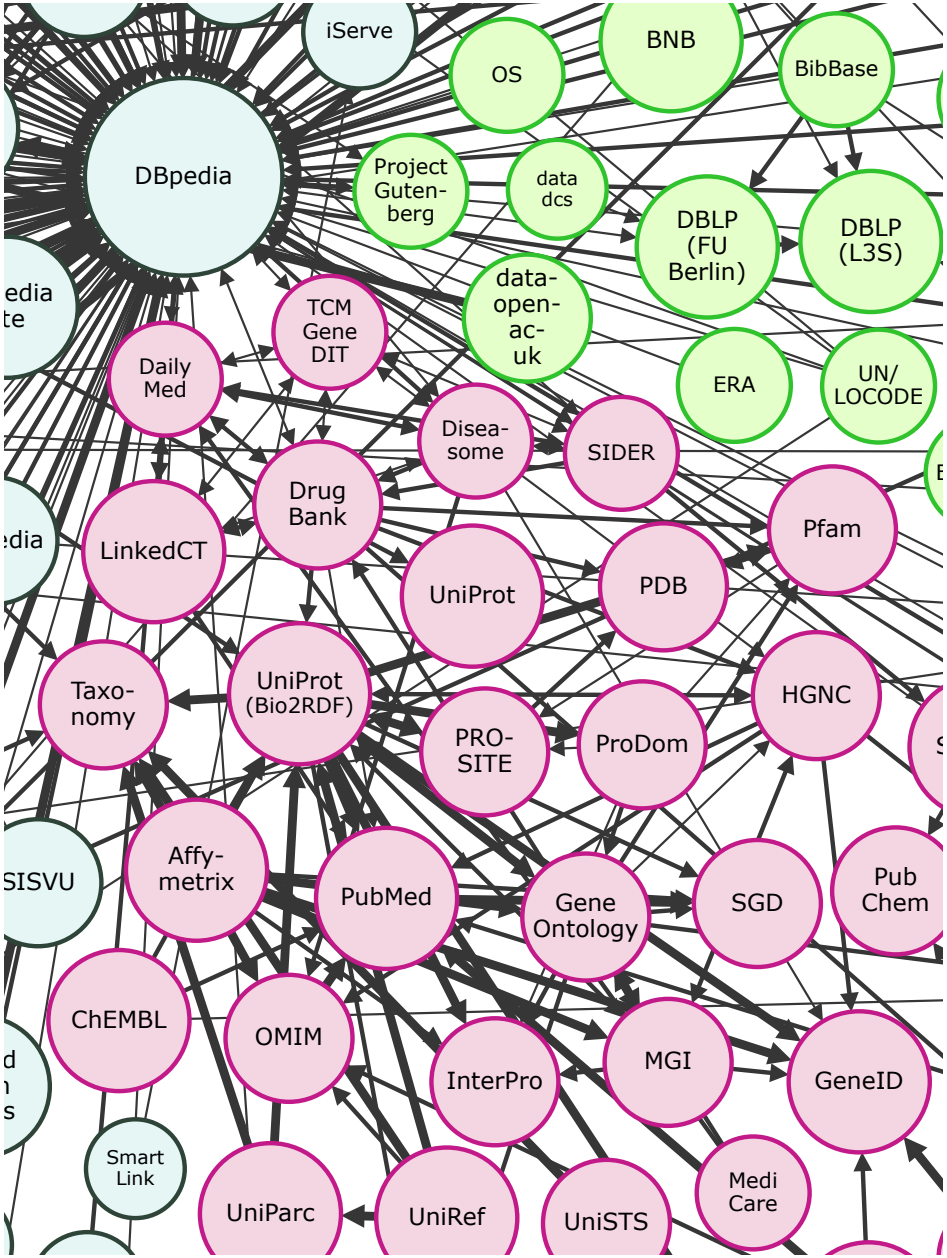
Google select ?book where ?book author ?x sex Female, ?book award Nebula, Hugo

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.

Existing data



- Many independent information sources.
- Links between these sources.

→ Linked Data Cloud.

Google Knowledge Graph

École normale supérieure - Paris
www.**ens**.fr/ - Translate this page
Etablissement 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
November 23 - 25. Salon du livre de Sciences humaines. November 26. Les lundis de la philosophie (Claudine Tiercelin-Collège de France). November 27 **...**
Formations prédoctorales - Étudiants internationaux - Contacts and Maps - DMA

E.N.S.A.D.
www.**ens**ad.fr/ - Translate this page
Sous l'égide de la Fondation Paris Sciences Lettres, SACRe est le fruit d'une collaboration entre l'**Ecole normale supérieure** et les **Ecoles** d'art et de **...**
Score: **23** / 30 - 13 Google reviews

31 Rue d'Ulm 75005 Paris, France
01 42 34 97 00

École Normale Supérieure - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**École_Normale_Supérieure**
The **École normale supérieure** (French pronunciation: [ekɔl nɔʁmal sypeʁjœʁ]; 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

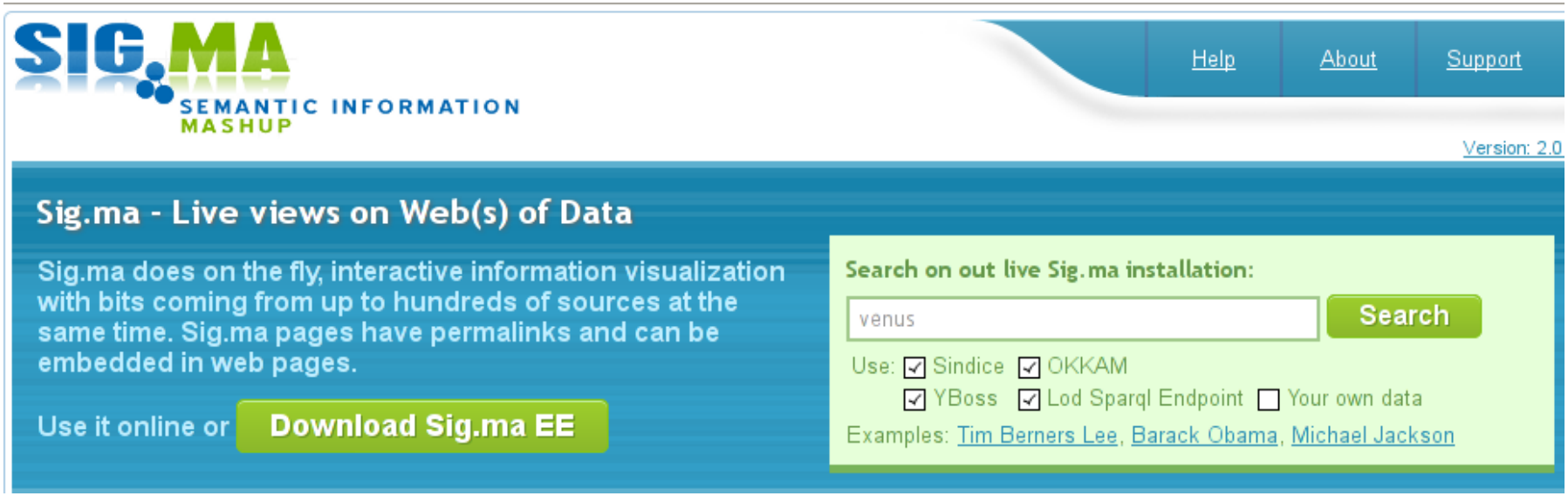
École Normale Supérieure

Directions

The École normale supérieure is a French grande école. The ENS was initially conceived during the French Revolution, and it was intended to provide the Republic with a new body of teachers, trained ...
Wikipedia

Address: 31 Rue d'Ulm, 75005 Paris, France
Phone: 01 42 34 97 00
Enrollment: 2,300 (2011)
Hours: Mon-Fri 9am–9pm
Sat-Sun Closed
Founded: 1794
Colors: Yellow, Purple

The perfect solution...



... or not!

header align:	center [18]
human relationship hyperlink:	http://www.asigurari-auto-rca.ro [13] http://floria.ro [13] http://www.viajoa.ro [13] Laptop News - Stiri, Review-uri, Sfaturi, Laptopuri [13] Sport Local - Primul loc pentru sportul local [13] http://www.meritacitit.ro [13] Social Media Marketing. ROI? [13]
is human relationship hyperlink of:	http://www.asigurari-auto-rca.ro [13] http://floria.ro [13] http://www.viajoa.ro [13] Laptop News - Stiri, Review-uri, Sfaturi, Laptopuri [13] Sport Local - Primul loc pentru sportul local [13] Social Media Marketing. ROI? [13] http://www.meritacitit.ro [13]
hypernym:	mollusk genus [12] inferior planet [14]
is holonym of:	quahog [12]
is hyponym of:	mollusk genus [12] inferior planet [14]
holonym:	Veneridae [12] solar system [14]
height:	355 [5,7] 510 [6,8,9,10]
is in synset of:	genus Venus [12] Venus [14,12]

Sources (20) Approved (0) Rejected (0)

1 About: Venus Kallipygos 34 facts | 2010-04-26
indice http://dbpedia.org/page/Venus_Kallipygos

2 Escape on Venus 43 facts | 2012-06-05
indice http://dbpedia.org/resource/Escape_on_Ven...

3 Venus 11 facts | 2010-06-16
indice http://scentscribbles.com/notes/Special:U...

4 Venus (The Grand Tour) 27 facts | 2010-08-26
indice http://rdf.freebase.com/ns/en.venus_the_g...

5 Venus 17 facts | 2010-10-17
indice http://www.slideshare.net/conanillo/venus...

6 Venus 19 facts | 2010-11-30
indice http://www.slideshare.net/telesatellitear...

7 Venus 19 facts | 2010-11-26
indice http://www.slideshare.net/andisimon/venus...

8 Venus 18 facts | 2010-11-29
indice http://www.slideshare.net/telesatellitear...

9 Venus 19 facts | 2010-11-30
indice http://www.slideshare.net/telesatellitear...

<- 1 2 3 -> reject all approve all

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!).

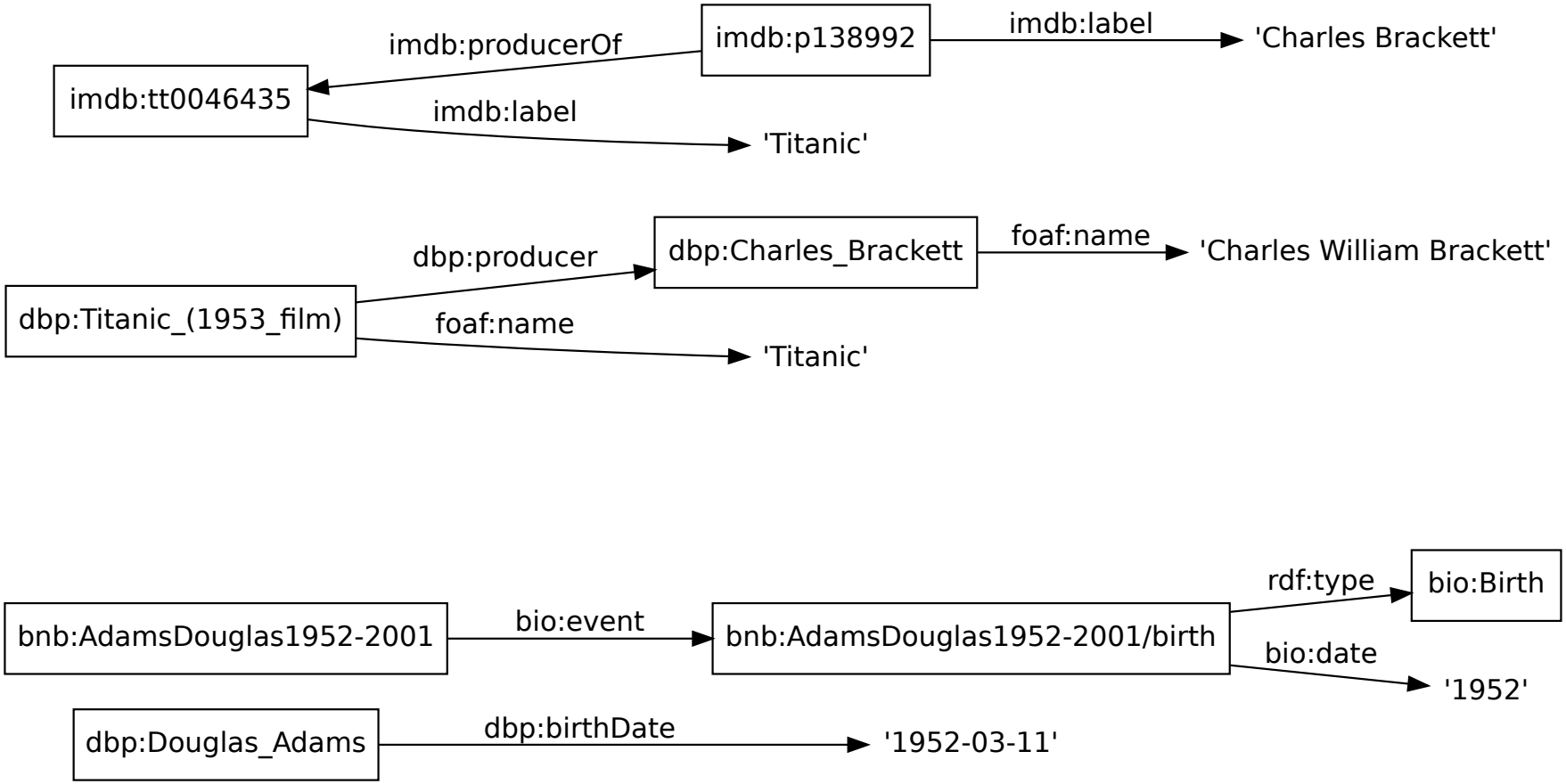
```
<div itemscope itemtype="http://schema.org/Person">
  <span itemprop="name">Jane Doe</span>
  <img src="janedoe.jpg" itemprop="image" />

  <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:jane-doe@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



Information extraction: Wikipedia

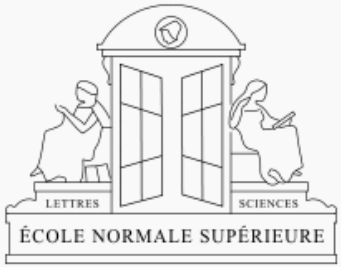
Wikipedia, a centralized island in the decentralized Web.

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

→ DBpedia, `http://dbpedia.org/`.

→ J. Hoffart, F. M. Suchanek, K. Berberich, G. Weikum. *YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia*. Special issue of the AI journal.

École normale supérieure



Established	1794
Type	EPCSCP
Director	Marc Mézard
Students	2700 [1]
Undergraduates	250 [1]
Postgraduates	Masters, agrégation, Ph.D
Location	Paris, France
Colours	Yellow, Purple
Nickname	ENS Ulm, Normale Sup'
Affiliations	Paris Sciences et Lettres - Quartier latin, Atomium Culture
Website	ens.fr

Information extraction: Hearst patterns

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 ia a celebrity
avenal is a city located in the state or province california

[Read the Web, Never Ending Language Learner, Carnegie Mellon University. `http://rtw.ml.cmu.edu/rtw/`]

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 Secretaire
- 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 Secretaire
- LIN Isabelle Étudiant
- MISTRAL Isabelle
- MOTTA 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

Information extraction: the Deep Web

Isabelle DELAIS

Fonctions :

- Secretaire

Affectations :

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

Adresse électronique : isabelle.delais@ens.fr

Adresses postales :

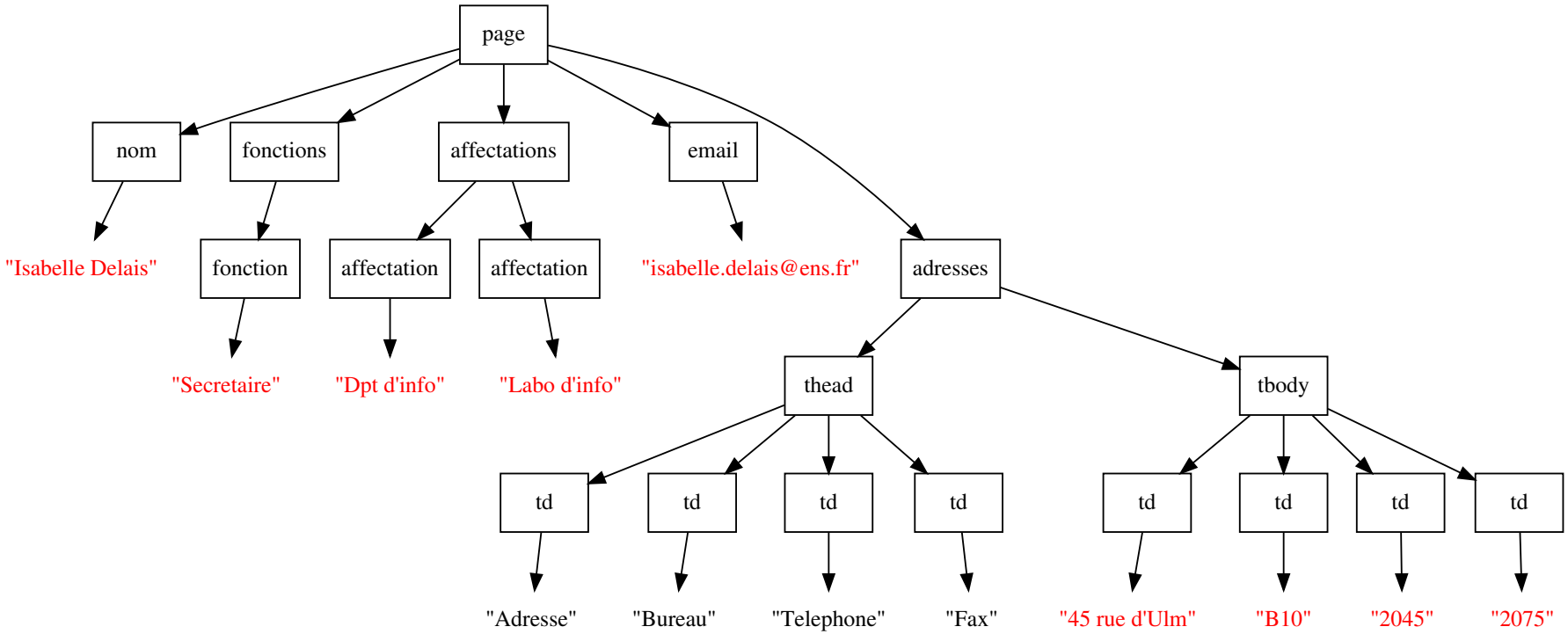
Adresse	Étage/Bureau	Téléphone	Fax
45, rue d'Ulm 75230 Paris cedex 05	Aile Rataud RDC, bureau B10	20 45	20 75

Mots-clés :

ok

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.
- M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, G. Weikum (MPI). *AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables*. VLDB'11 demo. `https://d5gate.ag5.mpi-sb.mpg.de/webaida/`

Corroboration

Corroboration

The **best text editor in the world** is vim[1][2], emacs[3][4], and notepad.exe[5].

- 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.
- A. Galland, S. Abiteboul, A. Marian, and P. Senellart, *Corroborating Information from Disagreeing Views*. WSDM'11.
- B. Zhao, B. I. P. Rubinstein, J. Gemmell, J. Han. *A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration*. VLDB'12.

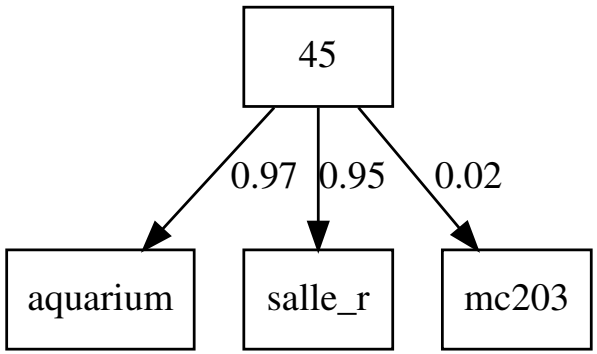
Uncertainty

- The two previous steps lead to uncertainty.
- Different ways to model this uncertainty:

Probabilistic databases

container	containee	p
aquarium	45	0.97
salle_r	45	0.95
mc203	45	0.02

Probabilistic XML



- Interesting problems: optimal representation, expressiveness, running time, etc.
- P. Senellart. *Probabilistic XML: A Data Model for the Web*. HDR thesis, 2012.

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.).
- M. Benedikt, G. Gottlob, P. Senellart. *Determining Relevance of Accesses at Runtime*. PODS'11.

"Find two pupils with two classes in common."

IsIn(*pupil*, *class*)

Pupil:
fechant
Submit

pupil	class
fechant	physique3
fechant	physique42

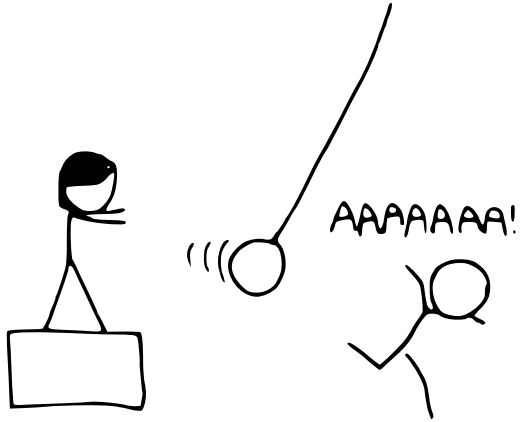
IsIn(*pupil*, *class*)

Class:
algoprog
Submit

pupil	class
bourgeat	algoprog
delpeuch	algoprog
forest	algoprog

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...

- S. Abiteboul. Bases de données, L3, 2nd semester. Covers the basics of relational databases.
- S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset, P. Senellart. Web data management. For those who want to find out more on their own: `http://webdam.inria.fr/Jorge/`
- S. Abiteboul, P. Rigaux, P. Senellart. Web data management, MPRI level 2. Covers the above book.

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
Frame 4: http://www.dataenthusiast.com/2011/05/85-unstructured-data-15-what-the-hell-is-going-on/
Frames 8, 9, 11, 14: http://google.com/
Frame 10: http://wolframalpha.com/
Frame 13: Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/
Frame 15: http://sig.ma/
Frame 18: http://schema.org/
Frame 21: https://en.wikipedia.org/wiki/Ecole_Normale_Superieure
Frame 22: http://well.blogs.nytimes.com/2012/11/21/the-love/, http://rtw.ml.cmu.edu/rtw/
Frames 23 to 25: http://annuaireweb.ens.fr/
Frames 27 and 28: http://en.wikipedia.org/
Frame 31: https://xkcd.com/755/