The Hidden Web, XML, and the Semantic Web: A Scientific Data Management Perspective

3h Tutorial at EDBT 2011

Fabian M. Suchanek,
Aparna Varde,
Richi Nayak,
Pierre Senellart

INRIA

MONTCLAIR STATE UNIVERSITY

QUT Queensland University of Technology
Brisbane Australia

TELECOM ParisTech

Overview

- Introduction
- The Hidden Web
- XML

Lunch

- DSML
- The Semantic Web
- Conclusion

All slides are available at
http://suchanek.name/work/publications/edbt2011tutorial

Motivation

UPPSALA
UNIVERSITET

Uppsala Universitet - Firefox

Job advertisements

Professors | PhD Students | Other

Application letter

Cedric Villani

Motivation

Google

Should we hire Cedric Villani?

Math News
"Certainly, **we should** treat people who need it", said **Cedric Villani**
www.dm.unito.it/

Cedric Villani
Born: 1973
Notable Awards: Fields Medal
Publications: ...
Scientific reputation: ...

Motivation

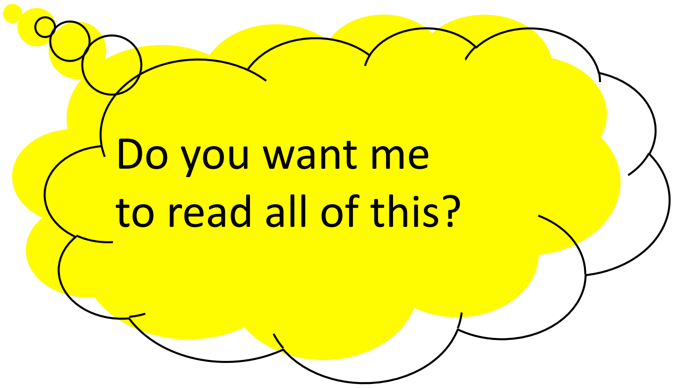
Google

Cedric Villani

About 198,000 results (0.18 seconds)

Cedric Villani’s homepage
Cedric Villani - Pierre et Marie Curie
villani.org

Cedric Villani - Wikipedia
Cedric Villani is a French mathematician...
en.wikipedia.org/wiki/Cedric_Villani


Do you want me to read all of this?

Cedric Villani – International Congress of Mathematicians
Cedric Villani worked on non-linear Landau damping
www.icm.org/2010

Interview with Cedric Villani
Cedric Villani : “I think world peace can still be achieved if we all work together.”
www.tabloid.com/news

Motivation

Google

Dear Larry, you are getting me wrong. I just want to know

3quarksdaily: August 2010

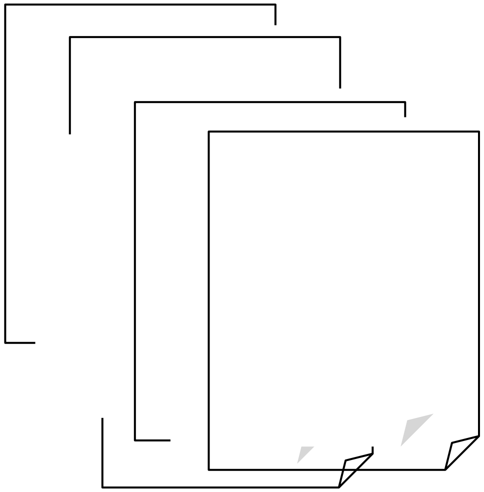
If you want good things to happen, be a good person**.**

3quarksdaily.com

Current trends on the Web

Fortunately, the Web consists not just of HTML pages...

This tutorial is about other types of data on the Web:



- The Hidden Web

  everything that is hidden behind Web forms

  What did he publish? Who are his co-authors?

- XML and DSML

  the clandestine lingua franca of the Web

  What is his research about?

- the Semantic Web

  defining semantics for machines

  When was he born? Who did he study with? What prizes was he awarded?

Not just about recruiting scientists

- General techniques for:
  - Discovering data sources of interest
  - Retrieving meaningful data
  - Mining information of interest

- … on “new” forms of Web information, underexploited by current search and retrieval systems

- Example of scientific data management, and more specifically Cedric Villani's works

Overview

- Introduction ✔
- The Hidden Web
- XML
- DSML
- The Semantic Web
- Conclusion

The Hidden Web

Pierre Senellart

INRIA Saclay & Télécom ParisTech

Paris, France

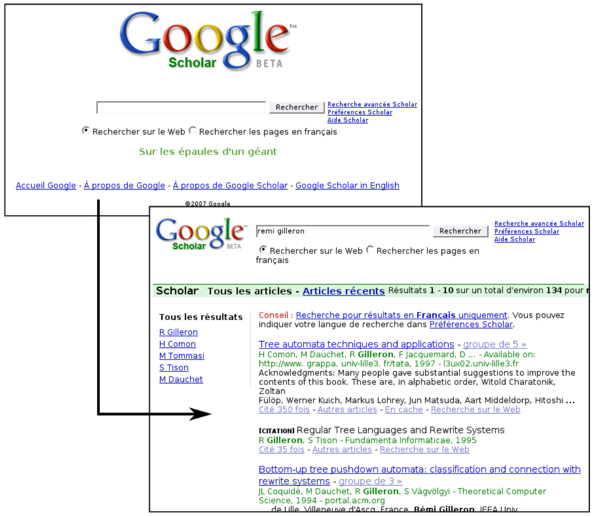
(pierre@senellart.com)

Outline: the hidden Web

- The Hidden Web
- Extensional and Intensional Approaches
- Understanding Web Forms
- Understanding Response Pages
- Perspectives

The Hidden Web

Definition (Hidden Web, Deep Web)
All the content of the Web that is not directly accessible through hyperlinks. In particular: HTML forms, Web services.


Google Scholar BETA
Rechercher
Recherche avancée Scholar
Préférences Scholar
Aide Scholar
● Rechercher sur le Web ○ Rechercher les pages en français
Sur les épaules d'un géant
Accueil Google - À propos de Google - À propos de Google Scholar - Google Scholar in English
©2007 Google
Google Scholar BETA
remi gilleron
Rechercher
Recherche avancée Scholar
Préférences Scholar
Aide Scholar
● Rechercher sur le Web ○ Rechercher les pages en français
Scholar Tous les articles - Articles récents
Tous les résultats
R Gilleron
H Comon
M Tommasi
S Tison
M Dauchet

Résultats **1** - **10** sur un total d'environ **134** pour r

Conseil : Recherche pour résultats en **Français** uniquement. Vous pouvez indiquer votre langue de recherche dans Préférences Scholar.

Tree automata techniques and applications - groupe de 5 »
H Comon, M Dauchet, R **Gilleron**, F Jacquemard, D ... - Available on: http://www. grappa. univ-lille3. fr/tata, 1997 - l3ux02.univ-lille3.fr
Acknowledgments: Many people gave substantial suggestions to improve the contents of this book. These are, in alphabetic order, Witold Charatonik, Zoltan
Fülöp, Werner Kuich, Markus Lohrey, Jun Matsuda, Aart Middeldorp, Hitoshi **...**
Cité 350 fois - Autres articles - En cache - Recherche sur le Web

[CITATION] Regular Tree Languages and Rewrite Systems
R **Gilleron**, S Tison - Fundamenta Informaticae, 1995
Cité 35 fois - Autres articles - Recherche sur le Web

Bottom-up tree pushdown automata: classification and connection with rewrite systems - groupe de 3 »
JL Coquidé, M Dauchet, R **Gilleron**, S Vágvölgyi - Theoretical Computer Science, 1994 - portal.acm.org
de Lille, Villeneuve d'Ascq, France. **Rémi Gilleron**, IEEA Univ.

Size estimate

- [Bri00] 500 times more content than on the surface Web! Dozens of thousands of databases.
- [HPWC07] ~ 400 000 deep Web databases.

Sources of the Deep Web

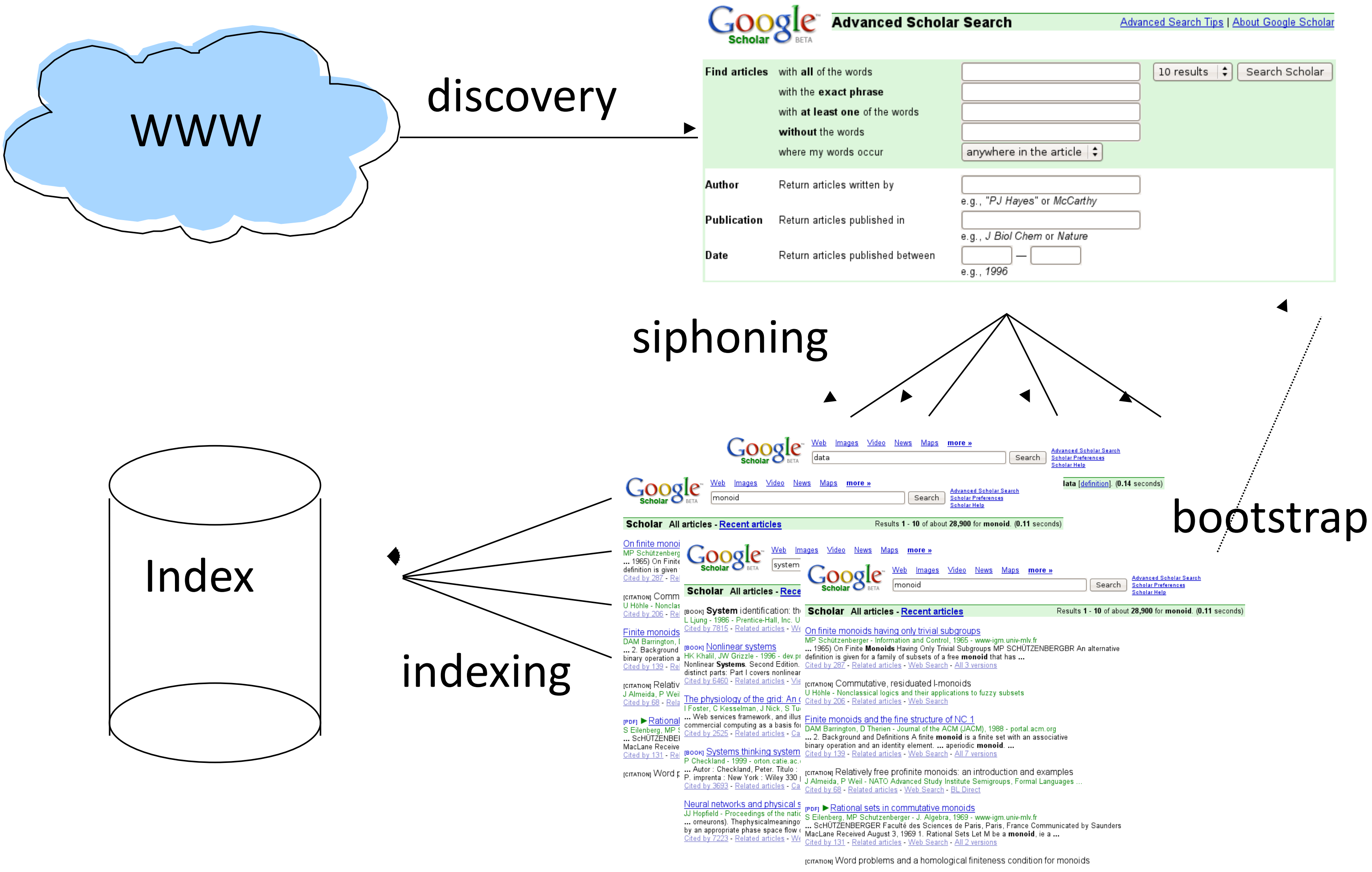
Examples

- Publication databases;
- Library catalogs;
- *Yellow Pages* and other directories;
- Weather services;
- Geolocalization services;
- US Census Bureau data;
- etc.

Discovering Knowledge from the Deep Web

- Content of the deep Web hidden to classical Web search engines (they just follow links)
- But very valuable and high quality!
- Even services allowing access through the surface Web (e.g., DBLP, e-commerce) have more semantics when accessed from the deep Web
- How to benefit from this information?
- How to do it automatically, in an unsupervised way?

Extensional Approach



Advanced Scholar Search | Advanced Search Tips | About Google Scholar

Find articles	with **all** of the words	10 results · Search Scholar
	with the **exact phrase**	
	with **at least one** of the words	
	without the words	
	where my words occur	anywhere in the article
Author	Return articles written by	e.g., *"PJ Hayes"* or *McCarthy*
Publication	Return articles published in	e.g., *J Biol Chem* or *Nature*
Date	Return articles published between	— e.g., *1996*

monoid · Search · Advanced Scholar Search · Scholar Preferences · Scholar Help

Scholar All articles - Recent articles · Results **1** - **10** of about **28,900** for **monoid**. (**0.11** seconds)

On finite monoids having only trivial subgroups
MP Schützenberger - Information and Control, 1965 - www-igm.univ-mlv.fr
... 1965) On Finite **Monoids** Having Only Trivial Subgroups MP SCHÜTZENBERGBR An alternative definition is given for a family of subsets of a free **monoid** that has ...
Cited by 287 - Related articles - Web Search - All 3 versions

[CITATION] Commutative, residuated l-monoids
U Höhle - Nonclassical logics and their applications to fuzzy subsets
Cited by 206 - Related articles - Web Search

Finite monoids and the fine structure of NC 1
DAM Barrington, D Therien - Journal of the ACM (JACM), 1988 - portal.acm.org
... 2. Background and Definitions A finite **monoid** is a finite set with an associative binary operation and an identity element. ... aperiodic **monoid**. ...
Cited by 139 - Related articles - Web Search - All 7 versions

[CITATION] Relatively free profinite monoids: an introduction and examples
J Almeida, P Weil - NATO Advanced Study Institute Semigroups, Formal Languages ...
Cited by 68 - Related articles - Web Search - BL Direct

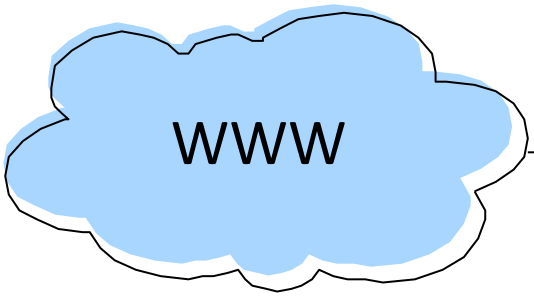
[PDF] ▶ Rational sets in commutative monoids
S Eilenberg, MP Schutzenberger - J. Algebra, 1969 - www-igm.univ-mlv.fr
... SCHÜTZENBERGER Faculté des Sciences de Paris, Paris, France Communicated by Saunders MacLane Received August 3, 1969 1. Rational Sets Let M be a **monoid**, ie a ...
Cited by 131 - Related articles - Web Search - All 2 versions

[CITATION] Word problems and a homological finiteness condition for monoids

Notes on the Extensional Approach

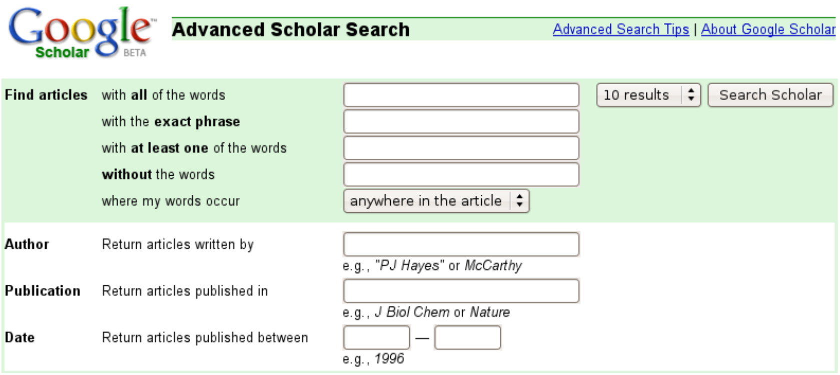
- Main issues:
  - Discovering services
  - Choosing appropriate data to submit forms
  - Use of data found in result pages to bootstrap the siphoning process
  - Ensure good coverage of the database
- Approach favored by Google [MHC+06], used in production [MAAH09]
- Not always feasible (huge load on Web servers)
- Does not help in getting structured information!

Intensional Approach



WWW

discovery



Advanced Scholar Search Advanced Search Tips | About Google Scholar

Find articles with **all** of the words □ 10 results Search Scholar

with the **exact phrase** □

with **at least one** of the words □

without the words □

where my words occur anywhere in the article

Author Return articles written by □

e.g., *"PJ Hayes"* or *McCarthy*

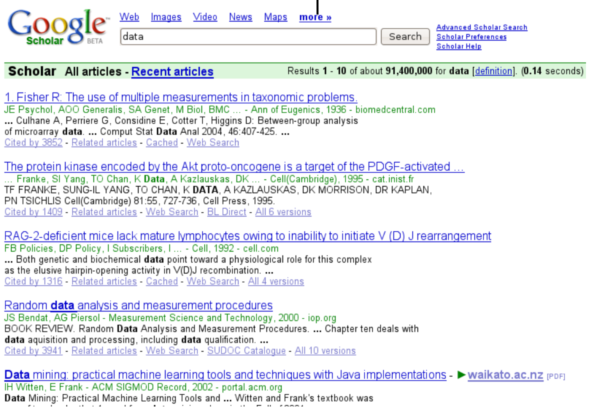
Publication Return articles published in □

e.g., *J Biol Chem* or *Nature*

Date Return articles published between □ — □

e.g., *1996*

probing



data

Scholar All articles - **Recent articles** Results **1** - **10** of about **91,400,000** for **data** [definition]. (**0.14** seconds)

1. Fisher R: The use of multiple measurements in taxonomic problems.
JE Psychol, AOO Generalis, SA Genet, M Biol, BMC ... - Ann of Eugenics, 1936 - biomedcentral.com
... Culhane A, Perriere G, Considine E, Cotter T, Higgins D: Between-group analysis of microarray **data**. **...** Comput Stat **Data** Anal 2004, 46:407-425. **...**
Cited by 3852 - Related articles - Cached - Web Search

The protein kinase encoded by the Akt proto-oncogene is a target of the PDGF-activated ...
... Franke, SI Yang, TO Chan, K **Data**, A Kazlauskas, DK ... - Cell(Cambridge), 1995 - cat.inist.fr
TF FRANKE, SUNG-IL YANG, TO CHAN, K **DATA**, A KAZLAUSKAS, DK MORRISON, DR KAPLAN, PN TSICHLIS Cell(Cambridge) 81:55, 727-736, Cell Press, 1995.
Cited by 1409 - Related articles - Web Search - BL Direct - All 6 versions

RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V (D) J rearrangement
FB Policies, DP Policy, I Subscribers, I ... - Cell, 1992 - cell.com
... Both genetic and biochemical **data** point toward a physiological role for this complex as the elusive hairpin-opening activity in V(D)J recombination. **...**
Cited by 1316 - Related articles - Cached - Web Search - All 4 versions

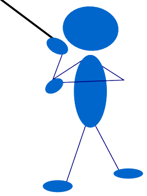
Random **data** analysis and measurement procedures
JS Bendat, AG Piersol - Measurement Science and Technology, 2000 - iop.org
BOOK REVIEW. Random **Data** Analysis and Measurement Procedures. **...** Chapter ten deals with **data** aquisition and processing, including **data** qualification. **...**
Cited by 3941 - Related articles - Web Search - SUDOC Catalogue - All 10 versions

Data mining: practical machine learning tools and techniques with Java implementations - ► **waikato.ac.nz** [PDF]
IH Witten, E Frank - ACM SIGMOD Record, 2002 - portal.acm.org
Data Mining: Practical Machine Learning Tools and **...** Witten and Frank's textbook was

analyzing

Form wrapped as a Web service

query



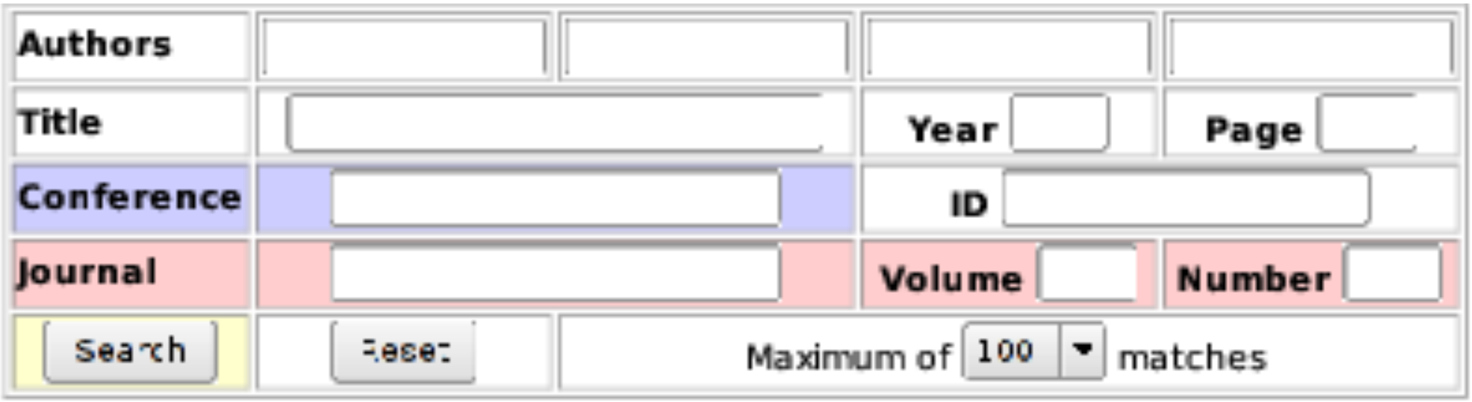
Notes on the Intensional Approach

- More ambitious [CHZ05, SMM+08]
- Main issues:
  - Discovering services
  - Understanding the structure and semantics of a form
  - Understanding the structure and semantics of result pages (wrapper induction)
  - Semantic analysis of the service as a whole
- No significant load imposed on Web servers

Discovering deep Web forms

- Crawling the Web and selecting forms
- But not all forms!
  - Hotel reservation
  - Mailing list management
  - Search within a Web site
- **Heuristics**: prefer GET to POST, no password, no credit card number, more than one field, etc.
- Given domain of interest (e.g., scientific publications): use focused crawling to restrict to this domain

Web forms



Maximum of 100 matches

- **Simplest case**: associate each form field with some domain concept
- **Assumption:** fields independent from each other (not always true!), can be queried with words that are part of a domain instance

Structural analysis of a form (1/2)

- Build a context for each field:
  - label tag;
  - id and name attributes;
  - text immediately before the field.
- Remove stop words, stem
- Match this context with concept names or concept ontology
- Obtain in this way candidate annotations

Structural analysis of a form (2/2)

For each field annotated with concept *c*:

- Probe the field with nonsense word to get an error page
- Probe the field with instances of concept *c*
- Compare pages obtained by probing with the error page (e.g., clustering along the DOM tree structure of the pages), to distinguish error pages and result pages
- Confirm the annotation if enough result pages are obtained

Bootstrapping the siphoning

- Siphoning (or probing) a deep Web database requires many relevant data to submit the form with
- **Idea**: use most frequent words in the content of the result pages
- Allows bootstrapping the siphoning with just a few words!

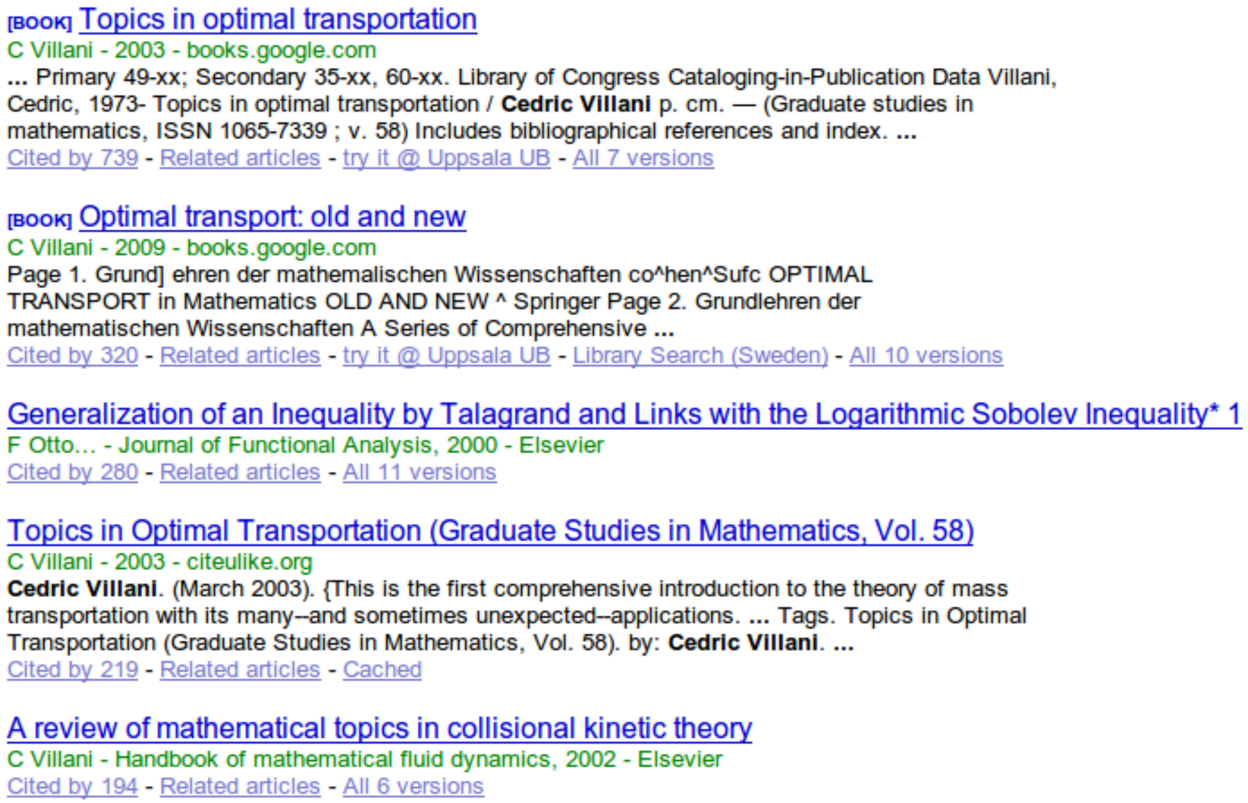
Inducing wrappers from result pages

Pages r

- share
- set o
- unkn

Goal

Buildin s, in a fully au



[BOOK] Topics in optimal transportation
C Villani - 2003 - books.google.com
... Primary 49-xx; Secondary 35-xx, 60-xx. Library of Congress Cataloging-in-Publication Data Villani, Cedric, 1973- Topics in optimal transportation / **Cedric Villani** p. cm. — (Graduate studies in mathematics, ISSN 1065-7339 ; v. 58) Includes bibliographical references and index. ...
Cited by 739 - Related articles - try it @ Uppsala UB - All 7 versions

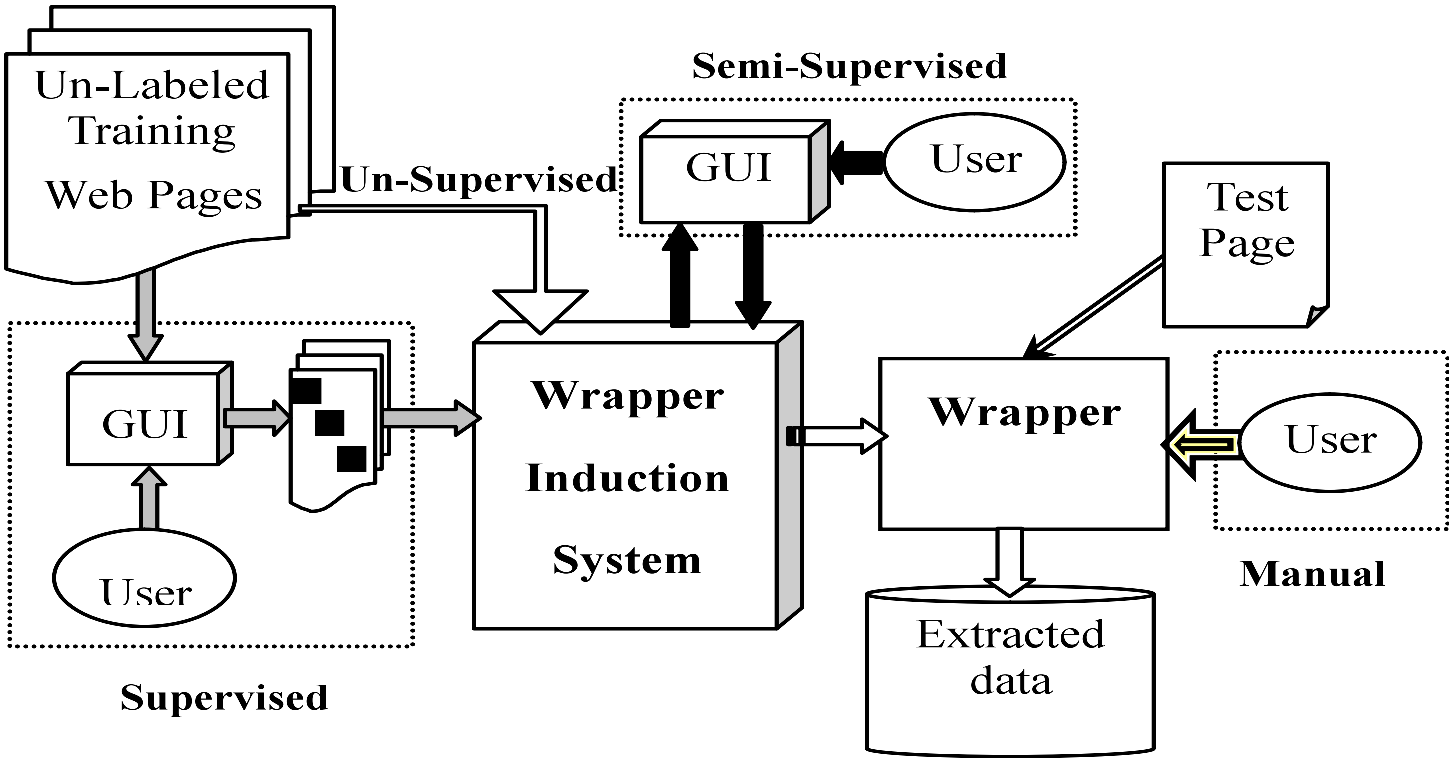
[BOOK] Optimal transport: old and new
C Villani - 2009 - books.google.com
Page 1. Grund] ehren der mathemalischen Wissenschaften co^hen^Sufc OPTIMAL TRANSPORT in Mathematics OLD AND NEW ^ Springer Page 2. Grundlehren der mathematischen Wissenschaften A Series of Comprehensive ...
Cited by 320 - Related articles - try it @ Uppsala UB - Library Search (Sweden) - All 10 versions

Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality* 1
F Otto... - Journal of Functional Analysis, 2000 - Elsevier
Cited by 280 - Related articles - All 11 versions

Topics in Optimal Transportation (Graduate Studies in Mathematics, Vol. 58)
C Villani - 2003 - citeulike.org
Cedric Villani. (March 2003). {This is the first comprehensive introduction to the theory of mass transportation with its many--and sometimes unexpected--applications. ... Tags. Topics in Optimal Transportation (Graduate Studies in Mathematics, Vol. 58). by: **Cedric Villani**. ...
Cited by 219 - Related articles - Cached

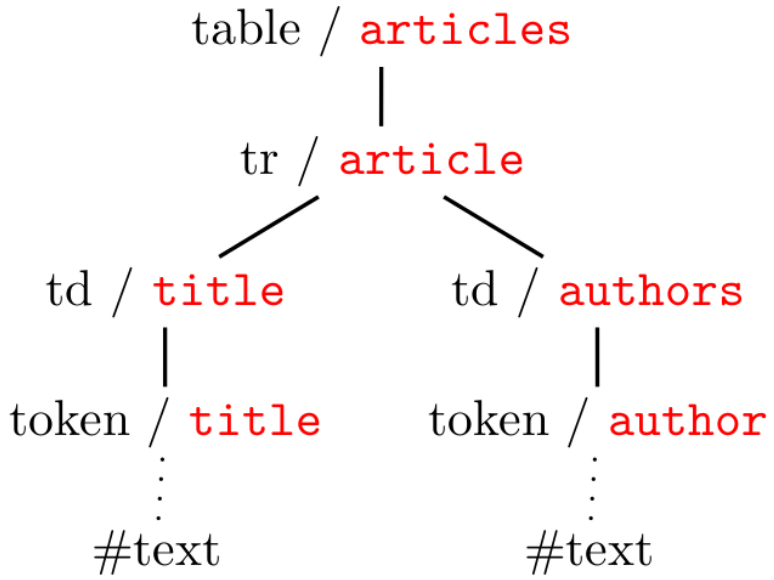
A review of mathematical topics in collisional kinetic theory
C Villani - Handbook of mathematical fluid dynamics, 2002 - Elsevier
Cited by 194 - Related articles - All 6 versions

Information extraction systems [CKGS06]

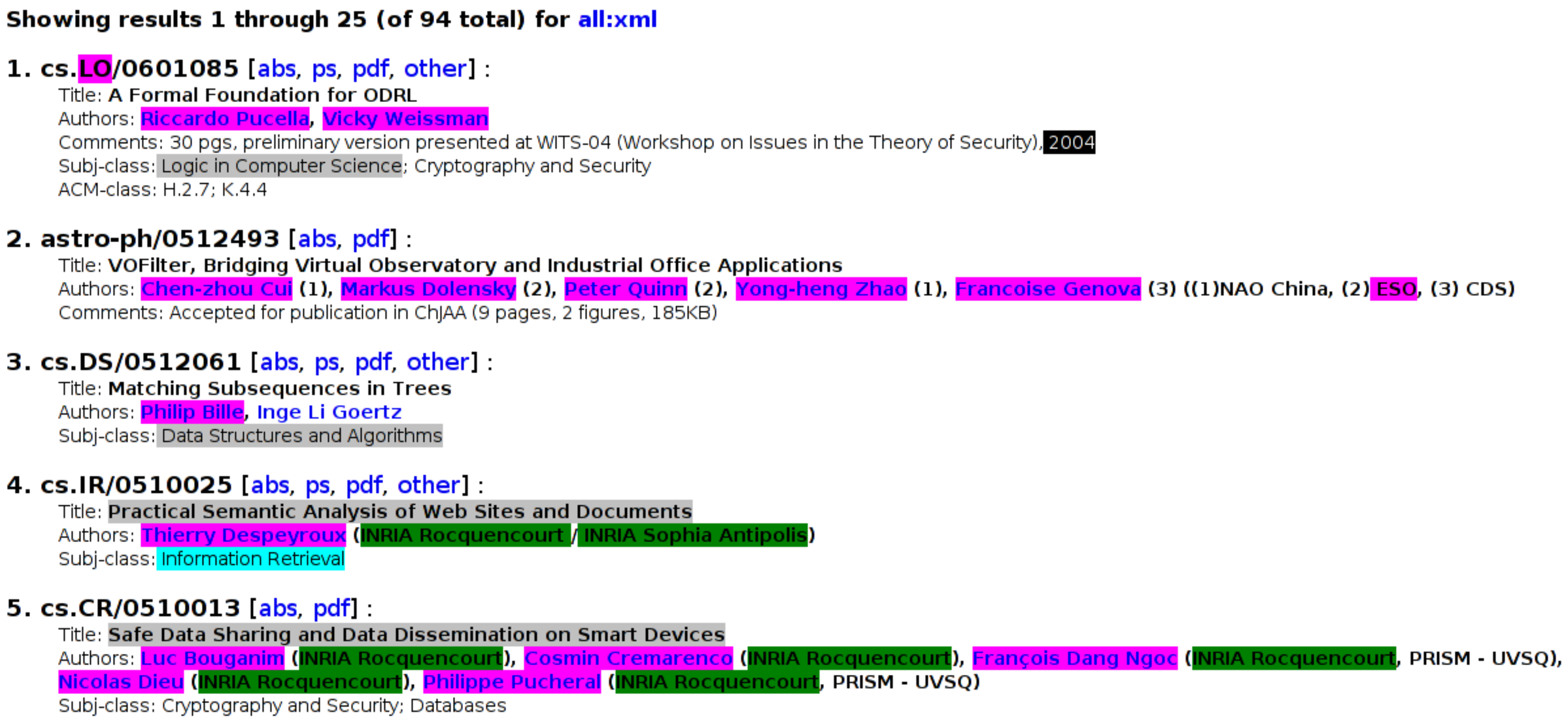


Unsupervised Wrapper Induction

- Use the (repetitive) structure of the result pages to infer a wrapper for all pages of this type
- Possibly: use in parallel with annotation by recognized concept instances to learn with both the structure and the content



Annotating with domain instances [SMM+08]



Showing results 1 through 25 (of 94 total) for all:xml

1. cs.LO/0601085 [abs, ps, pdf, other] :
Title: **A Formal Foundation for ODRL**
Authors: **Riccardo Pucella**, **Vicky Weissman**
Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004
Subj-class: Logic in Computer Science; Cryptography and Security
ACM-class: H.2.7; K.4.4

2. astro-ph/0512493 [abs, pdf] :
Title: **VOFilter, Bridging Virtual Observatory and Industrial Office Applications**
Authors: **Chen-zhou Cui (1), Markus Dolensky (2), Peter Quinn (2), Yong-heng Zhao (1), Francoise Genova (3) ((1)NAO China, (2) ESO, (3) CDS)**
Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

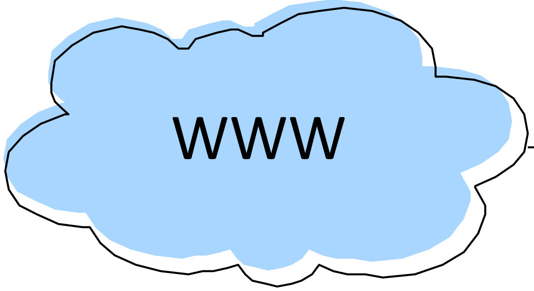
3. cs.DS/0512061 [abs, ps, pdf, other] :
Title: **Matching Subsequences in Trees**
Authors: **Philip Bille**, **Inge Li Goertz**
Subj-class: Data Structures and Algorithms

4. cs.IR/0510025 [abs, ps, pdf, other] :
Title: **Practical Semantic Analysis of Web Sites and Documents**
Authors: **Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)**
Subj-class: Information Retrieval

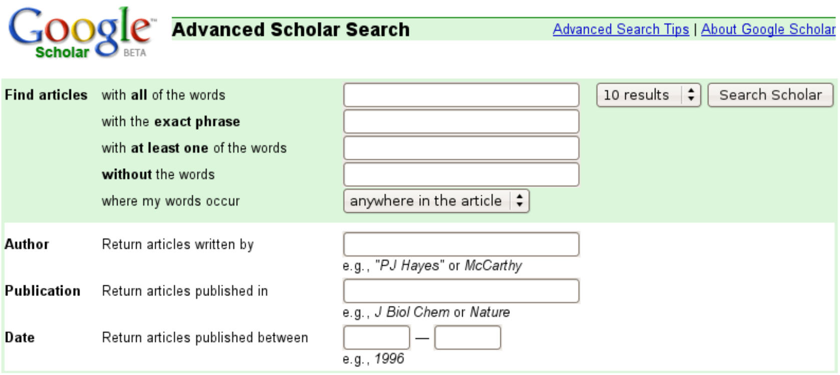
5. cs.CR/0510013 [abs, pdf] :
Title: **Safe Data Sharing and Data Dissemination on Smart Devices**
Authors: **Luc Bouganim (INRIA Rocquencourt), Cosmin Cremarenco (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dieu (INRIA Rocquencourt), Philippe Pucheral (INRIA Rocquencourt, PRISM - UVSQ)**
Subj-class: Cryptography and Security; Databases

And generalizing from that!

Recap: what does work?



discovery



Advanced Scholar Search

Find articles with **all** of the words

with the **exact phrase**

with **at least one** of the words

without the words

where my words occur: anywhere in the article

10 results

Author Return articles written by

e.g., *"PJ Hayes"* or *McCarthy*

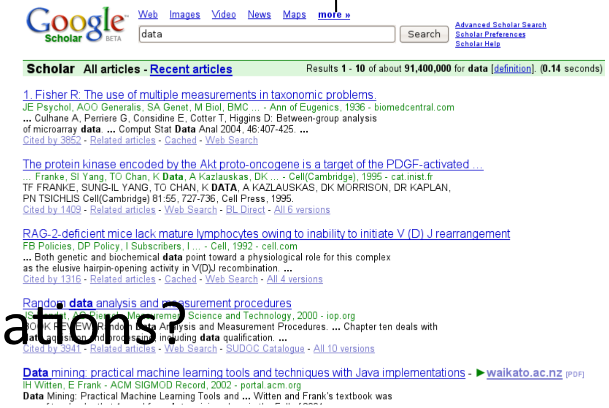
Publication Return articles published in

e.g., *J Biol Chem* or *Nature*

Date Return articles published between —

e.g., *1996*

probing



Search box: data

Scholar All articles - Recent articles — Results **1** - **10** of about **91,400,000** for **data** [definition]. (**0.14** seconds)

1. Fisher R: The use of multiple measurements in taxonomic problems.
JE Psychol, AOO Generalis, SA Genet, M Biol, BMC ... - Ann of Eugenics, 1936 - biomedcentral.com
... Culhane A, Perriere G, Considine E, Cotter T, Higgins D: Between-group analysis of microarray **data**. **...** Comput Stat **Data** Anal 2004, 46:407-425. **...**
Cited by 3852 - Related articles - Cached - Web Search

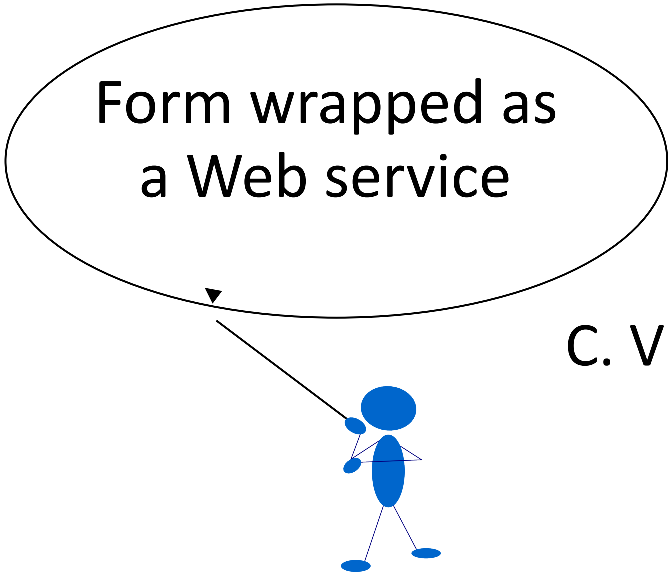
The protein kinase encoded by the Akt proto-oncogene is a target of the PDGF-activated ...
... Franke, SI Yang, TO Chan, K **Data**, A Kazlauskas, DK ... - Cell(Cambridge), 1995 - cat.inist.fr
TF FRANKE, SUNG-IL YANG, TO CHAN, K **DATA**, A KAZLAUSKAS, DK MORRISON, DR KAPLAN, PN TSICHLIS Cell(Cambridge) 81:55, 727-736, Cell Press, 1995.
Cited by 1409 - Related articles - Web Search - BL Direct - All 6 versions

RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V (D) J rearrangement
FB Policies, DP Policy, I Subscribers, I ... - Cell, 1992 - cell.com
... Both genetic and biochemical **data** point toward a physiological role for this complex as the elusive hairpin-opening activity in V(D)J recombination. **...**
Cited by 1316 - Related articles - Cached - Web Search - All 4 versions

Random **data** analysis and measurement procedures
JS Bendat, AG Piersol - Measurement Science and Technology, 2000 - iop.org
BOOK REVIEW: Random **Data** Analysis and Measurement Procedures. **...** Chapter ten deals with **data** acquisition and processing, including **data** qualification. **...**
Cited by 3941 - Related articles - Web Search - SUDOC Catalogue - All 10 versions

Data mining: practical machine learning tools and techniques with Java implementations - ▶ waikato.ac.nz [PDF]
IH Witten, E Frank - ACM SIGMOD Record, 2002 - portal.acm.org
Data Mining: Practical Machine Learning Tools and **...** Witten and Frank's textbook was

analyzing



Form wrapped as a Web service

C. Villani's publications?

Some perspectives

- Processing complex (relational) queries over deep Web sources [CM10]
- Dealing with complex forms (fields allowing Boolean operators, dependencies between fields, etc.)
- Static analysis of JavaScript code to determine which fields of a form are required, etc.
- A lot of this is also applicable to Web 2.0/AJAX applications

References

[Bri00] BrightPlanet. **The deep Web: Surfacing hidden value**. White paper, 2000.

[CHZ05] K. C.-C. Chang, B. He, and Z. Zhang. **Towards large scale integration: Building a metaquerier over databases on the Web**. In *Proc. CIDR*, 2005.

[CKGS06] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan. **A survey of Web information extraction systems**. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411-1428, 2006.

[CMM01] V. Crescenzi, G. Mecca, and P. Merialdo. **Roadrunner: Towards automatic data extraction from large Web sites**. In *Proc. VLDB*, Roma, Italy, Sep. 2001.

[CM10] A. Calì, D. Martinenghi, **Querying the deep Web**. In *Proc. EDBT*, 2010.

[HPWC07] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang. **Accessing the deep Web: A survey**. *Communications of the ACM*, 50(2):94–101, 2007.

[MAAH06] J. Madhavan, L. Afanasiev, L. Antova, and A. Y. Halevy, **Harnessing the Deep Web: Present Future**. In *Proc. CIDR*, 2009.

[MHC+06] J. Madhavan, A. Y. Halevy, S. Cohen, X. Dong, S. R. Jeffery, D. Ko, and C. Yu. **Structured data meets the Web: A few observations**. *IEEE Data Engineering Bulletin*, 29(4):19–26, 2006.

[SMM+08] P. Senellart, A. Mittal, D. Muschick, R. Gilleron et M. Tommasi, **Automatic Wrapper Induction from Hidden-Web Sources with Domain Knowledge**. In *Proc. WIDM*, 2008.

Overview

- Introduction ✔
- The Hidden Web ✔
- XML
- DSML
- The Semantic Web
- Conclusion

XML: Data Modeling and Mining

Richi Nayak
Computer Science Discipline
Queensland University of Technology
Brisbane, Australia

r.nayak@qut.edu.au

XML: An Example

- XML is a semi structured language

```
<Book Id= “B105”>
    <Title> Topics in Optimal Transportation </Title>
    <Author>
            <Name> Cedric Villani </Name>
    </Author>
    <Publisher>
            <Name> American Mathematical Society </Name>
            <Place> NewYork</Place>
     </Publisher>
 </Book>
```

Outline

- XML: Introduction
- XML Mining for Data Management
  - Challenges and Process
- XML Clustering
  - Handling XML Features
- XML Frequent Pattern Mining
  - Types of Patterns
- Future directions

XML (eXtensible Markup Language)

- Standard for information and exchange

- XML v. HTML
  - HTML: restricted set of tags, e.g. <TABLE>, <H1>, <B>, etc.
  - XML: you can create your own tags

- Selena Sol (2000) highlights the four major benefits of using XML language:
  - XML separates data from presentation which means making changes to the display of data does not affect the XML data;
  - Searching for data in XML documents becomes easier as search engines can parse the description-bearing tags of the XML documents;
  - XML tag is human readable, even a person with no knowledge of XML language can still read an XML document;
  - Complex structures and relations of data can be encoded using XML.

XML: Usage

- Supports wide-variety of applications
  - Handle summaries of facts or events
    - RSS news feeds, Legal decisions, Company balance sheets
  - Scientific literature
    - Research articles, Medical reports, Book reviews
  - Technical documents
    - Data sheets, Product feature reviews, Classified advertisements
- More than 50 domain specific languages based on XML
- Wikipedia with over 3.4 M XML documents in English.

In essence – XML is anywhere and everywhere

Challenges in XML Management and Mining

```
<Book Id="B105">
    <Title> Topics in Optimal Transportation </Title>
    <Author>
       <Name>Cedric Villani</Name>
    </Author>
    <Publisher>
         <Name> American Mathematical Society </Name>
         <Place> NewYork</Place>
      </Publisher>
 </Book>
```

❑ Semi-structured

❑ Two features

- Structure
- Content

❑ Hierarchical relationship

```
<Author>
   <Name>Cedric Villani</Name>
</Author>
```

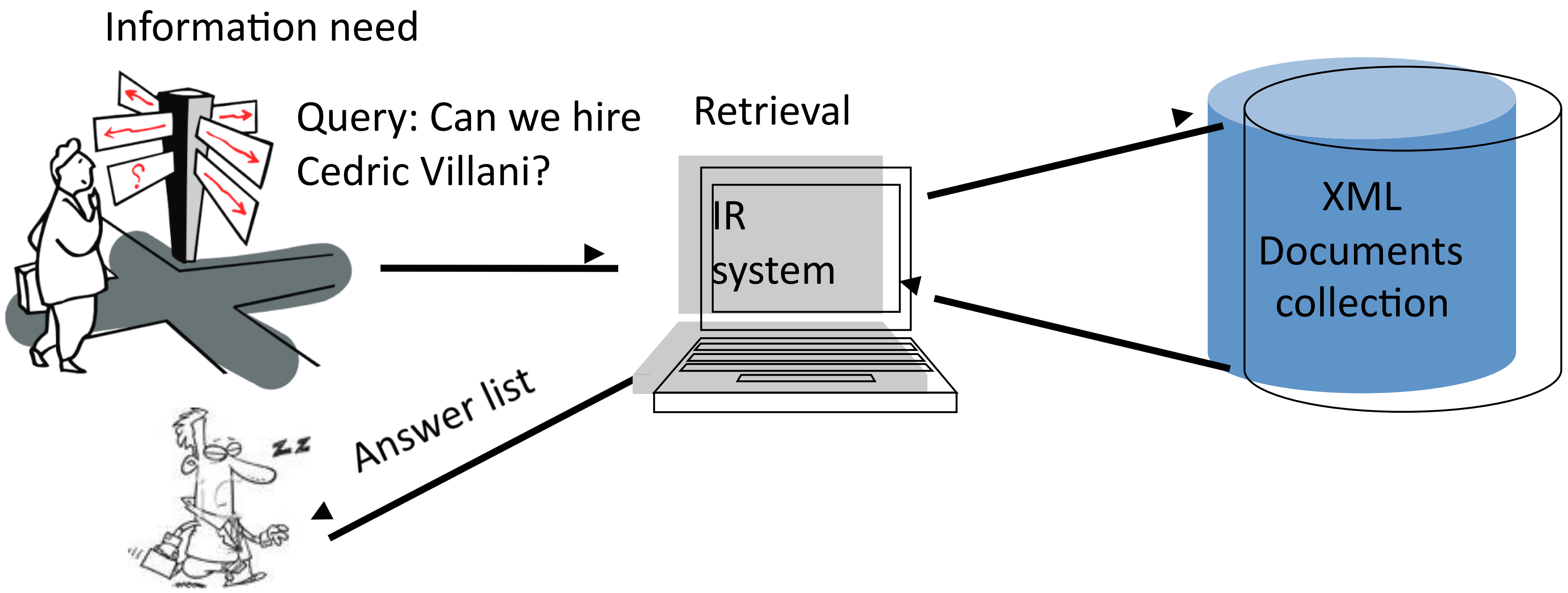
```
<Publisher>
   <Name>American Mathematical Society</Name>
</Publisher>
```

❑ Unbounded nesting

❑ User-defined tags – polysemy problems

❑ XML Data mining track in Initiative for Evaluation of XML documents (INEX) forum

Scenario : Searching XML documents collection

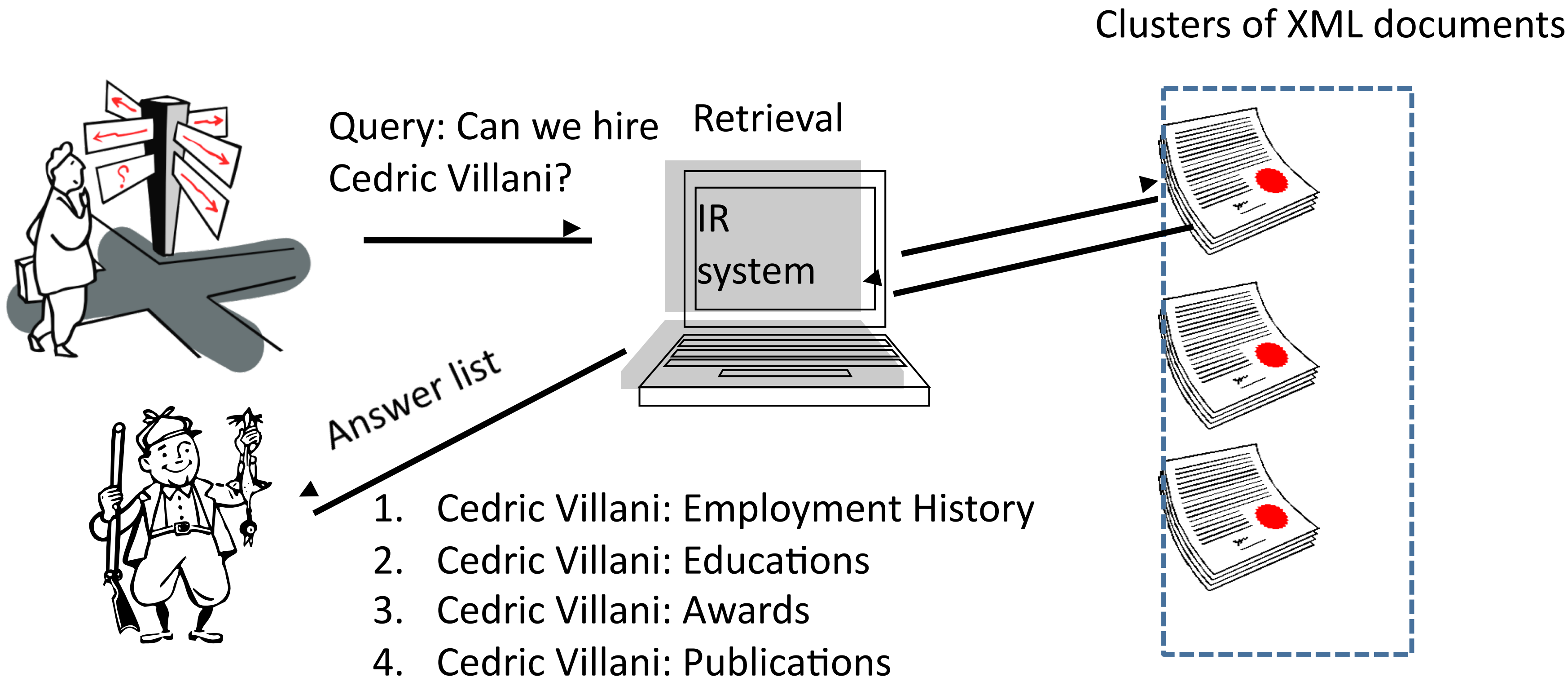


Problems:

1. Searches all the documents.
2. Computationally expensive.
3. Time consuming task.
4. Difficult to manage.

How to effectively manage the XML documents collection?

Querying XML Collections Using Clustering

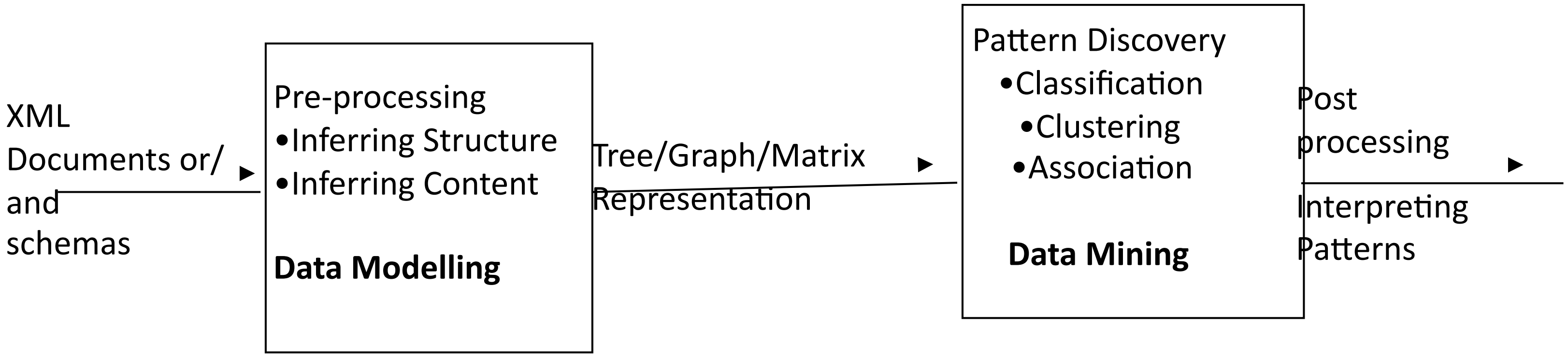


1. Cedric Villani: Employment History
2. Cedric Villani: Educations
3. Cedric Villani: Awards
4. Cedric Villani: Publications

Clustering of XML documents helps to:

1. Reduce the search space for querying
2. Reduce the time taken to respond to a query
3. Easy management of XML documents

XML Mining Process


XML Documents or/ and schemas

Pre-processing
- Inferring Structure
- Inferring Content

Data Modelling

Tree/Graph/Matrix Representation

Pattern Discovery
- Classification
- Clustering
- Association

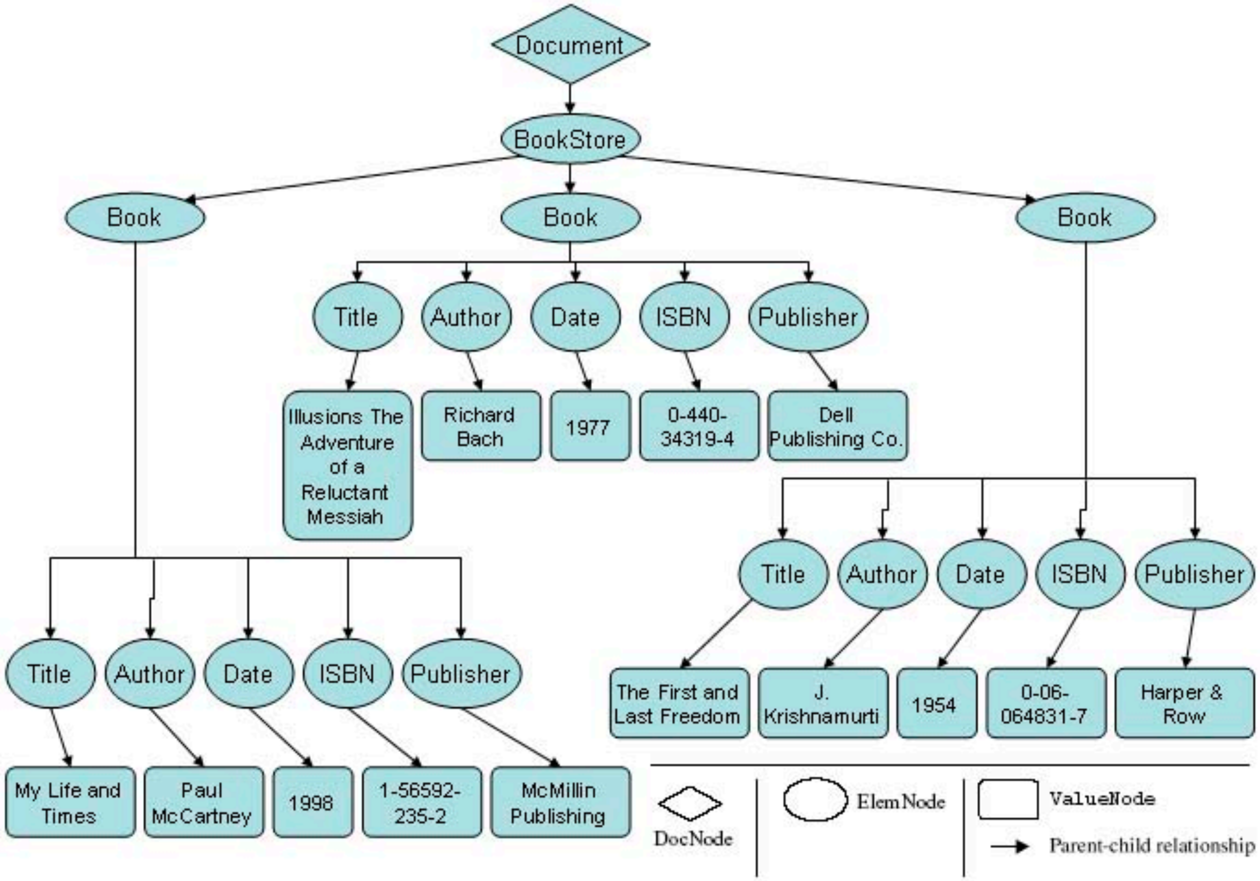
Data Mining

Post processing

Interpreting Patterns

XML: Data Model

XML can be represented as a matrix or a tree or a graph oriented data model.



XML Data Models: Matrix and Tree

d_1

<R>
<E$_1$>t_1, t_2, t_3
<E$_2$>t_4, t_3, t_6
<E$_3$>t_5, t_4, t_7
<E$_{3.1}$>t_5, t_2, t_1
<E$_{3.2}$>t_7, t_9

d_2

<R>
<E$_1$>t_1, t_4
<E$_2$>t_3, t_3
<E$_3$>t_4, t_7
<E$_{3.1}$>t_2, t_9
<E$_{3.2}$>t_2, t_7,
t_8, t_{10}

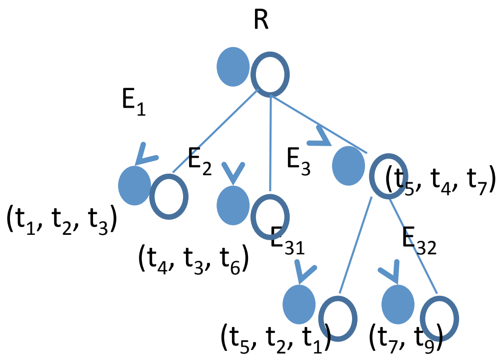
d_3

<R>
<E$_1$>t_1, t_2
<E$_2$>t_3, t_3
<E$_3$>t_5, t_4, t_7
<E$_{3.1}$>t_5, t_2, t_1
<E$_{3.2}$>t_7, t_9

d_4

<R>
<E$_1$>t_1, t_4
<E$_3$>t_4, t_7
<E$_3$>t_4, t_8
<E$_1$>t_1, t_4

Four Example XML Documents



Equivalent Tree Representation

	d_1	d_2	d_3	d_4
t_1	2	1	2	2
t_2	2	2	2	0
t_3	2	2	2	0
t_4	2	2	1	4
t_5	2	0	2	0
t_6	1	0	0	0
t_7	2	2	2	1
t_8	0	1	0	1
t_9	1	1	1	0
t_{10}	0	1	0	0

Equivalent Content Matrix Representation

	d_1	d_2	d_3	d_4
R/E$_1$	1	1	1	2
R/E$_2$	1	1	1	0
R/E$_3$/ E$_{3.1}$	1	2	1	0
R/E$_3$/ E$_{3.2}$	1	0	1	0
R/E$_3$	1	1	1	2

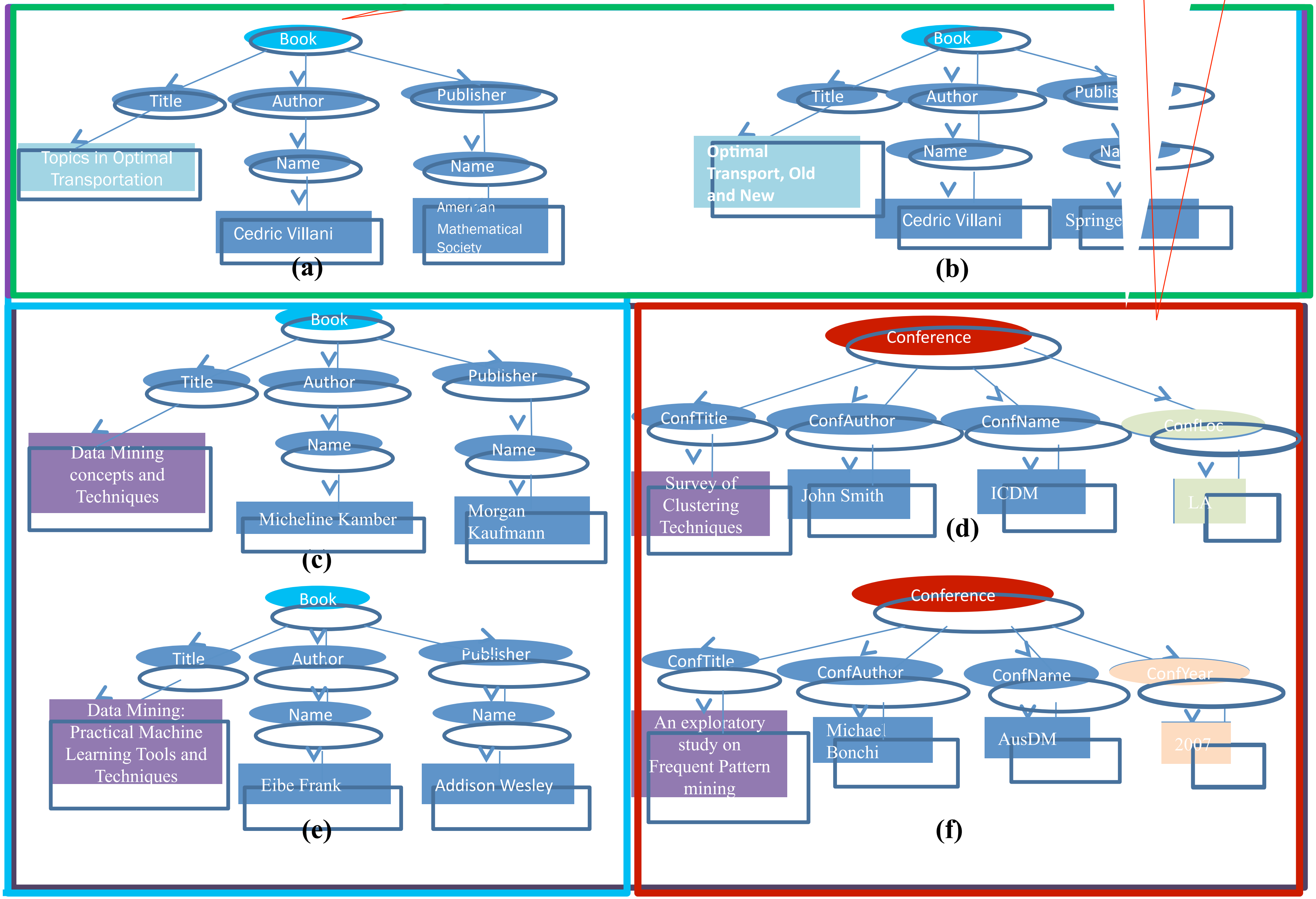
Equivalent Structure Matrix Representation

Some Mining Examples

- Grouping and classifying documents/schemas
- Mining frequent tree patterns
- Schema discovery
- Mining association rules
- Mining XML queries

Structure and Content-based cluster

Large-sized cluster on data mining



Implicit combination

- ❑Using Vector Space Model (VSM)

```
<Book Id="B105">
   <Title> Topics in Optimal Transportation
   </Title>
   <Author>
      <Name> Cedric Villani </Name>
   </Author>
   <Publisher>
        <Name> American Mathematical
                 Society </Name>
       <Place> NewYork</Place>
    </Publisher>
 </Book>
```

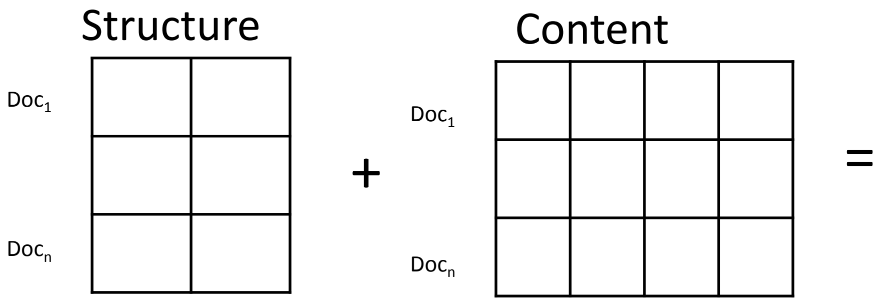
Topic	Optimal	Transport	Cedric	Villani	American	Mathematical	Society	NewYork

Book/Title	Book/Author/Name	Book/Publisher/Name	Book/Publisher/Place

XML clustering methods based on structure and content features

- Using linear combination (Tran & Nayak,2008, Yanming et al.,2008)

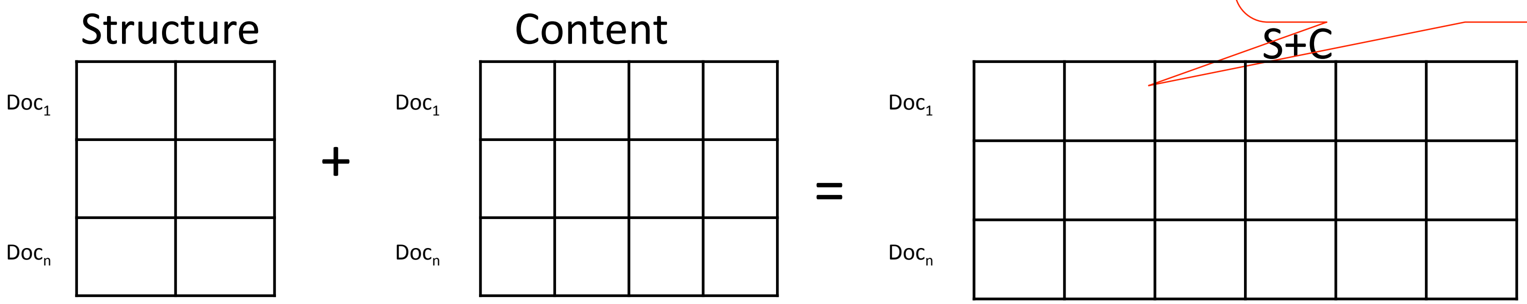
How to choose **α** and **β**?



$\boldsymbol{\alpha}\mathit{Sim}(\text{Structure}) + \boldsymbol{\beta}\mathit{Sim}(\text{Content})$

- Using Structure and Content Matrix concatenation (SCVM- Zhang et al.,2010)

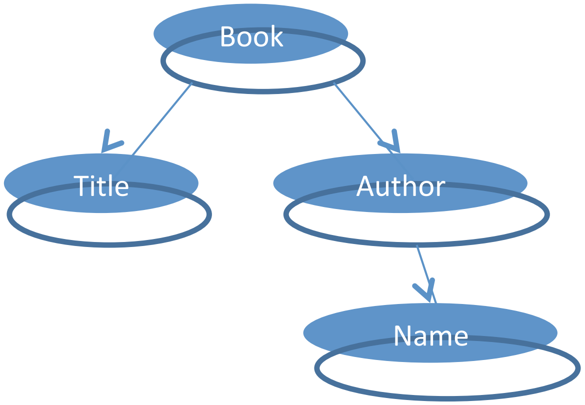
1.Large-sized matrix
2. No relationship between structure and content

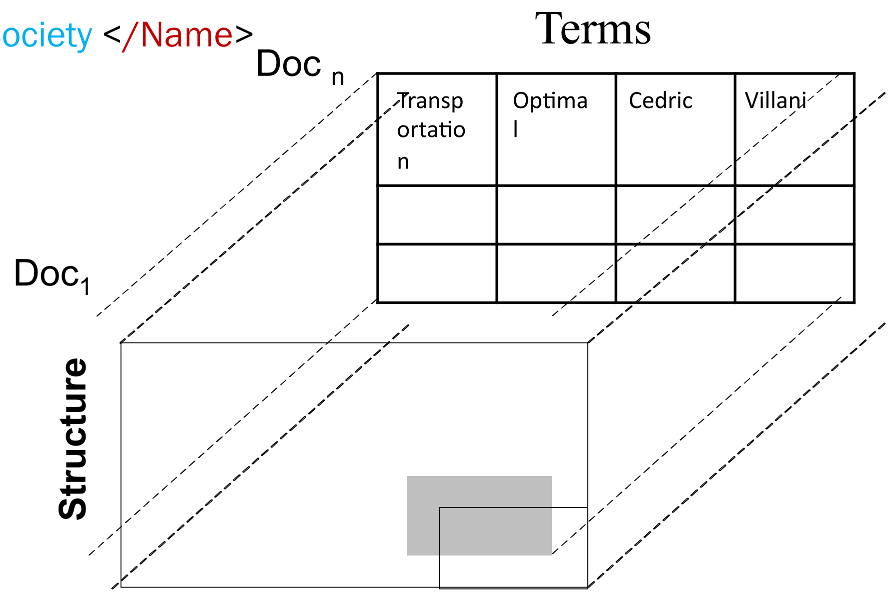


Explicit Combination

- Using Tensor Space Model (TSM)

```
<Book Id="B105">
   <Title> Topics in Optimal Transportation </Title>
   <Author>
      <Name>Cedric Villani</Name>
   </Author>
   <Publisher>
        <Name> American Mathematical Society </Name>
       <Place> NewYork</Place>
    </Publisher>
 </Book>
```





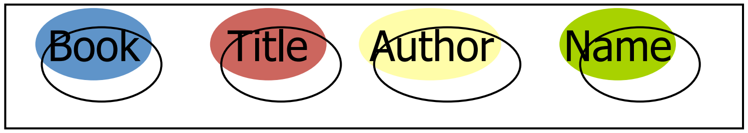
XML Frequent pattern mining

- Involves identifying the common or frequent patterns.
- Frequent patterns in XML documents based on the structure.
- Frequent pattern mining can be used as kernel functions for different data mining tasks:
  - Clustering
  - Link analysis
  - Classification

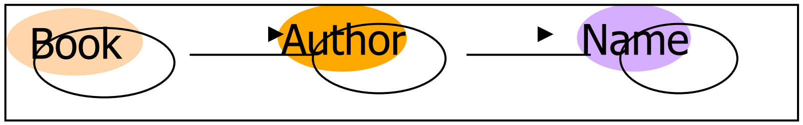
What is meant by frequent patterns

- Common patterns based on an user-defined support threshold (min_supp)
- Provide summaries of the data
- Patterns could be itemsets, subpaths, **subtrees**, subgraphs

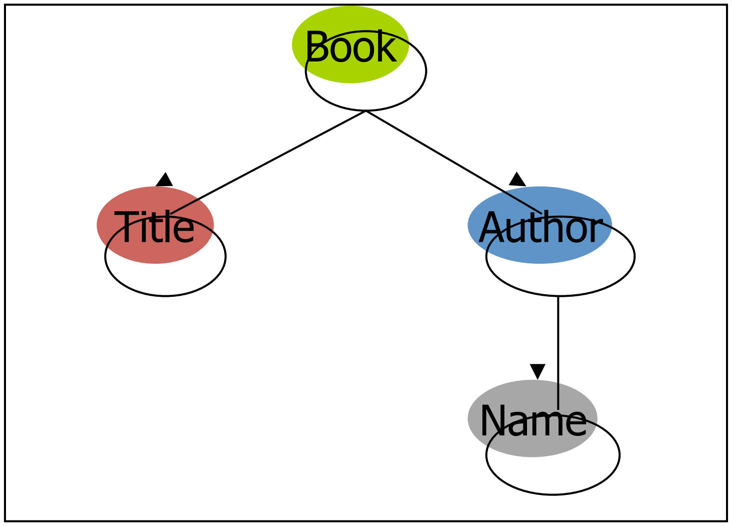
Itemset



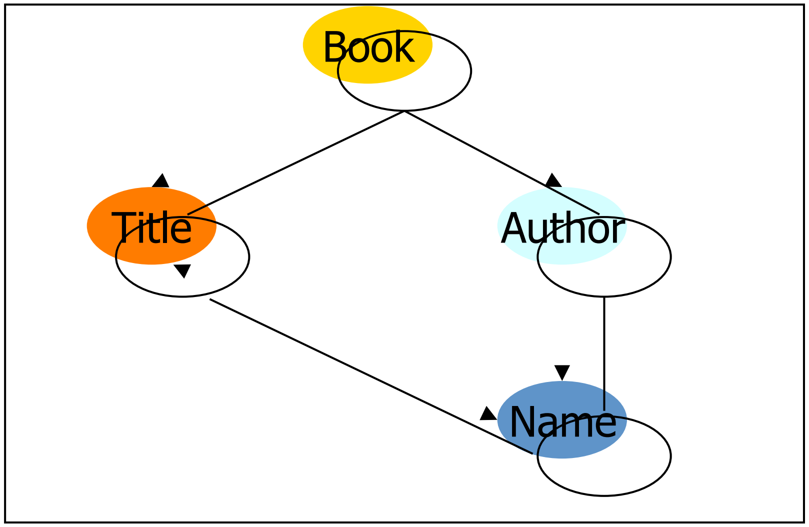
Subpath



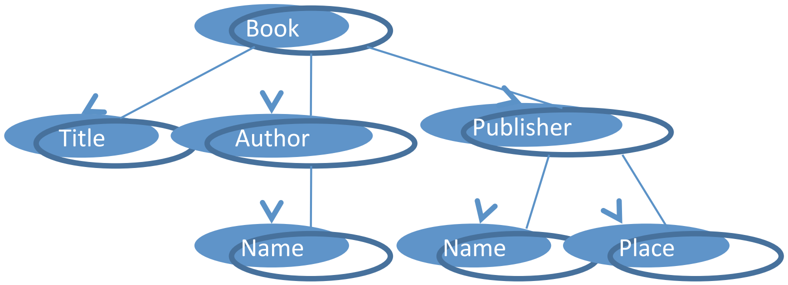
Subtree



Subgraph



Types of subtrees

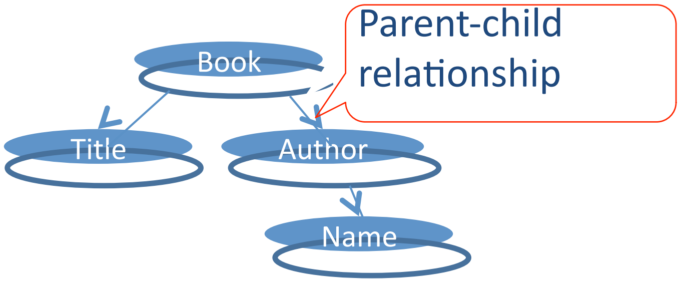


- □ **On node relationship**
- □ **On conciseness**

On node relationship

Induced subtree

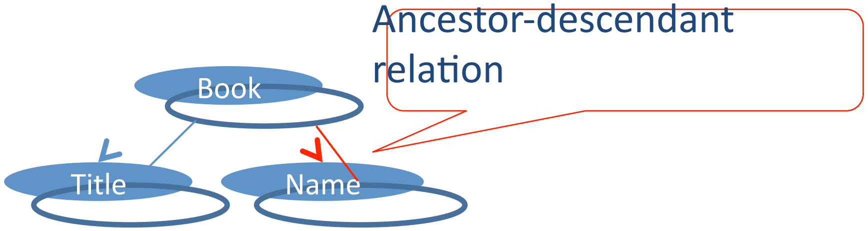
- Preserves **parent-child** relationship



Parent-child relationship

Embedded subtree

-Preserves **ancestor-descendant** relationship



Ancestor-descendant relation

On conciseness

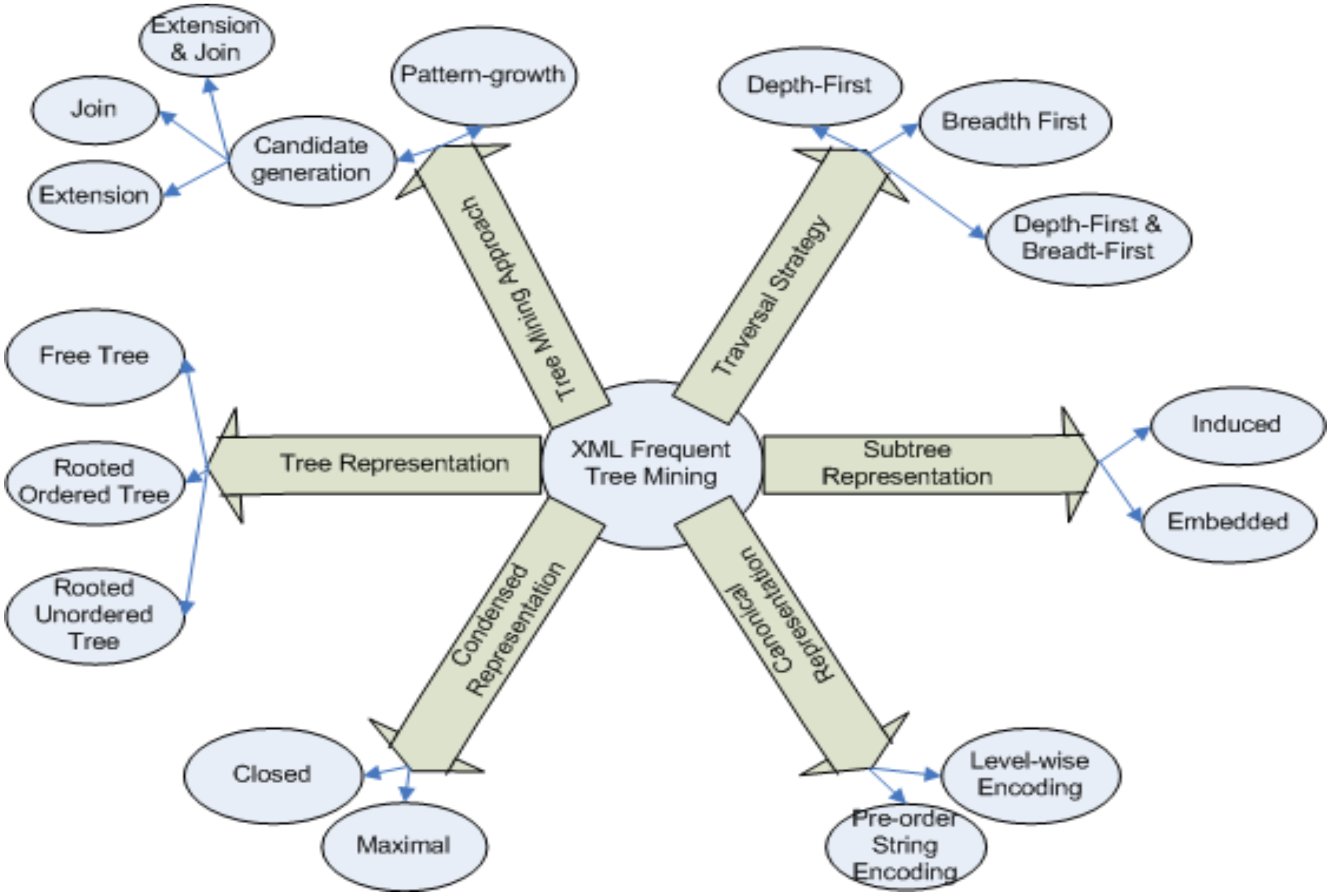
- **Maximal frequent subtrees**

  In a given document tree dataset, $DT = \{DT_1, DT_2, DT_3, \ldots, DT_n\}$, if there exists two frequent subtrees DT' and DT'', DT' is said to be maximal of DT'' iff $DT' \supset_t DT''$, $supp(DT') \leq supp(DT'')$;

- **Closed frequent subtrees**

  A frequent subtree DT' is closed of DT'' iff for every $DT' \supset_t DT''$, $supp(DT') = supp(DT'')$

Frequent Tree Mining: Methods Status



Future Directions: XML Mining

- Scalability
  - Incremental Approaches
- Combining structure and content efficiently
  - Advanced data representational models and mining methods
- Application Context

Reading Articles

- R. Nayak (2008) “XML Data Mining: Process and Applications”, Chapter 15 in “Handbook of Research on Text and Web Mining Technologies”, Ed: Min Song and Yi-Fang Wu. Publisher: Idea Group Inc., USA. PP. 249 -271.
- S. Kutty and R. Nayak (2008) “Frequent Pattern Mining on XML documents”, Chapter 14 in “Handbook of Research on Text and Web Mining Technologies”, Ed: Min Song and Yi-Fang Wu. Publisher: Idea Group Inc., USA. PP. 227 -248.
- R. Nayak (2008) “Fast and Effective Clustering of XML Data Utilizing their Structural Information”. Knowledge and Information Systems (KAIS). Volume 14, No. 2, February 2008 pp 197-215.
- C. C. Aggarwal, N. Ta, J. Wang, J. Feng, and M. Zaki, "Xproj: a framework for projected structural clustering of xml documents," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining San Jose, California, USA: ACM, 2007, pp. 46-55.
- Nayak, R., & Zaki, M. (Eds.). (2006). Knowledge Discovery from XML documents: PAKDD 2006 Workshop Proceedings (Vol. 3915): Springer-Verlag Heidelberg.
- NAYAK, R. AND TRAN, T. 2007. A progressive clustering algorithm to group the XML data by structural and semantic similarity. *International Journal of Pattern Recognition and Artificial Intelligence 21, 4, 723–743.*
- Y. Chi, S. Nijssen, R. R. Muntz, and J. N. Kok, "Frequent Subtree Mining- An Overview," in Fundamenta Informaticae. vol. 66: IOS Press, 2005, pp. 161-198.
- L. Denoyer and P. Gallinari, "Report on the XML mining track at INEX 2005 and INEX 2006: categorization and clustering of XML documents," SIGIR Forum, vol. 41, pp. 79-90, 2007.
- BERTINO, E., GUERRINI, G., AND MESITI, M. 2008. Measuring the structural similarity among XML documents and DTDs. Intelligent Information Systems 30, 1, 55–92.
- BEX, G. J., NEVEN, F., AND VANSUMMEREN, S. 2007. Inferring XML schema definitions from XML data. In Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, Austria, 998–1009.
- BILLE, P. 2005. A survey on tree edit distance and related problems. Theoretical Computer Science 337, 1-3, 217–239.
- BONIFATI, A., MECCA, G., PAPPALARDO, A., RAUNICH, S., AND SUMMA, G. 2008. Schema mapping verification:the spicy way. In EDBT. 85–96.
- A. Algergawy, M. Mesiti and R. Nayak (forthcoming) “XML Data Clustering: An Overview”, ACM Computing Surveys, Accepted 25th October, 2009, (42 pages) Tentatively assigned to appear in Vol. 44, issue # 2 (June 2012).
- A. Algergawy, R. Nayak, Gunter Saake (2010) Element Similarity Measures in XML Schema Matching. Information Sciences, 180 (2010), 4975-4998.
- Kutty, S., R. Nayak, and Y. Li. (2011) XML documents clustering using tensor space model, in proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2011), Shenzen,China

Related Publications

- BOUKOTTAYA, A. AND VANOIRBEEK, C. 2005. Schema matching for transforming structured documents. In *DocEng'05. 101–110.*
- FLESCA, S., MANCO, G., MASCIARI, E., PONTIERI, L., AND PUGLIESE, A. 2005. Fast detection of XML structural similarity. *IEEE Trans. on Knowledge and Data Engineering 17, 2, 160–175.*
- GOU, G. AND CHIRKOVA, R. 2007. Efficiently querying large XML data repositories: A survey. *IEEE Trans. on Knowledge and Data Engineering 19, 10, 1381–1403.*
- NAYAK, R. AND IRYADI,W. 2007. XML schema clustering with semantic and hierarchical similarity measures. *Knowledge-based Systems 20, 336–349.*
- Kutty, S., Nayak, R., & Li, Y. (2007). *PCITMiner- Prefix-based Closed Induced Tree Miner for finding closed induced frequent subtrees.* Paper presented at the the Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia.
- TAGARELLI, A. AND GRECO, S. 2006. Toward semantic XML clustering. In *SDM 2006. 188–199.*
- Rusu, L. I., Rahayu, W., & Taniar, D. (2007). Mining Association Rules from XML Documents. In A. Vakali & G. Pallis (Eds.), *Web Data Management Practices:*
- Li, H.-F., Shan, M.-K., & Lee, S.-Y. (2006). Online mining of frequent query trees over XML data streams. In *Proceedings of the 15th international conference on World Wide Web* (pp. 959-960). Edinburgh, Scotland: ACM Press.
- Zaki, M. J.:(2005):Efficiently mining frequent trees in a forest: algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering*, 17 (8): 1021-1035
- Wan, J. W. W. D., G. (2004). Mining Association rules from XML data mining query. *Research and practice in Information Technology, 32*, 169-174.

Overview

- Introduction ✔
- The Hidden Web ✔
- XML ✔
- DSML
- The Semantic Web
- Conclusion

Domain-Specific Markup Languages: Development and Applications

Aparna Varde
Department of Computer Science
Montclair State University
Montclair, NJ, USA

(vardea@mail.montclair.edu)

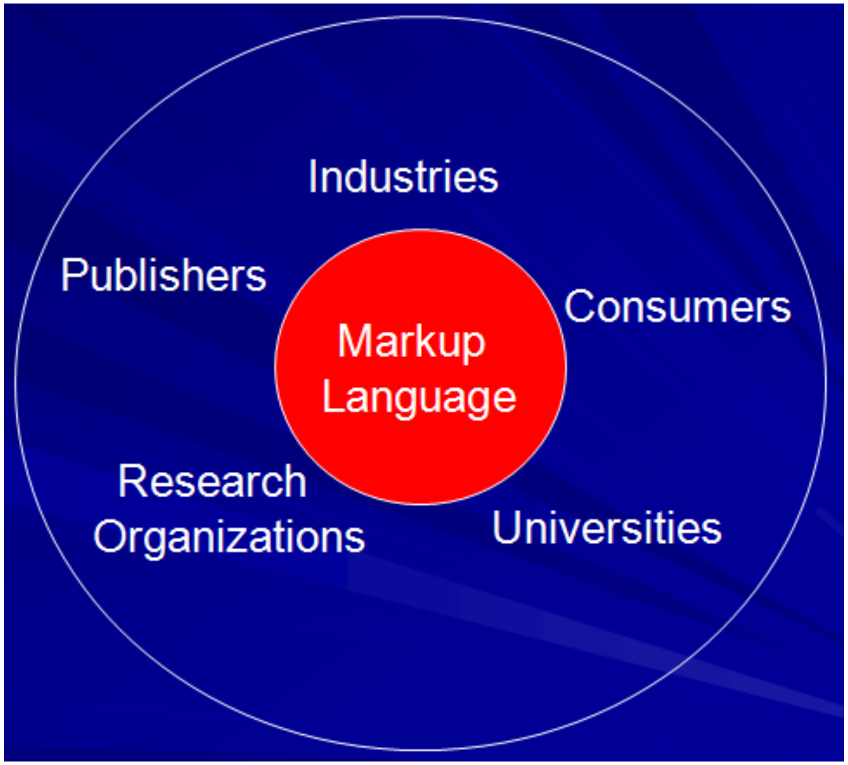
Presented by Richi Nayak

What is a Domain-Specific Markup Language (DSML)

- Medium of communication for users of the domain

- Follows XML syntax

- Encompasses the semantics of the domain



DSML users

Examples of DSMLs

- MML: Medical Markup Language
- CML: Chemical Markup Language
- MatML: Materials Markup Language
- WML: Wireless Markup Language
- MathML: Mathematics Markup Language

Need for DSMLs in scientific data management

- Help to capture semantics from a domain perspective
- Serve as worldwide standards for communication in the given scientific domain
- Facilitate information retrieval using XML based standards
- Assist in mining scientific data by guiding the discovery of knowledge as a domain expert would

MathML: Cedric Villani

- Consider the works of Cedric Villani, following the example used earlier in the tutorial
- An equation $H = \int \rho \log \rho \, dV$ is used in Villani's works in optimal transportation and curvature
- In this equation ρ is the density, V is the volume, such that $\mu = \rho V$, and H, denoting $H(\mu)$, is the information, i.e.,negative of the entropy

MathML: Presentation Markup in Villani’s works

```
<mrow>
    <mi> H </mi>
    <mo> = </mo>
    <mo> ∫ </mo>
    <mi> ρ </mi>
     <mo> log </mo>
    <mi> ρ </mi>
    <mo> d</mo>
    <mi> V <mi>
</mrow>
```

Interesting issues in DSMLs

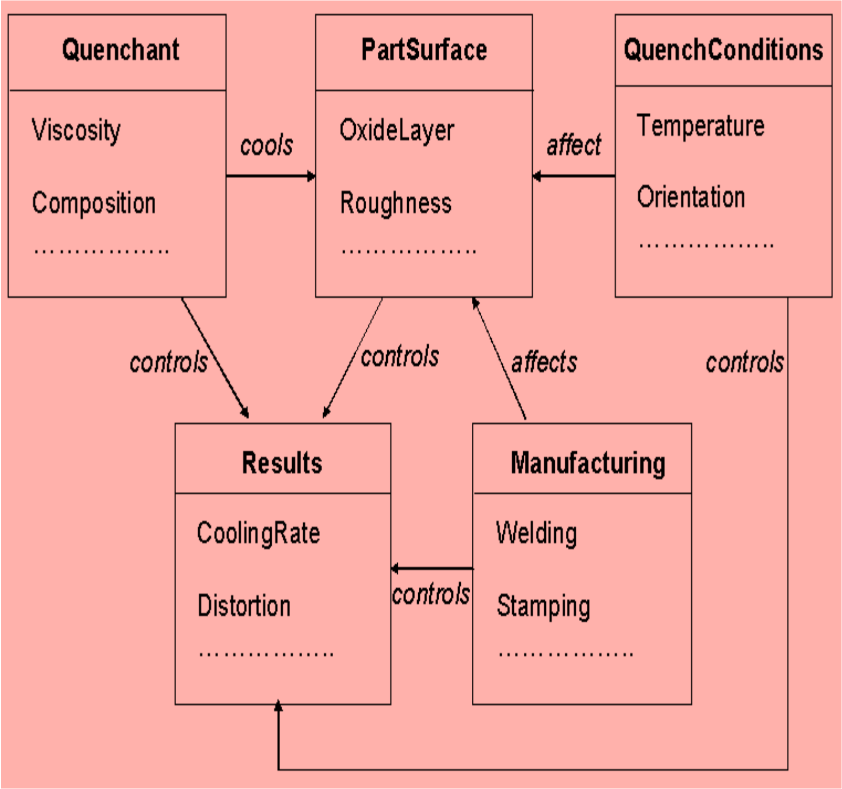
- DSML developmental steps with a view to aid scientific data management
- Application of XML constraints to preserve semantics
- XQuery for Information retrieval
- Mining DSML documents

DSML developmental steps

1. Data Modeling
2. Ontology Creation
3. Schema Development

Data Modeling

- Tools such as ER models are useful in modeling the data
- This helps create a picture of entities in the domain, view their attributes and understand their relationships
- Figure shows an example of an ER diagram in a Materials Science process called Quenching or rapid cooling during heat treatment
- ER modeling provides good mapping with real-world scenarios helpful in scientific data management
- E.g., attributes here represent features of interest in data mining techniques useful in discovering knowledge from data



Example of ER model a Materials Science process

Ontology Creation

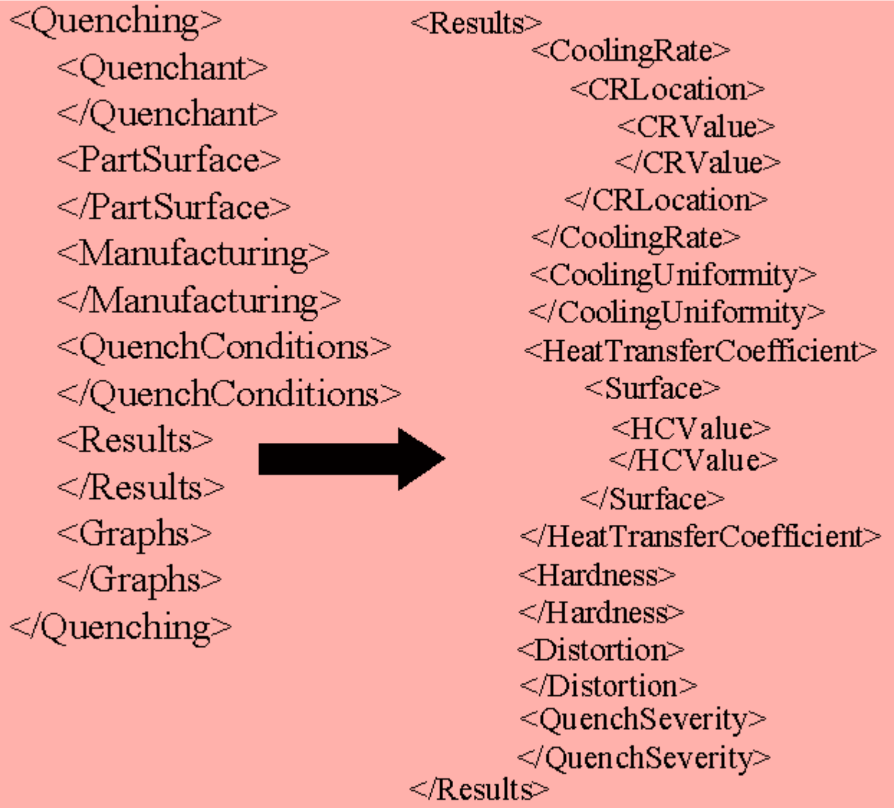
- Ontology is a formal manner of knowledge representation
- Should be formalized using standards: RDF, OWL
- E.g., Synonyms depicted using "sameAs" in OWL as shown in the figure (Quenchant also called cooling medium etc.)
- Ontology creation is useful in preserving semantics in scientific data management
- In knowledge discovery from scientific data, it is important to capture the domain-specific meaning of terms w. r. t. context, for correct interpretation of results

```
<Quenchant rdf:ID="Quenchant">
<owl:sameAs rdf:resource="#CoolingMedium" />
</Quenchant>
<PartSurface rdf:ID="PartSurface">
<owl:sameAs rdf:resource="#ProbeSurface" />
<owl:sameAs rdf:resource="#WorkpieceSurface" />
</PartSurface>
<Manufacturing rdf:ID="Manufacturing">
<owl:sameAs rdf:resource="#Production" />
</Manufacturing>
```

Partial Snapshot of Ontology in Materials Science

Schema Development

- Schema provides the structure of the markup language
- E-R model, requirements specification and ontology serve as the basis for schema design
- Schema development can involve several iterations, which can include discussions with standards bodies
- A good schema implies more systematic data storage capturing domain semantics which is useful in scientific data management
- XML constraints help preserve semantic restrictions



```
<Quenching>
    <Quenchant>
    </Quenchant>
    <PartSurface>
    </PartSurface>
    <Manufacturing>
    </Manufacturing>
    <QuenchConditions>
    </QuenchConditions>
    <Results>
    </Results>
    <Graphs>
    </Graphs>
</Quenching>
```

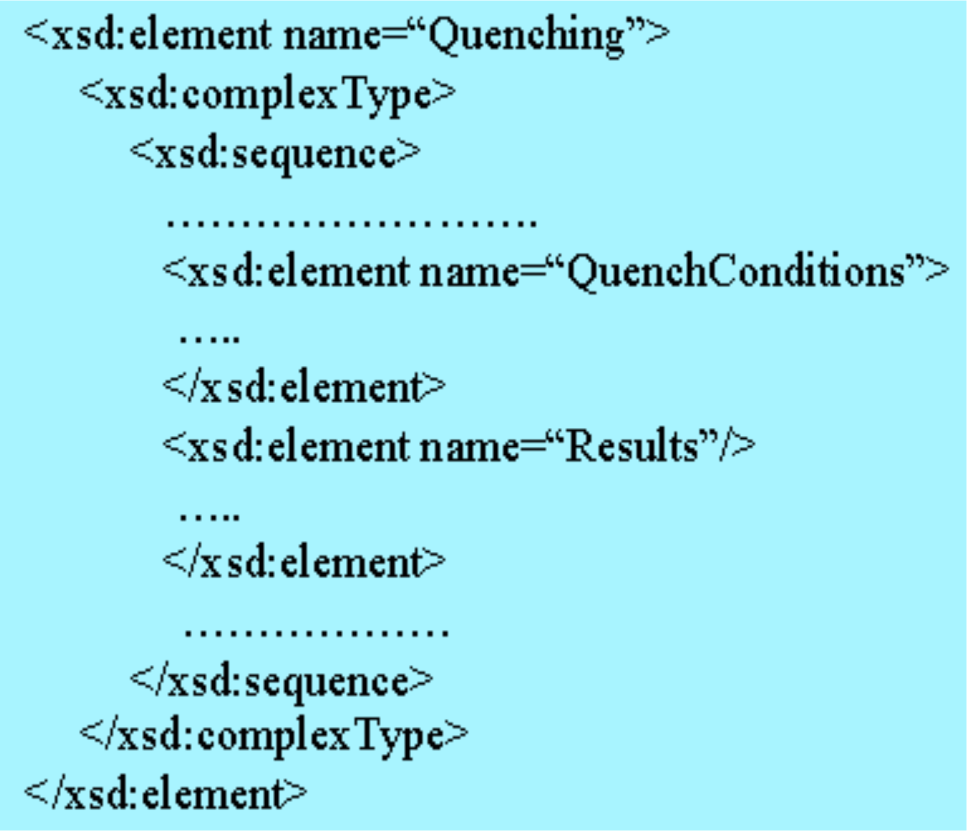
```
<Results>
        <CoolingRate>
            <CRLocation>
                <CRValue>
                </CRValue>
            </CRLocation>
        </CoolingRate>
        <CoolingUniformity>
        </CoolingUniformity>
        <HeatTransferCoefficient>
            <Surface>
                <HCValue>
                </HCValue>
            </Surface>
        </HeatTransferCoefficient>
        <Hardness>
        </Hardness>
        <Distortion>
        </Distortion>
        <QuenchSeverity>
        </QuenchSeverity>
</Results>
```

Example Partial Snapshot of Schema in Materials Science

Application of XML Constraints in DSMLs

1. Sequence Constraint
2. Choice Constraint
3. Key Constraint
4. Occurrence Constraint

Sequence Constraint

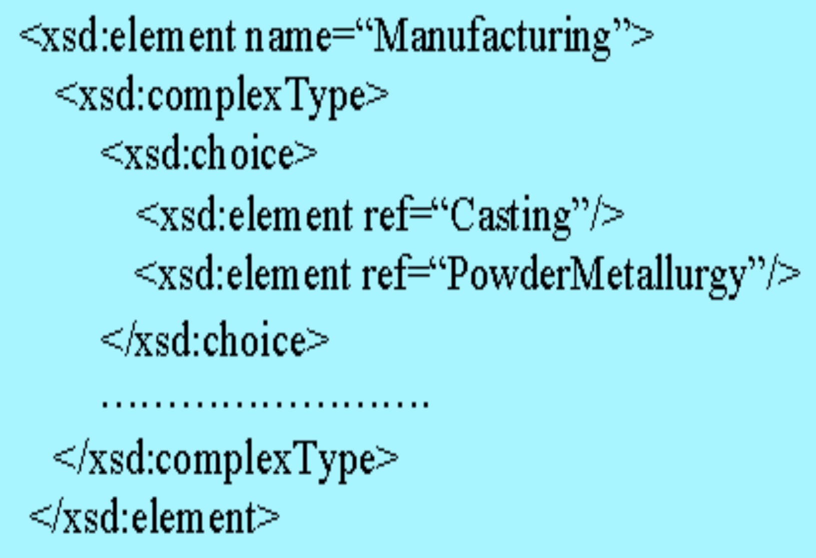


```
<xsd:element name=“Quenching”>
    <xsd:complexType>
        <xsd:sequence>
          ........................
          <xsd:element name=“QuenchConditions”>
           .....
          </xsd:element>
          <xsd:element name=“Results”/>
           .....
          </xsd:element>
            .................
        </xsd:sequence>
    </xsd:complexType>
</xsd:element>
```

Sequence Constraint example in a scientific domain

- Used to declare elements to occur in a certain order as recommended in a given domain
- Examples:
  - Storing the input conditions of a Materials Science experiment before its results
  - Storing details of a medical diagnostic process before its observations

Choice Constraint

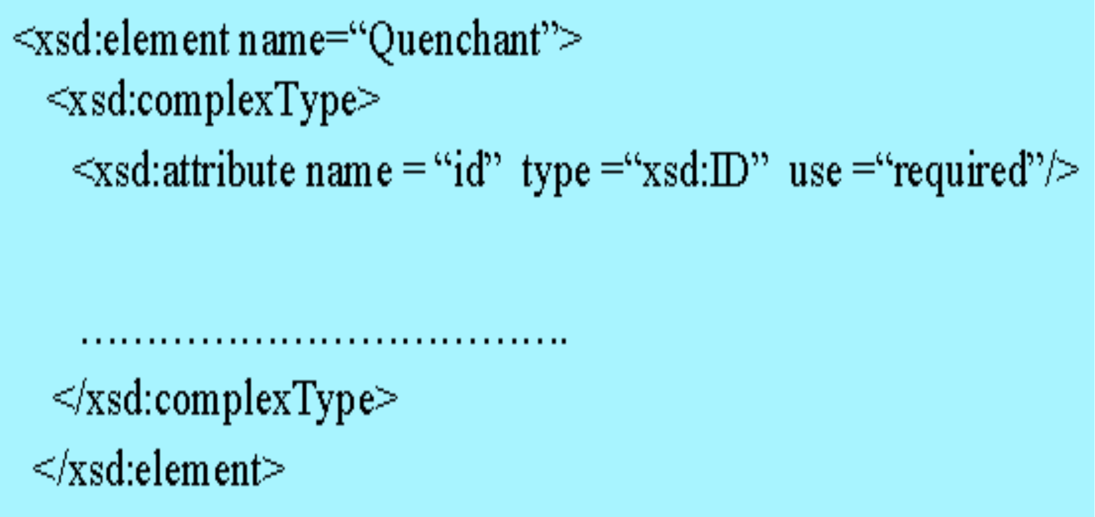


```
<xsd:element name="Manufacturing">
  <xsd:complexType>
     <xsd:choice>
        <xsd:element ref="Casting"/>
        <xsd:element ref="PowderMetallurgy"/>
     </xsd:choice>
     ........................
  </xsd:complexType>
 </xsd:element>
```

Choice Constraint example in a scientific domain

- Used to declare domain-specific mutually exclusive elements, i.e., only one of them can exist
- Examples
  - In Materials Science, a part can be manufactured by either *Casting* or *Powder Metallurgy*, not both
  - In Medicine, a tumor can be *malignant* or *benign*, not both

Key Constraint



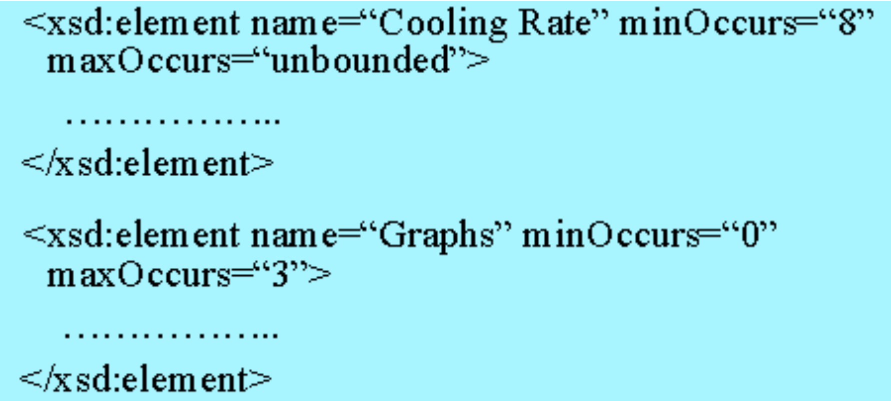
```
<xsd:element name=“Quenchant”>
  <xsd:complexType>
    <xsd:attribute name = “id”  type =“xsd:ID”  use =“required”/>


     …………………………………..
  </xsd:complexType>
 </xsd:element>
```

Key Constraint example in a scientific domain

- Used to declare an attribute to be a unique identifier as required in the domain
- Example:
  - In Heat Treating, ID of Quenchant, for a given quenching (rapid cooling) process
  - In Medicine, name of patient for a given diagnosis

Occurrence Constraint



```
<xsd:element name=“Cooling Rate” minOccurs=“8”
  maxOccurs=“unbounded”>
    ……………...
</xsd:element>
<xsd:element name=“Graphs” minOccurs=“0”
  maxOccurs=“3”>
    ……………...
</xsd:element>
```

Occurrence Constraint example in a scientific domain

- Used to declare minimum and maximum permissible occurrences of an element with respect to the domain
- Example:
  - In Materials, Cooling Rate must be recorded for at least 8 points, no upper bound
  - In same context, at most 3 Graphs are stored, no lower bound
  - In medicine, an upper and lower bound can be imposed on number of diagnoses per patient w.r.t. the application

Information Retrieval using XQuery

- XQuery (XML Query Language) developed by the World Wide Web Consortium (W3C)
- XQuery can retrieve information stored using domain-specific markup languages designed with XML tags
- DSMLs facilitate this by allowing additional tags to be used for storage to enhance querying efficiency, by anticipating typical user queries
- Example: In Medicine, place additional tags within the details of <Patient> to separate their <PersonalData> from their <DiagnosticData> because more queries are likely to be executed on the patients' diagnosis

Mining DSML documents

- Using DSMLs for data mining enhances the effectiveness of results using techniques such as association rules and clustering
- This is because the domain-specific tags guide the mining process as a domain expert would
- This applies to semi-structured XML-based data and also plain text documents in the domain that can be converted to XML format using the DSML tags

Association Rule Mining

- Association Rules are of the type A => B
  - Example: fever => flu
- Interestingness measures
  - Rule confidence : P(B/A)
  - Rule support: P(AUB)
- Rules derived as shown in example
- Data stored using DSMLs facilitates rule derivation over semi-structured text
- This is also useful for plain text sources converted to semi-structured format by capturing relevant data using the tags
- In the absence of such tags, if we mined rules from plain text, we could get rules such as patient => diagnosis because these terms co-occur frequently, but such rules are not meaningful
- Thus DSMLs capture semantics in mining

❑ <fever> yes </fever> in 90/100 instances

❑ <flu> yes </flu> in 70/100 instances

  ❑ 60 of these in common with fever

❑ Association Rule

  fever = yes => flu = yes

❑ Rule confidence: 60/90 = 67%

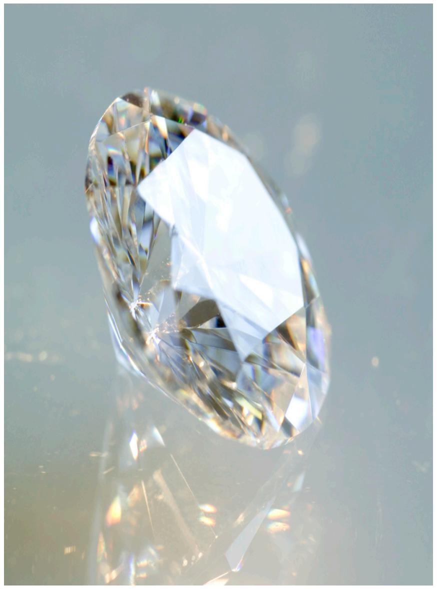
❑ Rule support: 60/100 = 60%

Challenges in scientific data management with XML and DSMLs

1. Effectively modeling both structure and content features for XML documents to adequately represent scientific data and investigating how DSMLs can be useful here
2. Combining structure and content features in different types of data models which do not affect the scalability of the mining process
3. Integrating background knowledge of scientific processes in XML mining algorithms and harnessing DSMLs here
4. Developing procedures to enhance a document representation to reflect the semantic structure embedded in the scientific data
5. Developing new standards as needed especially to foster knowledge discovery by synergizing XML and DSMLs

Summary: XML and DSML

- Applications with large amounts of raw strategic data in XML will be there.
- XML data mining techniques will be a plus for the adoption of XML as a data model for modern applications.
- XML mining, in order to be more than a temporary fade, must deliver useful solutions for practical applications.



Overview

- Introduction ✓
- The Hidden Web ✓
- XML ✓
- The Semantic Web
- Conclusion

Overview

- Introduction ✓
- The Hidden Web ✓
- XML ✓
- DSML ✓
- The Semantic Web
- Conclusion

The Semantic Web

Fabian M. Suchanek

INRIA Saclay

Paris, France

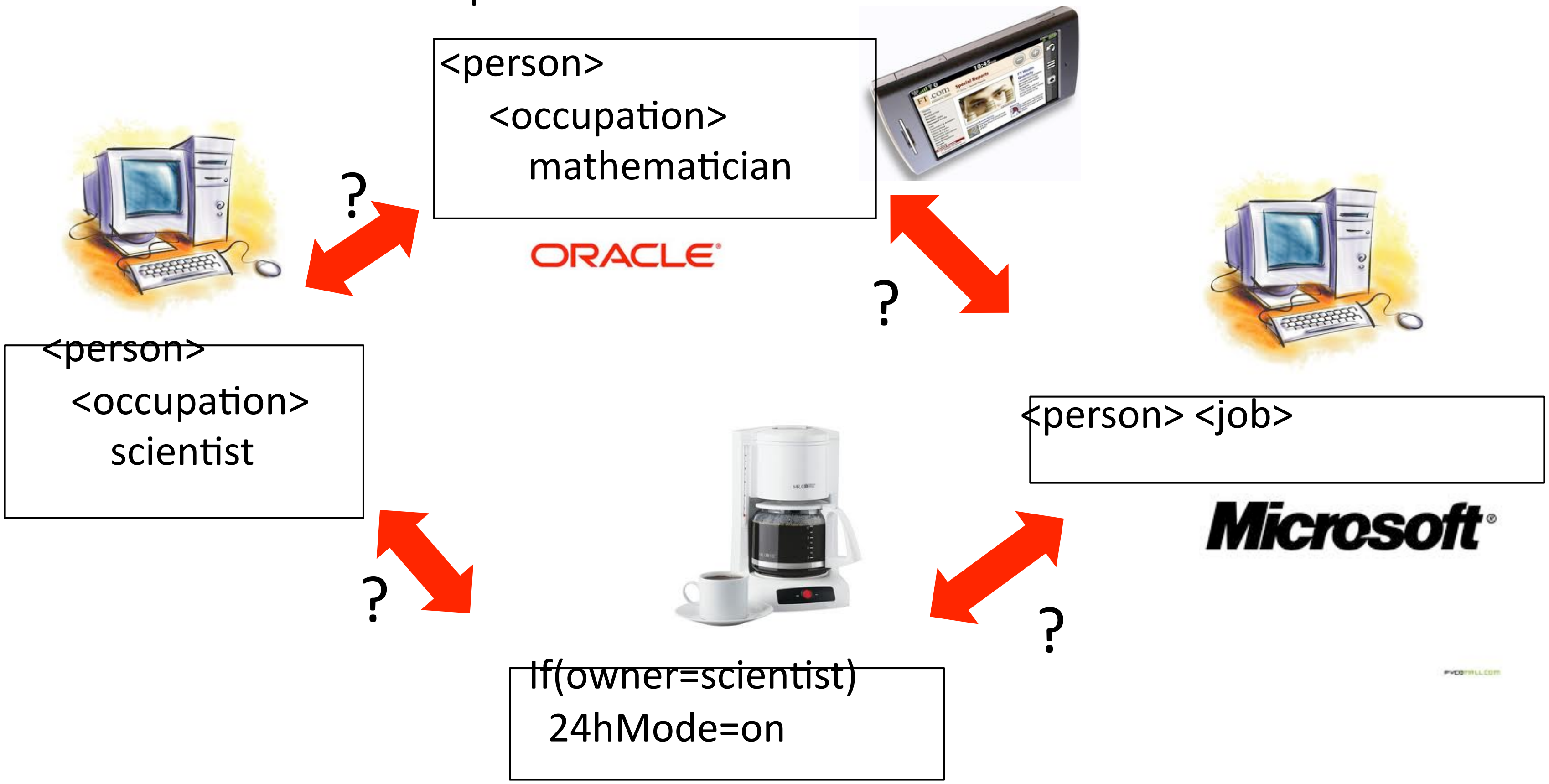
http://suchanek.name

SW: Motivation

We just saw how to express structured data in a standardized format, XML.
We also saw how DSMLs can provide semantic standards.

But even for XML documents in a DSML, data exchange is not trivial, in particular

- if the data resides on different devices
- if the domains are modeled by different people
- if we need taxonomic structure
- if we need more complex constraints



<person>
<occupation>
mathematician

<person>
<occupation>
scientist

<person> <job>

If(owner=scientist)
24hMode=on

SW: Use cases

Examples:

- Booking a flight

  Interaction between office computer, flight company, travel agency, shuttle services, hotel, my calendar

- Finding a restaurant

  Interaction between mobile device, map service, recommendation service, restaurant reservation service

- Intelligent home

  Fridge knows my calendar, orders food if I am planning a dinner

- Intelligent cars

  Car knows my schedule, where and when to get gas, how not to hit other cars, what are the legal regulations

- Web search

  Combining information from different sources to figure out whether to hire Cedric Villani

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

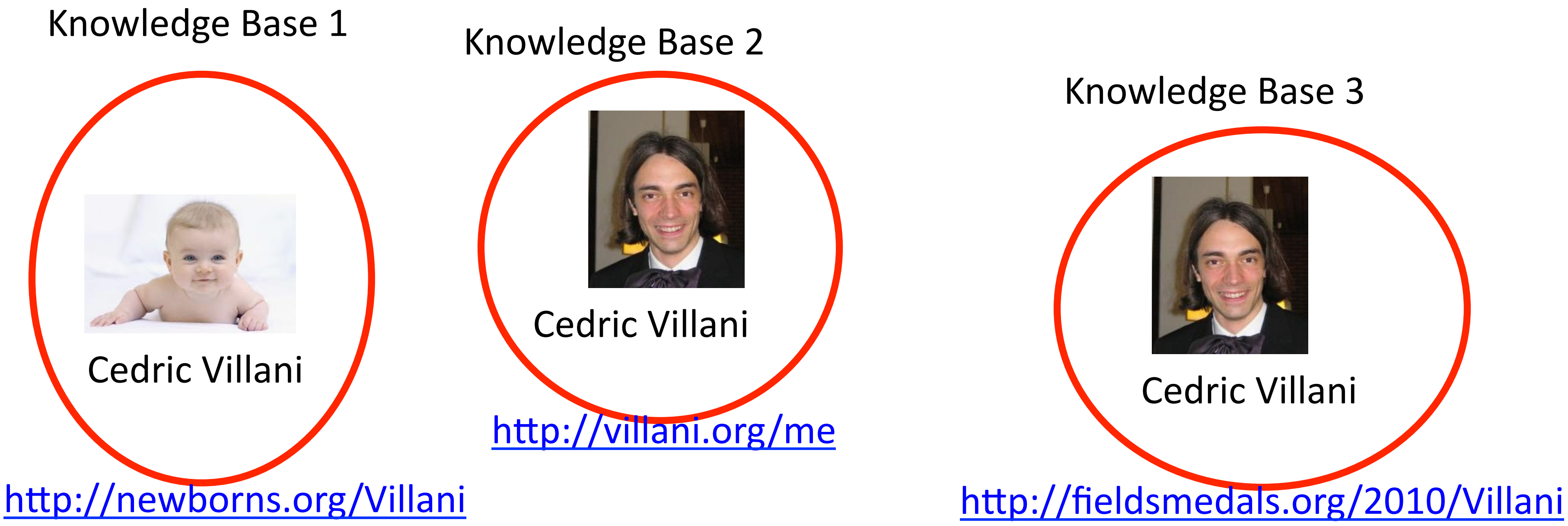
- make computers „understand" the data they store
- allow them to reason about information
- allow them to share information across different systems

For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

- identifying entities in a globally unique way (URIs)
- defining semantics in a machine-readable way (RDF)
- defining taxonomies (RDFS)
- defining logical consistency in a uniform way (OWL)
- storing ontologies (N3, XML, RDFa)
- sharing ontologies (Cool URIs)

SW: URIs

A **Uniform Resource Identifier** (URI) is a string of characters used to identify an entity on the Internet



Knowledge Base 1

Cedric Villani

http://newborns.org/Villani

Knowledge Base 2

Cedric Villani

http://villani.org/me

Knowledge Base 3

Cedric Villani

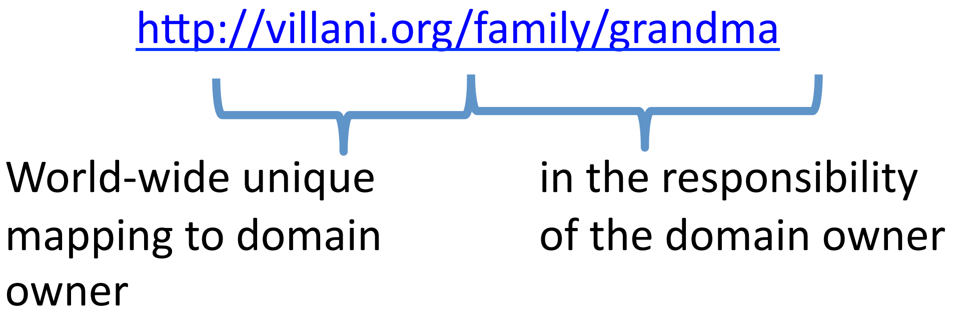
http://fieldsmedals.org/2010/Villani

The same thing
can have different URIs,
but different things
always have
different URIs

[URI]

SW: URIs

A **Uniform Resource Identifier** (URI) is a string of characters used to identify an entity on the Internet



http://villani.org/family/grandma

World-wide unique mapping to domain owner

in the responsibility of the domain owner

⇒ There should be no URI with two meanings

⇒ People can invent all kinds of URIs

- a company can create URIs to identify its products
- an organization can assign sub-domains and each sub-domain can define URIs
- individual people can create URIs from their homepage
- people can create URIs from any URL for which they have exclusive rights to create URIs

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

- make computers „understand“ the data they store
- allow them to reason about information
- allow them to share information across different systems

For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

- identifying entities in a globally unique way (URIs) ✔
- defining semantics in a machine-readable way (RDF)
- defining taxonomies (RDFS)
- defining logical consistency in a uniform way (OWL)
- storing ontologies (N3, XML, RDFa)
- sharing ontologies (Cool URIs)

SW: RDF

The **Resource Description Framework** (RDF) is a knowledge representation formalism that is very similar to the entity-relationship model.

Assume we have the following URIs:

A URI for Villani:	http://villani.org/me
A URI for "winning a prize":	http://inria.fr/rdf/dta#won
A URI for the Fields medal:	http://mathunion.com/FieldsMedal

An **RDF statement** is a triple of 3 URIs: The subject, the predicate and the object.

http://villani.org/me http://inria.fr/rdf/dta#won http://mathunion.com/FieldsMedal

We can understand an RDF statement as a First Order Logic statement with a binary predicate

won(Villani, FieldsMedal)

[RDF]

SW: Namespaces

A **namespace** is an abbreviation for the prefix of a URI.

```
@prefix  v:         http://villani.org/
@prefix  inria:     http://inria.fr/rdf/dta#
@prefix  m:         http://mathunion.com/
```

An **RDF statement** is a triple of 3 URIs: The subject, the predicate and the object.

```
http://villani.org/me      http://inria.fr/rdf/dta#won       http://mathunion.com/FieldsMedal
```

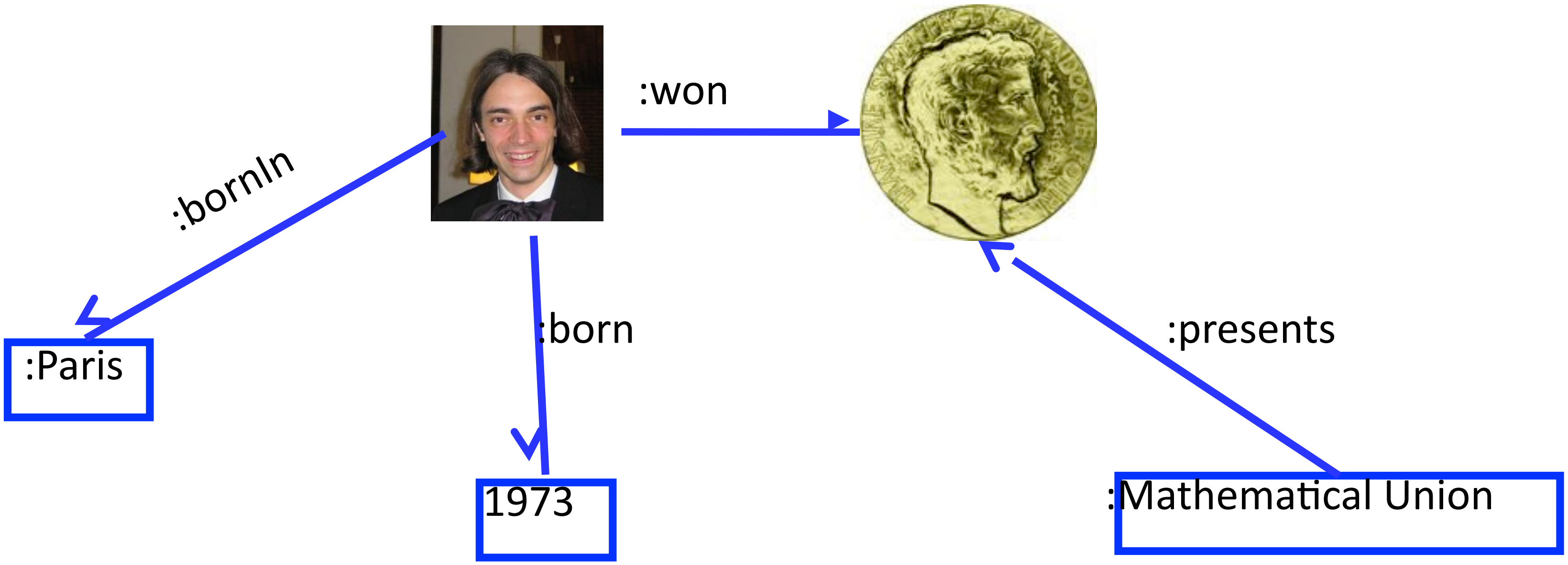
... with the above namespaces, this becomes...

```
v:me                       inria:won                         m:prize
```

The **default name space** is indicated by “:”

SW: Ontologies

Example RDF-graph:



:won

:bornIn

:born

:presents

:Paris

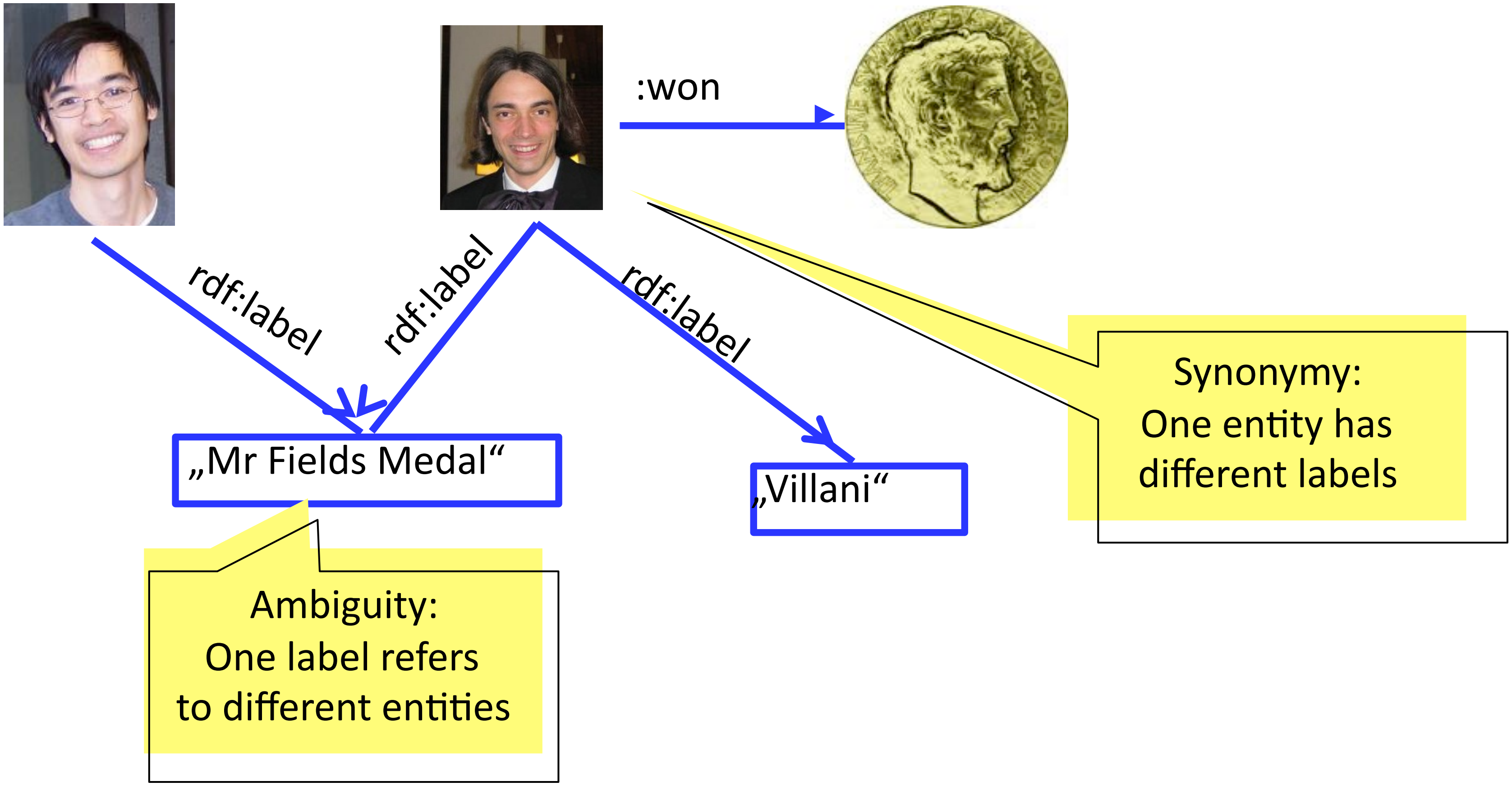
1973

:Mathematical Union

We call such a
graph an
ontology

SW: Labels

RDF distinguishes between the entities and their labels.



Synonymy:
One entity has
different labels

Ambiguity:
One label refers
to different entities

The fact that an entity has a label is expressed by the **label** predicate from the standard namespace rdf (http://w3c.org/...).

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

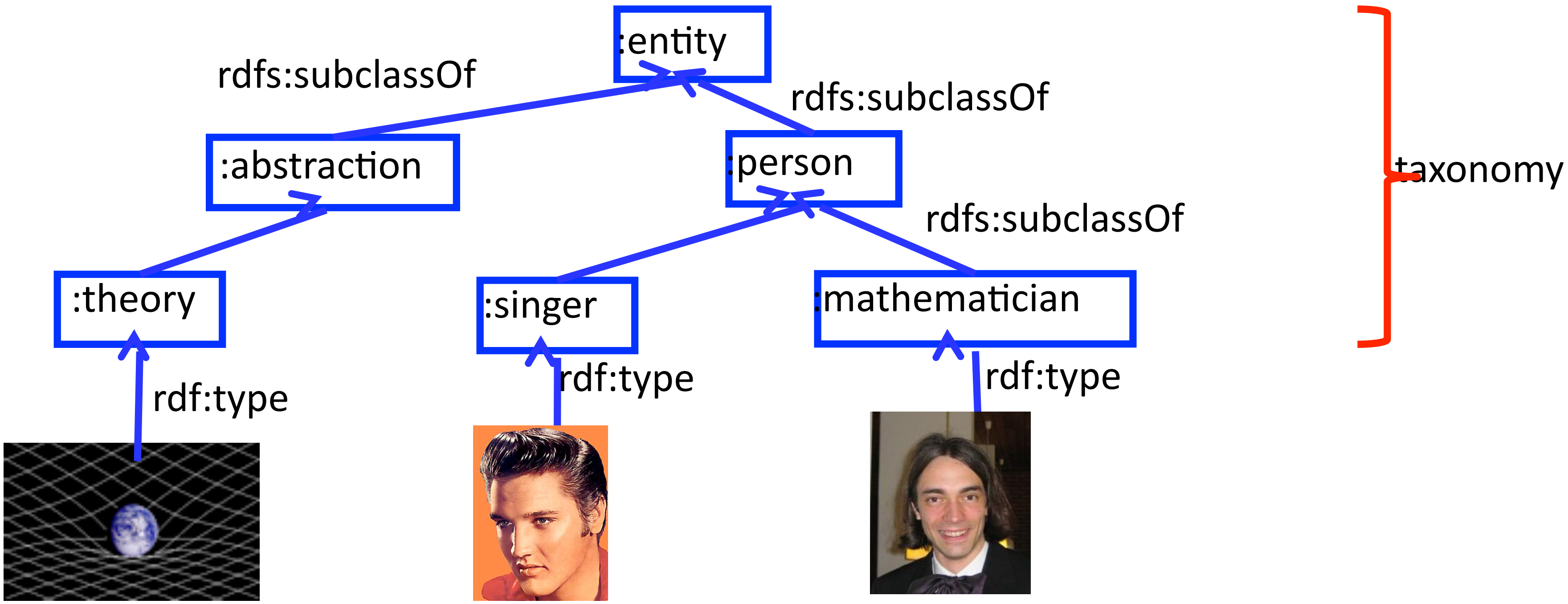
- make computers „understand“ the data they store
- allow them to reason about information
- allow them to share information across different systems

For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

- identifying entities in a globally unique way (URIs) ✔
- defining semantics in a machine-readable way (RDF) ✔
- defining taxonomies (RDFS)
- defining logical consistency in a uniform way (OWL)
- storing ontologies (N3, XML, RDFa)
- sharing ontologies (Cool URIs)
- querying ontologies (SPARQL)

SW: Classes

A **class** (also called concept) can be understood as a set of similar entities.



:entity

rdfs:subclassOf

rdfs:subclassOf

:abstraction

:person

rdfs:subclassOf

:theory

:singer

:mathematician

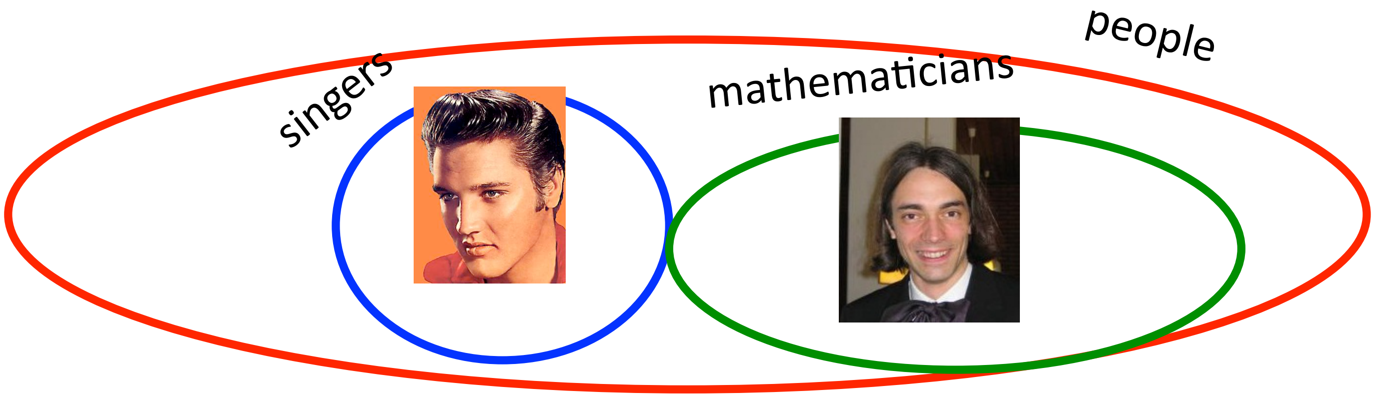
rdf:type

rdf:type

rdf:type

taxonomy

A **super-class** of a class is a class that is more general than the first class (like a super-set).



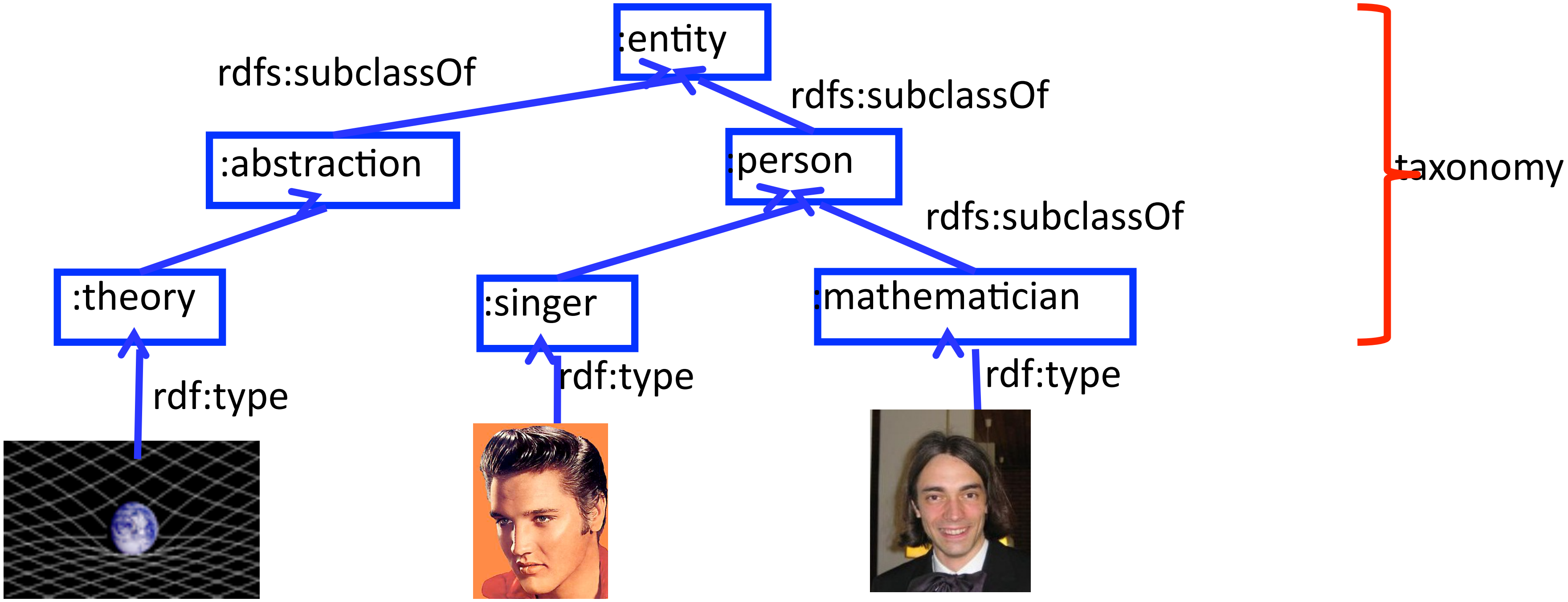
people

singers

mathematicians

SW: Classes

A **class** (also called concept) can be understood as a set of similar entities.



The fact that an entity belongs to a class is expressed by the **type** predicate from the standard namespace rdf (http://w3c.org/...).

The fact that a class is a sub-class of another class is expressed by the **subclassOf** predicate from the standard namespace rdfs (http://w3c.org/...).

For the other entities, we are using the default namespace ":" here. [RDFS]

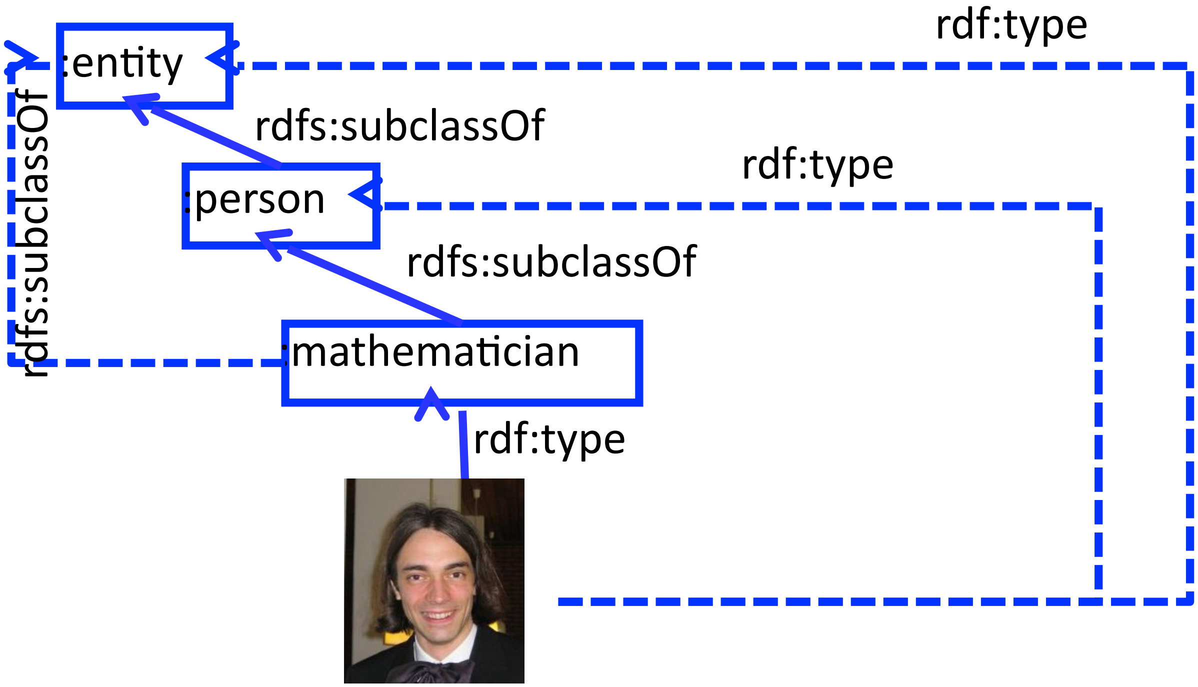
SW: Entailment

RDFS defines a set of 44 **entailment rules**.
Each entailment rule is of the form

If the ontology
contains such and such
triples

---

then add this triple



The entailment rules are applied recursively until the graph does not change any more.

This can be done in polynomial time. Whether this is done physically or deduced at query time is an implementation issue.

∀ x, y, z: subclassOf(x,y) /\ subclassOf(y,z) => subclassOf(x,z)

∀ x, y, z: type(x,y) /\ subclassOf(y,z) => type(x,z)

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

- make computers „understand“ the data they store
- allow them to reason about information
- allow them to share information across different systems

For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

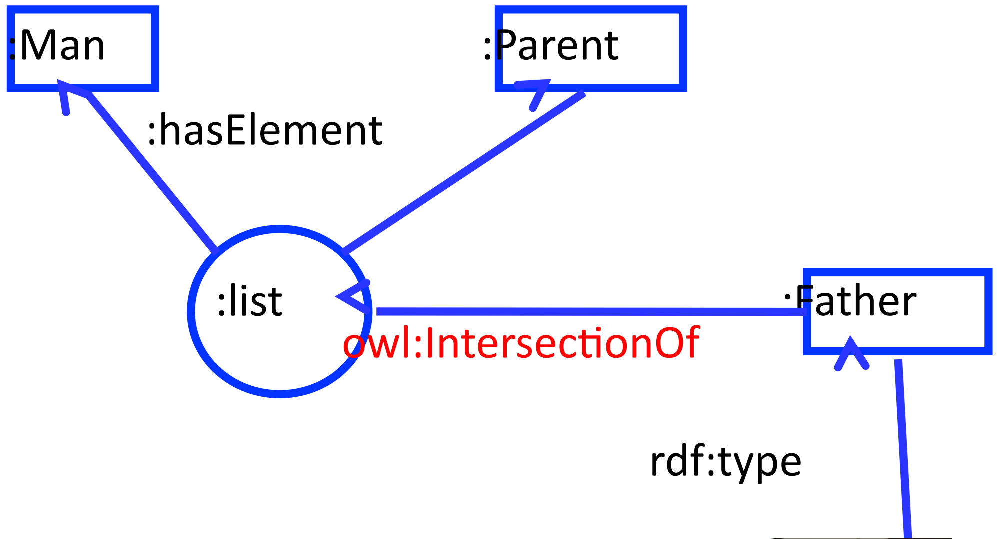
- identifying entities in a globally unique way (URIs) ✔
- defining semantics in a machine-readable way (RDF) ✔
- defining taxonomies (RDFS) ✔
- defining logical consistency in a uniform way (OWL)
- storing ontologies (N3, XML, RDFa)
- sharing ontologies (Cool URIs)
- querying ontologies (SPARQL)

SW: OWL

The **Web Ontology Language** (OWL) is a namespace that defines more predicates with semantic rules.

X rdf:type C
C owl:intersectionOf LIST
LIST hasElement Z

X rdf:type Z



owl:reflexiveIntersectionOf

owl:hyperSymmetricProperty

owl:twoOf

owl:oneOf

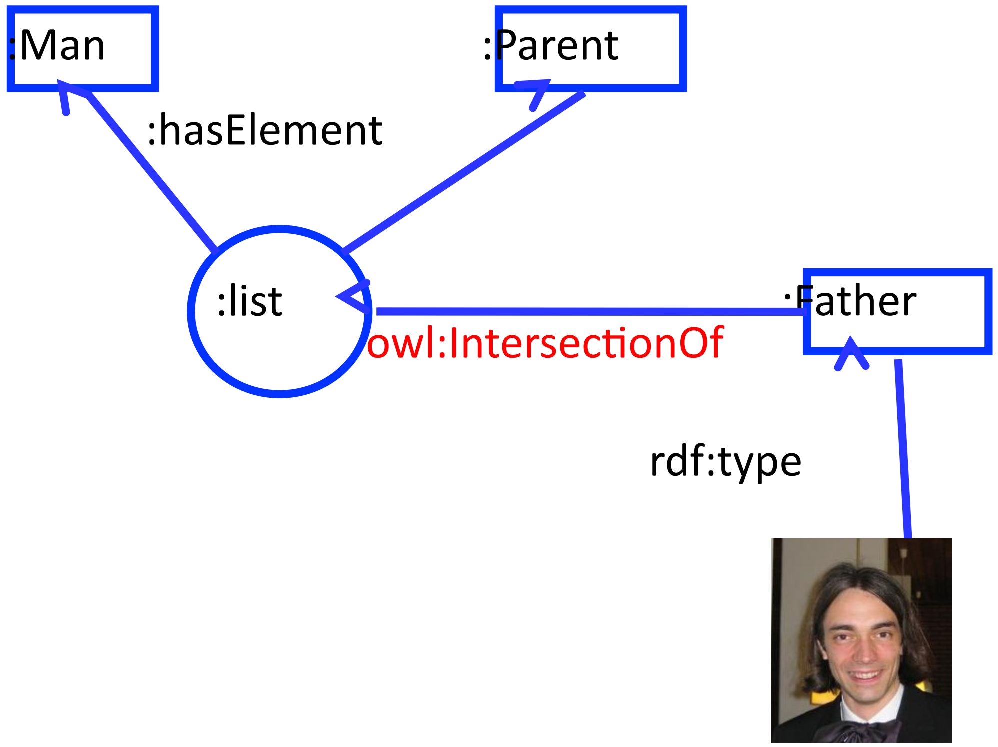
owl:complicatedCombinationOf

=> OWL is undecideable

The “list” is an RDF list with predicates defined there

SW: OWL-DL

The **Web Ontology Language** (OWL) is a namespace that defines more predicates with semantic rules.



OWL comes with the following decideable sub-sets (**profiles**)

- OWL-EL
- OWL-RL
- OWL-QL
- OWL-DL → Description Logic

OWL-DL comes with a special notation:

father = parent |┼ man

[OWL]

OWL: OWL-DL

Class constructors:

$X \sqcap Y$	The class of things that are in both X and Y
$X \sqcup Y$	The class of things that are in X or in Y
~X	The class of things that are not in X
$\forall$ R.C	The class of things where all R-links lead to a C
$\exists$ R.C	The class of things where there is a R-link to a C

Assertions:

$X \sqsubseteq Y$	X is a subclass of Y (everything in X is also in Y)
a:C	a is a thing in the class C
(a,b):R	a and b stand in the relation R, i.e., R(a,b)

villani: person $\sqcap$ $\exists$ hasChild.happyPerson

mathematician $\sqsubseteq$ theoreticalMathematician $\sqcup$ appliedMathematician

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

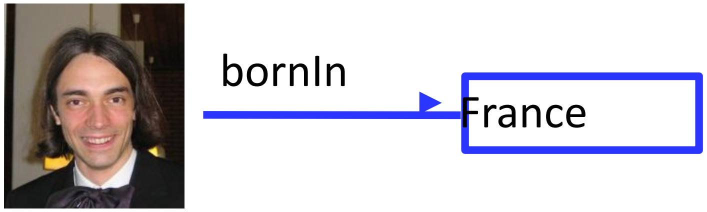
- make computers „understand" the data they store
- allow them to reason about information
- allow them to share information across different systems

For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

- identifying entities in a globally unique way (URIs) ✔
- defining semantics in a machine-readable way (RDF) ✔
- defining taxonomies (RDFS) ✔
- defining logical consistency in a uniform way (OWL) ✔
- storing ontologies (N3, XML, RDFa)
- sharing ontologies (Cool URIs)
- querying ontologies (SPARQL)

SW: Storage

There are multiple standard notations for RDF data



bornIn

France

```
@prefix   v:  http://villani.org/
@prefix   inria:  http://inria.fr/dta#
 v:Myself    inria:bornIn  <http://france.fr>  .
….
```

Notation 3 (N3):
space-separated triples
Similar: Turtle

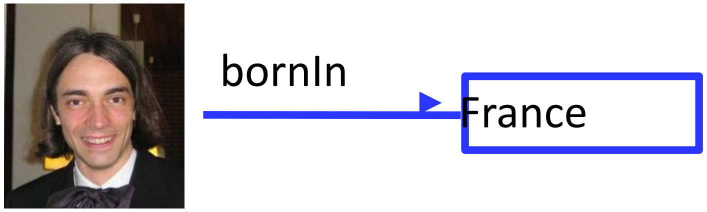
```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf=" http://www.w3.org/1999/02/22-rdf-syntax-ns# "
             xmlns:inria="http://inria.fr/dta#  ">

  <rdf:Description rdf:about=" http://villani.org/Myself ">
     <inria:bornIn rdf:resource=" http://france.fr " />
  </rdf:Description>
```

XML notation:
Uses XML namespaces

SW: Storage

There are multiple standard notations for RDF data



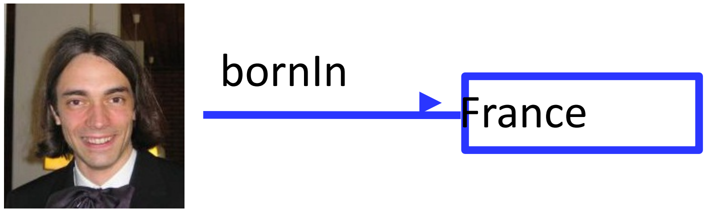
SQL database:
Usually one big table
of triples

Subject	Predicate	Object
http://villani.org/Myself	http://inria.fr/dta#bornIn	http://france.fr
…	…	…

Specifically tuned databases:
RDF 3X
OpenLink Software Virtuoso

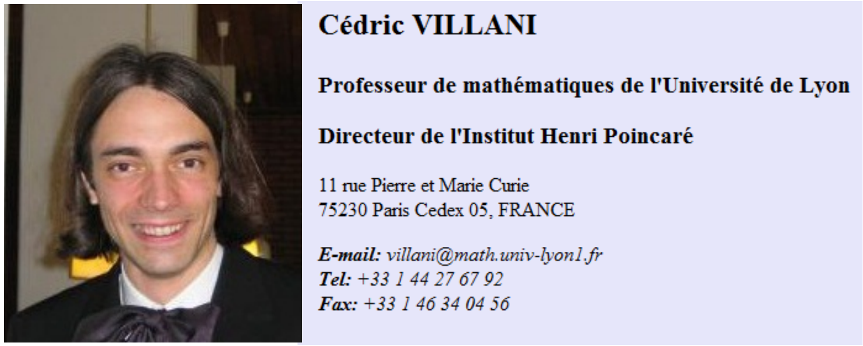
SW: Storage: RDFa

There are multiple standard notations for RDF data



RDF can be embedded into an HTML document

```
<div xmlns:v=”http://villani.org/" typeof="v:Person” about=“v:Villani” >
   I was born in <a rel="v:bornIn” href=“http://france.fr“>France</a>
    ...
</div>
```



Cédric VILLANI

Professeur de mathématiques de l'Université de Lyon

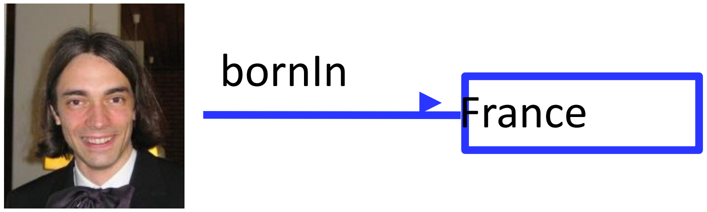
Directeur de l'Institut Henri Poincaré

11 rue Pierre et Marie Curie
75230 Paris Cedex 05, FRANCE

E-mail: *villani@math.univ-lyon1.fr*
Tel: *+33 1 44 27 67 92*
Fax: *+33 1 46 34 04 56*

SW: Storage

There are multiple standard notations for RDF data



RDF ontologies can live

- in text files („Notation 3“)
- in XML files
- in SQL databases
- in specifically tuned database systems (eg., RDF 3X or OpenLink Virtuoso)
- embedded in HTML pages („RDFa“)

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

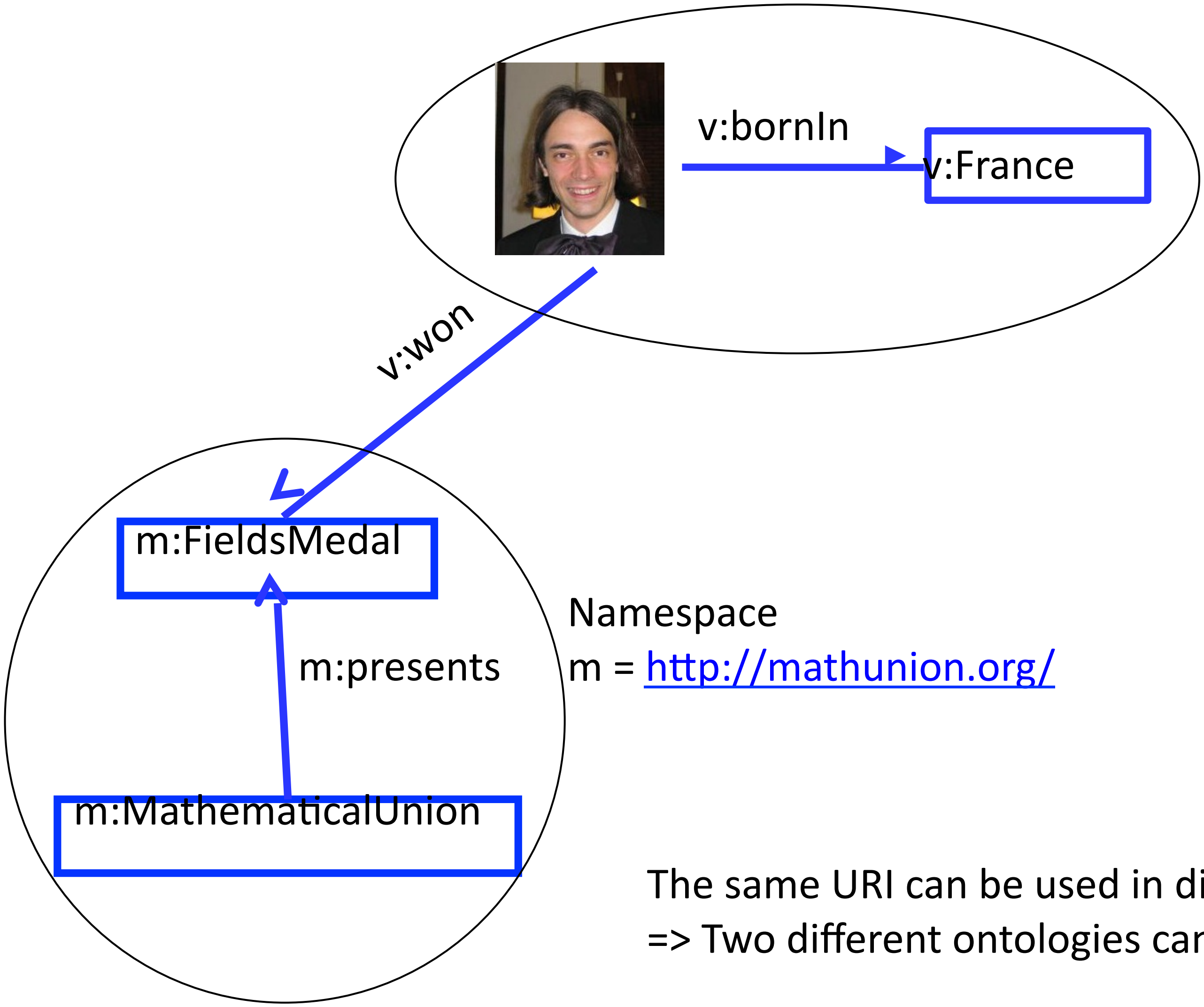
- make computers „understand“ the data they store
- allow them to reason about information
- allow them to share information across different systems

For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

- identifying entities in a globally unique way (URIs) ✔
- defining semantics in a machine-readable way (RDF) ✔
- defining taxonomies (RDFS) ✔
- defining logical consistency in a uniform way (OWL) ✔
- storing ontologies (N3, XML, RDFa) ✔
- sharing ontologies (Cool URIs)
- querying ontologies (SPARQL)

SW: Sharing

If two RDF graphs share one node, they are actually one RDF graph.



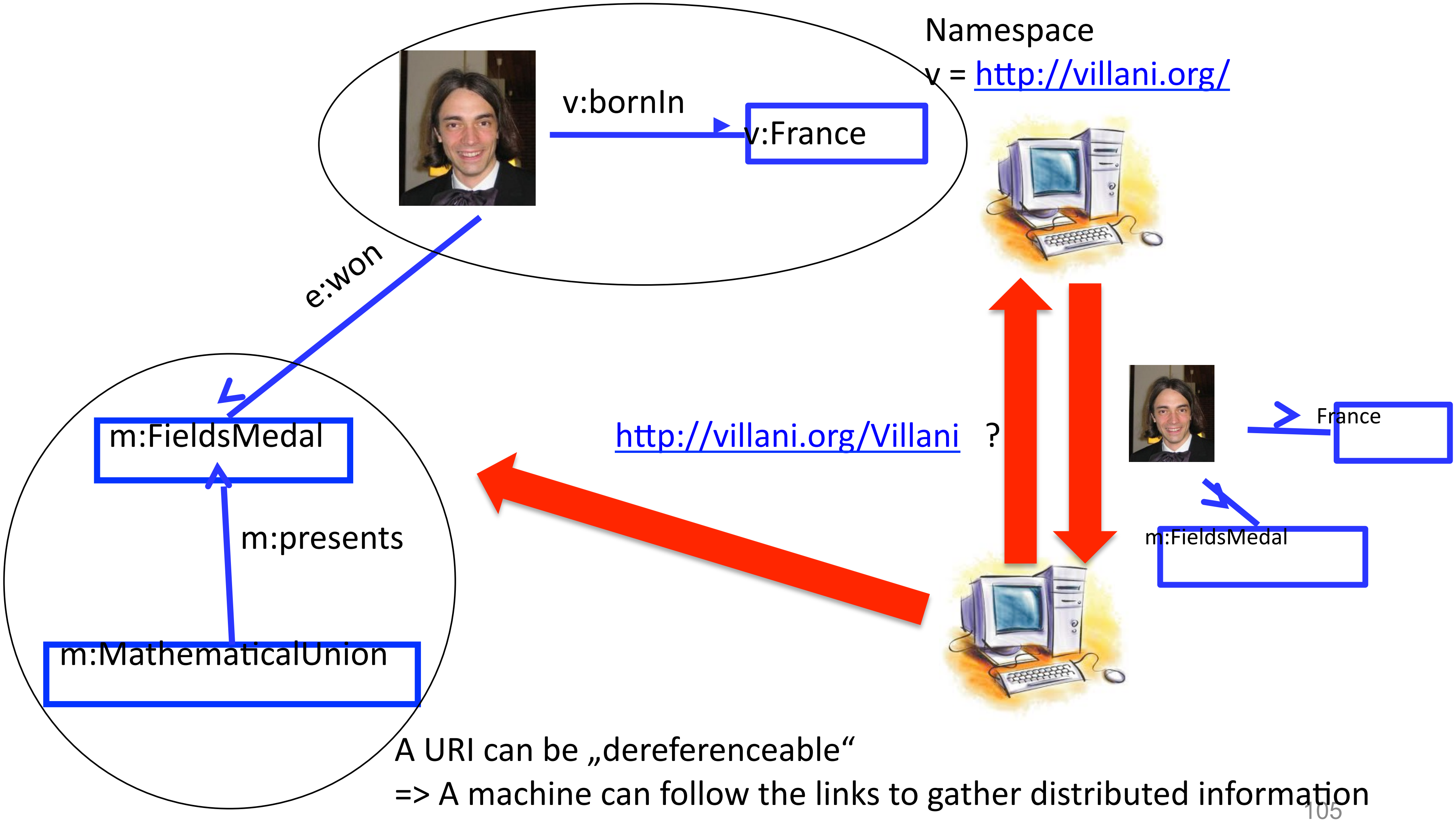
Namespace
v = http://villani.org/

Namespace
m = http://mathunion.org/

The same URI can be used in different data sets
=> Two different ontologies can talk about an identical thing

SW: Cool URIs

The “Cool URI protocol” allows a machine to access an ontological URI.
(This assumes that the ontology is stored on an Internet-accessible server in the namespace.)



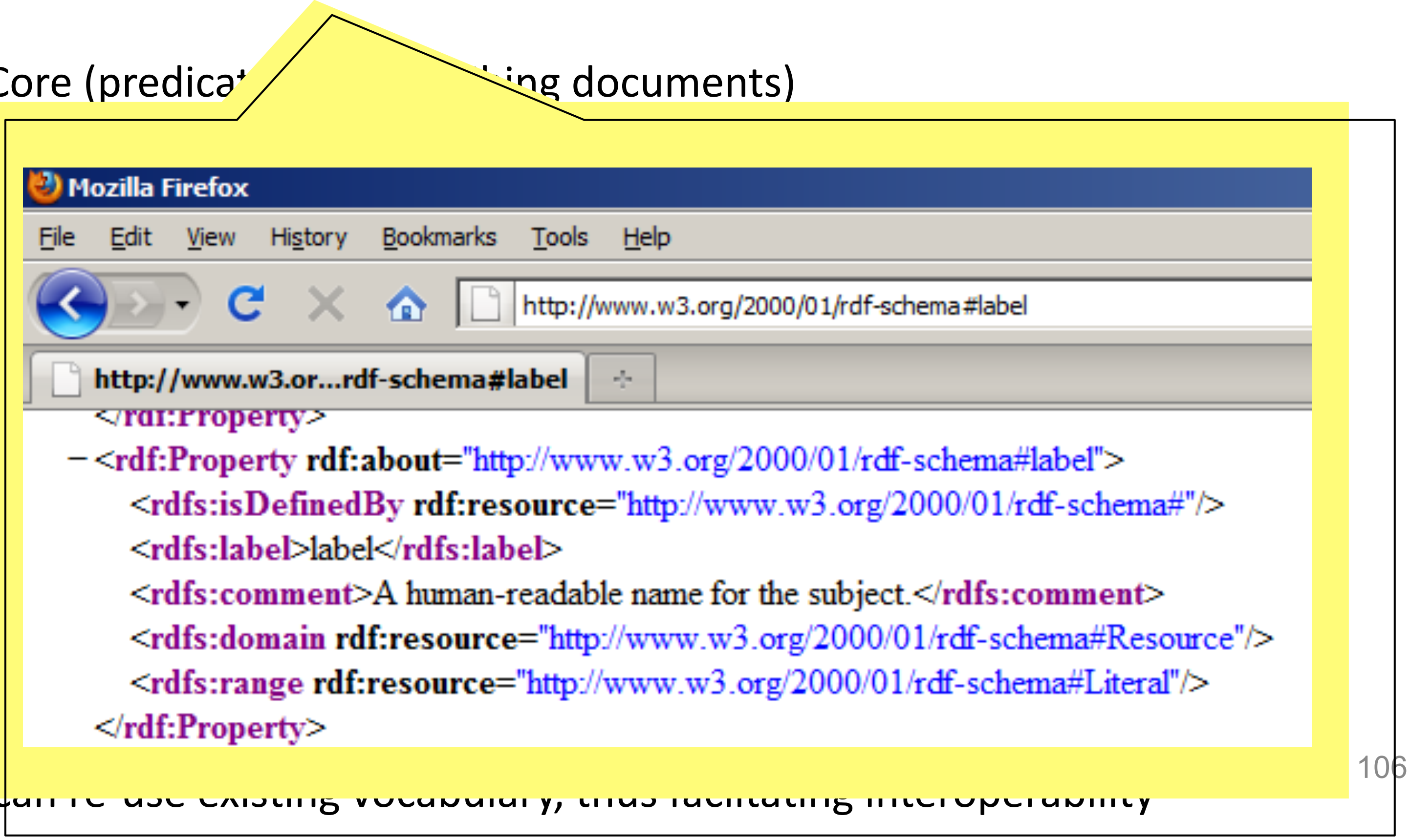
A URI can be „dereferenceable“
=> A machine can follow the links to gather distributed information

SW: Standard Vocabulary

A number of standard vocabularies have evolved

rdf: The basic RDF vocabulary
http://www.w3.org/1999/02/22-rdf-syntax-ns#

rdfs: RDF Schema vocabulary
http://www.w3.org/2000/01/rdf-schema#


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.w3.org/2000/01/rdf-schema#label
http://www.w3.or...rdf-schema#label

```
  </rdf:Property>
- <rdf:Property rdf:about="http://www.w3.org/2000/01/rdf-schema#label">
    <rdfs:isDefinedBy rdf:resource="http://www.w3.org/2000/01/rdf-schema#"/>
    <rdfs:label>label</rdfs:label>
    <rdfs:comment>A human-readable name for the subject.</rdfs:comment>
    <rdfs:domain rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
    <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
  </rdf:Property>
```

dc: Dublin Core (predicat ing documents)
http://p

foaf: Friend (
http://x

cc: Creative
http://c

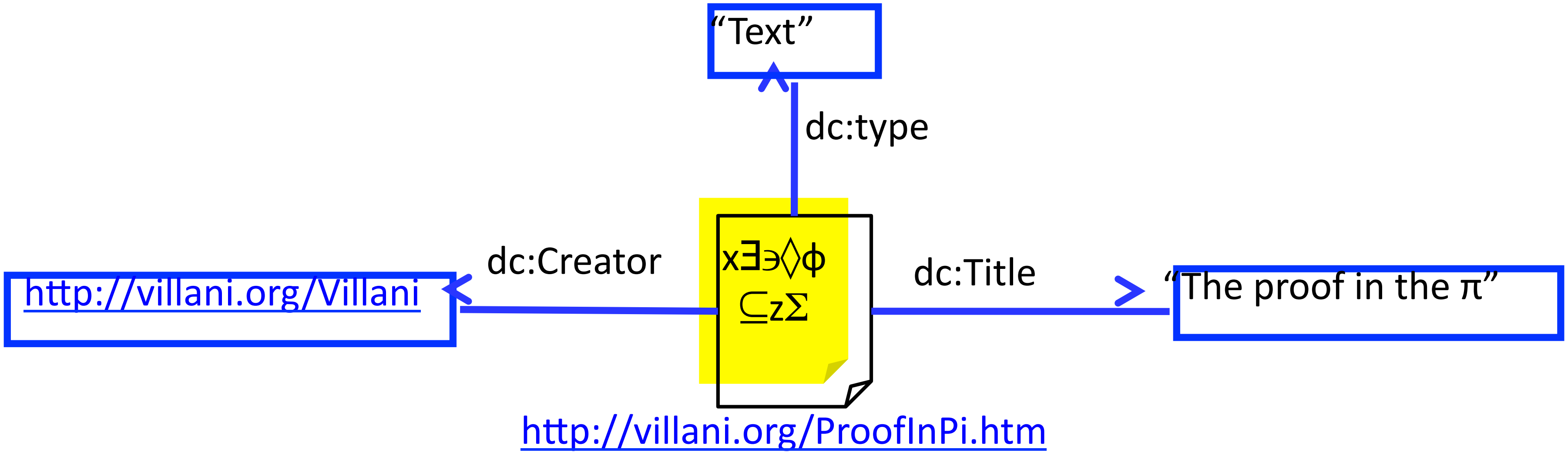
ogp: Open G
http://c

Standard vocab
=> Ontologies can re-use existing vocabulary, thus facilitating interoperability

SW: Dublin Core

A number of standard vocabularies have evolved

dc: Dublin Core (predicates for describing documents)
http://purl.org/dc/elements/1.1/



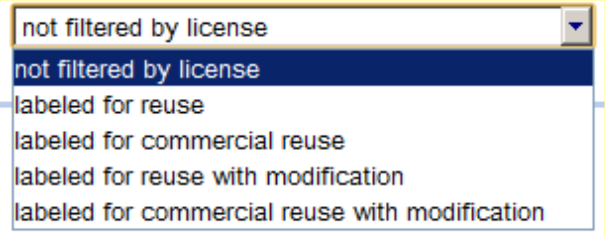
SW: Creative Commons

A number of standard vocabularies have evolved

cc: Creative Commons (types of licences)
http://creativecommons.org/ns#

Used in Google Image Search:

```
<div about="image.jpg">
  <a rel="cc:license" href="http://creativecommons.org/licenses/by”>CC-BY</a>
</div>
```



not filtered by license

not filtered by license
labeled for reuse
labeled for commercial reuse
labeled for reuse with modification
labeled for commercial reuse with modification

Creative Commons is a non-profit organization, which defines popular licenses, notably

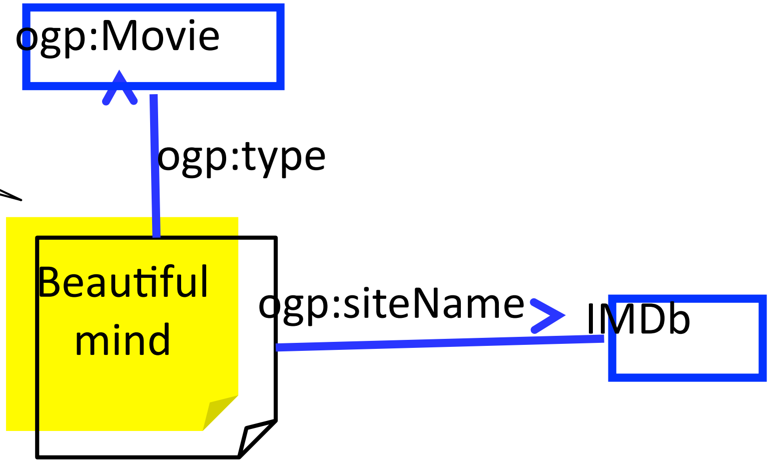
- CC-BY: Free for reuse, just give credit to the author
- CC-BY-NC: Free for reuse, give credit, non-commercial use only
- CC-BY-ND: Free for reuse, give credit, do not create derivative works

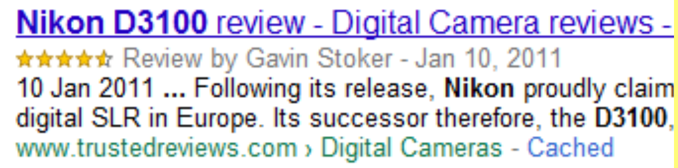
SW: Open Graph Protocol

```
www.imdb.com/title/tt0268978/
<html xmlns:og=http://ogp.me/ns# >
  …
  <meta property='og:type' content='movie' />
  <meta property='fb:app_id' content='123' />
  …
</html>
```

d

ıtions for Web pages)





Nikon D3100 review - Digital Camera reviews -
★★★★☆ Review by Gavin Stoker - Jan 10, 2011
10 Jan 2011 ... Following its release, **Nikon** proudly claim
digital SLR in Europe. Its successor therefore, the **D3100**,
www.trustedreviews.com › Digital Cameras - Cached

RDF da llowing the Open Graph Protocol is often embedded in HTML pages, thus all wing the Facebook LIKE button to work.

Google has defined its own namespace, which allows annotating HTML pages with meta-information that will show up in “rich snippets”.

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

- make computers „understand“ the data they store
- allow them to reason about information
- allow them to share information across different systems

For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

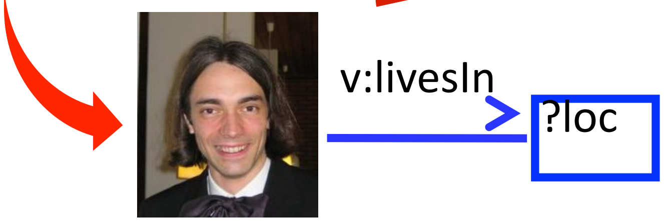
- identifying entities in a globally unique way (URIs) ✔
- defining semantics in a machine-readable way (RDF) ✔
- defining taxonomies (RDFS) ✔
- defining logical consistency in a uniform way (OWL) ✔
- storing ontologies (N3, XML, RDFa) ✔
- sharing ontologies (Cool URIs) ✔
- querying ontologies (SPARQL)

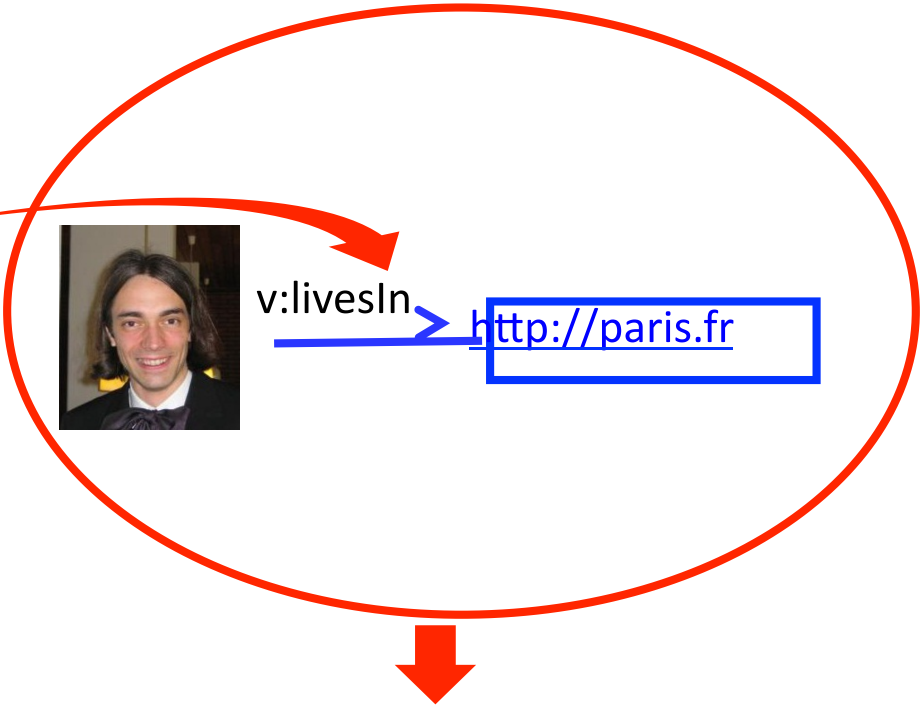
SW: SPARQL

SPARQL (SPARQL Protocol and RDF Query Language)
is the query language of the Semantic Web.

```
PREFIX v: <http://villani.org/>

SELECT ?loc
WHERE {
  v:villani   v:livesIn   ?loc.
}
```





?loc = http://paris.fr

SPARQL resembles SQL, adapted to the Semantic Web
Many ontologies provide a “SPARQL endpoint” where SPARQL queries can be asked.

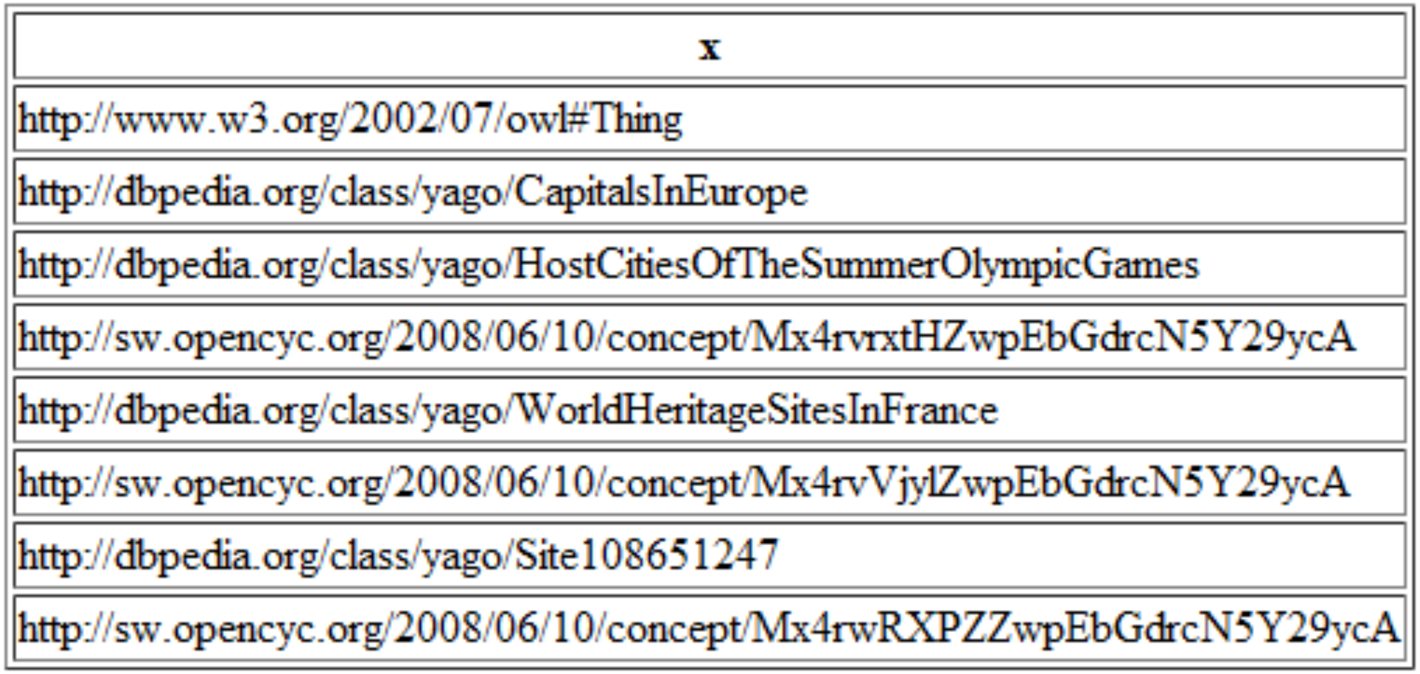
[SPARQL]

SW: SPARQL Example

Example at http://dbpedia-live.openlinksw.com/sparql/ :

Let‘s ask DBpedia, one of the major ontologies in the Semantic Web

```
select distinct ?x {
  <http://dbpedia.org/resource/Paris>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
  ?x
}
```



x
http://www.w3.org/2002/07/owl#Thing
http://dbpedia.org/class/yago/CapitalsInEurope
➡ http://dbpedia.org/class/yago/HostCitiesOfTheSummerOlympicGames
http://sw.opencyc.org/2008/06/10/concept/Mx4rvrxtHZwpEbGdrcN5Y29ycA
http://dbpedia.org/class/yago/WorldHeritageSitesInFrance
http://sw.opencyc.org/2008/06/10/concept/Mx4rvVjylZwpEbGdrcN5Y29ycA
http://dbpedia.org/class/yago/Site108651247
http://sw.opencyc.org/2008/06/10/concept/Mx4rwRXPZZwpEbGdrcN5Y29ycA

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

- make computers „understand“ the data they store
- allow them to reason about information
- allow them to share information across different systems

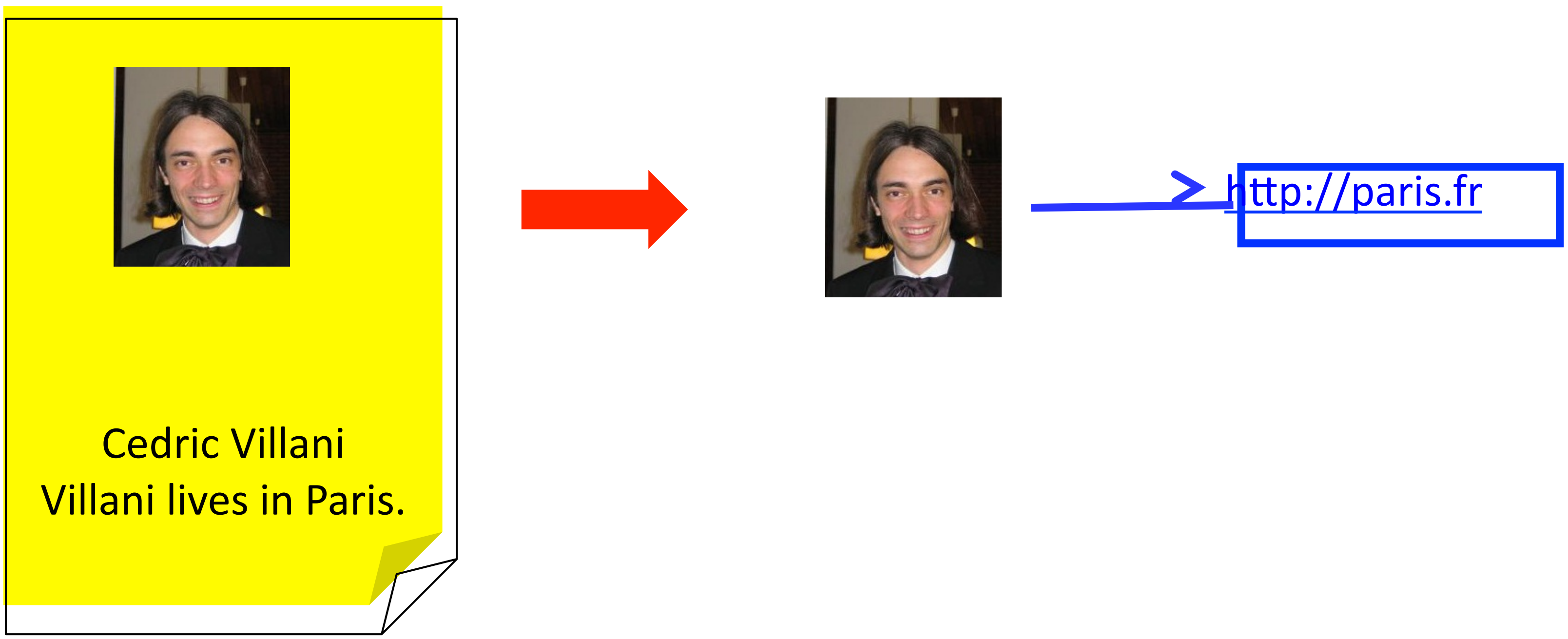
For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

- identifying entities in a globally unique way (URIs) ✔
- defining semantics in a machine-readable way (RDF) ✔
- defining taxonomies (RDFS) ✔
- defining logical consistency in a uniform way (OWL) ✔
- storing ontologies (N3, XML, RDFa) ✔
- sharing ontologies (Cool URIs) ✔
- querying ontologies (SPARQL) ✔

Great, now where do we get the data from?

SW: Information Extraction

The dream of **information extraction** is to make unstructured information (read: Web documents) available as structured information (here: ontologies).



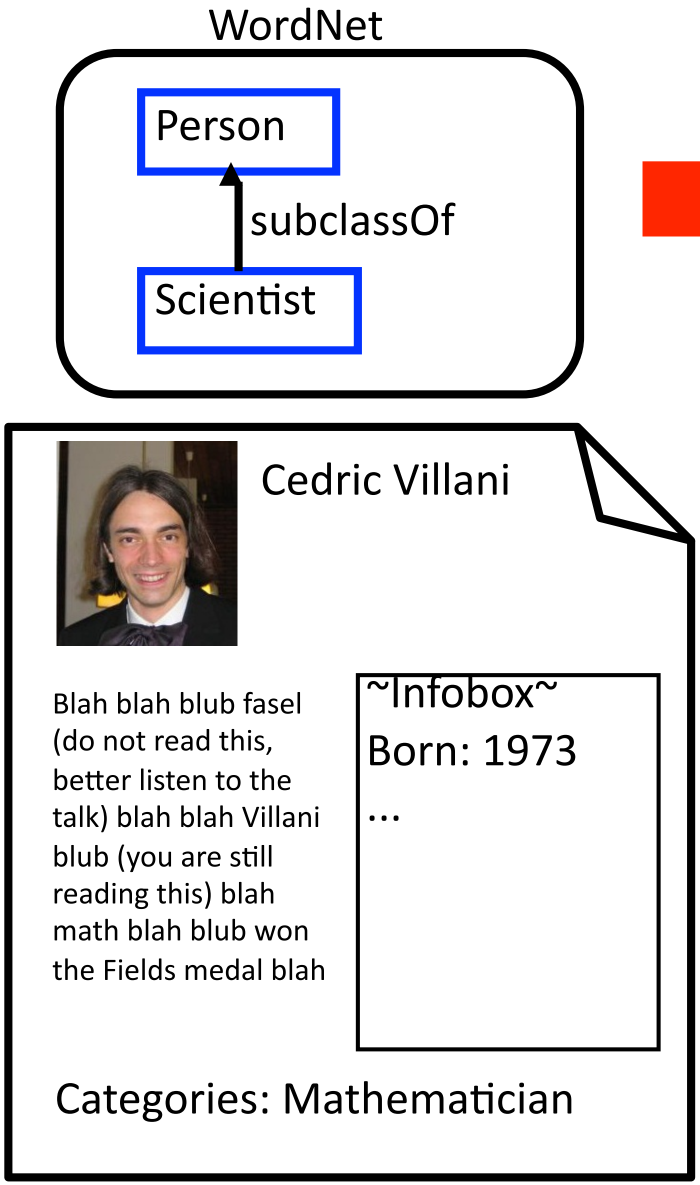
Cedric Villani

Villani lives in Paris.

http://paris.fr

SW: YAGO

For Information Extraction, let’s start from Wikipedia

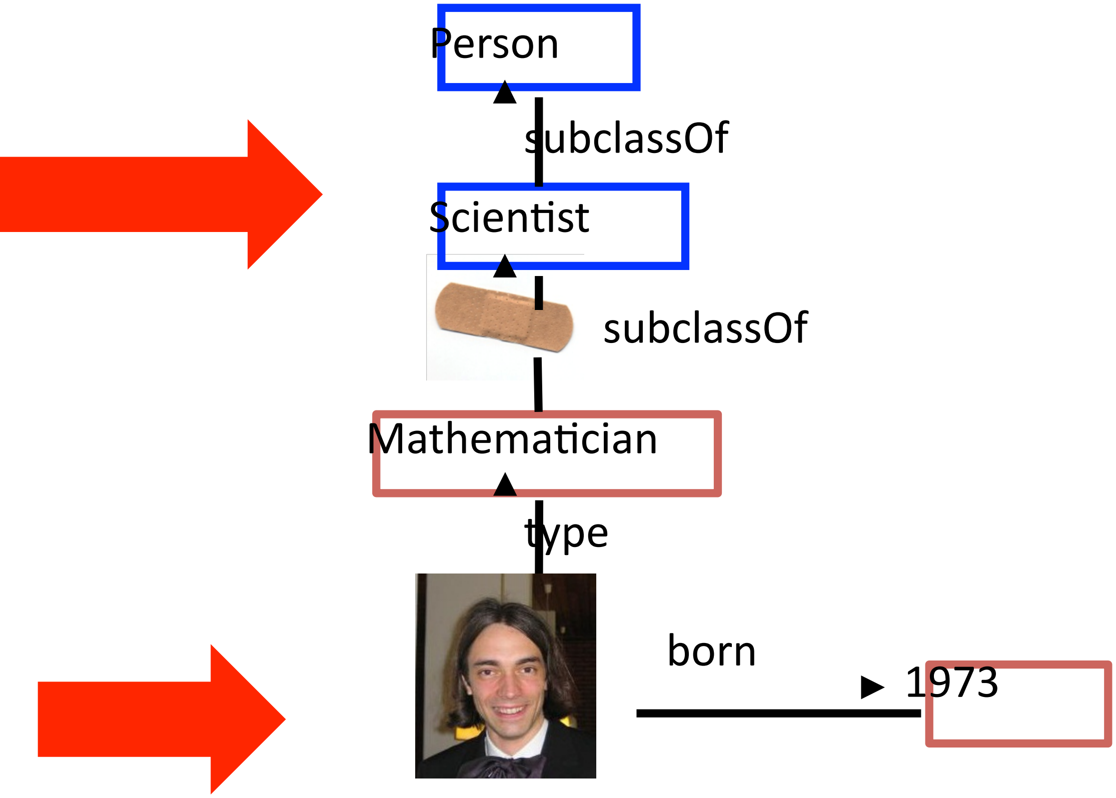


Cedric Villani

Blah blah blub fasel (do not read this, better listen to the talk) blah blah Villani blub (you are still reading this) blah math blah blub won the Fields medal blah

~Infobox~
Born: 1973
…

Categories: Mathematician



Exploit Infoboxes

Exploit conceptual categories

Add WordNet

SW: Ontologies from Wikipedia

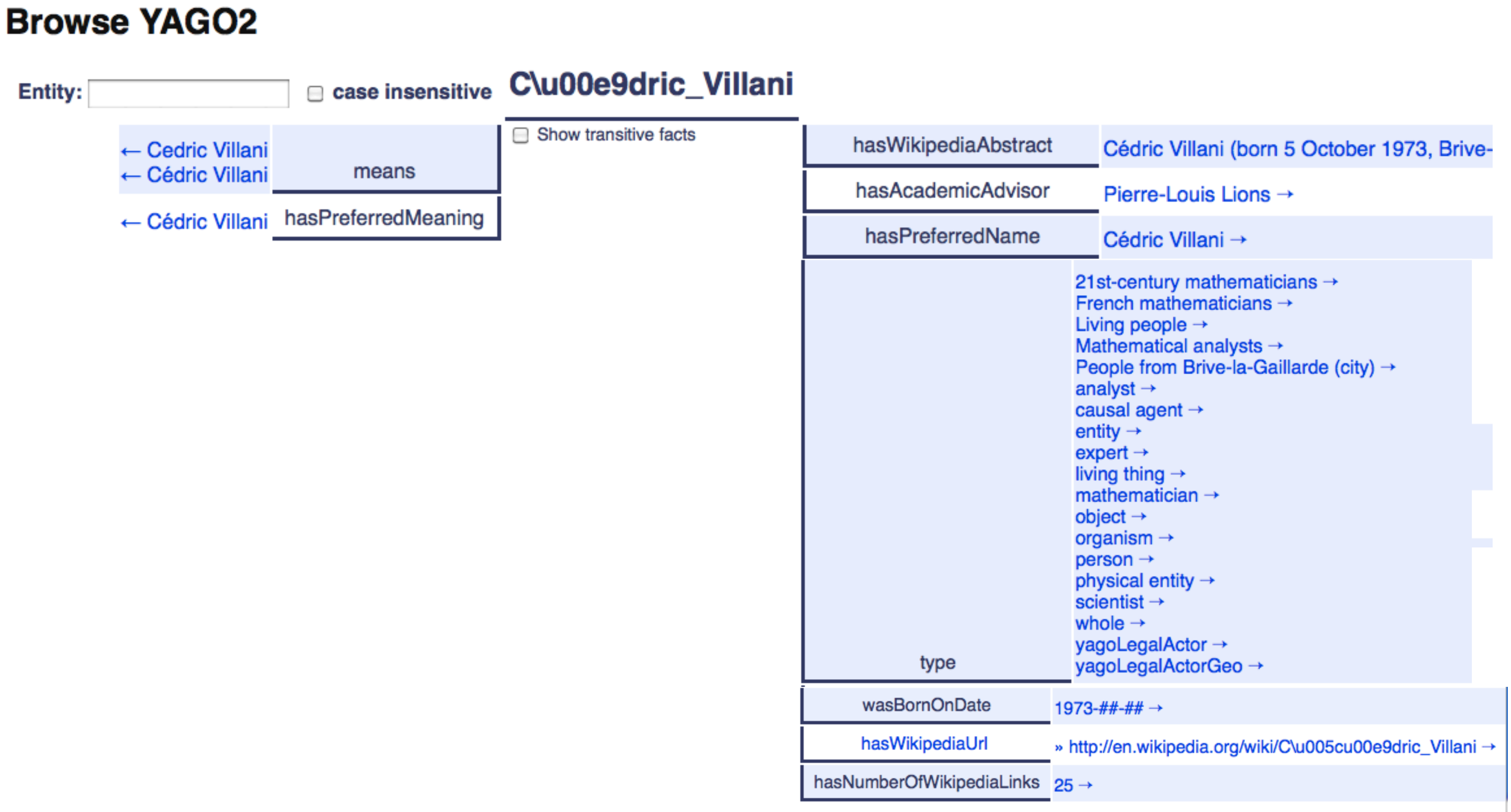
Information Extraction from Wikipedia has lead to several large ontologies:

- YAGO (http://mpii.d/yago , 10m entities, 80m facts, 95% accuracy) [YAGO, YAGO2]
- DBpedia (http://dbpedia.org/ , 3.5m entities, 670m facts) [DBpedia]
- Freebase (http://freebase.com , 20m entities)

These are huge knowledge bases,
which contain not just a class taxonomy,
but also instances and facts

SW: Example

Here is what the YAGO ontology (http://mpii.de/yago) knows about Cedric Villani:



Browse YAGO2

Entity: □ case insensitive

C\u00e9dric_Villani

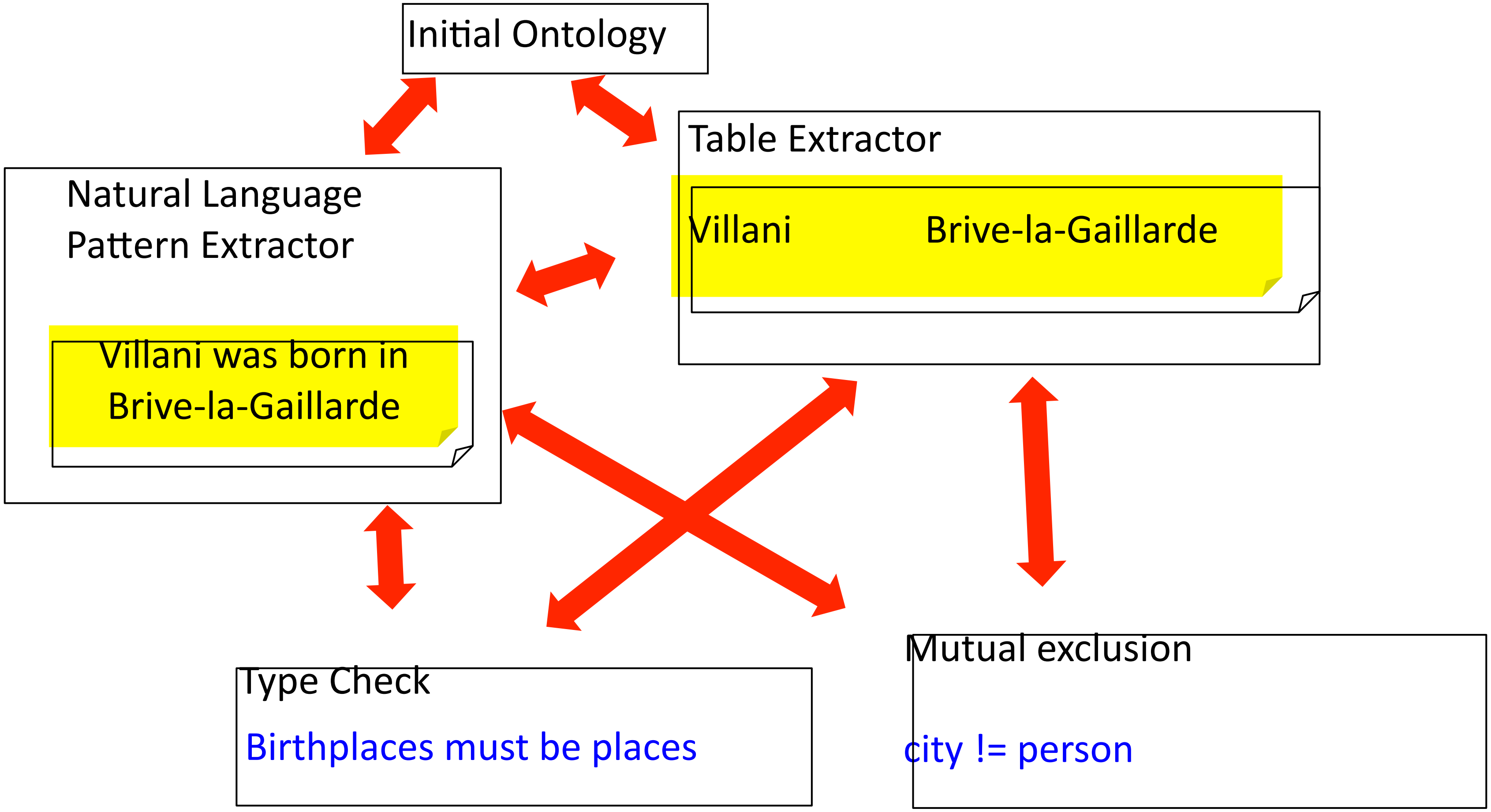
□ Show transitive facts

← Cedric Villani ← Cédric Villani	means
← Cédric Villani	hasPreferredMeaning

hasWikipediaAbstract	Cédric Villani (born 5 October 1973, Brive-
hasAcademicAdvisor	Pierre-Louis Lions →
hasPreferredName	Cédric Villani →
type	21st-century mathematicians → French mathematicians → Living people → Mathematical analysts → People from Brive-la-Gaillarde (city) → analyst → causal agent → entity → expert → living thing → mathematician → object → organism → person → physical entity → scientist → whole → yagoLegalActor → yagoLegalActorGeo →
wasBornOnDate	1973-##-## →
hasWikipediaUrl	» http://en.wikipedia.org/wiki/C\u005cu00e9dric_Villani →
hasNumberOfWikipediaLinks	25 →

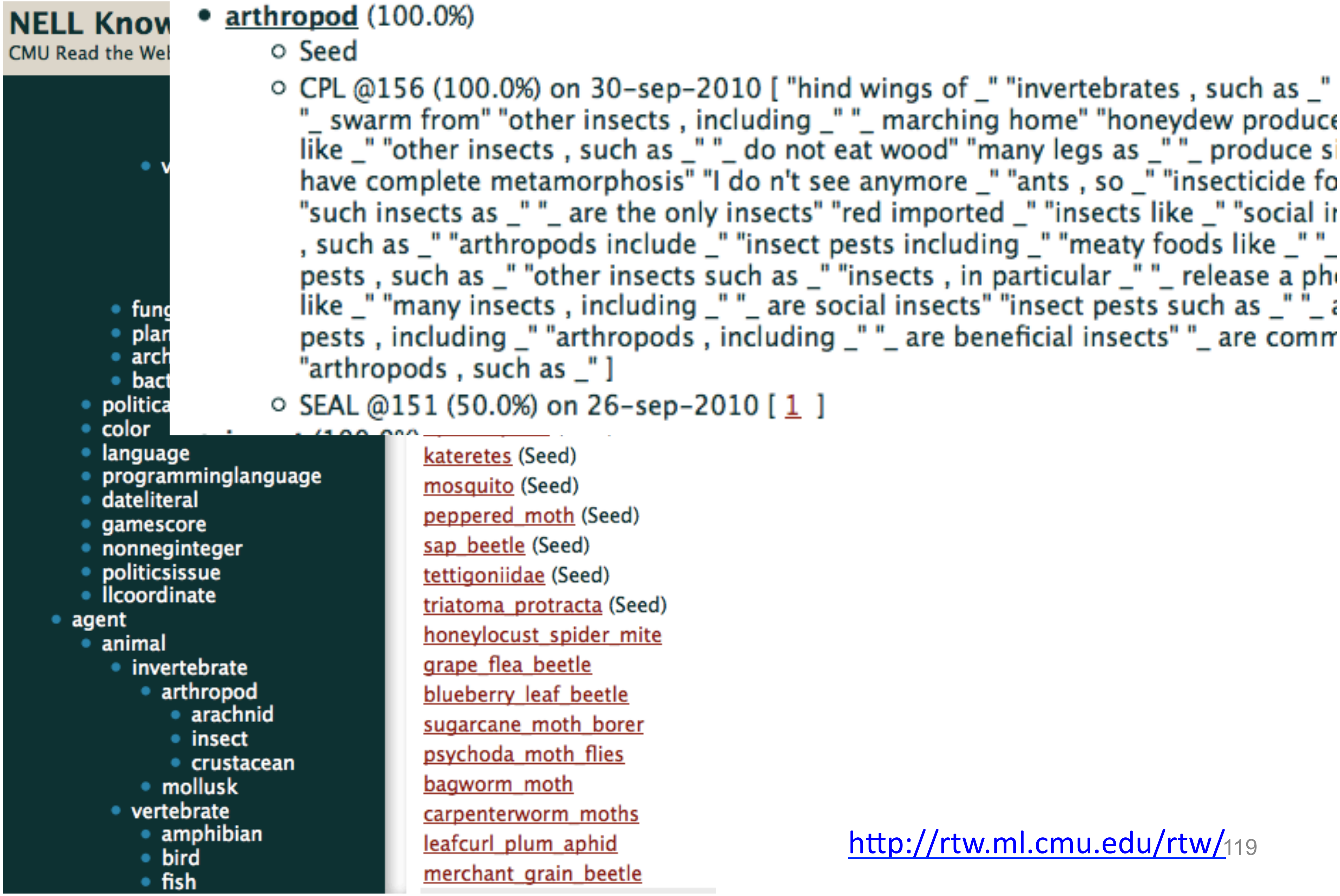
SW: NELL

Other projects extract the data from the “real Web”


Initial Ontology
Natural Language Pattern Extractor
Villani was born in Brive-la-Gaillarde
Table Extractor
Villani Brive-la-Gaillarde
Type Check
Birthplaces must be places
Mutual exclusion
city != person

http://rtw.ml.cmu.edu/rtw/

SW: NELL



NELL Know
CMU Read the Wel

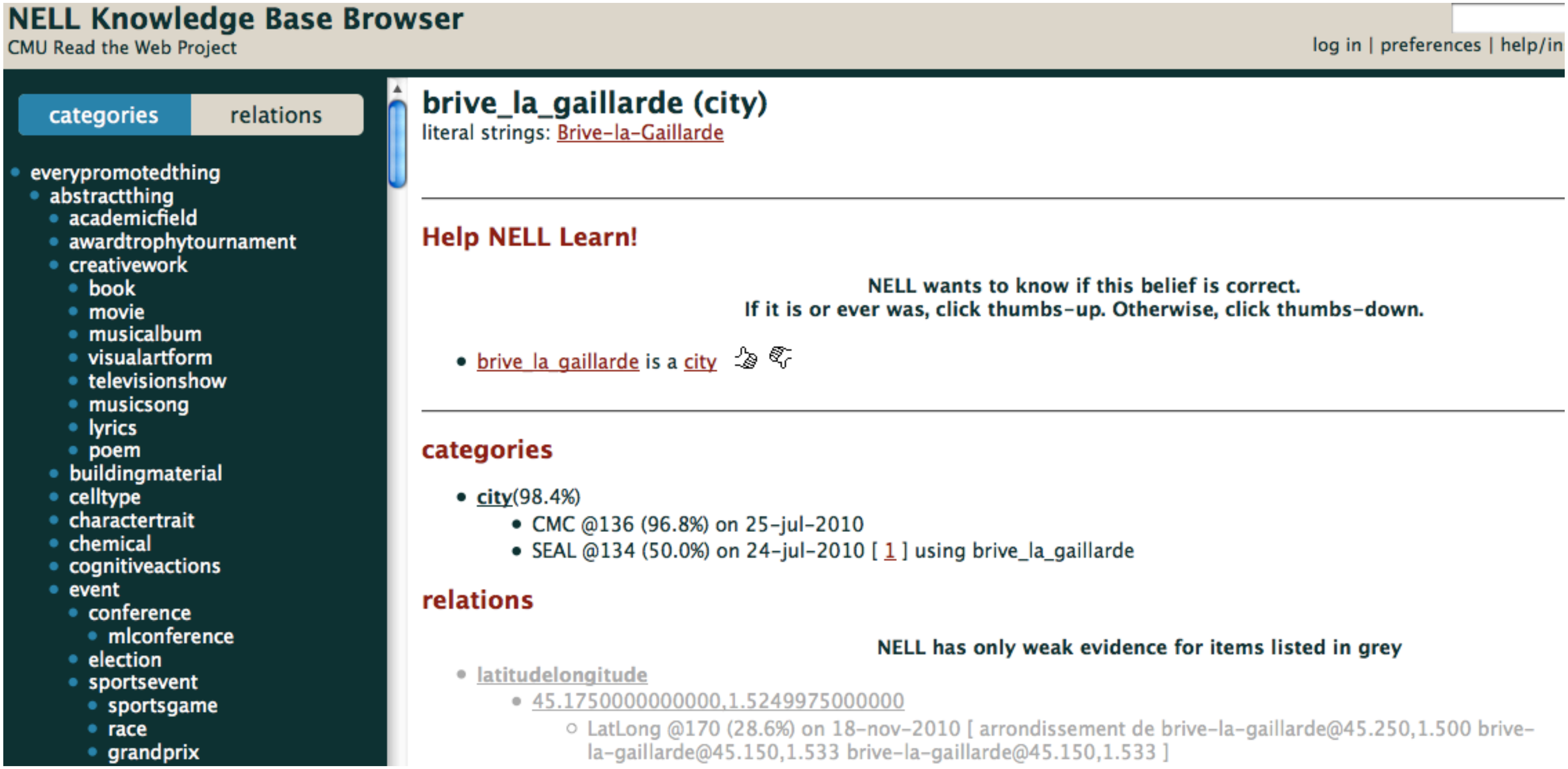
- arthropod (100.0%)
  - Seed
  - CPL @156 (100.0%) on 30-sep-2010 ["hind wings of _" "invertebrates , such as _"
    "_ swarm from" "other insects , including _" "_ marching home" "honeydew produce
    like _" "other insects , such as _" "_ do not eat wood" "many legs as _" "_ produce si
    have complete metamorphosis" "I do n't see anymore _" "ants , so _" "insecticide fo
    "such insects as _" "_ are the only insects" "red imported _" "insects like _" "social i
    , such as _" "arthropods include _" "insect pests including _" "meaty foods like _" "_
    pests , such as _" "other insects such as _" "insects , in particular _" "_ release a ph
    like _" "many insects , including _" "_ are social insects" "insect pests such as _" "_ a
    pests , including _" "arthropods , including _" "_ are beneficial insects" "_ are comm
    "arthropods , such as _"]
  - SEAL @151 (50.0%) on 26-sep-2010 [1]

- v
- fung
- plar
- arch
- bact
- politica
- color
- language
- programminglanguage
- dateliteral
- gamescore
- noneginteger
- politicsissue
- llcoordinate
- agent
  - animal
    - invertebrate
      - arthropod
        - arachnid
        - insect
        - crustacean
      - mollusk
    - vertebrate
      - amphibian
      - bird
      - fish

kateretes (Seed)
mosquito (Seed)
peppered_moth (Seed)
sap_beetle (Seed)
tettigoniidae (Seed)
triatoma_protracta (Seed)
honeylocust_spider_mite
grape_flea_beetle
blueberry_leaf_beetle
sugarcane_moth_borer
psychoda_moth_flies
bagworm_moth
carpenterworm_moths
leafcurl_plum_aphid
merchant_grain_beetle

http://rtw.ml.cmu.edu/rtw/

SW: NELL



brive_la_gaillarde (city)

literal strings: Brive-la-Gaillarde

Help NELL Learn!

NELL wants to know if this belief is correct.
If it is or ever was, click thumbs-up. Otherwise, click thumbs-down.

- brive_la_gaillarde is a city 👍 👎

categories

- city(98.4%)
  - CMC @136 (96.8%) on 25-jul-2010
  - SEAL @134 (50.0%) on 24-jul-2010 [1] using brive_la_gaillarde

relations

NELL has only weak evidence for items listed in grey

- latitudelongitude
  - 45.1750000000000,1.5249975000000
    - LatLong @170 (28.6%) on 18-nov-2010 [arrondissement de brive-la-gaillarde@45.250,1.500 brive-la-gaillarde@45.150,1.533 brive-la-gaillarde@45.150,1.533]

SW: Information Extraction

Other projects extract the data from the “real Web”.

- NELL (Never-Ending Language Learner, CMU; runs perpetually) [NELL]
- SOFIE & Prospera (Max-Planck-Institute; includes consistency checking) [SOFIE, PROSPERA]
- OntoUSP (University of Washington; uses deep linguistic processing) [OntoUSP]

These systems are designed to extract information from arbitrary Web documents on large scale.

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

- make computers „understand“ the data they store
- allow them to reason about information
- allow them to share information across different systems

For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

- identifying entities in a globally unique way (URIs) ✔
- defining semantics in a machine-readable way (RDF) ✔
- defining taxonomies (RDFS) ✔
- defining logical consistency in a uniform way (OWL) ✔
- storing ontologies (N3, XML, RDFa) ✔
- sharing ontologies (Cool URIs) ✔
- querying ontologies (SPARQL) ✔

Great, now where do we get the data from? ✔

And how does the Semantic Web look in practice?

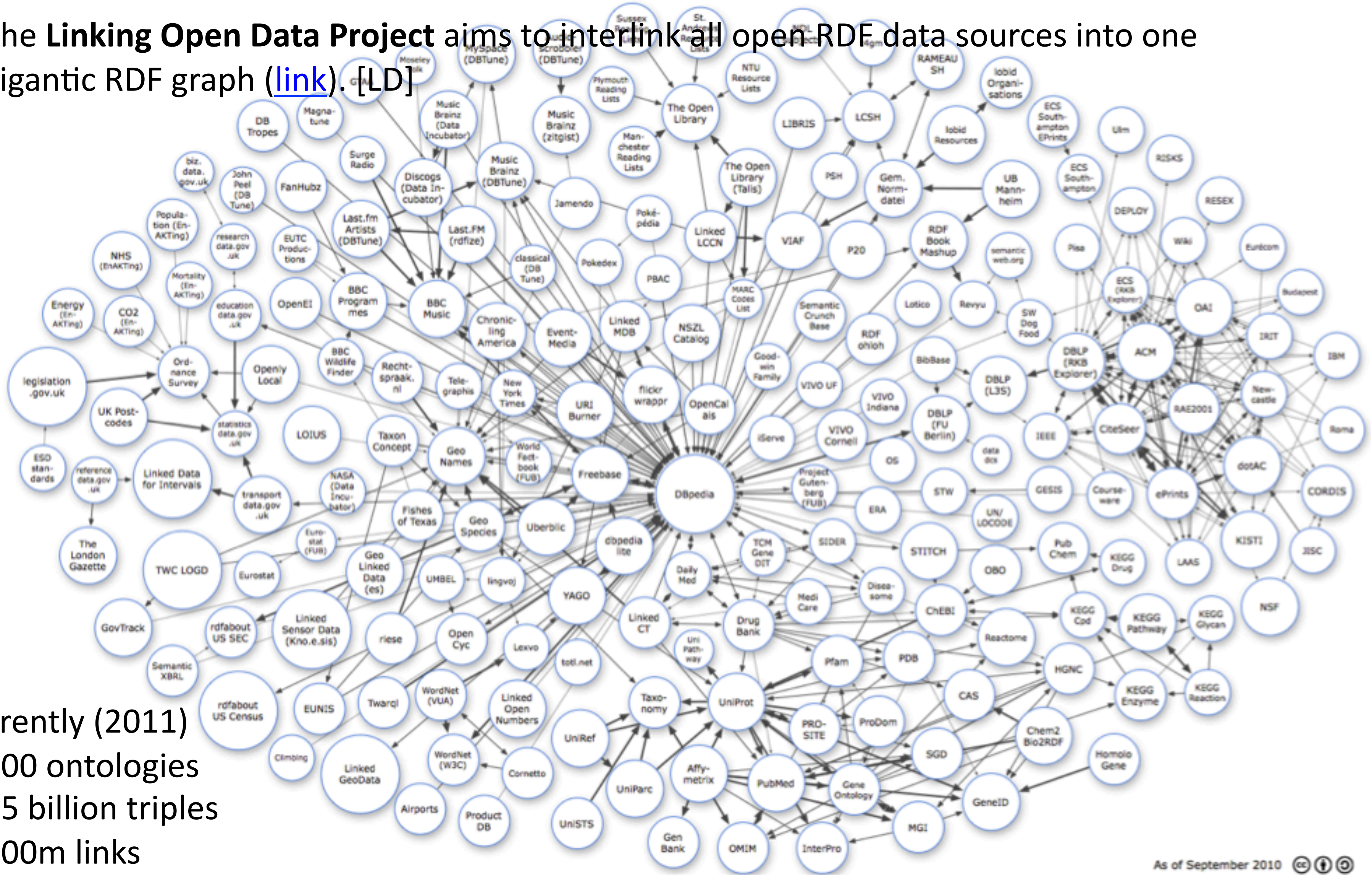
SW: Existing Ontologies

Hundreds of data sets are nowadays available in RDF
(http://www4.wiwiss.fu-berlin.de/lodcloud/)

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

SW: The Linked Data Cloud

The **Linking Open Data Project** aims to interlink all open RDF data sources into one gigantic RDF graph (link). [LD]



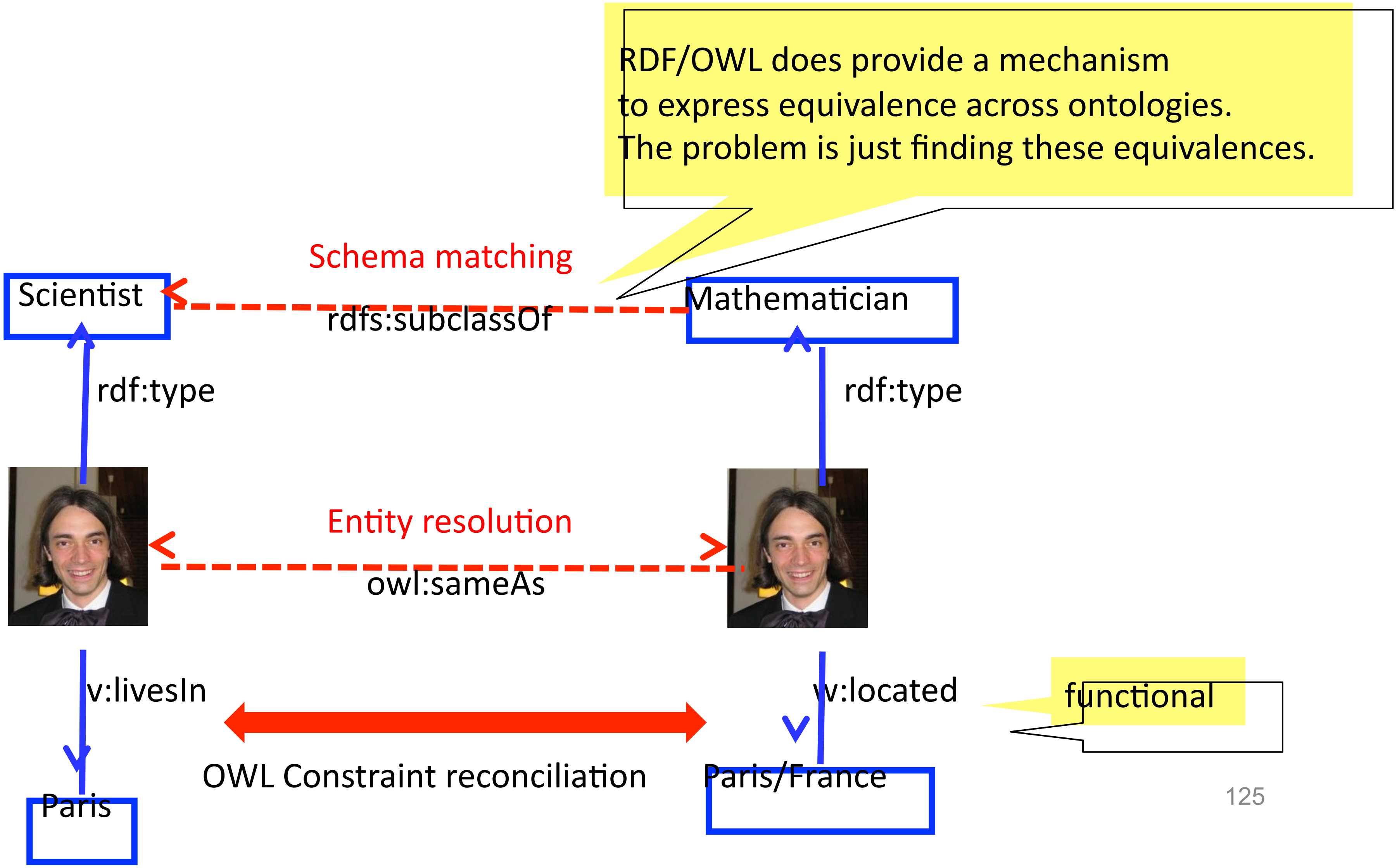
Currently (2011)

- 200 ontologies
- 25 billion triples
- 400m links

http://richard.cyganiak.de/2007/10/lod/imagemap.html

SW: Linking Data – the Challenge

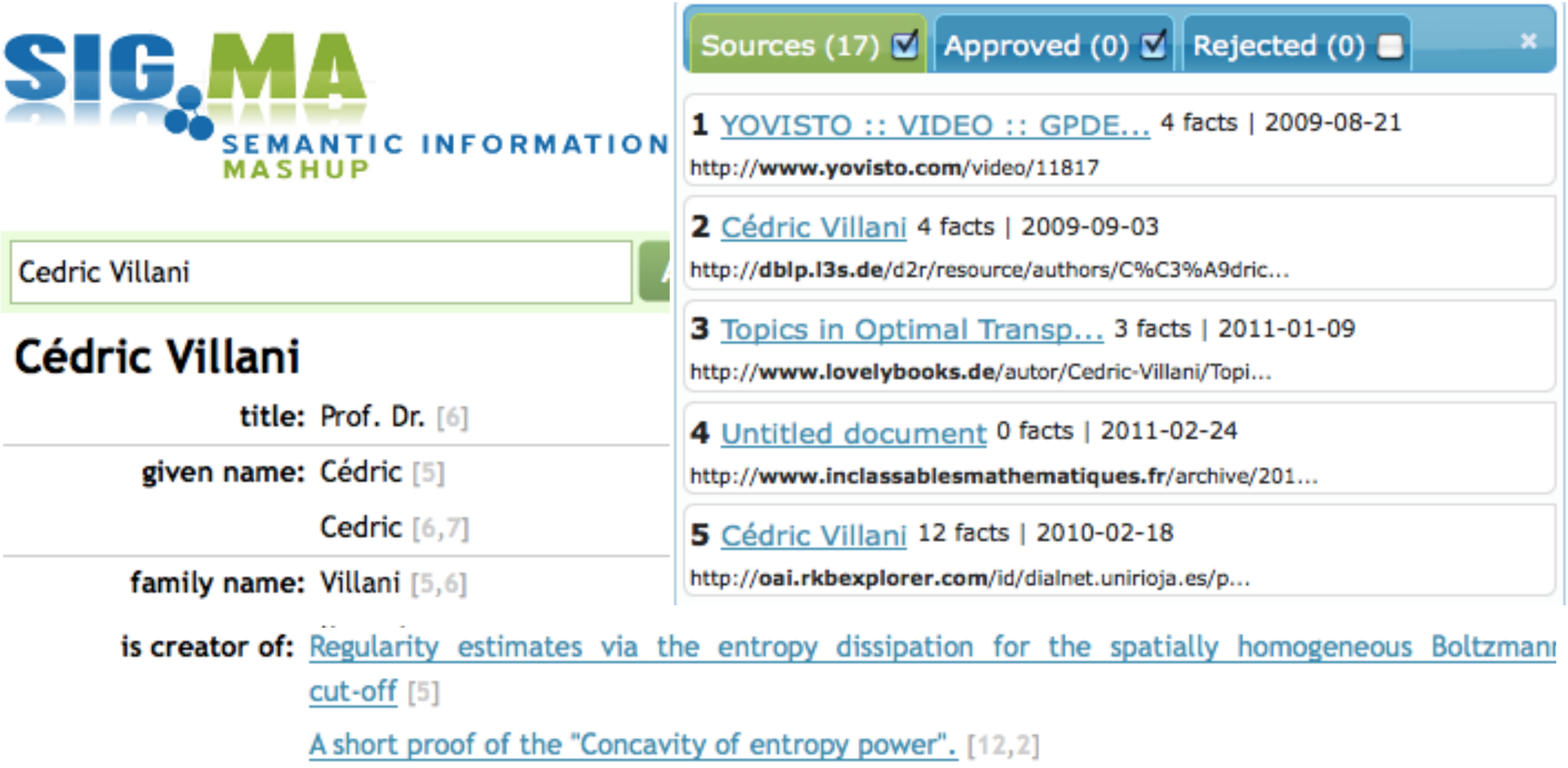
The **Linking Open Data Project** aims to interlink all open RDF data sources into one gigantic RDF graph.



RDF/OWL does provide a mechanism
to express equivalence across ontologies.
The problem is just finding these equivalences.

SW: SIGMA

The **SIGMA** engine (http://sig.ma) crawls the Semantic Web [SIGMA]



SIG.MA SEMANTIC INFORMATION MASHUP

Cedric Villani

Cédric Villani

title: Prof. Dr. [6]

given name: Cédric [5]

Cedric [6,7]

family name: Villani [5,6]

is creator of: Regularity estimates via the entropy dissipation for the spatially homogeneous Boltzmann cut-off [5]

A short proof of the "Concavity of entropy power". [12,2]

Sources (17) ☑ Approved (0) ☑ Rejected (0) □ ×

1 YOVISTO :: VIDEO :: GPDE... 4 facts | 2009-08-21
http://www.yovisto.com/video/11817

2 Cédric Villani 4 facts | 2009-09-03
http://dblp.l3s.de/d2r/resource/authors/C%C3%A9dric...

3 Topics in Optimal Transp... 3 facts | 2011-01-09
http://www.lovelybooks.de/autor/Cedric-Villani/Topi...

4 Untitled document 0 facts | 2011-02-24
http://www.inclassablesmathematiques.fr/archive/201...

5 Cédric Villani 12 facts | 2010-02-18
http://oai.rkbexplorer.com/id/dialnet.unirioja.es/p...

The Semantic Web

The **Semantic Web** is an evolving extension of the World Wide Web, with the aim to

- make computers „understand“ the data they store
- allow them to reason about information
- allow them to share information across different systems

For this purpose, the **Word Wide Web Consortium** (W3C) defines standards for

- identifying entities in a globally unique way (URIs) ✔
- defining semantics in a machine-readable way (RDF) ✔
- defining taxonomies (RDFS) ✔
- defining logical consistency in a uniform way (OWL) ✔
- storing ontologies (N3, XML, RDFa) ✔
- sharing ontologies (Cool URIs) ✔
- querying ontologies (SPARQL) ✔

Great, now where do we get the data from? ✔

And how does the Semantic Web look in practice? ✔

SW: References

[DBpedia] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann.
Dbpedia - a crystallization point for the web of data.
J. Web Semant., 7:154–165, September 2009.

[LD] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee.
Linked data on the Web. In WWW 2008, http://linkeddata.org

[NELL] Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., Tom M. Mitchell.
Coupled semi-supervised learning for information extraction. In WSDM 2010.

[OntoUSP] Hoifung Poon and Pedro Domingos.
Unsupervised ontology induction from text.
In ACL 2010.

[OWL] World Wide Web Consortium. OWL 2 Web Ontology Language,
W3C Recommendation,2009. http://www.w3.org/TR/owl2-overview/

[PROSPERA] Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum.
Scalable knowledge harvesting with high precision and high recall.
In WSDM 2011

[RDF] World Wide Web Consortium. RDF Primer, W3C Recommendation, 2004.
http://www.w3.org/TR/rdf-primer/

SW: References

[RDFS] World Wide Web Consortium. RDF Vocabulary Description Language 1.0: RDF Schema, W3CRecommendation, 2004. http://www.w3.org/TR/rdf-schema/

[SIGMA] Giovanni Tummarello, Richard Cyganiak, Michele Catasta, Szymon Danielczyk, Renaud Delbru, Stefan Decker. Sig.ma: Live views on the Web of Data Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 8, No. 4. (November 2010), pp. 355-364.

[SOFIE] Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. SOFIE: A Self-Organizing Framework for Information Extraction. In WWW 2009

[SPARQL] World Wide Web Consortium. SPARQL Query Language for RDF, W3C Recommendation,2008. http://www.w3.org/TR/rdf-sparql-query/

[URI] Network Working Group. Uniform Resource Identifier (URI): Generic Syntax, 2005. http://tools.ietf.org/html/rfc3986

[WordNet] C. Fellbaum, editor. WordNet: An Electronic Lexical Database. MIT Press, 1998.

[YAGO] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO - A Large Ontology from Wikipedia and WordNet. Elsevier Journal of Web Semantics, 6(3):203–217, September 2008.

[YAGO2] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis Kelham, Gerard de Melo, andGerhard Weikum. Yago2: Exploring and querying world knowledge in time, space, context, and many languages. In WWW, 2011.

Overview

- Introduction ✔
- The Hidden Web ✔
- XML ✔
- DSML ✔
- The Semantic Web ✔
- Conclusion

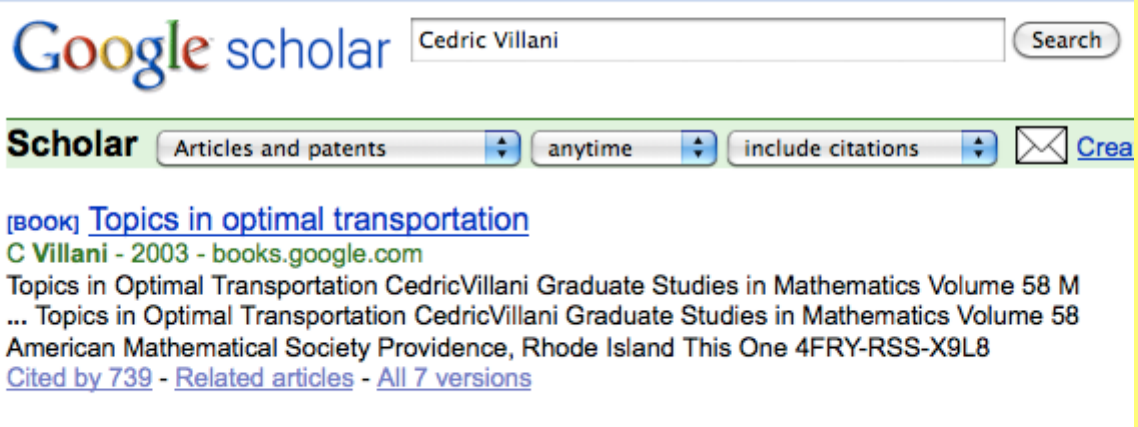
Conclusion

The Internet is not just Web pages.

There are

- the Hidden Web

The Hidden Web is the data available through forms.
It contains at least as much data as the surface Web



Cedric Villani

[BOOK] Topics in optimal transportation
C Villani - 2003 - books.google.com
Topics in Optimal Transportation CedricVillani Graduate Studies in Mathematics Volume 58 M ... Topics in Optimal Transportation CedricVillani Graduate Studies in Mathematics Volume 58 American Mathematical Society Providence, Rhode Island This One 4FRY-RSS-X9L8
Cited by 739 - Related articles - All 7 versions

This information can be exploited through

- intentional techniques („understanding“ the service)
- extensional techniques (crawling the service)

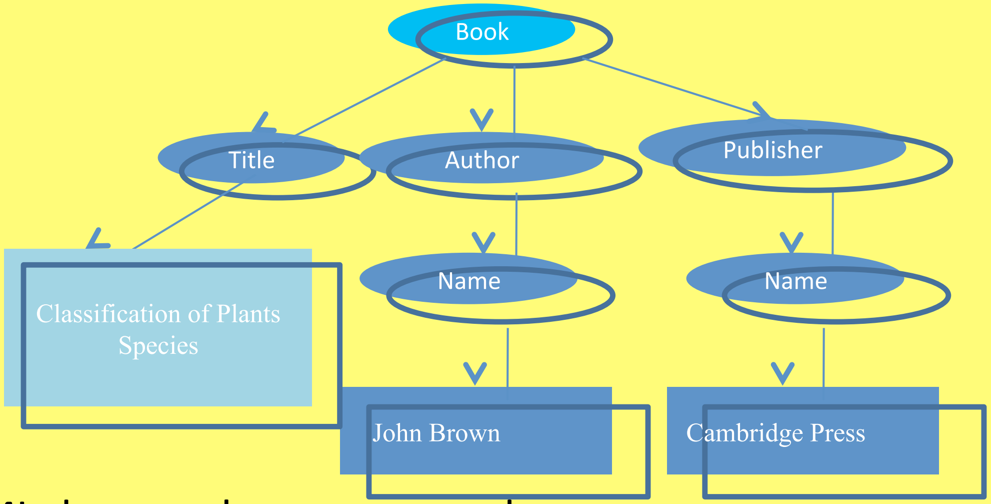
Conclusion

The Internet is not just Web pages.

There are

- the Hidden Web
- XML

XML is the lingua franca of information exchange.



XML data can be represented

- as trees
- as matrices
- as sequential text files

...which can serve different mining purposes.

The output of the mining helps in focused information retrieval.

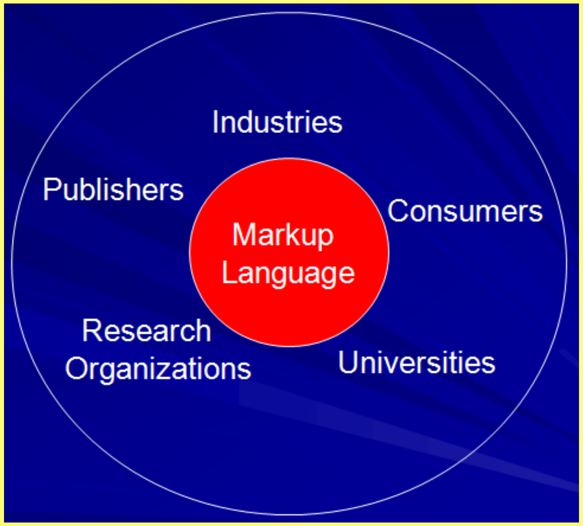
Conclusion

The Internet is not just Web pages.

There are

- the Hidden Web
- XML
- DSML

Domain specific markup languages give semantics to XML.



DSML design involves

- data modeling
- ontology creation
- schema development

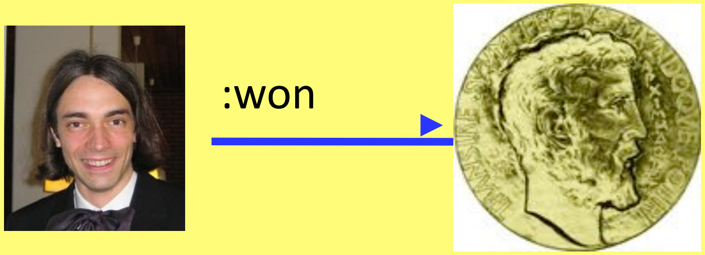
Conclusion

The Internet is not just Web pages.

There are

- the Hidden Web
- XML
- DSML
- the Semantic Web

The Semantic Web aims at standardizing the way semantic information is published.



The standards are

- URIs for identifying entities
- RDF for expressing facts
- OWL for reasoning

Conclusion

The Internet is not just Web pages.

There are

- the Hidden Web

How can we better guess the purpose of a Web service?
Howe can we understand the semantics of the form fields?

- XML

How can we scale up the mining process?
How can we find semantic tags for an XML document?

- DSML

How do we enforce consistency across DSMLs?
How do we use the semantics of DSMLs in retrieval?

- the Semantic Web

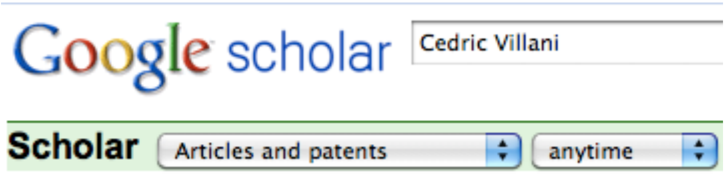
How can we grow ontologies automatically?
How can we interlink the existing ones?

These developments are by no means finalized, but active fields of research.

These developments also give us unprecedented sources of new information.

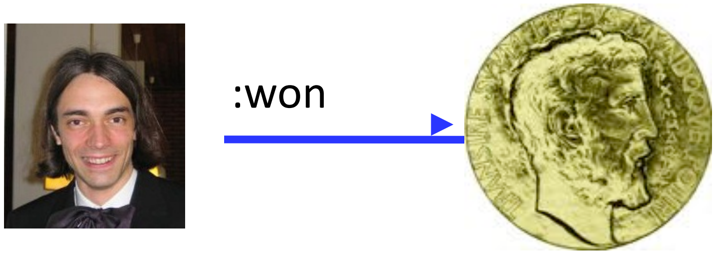
Conclusion

These developments give us unprecedented sources of new information, for example on the question of whether we should hire Cedric Villani...



Google scholar Cedric Villani

Scholar Articles and patents anytime



```
<math xmlns="http://www.w3.org/1998/Math/MathML">
     Σ x: ⋀ ⅔ ≈ ∞ × ⅝ Ω
</math>
```

... and the answer is probably YES

Thank you for your attention.