digiteo
Recherche en sciences & technologies de l'information

FORUM 2011
4E ÉDITION

Linked OpenData: Scientific Challenges and Applications

Ioana Manolescu-Goujot

Leo team

INRIA Saclay / U. Paris Sud-11 / CNRS

http://team.inria.fr/leo

Leo

Inria
INVENTEURS DU MONDE NUMÉRIQUE

LABORATOIRE DE RECHERCHE EN INFORMATIQUE

Plan

1. The original Web vision: short recall
2. First incarnation: XML
3. Second incarnation: RDF
4. Linked (Open) Data: the Web for the machines
5. More scientific problems around LOD
6. Conclusion

The perspective is quite influenced by the recently written "DigiWorlds" LabEx proposal

An early vision of the Web

A short history of the Web

Internet (the network)

- 1960: DARPA network
- 1986: TCP/IP
- 1989: Tim Berners-Lee proposal for an information system for the CERN (http://www.w3.org/History/1989/proposal.html)

A short history of the Web

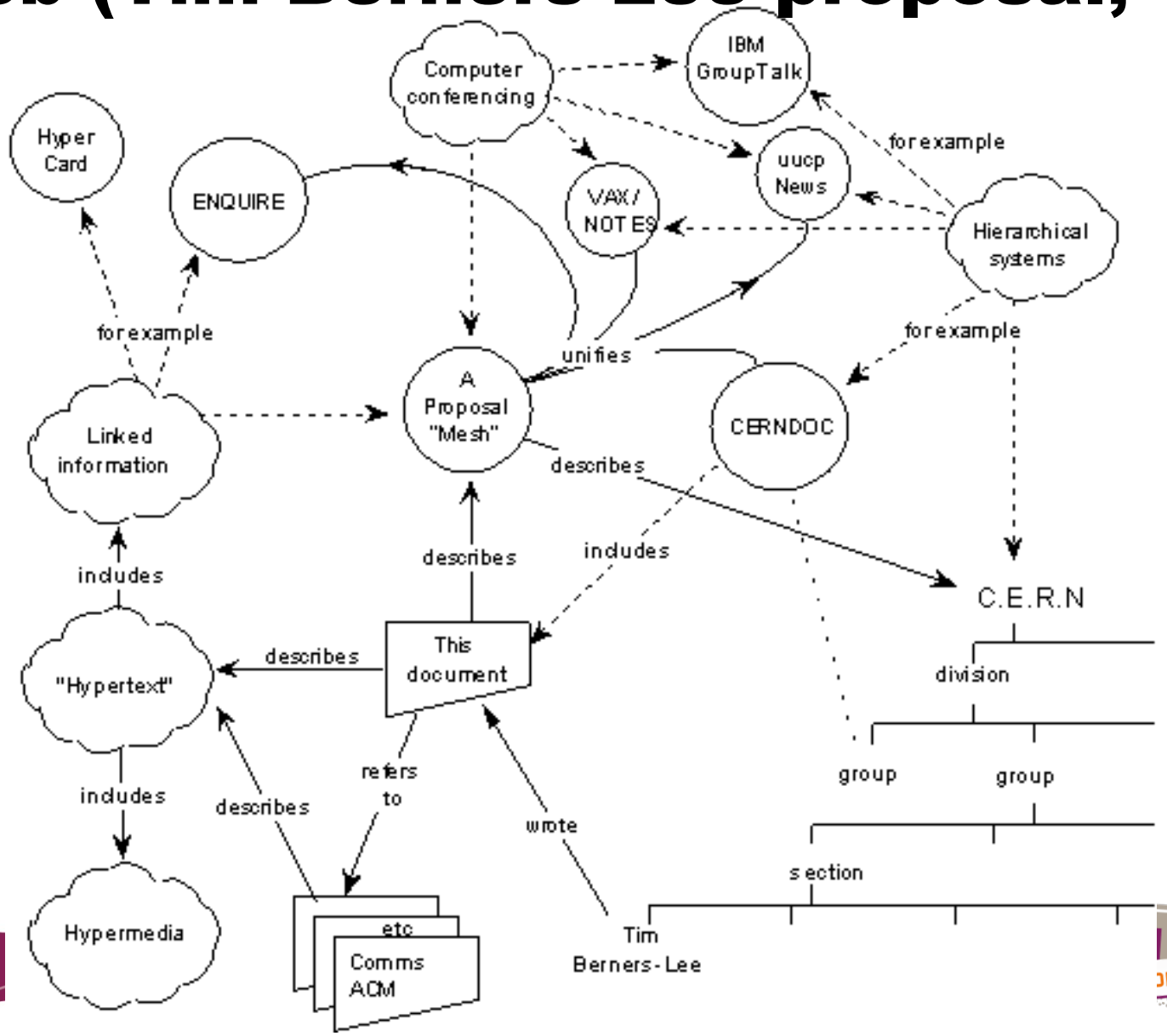
Internet (the network)

- 1960: DARPA network
- 1986: TCP/IP
- 1989: Tim Berners-Lee proposal for an information system for the CERN (http://www.w3.org/History/1989/proposal.html)

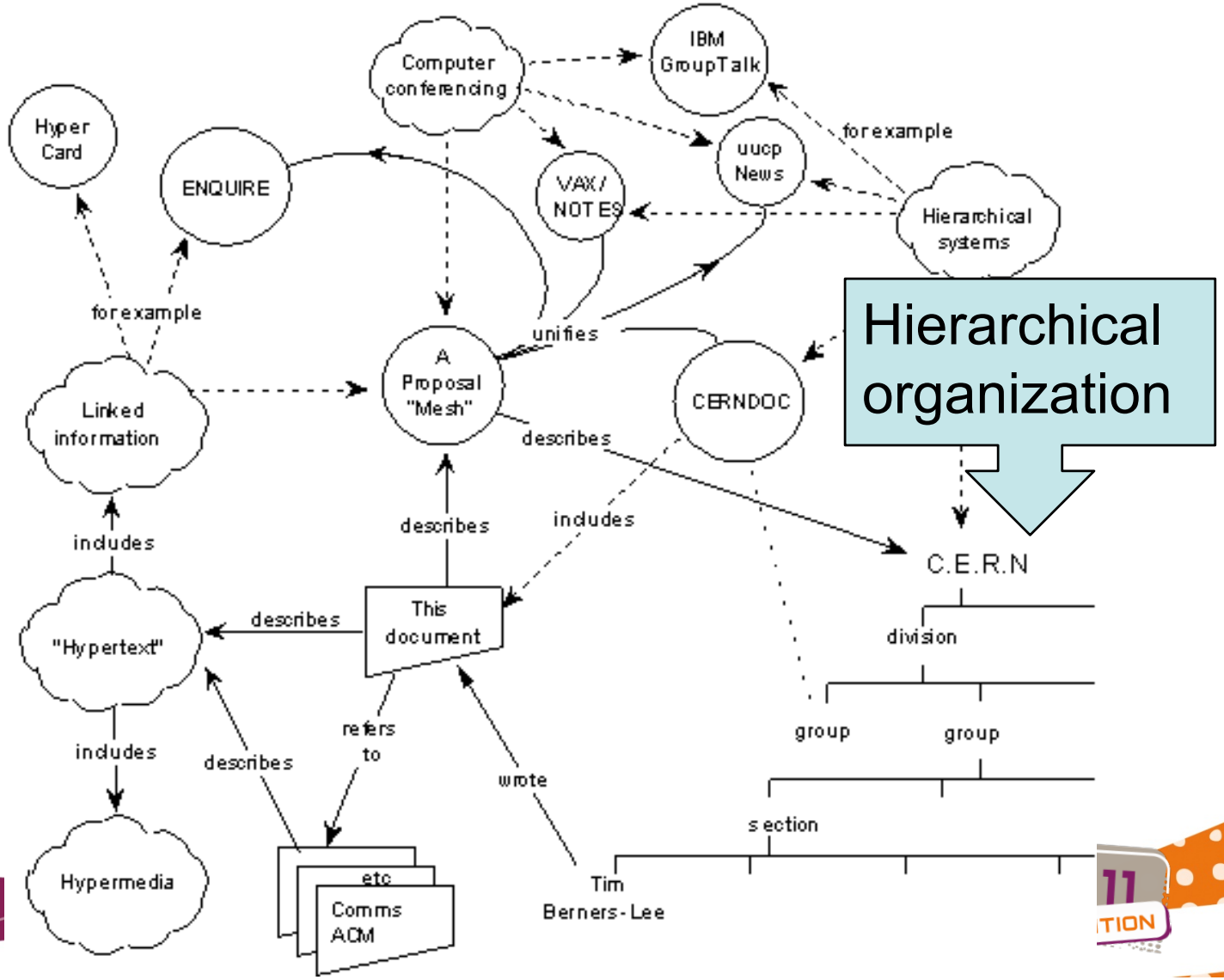
"*This proposal discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system. The sort of information we are discussing answers, for example, questions like:*

- *Where is this module used?*
- *Who wrote this code? Where does he work?*
- *What documents exist about that concept?*
- *Which laboratories are included in that project?*
- *Which systems depend on this device?*
- *What documents refer to this one?*"

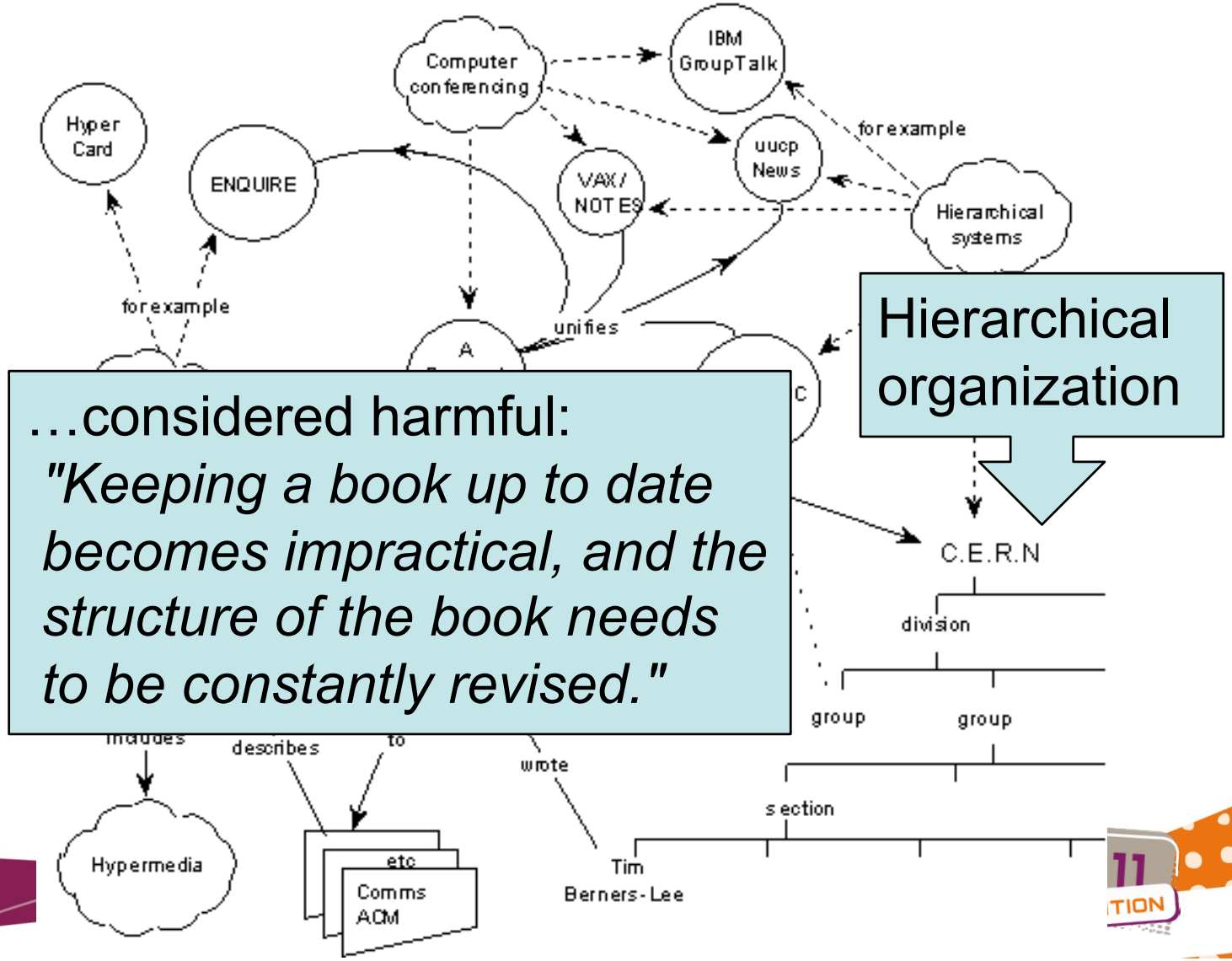
The Web (Tim Berners-Lee proposal, 1989)



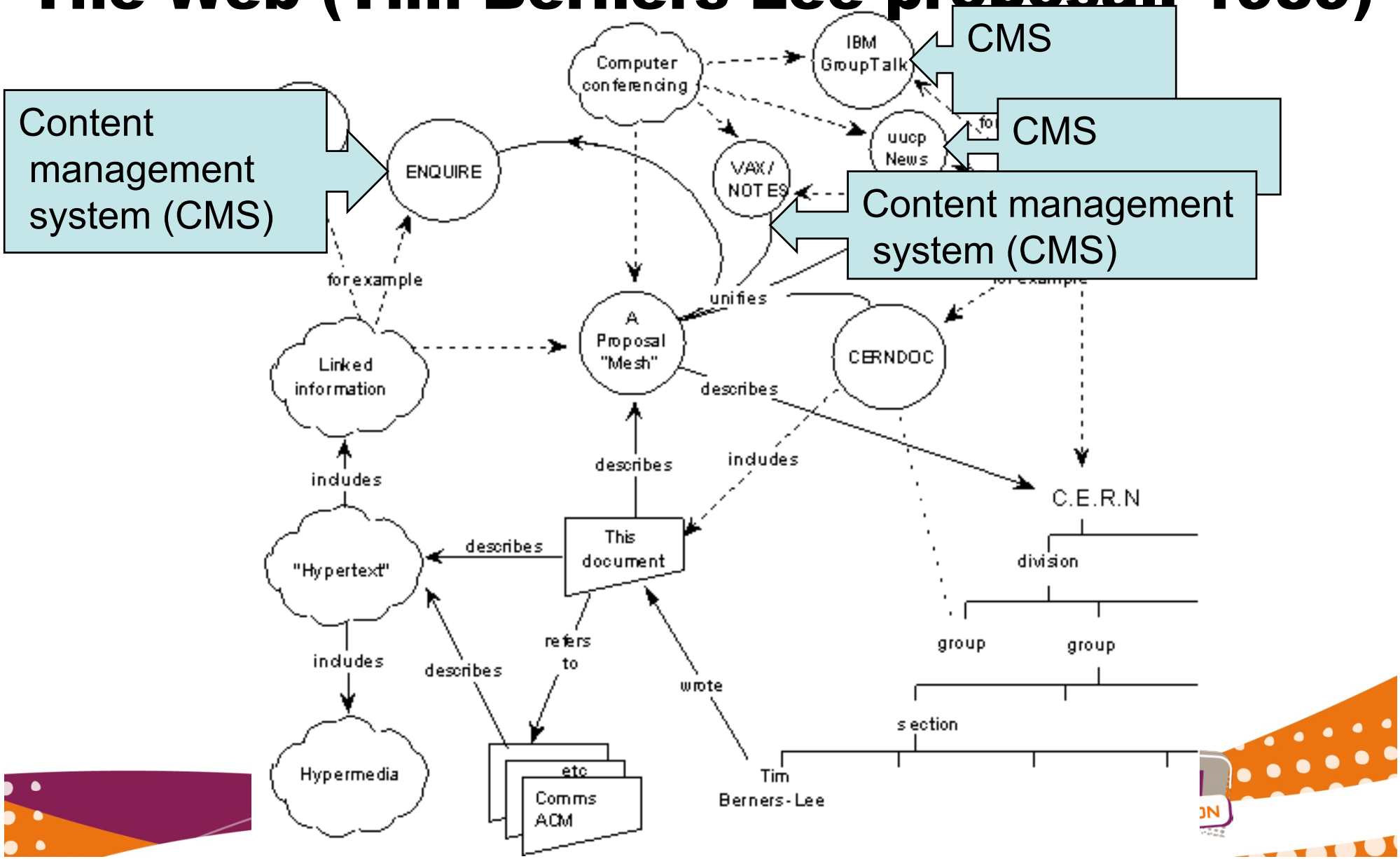
The Web (Tim Berners-Lee proposal, 1989)



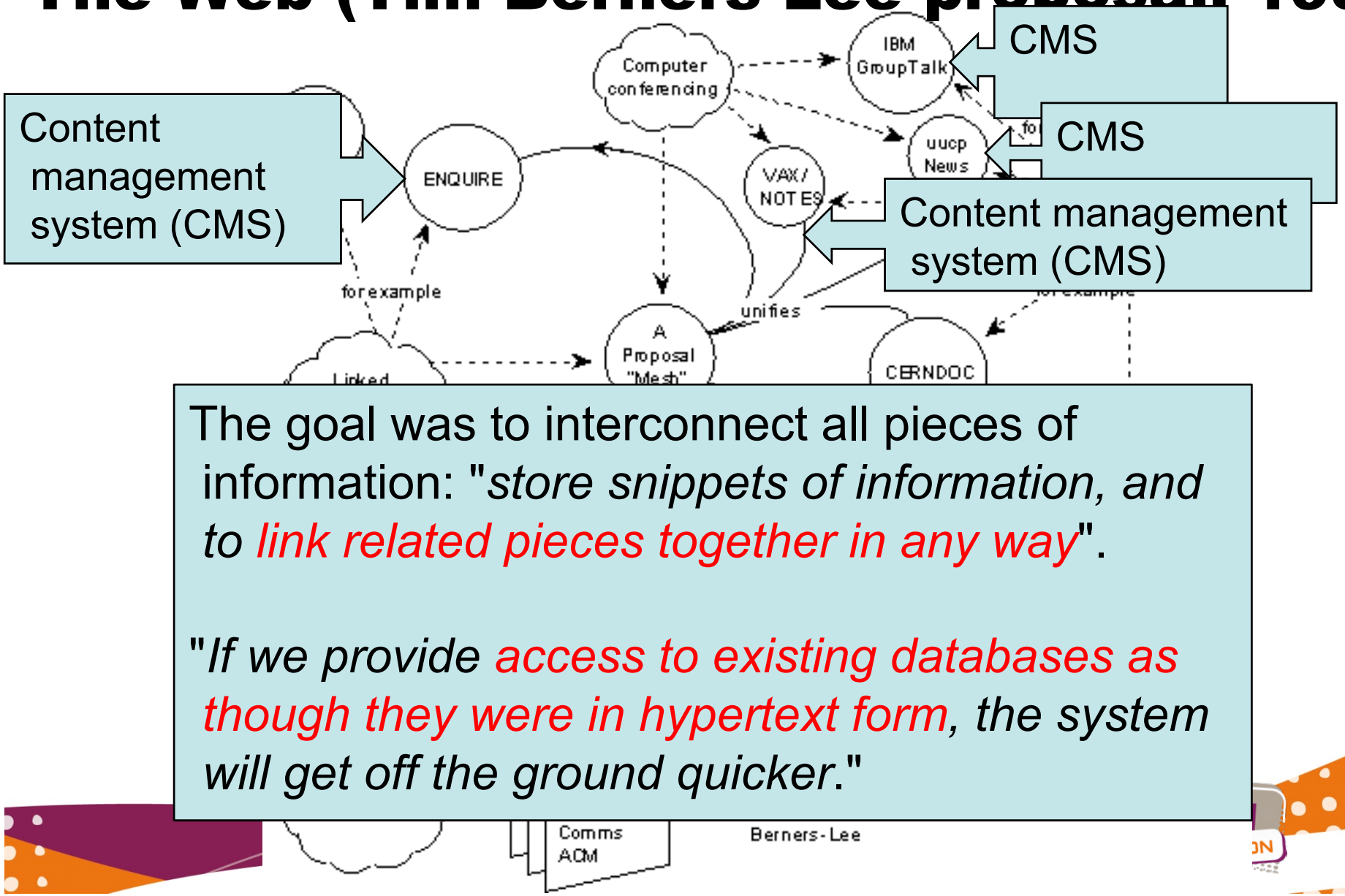
The Web (Tim Berners-Lee proposal, 1989)



The Web (Tim Berners-Lee proposal, 1989)



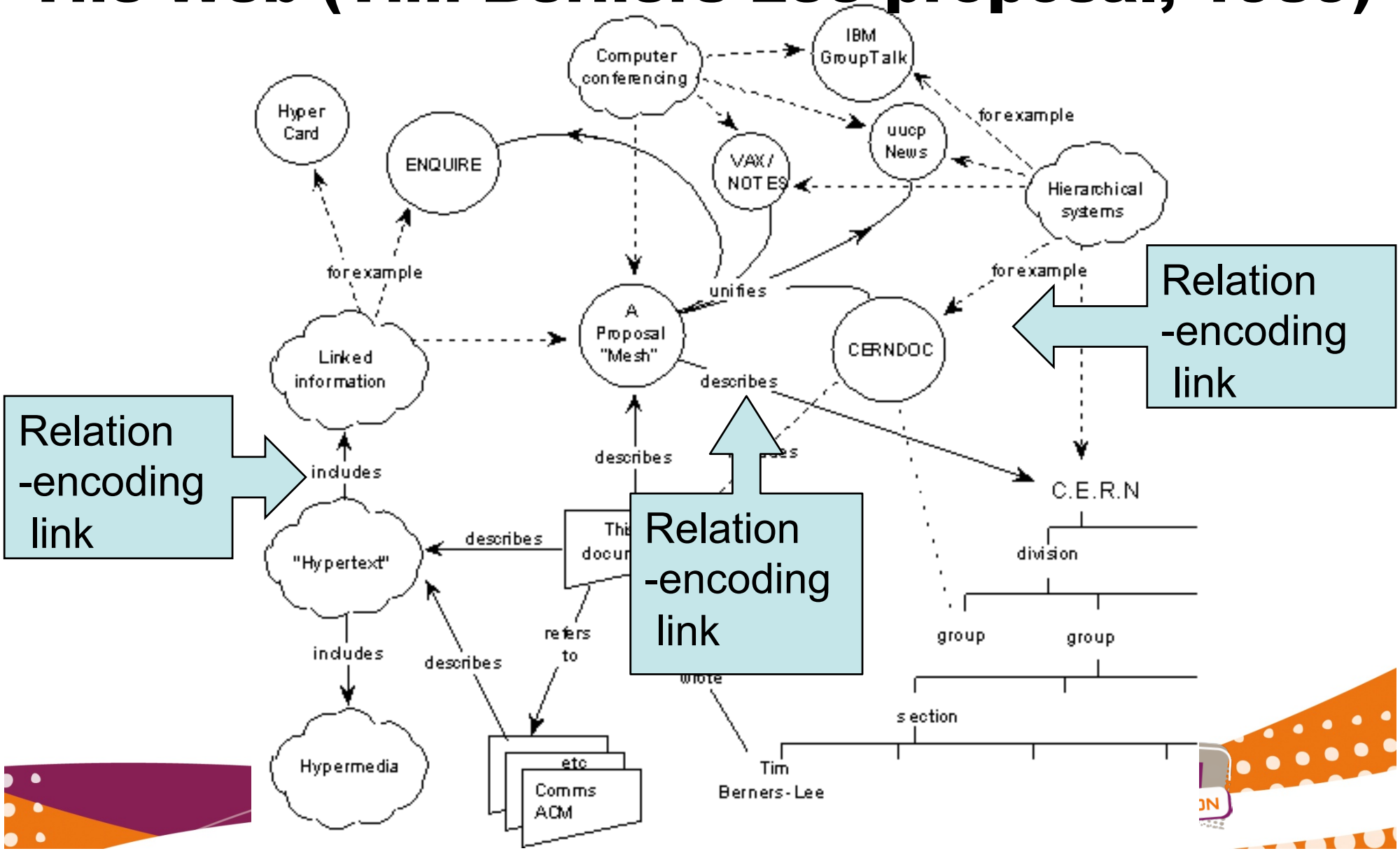
The Web (Tim Berners-Lee proposal, 1989)



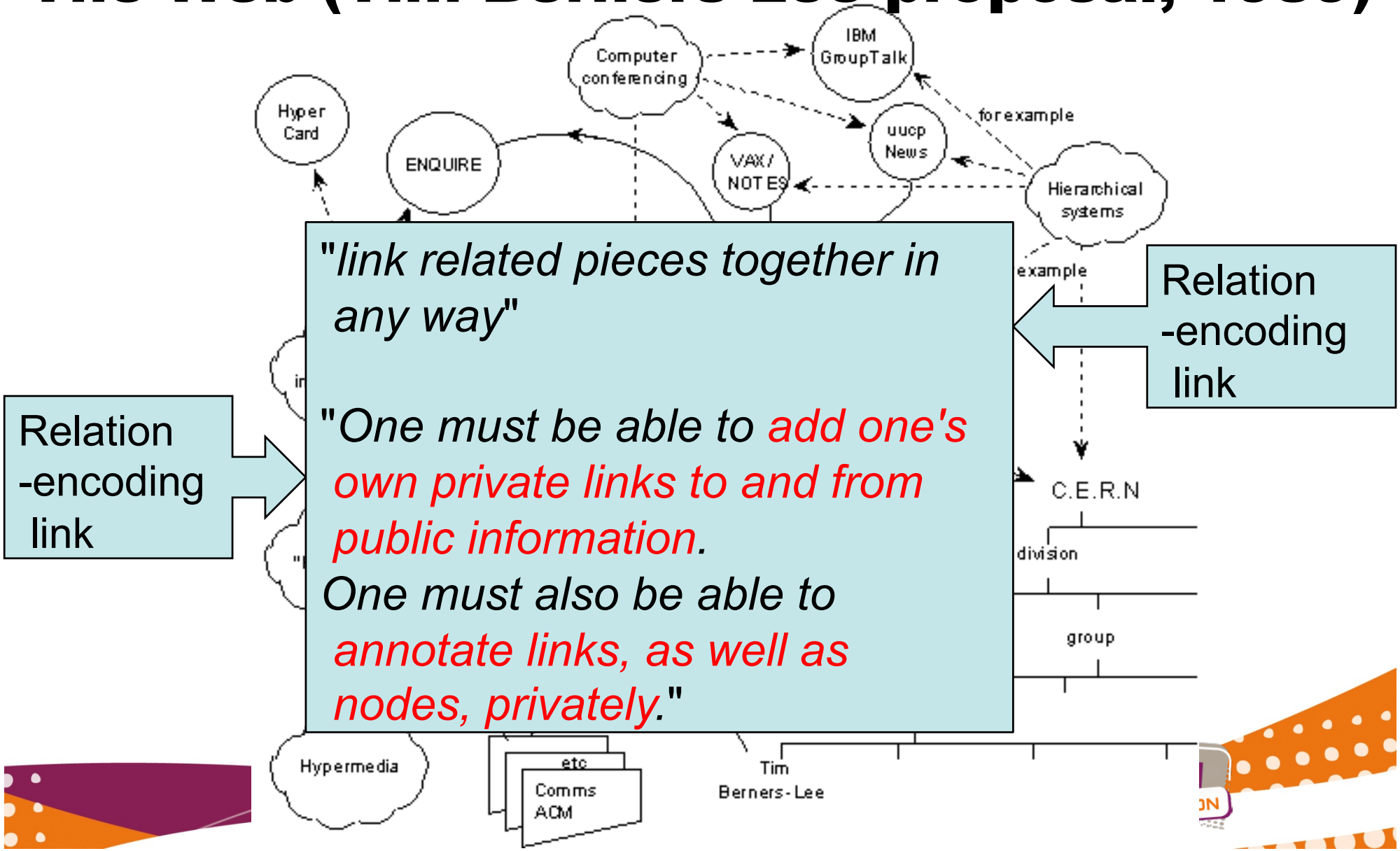
The goal was to interconnect all pieces of information: "*store snippets of information, and to link related pieces together in any way*".

"*If we provide access to existing databases as though they were in hypertext form, the system will get off the ground quicker.*"

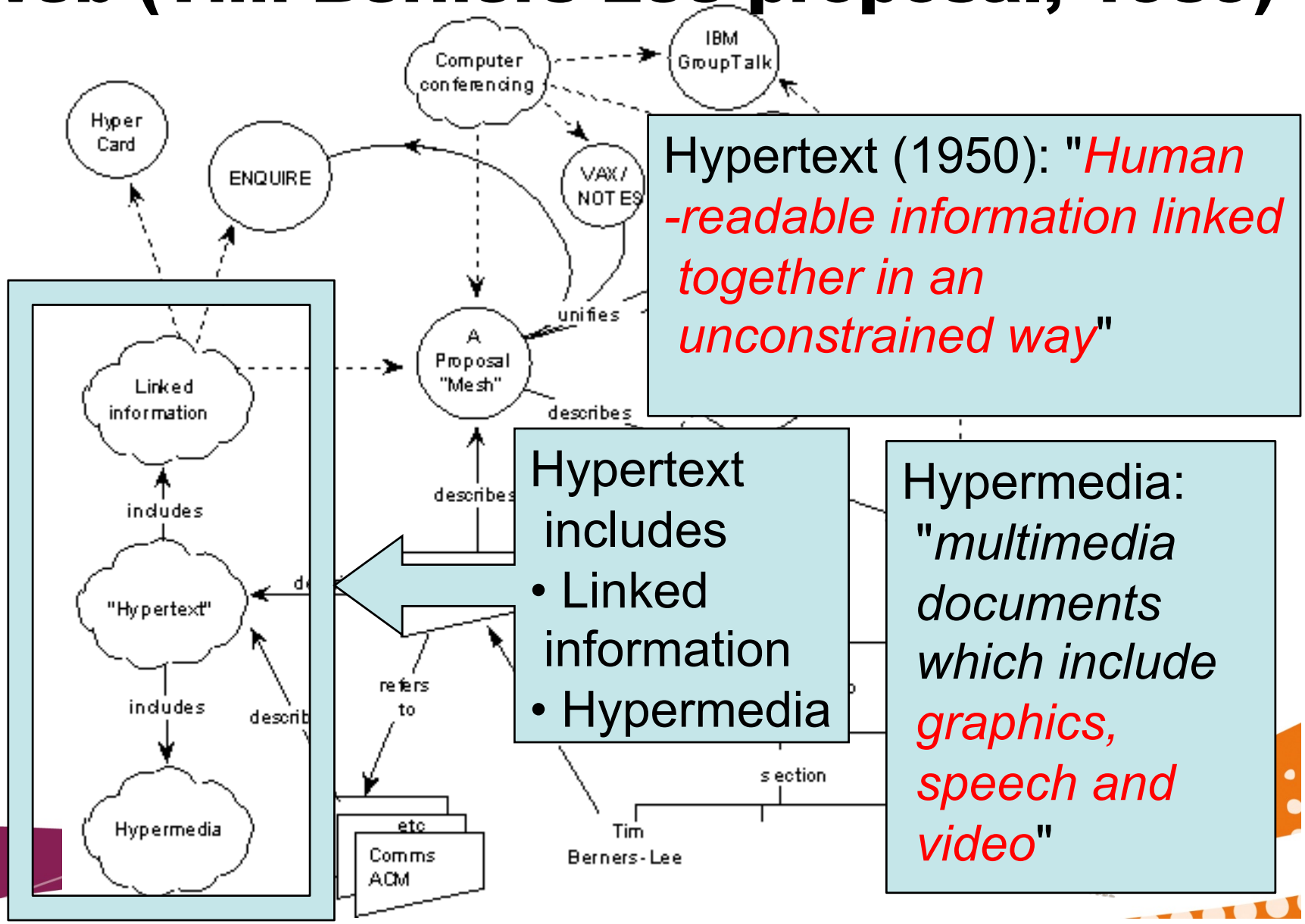
The Web (Tim Berners-Lee proposal, 1989)



The Web (Tim Berners-Lee proposal, 1989)



The Web (Tim Berners-Lee proposal, 1989)



Hypertext (1950): "*Human -readable information linked together in an unconstrained way*"

Hypertext includes
- Linked information
- Hypermedia

Hypermedia: "*multimedia documents which include graphics, speech and video*"

The Web, continued

- **Internet (the network)**
  - 1960: DARPA network
  - 1986: TCP/IP
  - 1989: Tim Berners-Lee proposal for an information system for the CERN
  - 1991: HTTP, 1995: commercial Internet
- **The Web as a database (first generation)**
  - Programs **exchange data over the Web**.
  - First applications: e-commerce sites (Junglee → Amazon, U. Stanford)
    - Many heterogeneous data sources → **self-describing data**
    - 1998: XML
      - Tree-structured, "loose" format for complex data
      - "Clean HTML": **separate content from presentation**

digiteo
Recherche en sciences & technologies de l'information

A first incarnation of the World Wide Web vision:

XML

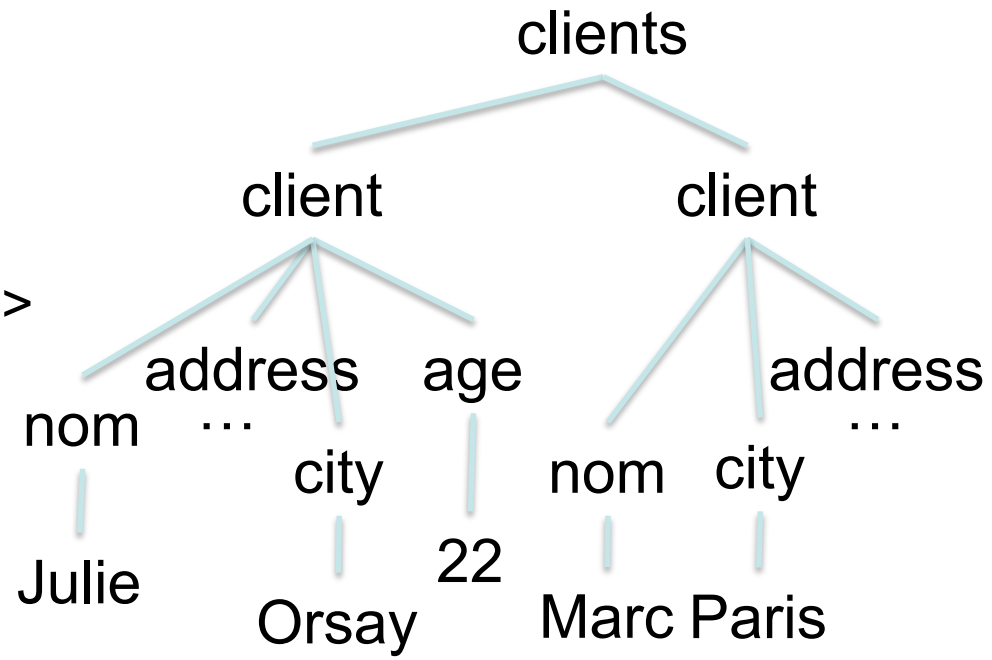
(World Wide Web Consortium, 1998)

FORUM 2011
4E ÉDITION

Self-describing data: XML

clients.xml:

```
<clients>
 <client><nom>Julie</nom>
  <address>1,rue Dugommier</address>
  <city>Paris</city><age>22</age>
 </client>
 <client><nom>Marc</nom>…
 </client>
</clients>
```



- Flexible
- Platform-independent
- Separate content from presentation
- Schema possible (not compulsory)

Applications enabled by XML

- All kinds of content management on the Web
  - Multiple presentation for the same information (XSL, CSS → mobile devices…)
  - Exporting structured (database) data through Web pages
  - News feeds

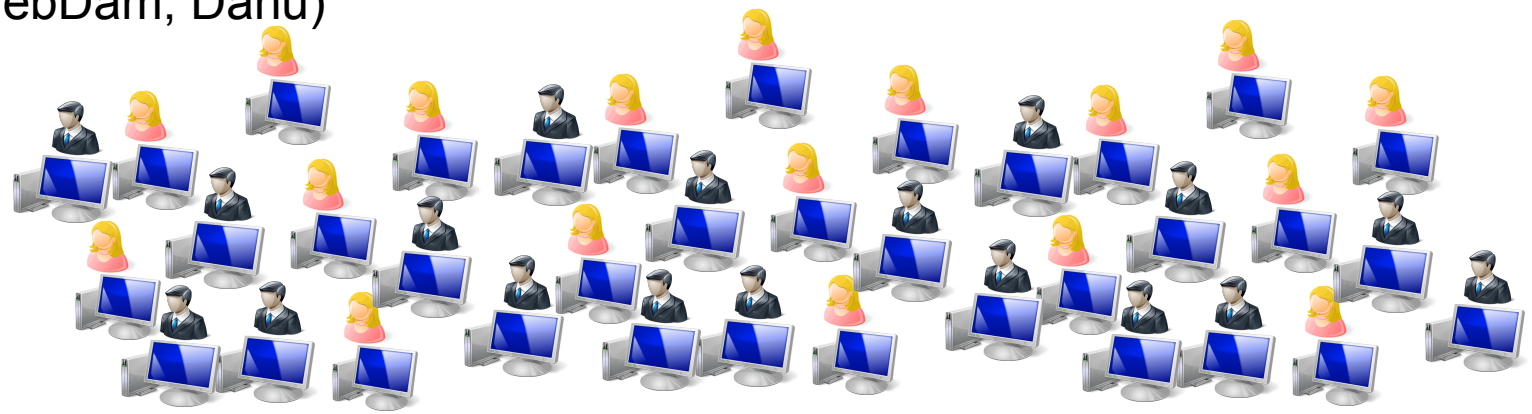
Applications enabled by XML

Source de : http://www.inria.fr/

```
'/institut/relations-internationales"><span>Relations internationales</span></a>


    <ul>
                <li><a href="/institut/relations-internationales/mot-d-helene-kirchner">Mot d'Hélène Kirchner</a></li>
                <li><a href="/institut/relations-internationales/partenariats-strategiques2">Partenariats stratégiques</a><
                <li><a href="/institut/relations-internationales/actions-dans-le-monde">Actions dans le monde</a></li>
                <li><a href="/institut/relations-internationales/appels-a-projets">Appels à projets</a></li>
                <li><a href="/institut/relations-internationales/contacts">Contacts</a></li>
        </ul>
                </li>
                        <li><a href="/institut/partenariats"><span>Partenariats</span></a>


    <ul>
                <li><a href="/institut/partenariats/partenariats-academiques">Partenariats académiques</a></li>
                <li><a href="/institut/partenariats/partenariats-industriels">Partenariats industriels</a></li>
                <li><a href="/institut/partenariats/partenariats-europeens">Partenariats européens</a></li>
        </ul>
                </li>
                        <li><a href="/institut/recrutement-metiers"><span>Recrutement &amp;amp; métiers</span></a>


    <ul>
                <li><a href="/institut/recrutement-metiers/mot-de-muriel-sinanides">Mot de Muriel Sinanidès</a></li>
                <li><a href="/institut/recrutement-metiers/diversite-de-nos-metiers">Diversité de nos métiers</a></li>
                <li><a href="/institut/recrutement-metiers/nous-rejoindre">Nous rejoindre</a></li>
                <li><a href="/institut/recrutement-metiers/offres">Offres</a></li>
        </ul>
                </li>
            </ul>
```

Applications enabled by XML

- All kinds of content management on the Web
  - Multiple presentation for the same information (XSL, CSS → mobile devices…)
  - Exporting structured (database) data through Web pages
  - News feeds
- Automated communication between programs on the Web
  - **Web services** → coordination, synchronisation, typing… INRIA/LRI Mexico, Fortesse, …
  - **Active XML**: XML including calls to Web services (INRIA Gemo/Leo → ERC WebDam, Dahu)



XML: some interesting problems

- Efficient processing
  - Large data volumes accumulating, complex query/update language XQuery
  - Database techniques: materialized views (Leo)
  - Static analysis, type-driven techniques (Leo, Proval)
  - Streaming (Mostrare@Lille + Innovimax)
  - Tree automata techniques for expressing XML computations (Proval)
  - Scaling up to the cloud through Map-Reduce extensions (Leo, with TU Berlin)
- Probabilistic XML (DBWeb @ Telecom ParisTech, ERC WebDam):
  - XML data may come with uncertainty (extracted from multiple Web sources, result of reconciliation, result of uncertain devices…)
  - Uncertainty is computed and preserve through query evaluation
  - Algorithmic complexity issues

CODEX project ANR-08-DEFIS-004

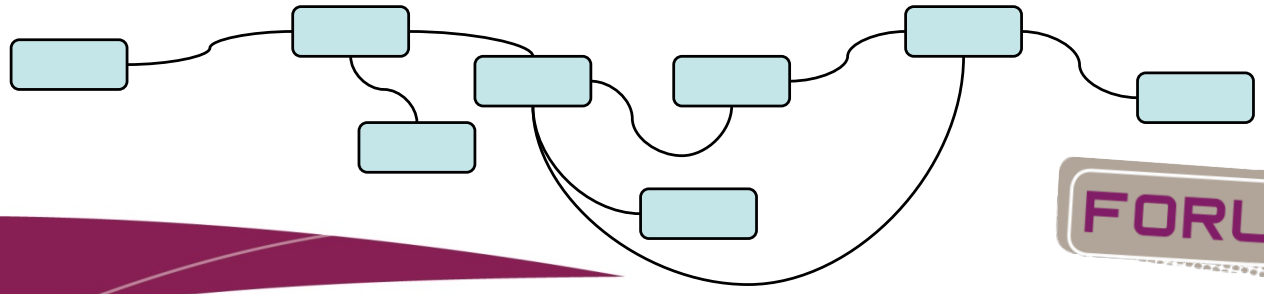
EIT ICT Labs "Europa" with TU Berlin, SICS, TU Delft, KTH etc ("Cloud Computing" Research Action line)

Critique of XML: each information can appear in only one place

- "Classification" applications do fine, also structured text
- Fundamentally restrictive for **data = real world**!

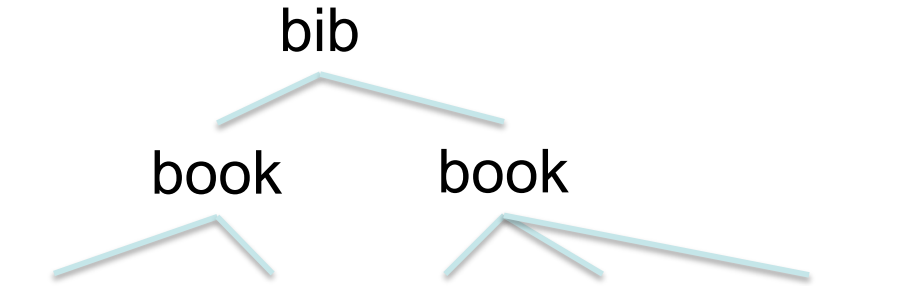
"*Many systems are organised hierarchically. A tree has the practical advantage of giving every node a unique name. However,* ***it does not allow the system to model the real world***."

(On newsgroups)*: "Typically, a discussion under one newsgroup will develop into a different topic, at which point* ***it ought to be in a different part of the tree***."

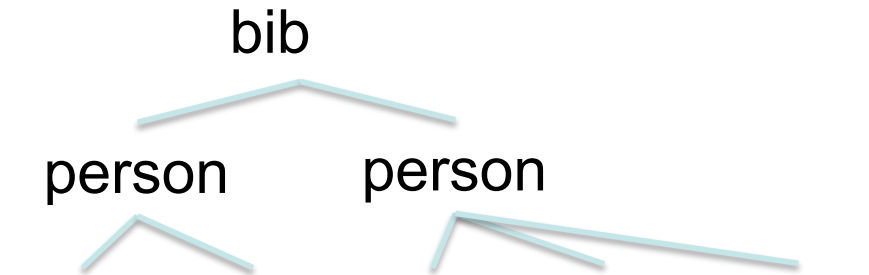


The librarian's dilemma

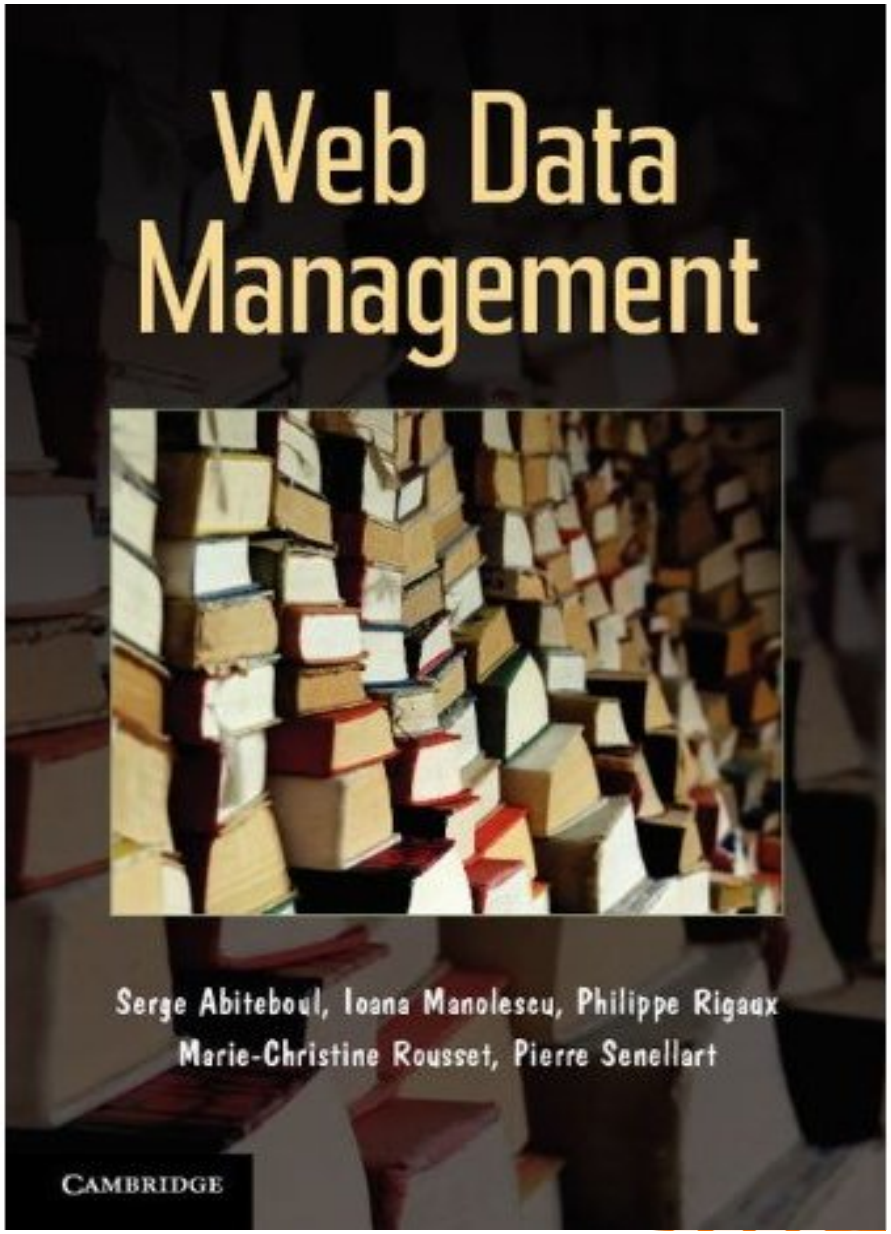
Organize by author or by book?



/bib/book[author="Serge"]



/bib/person[name="Serge"]/book



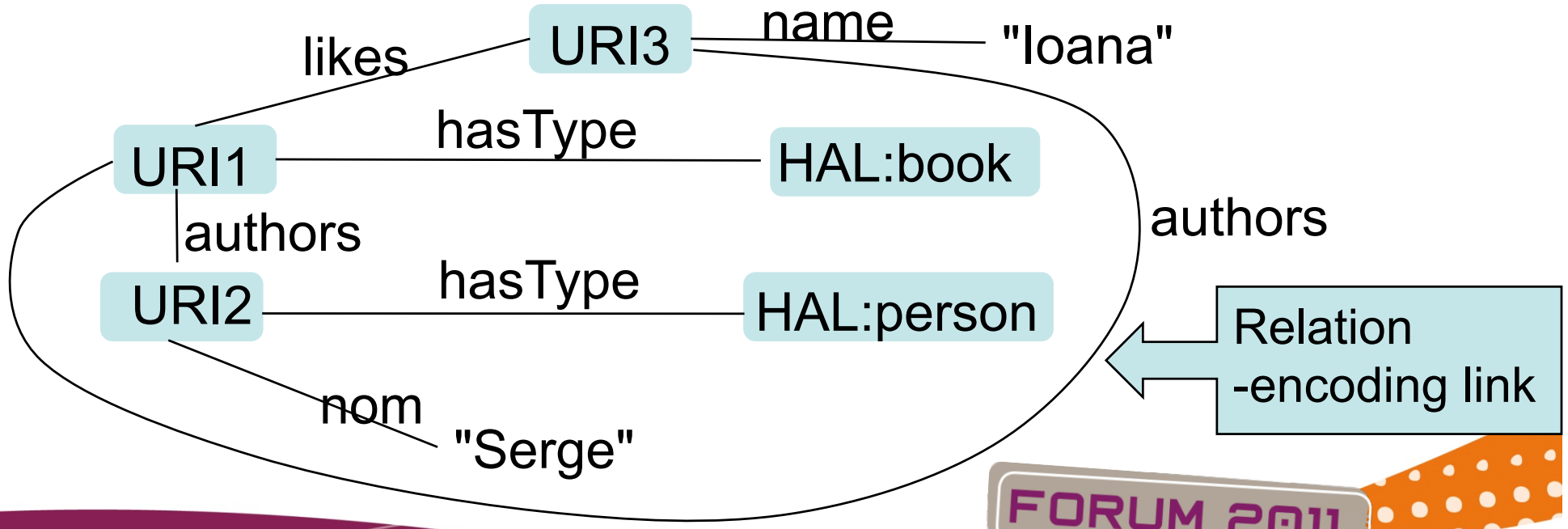
A second incarnation of the World Wide Web vision:

RDF

(World Wide Web Consortium, 2003)

Resource Description Framework (RDF)

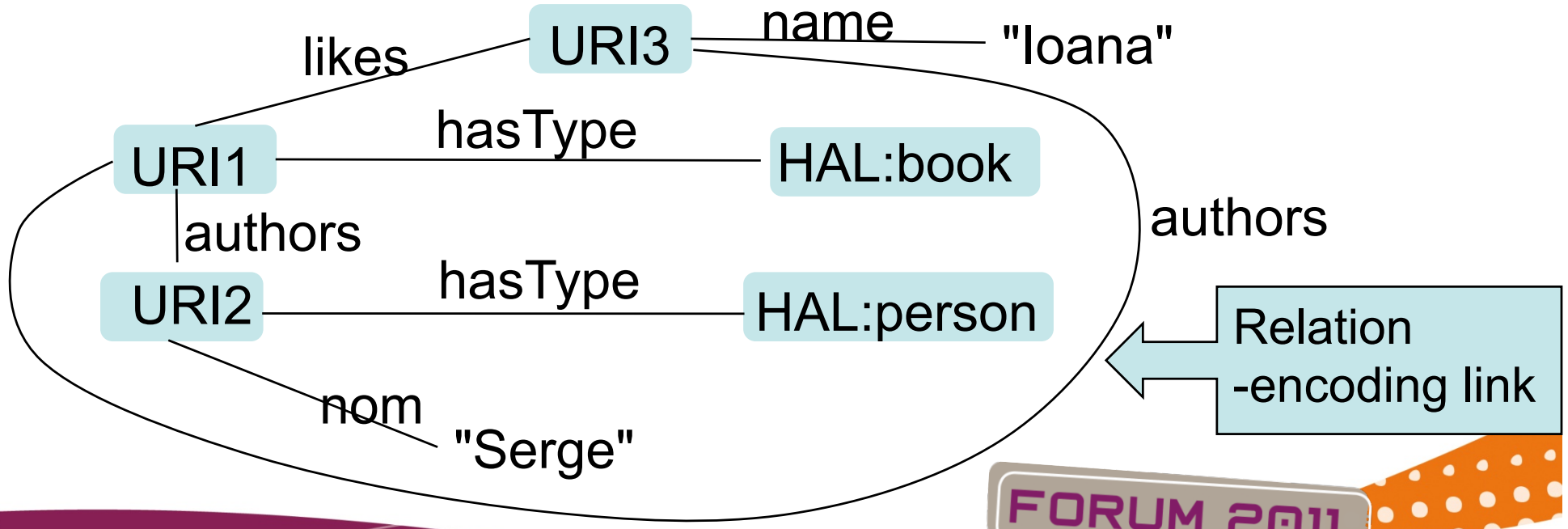
- Resources have properties.
- Resources have URIs (Universal Resource Identifiers)
- Properties have names (which are also URIs)
- An entity's property value is either a resource, or a simple value



Reasoning on RDF data

- Types are special properties
- They enable reasoning according to rules that are part of the RDF semantics

Ex: HAL:person subclassOf INSEE:person → URI2 hasType INSEE2:person



Improving RDF query performance through materialized views

Problem: RDF data has no regularity, no structure → query processing performance degrades

Input: RDF database D, RDF Schema S, workload $\{Q_1, Q_2, \ldots, Q_n\}$

Output: Set of views $\{V_1, V_2, \ldots, V_k\}$ to materialize in order to minimize cost (workload processing + view storage and maintenance)

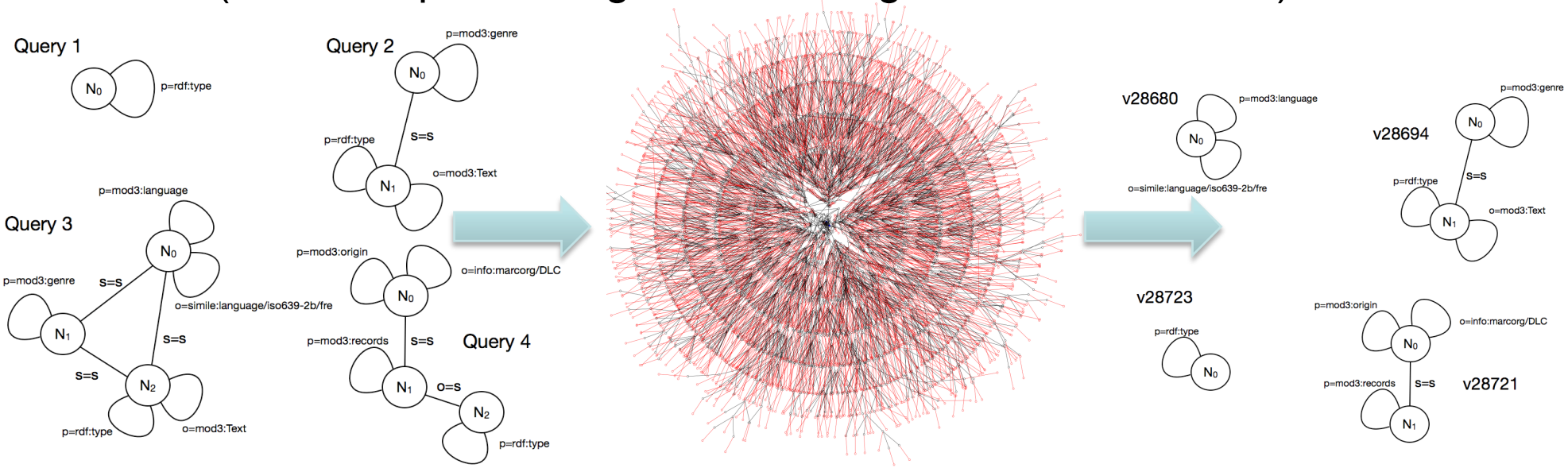
Difficulties: implicit RDF data, large queries

Leo paper @ PVLDB 2011

Improving RDF query performance through materialized views

Input: RDF database D, RDF Schema S, workload $\{Q_1, Q_2, \ldots, Q_n\}$

Output: Set of views $\{V_1, V_2, \ldots, V_k\}$ to materialize in order to minimize cost (workload processing + view storage and maintenance)



Linked (Open) Data:

the World Wide Web vision for the machines

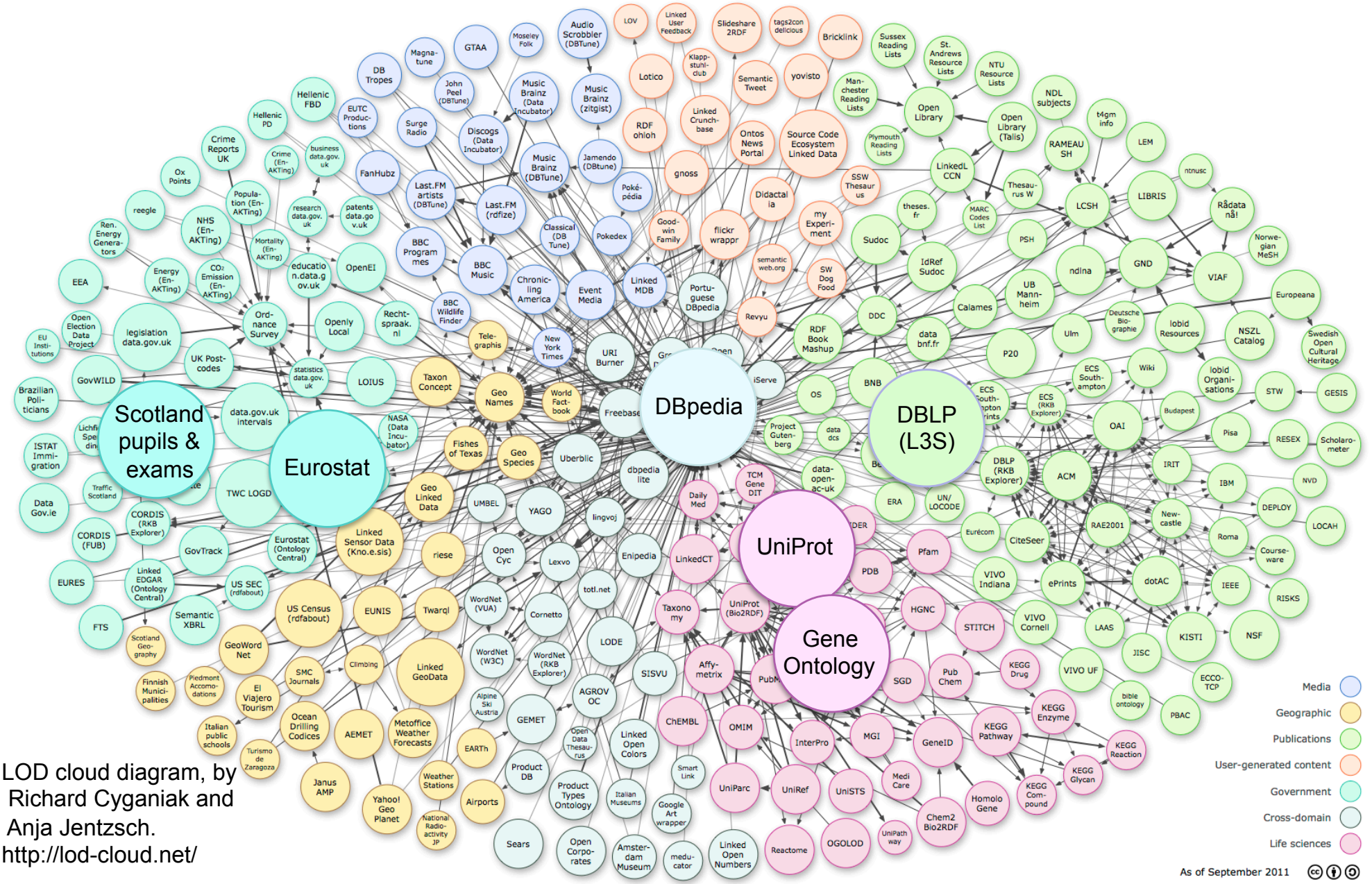
FORUM 2011
4E ÉDITION

Linked vs. Open Data

1. **Linked Data**:
   "recommended **best practice** for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF"
   - (Tim Berners-Lee) vision for the Web
2. **Open Data**:
   "i**dea** that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control"
   - In principle, orthogonal to the Linked aspect
   - In practice, Linked is a technical mean toward Open

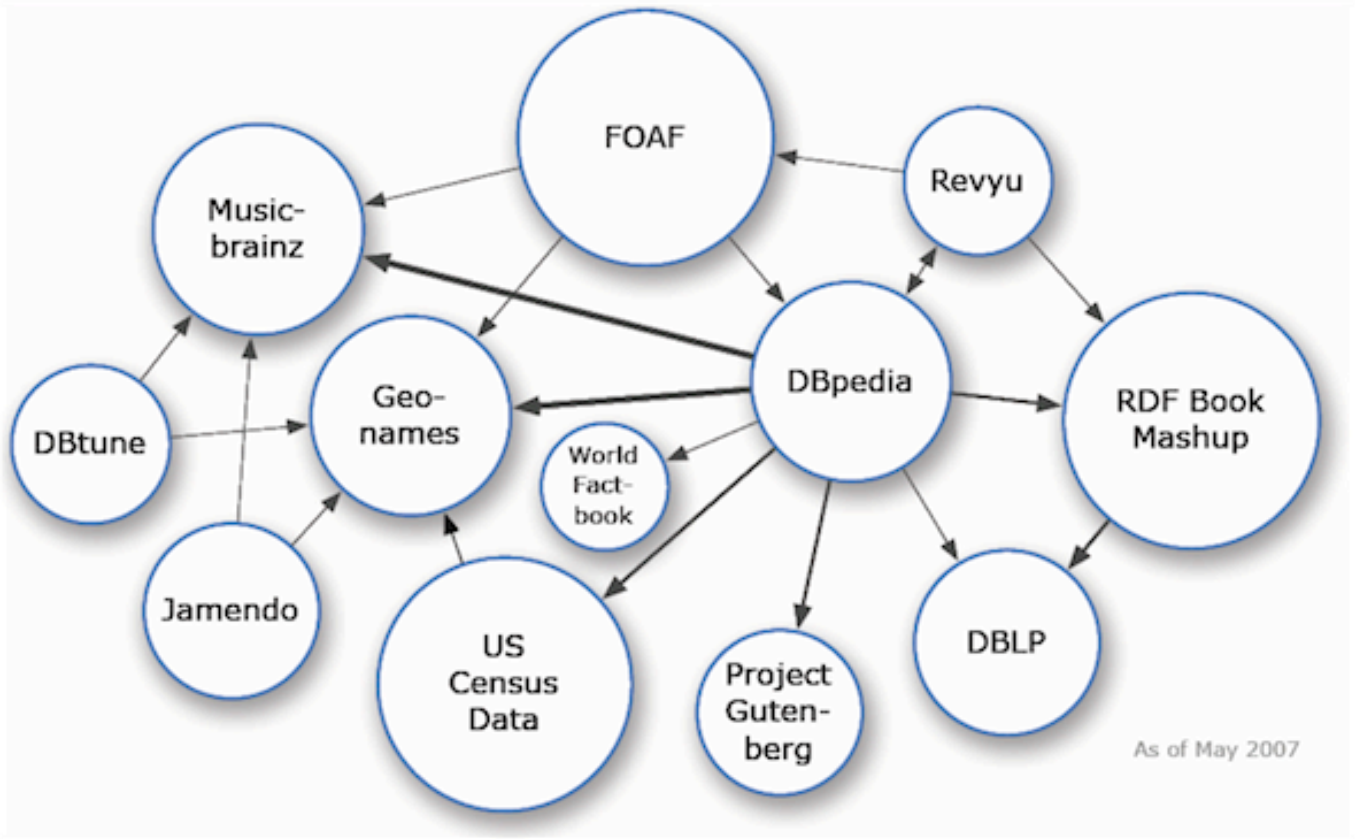
digiteo
Recherche en sciences & technologies de l'information

Linked Open Data Cloud



LOD cloud diagram, by Richard Cyganiak and Anja Jentzsch.
http://lod-cloud.net/

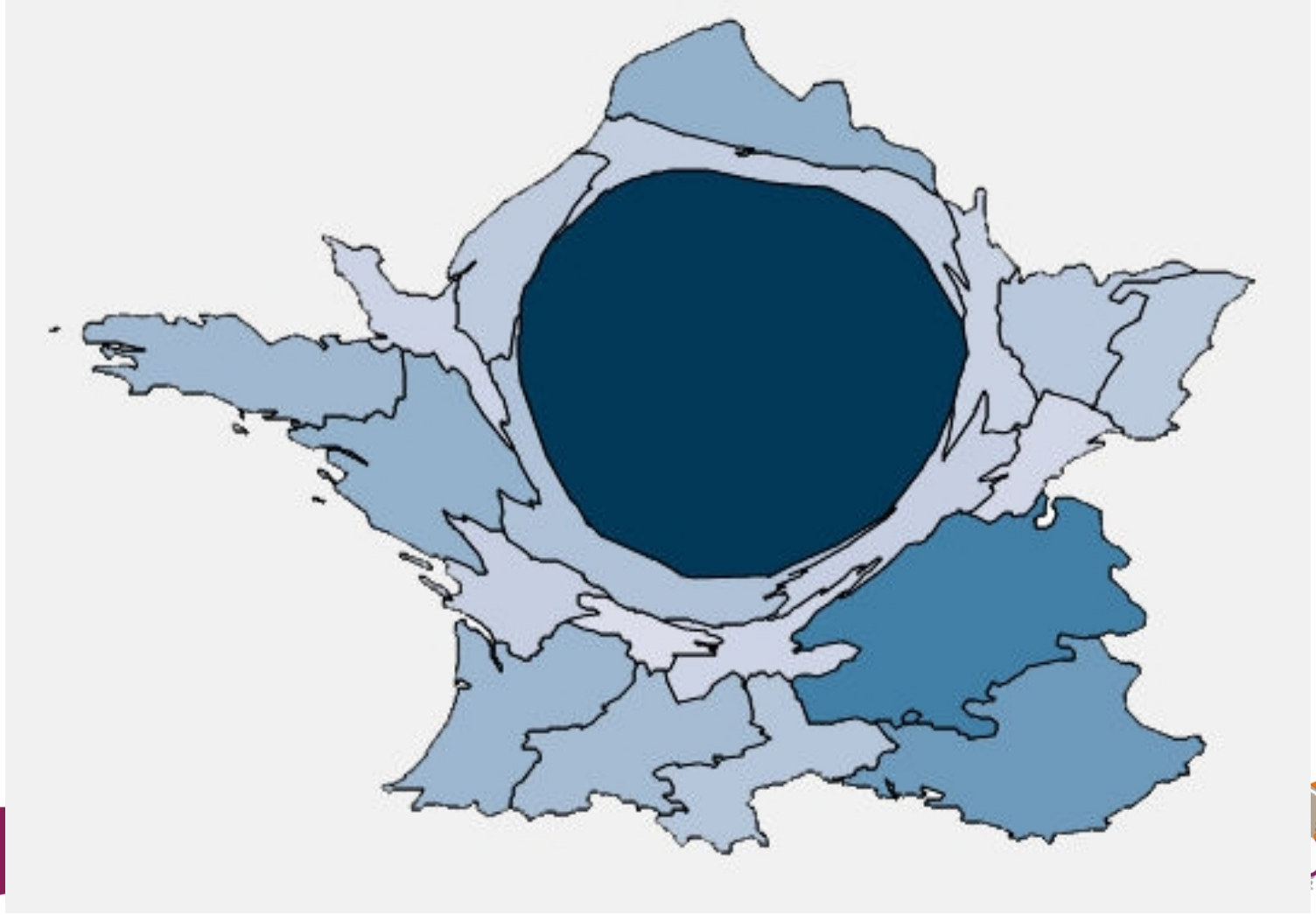
Linked Open Data Cloud (05/07)



More Open Data: data.gov (US)

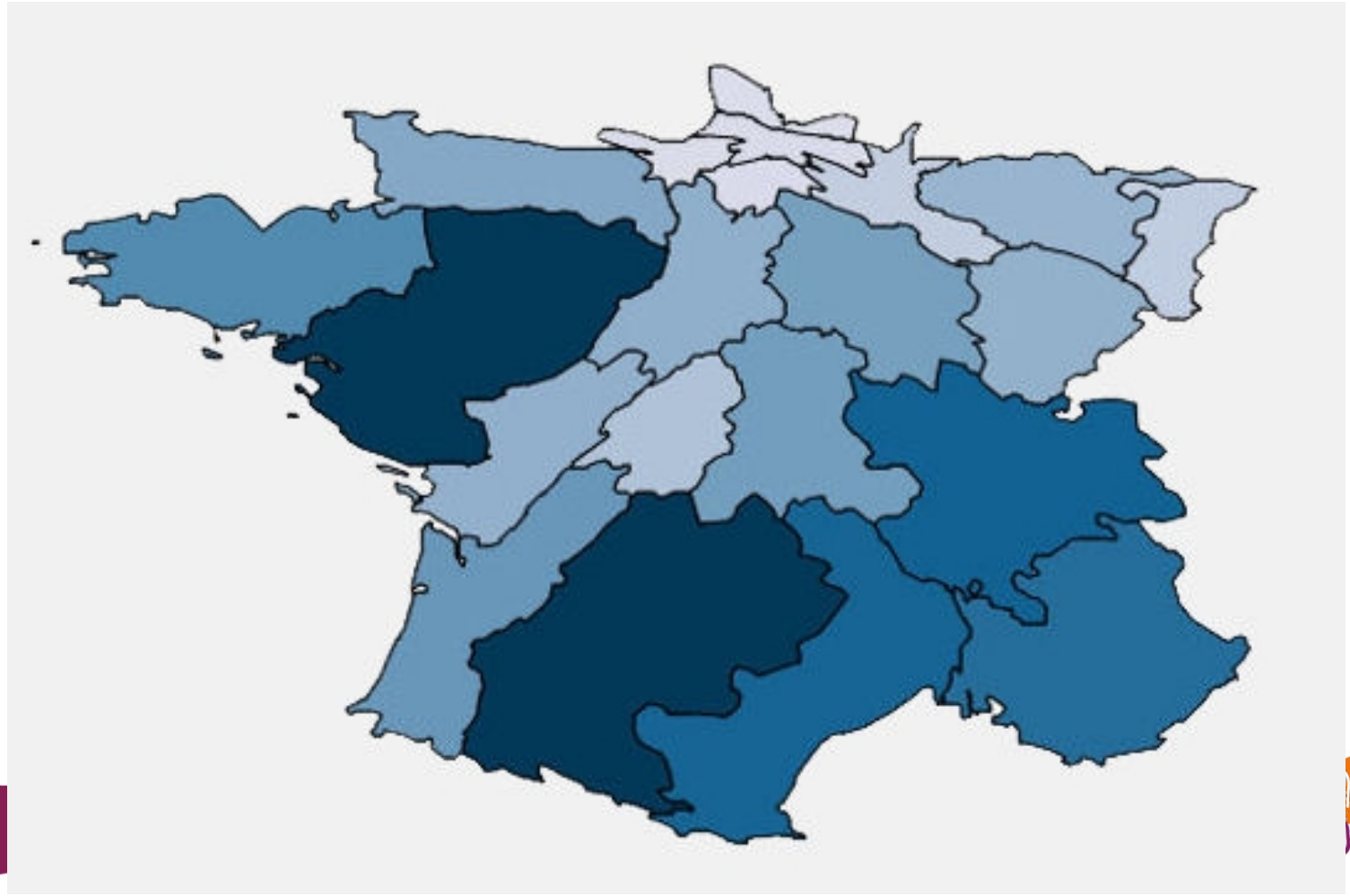
More OpenData: from Etalab (FR)

GDP per French region (Le Journal Du Net, 07/09/2011)



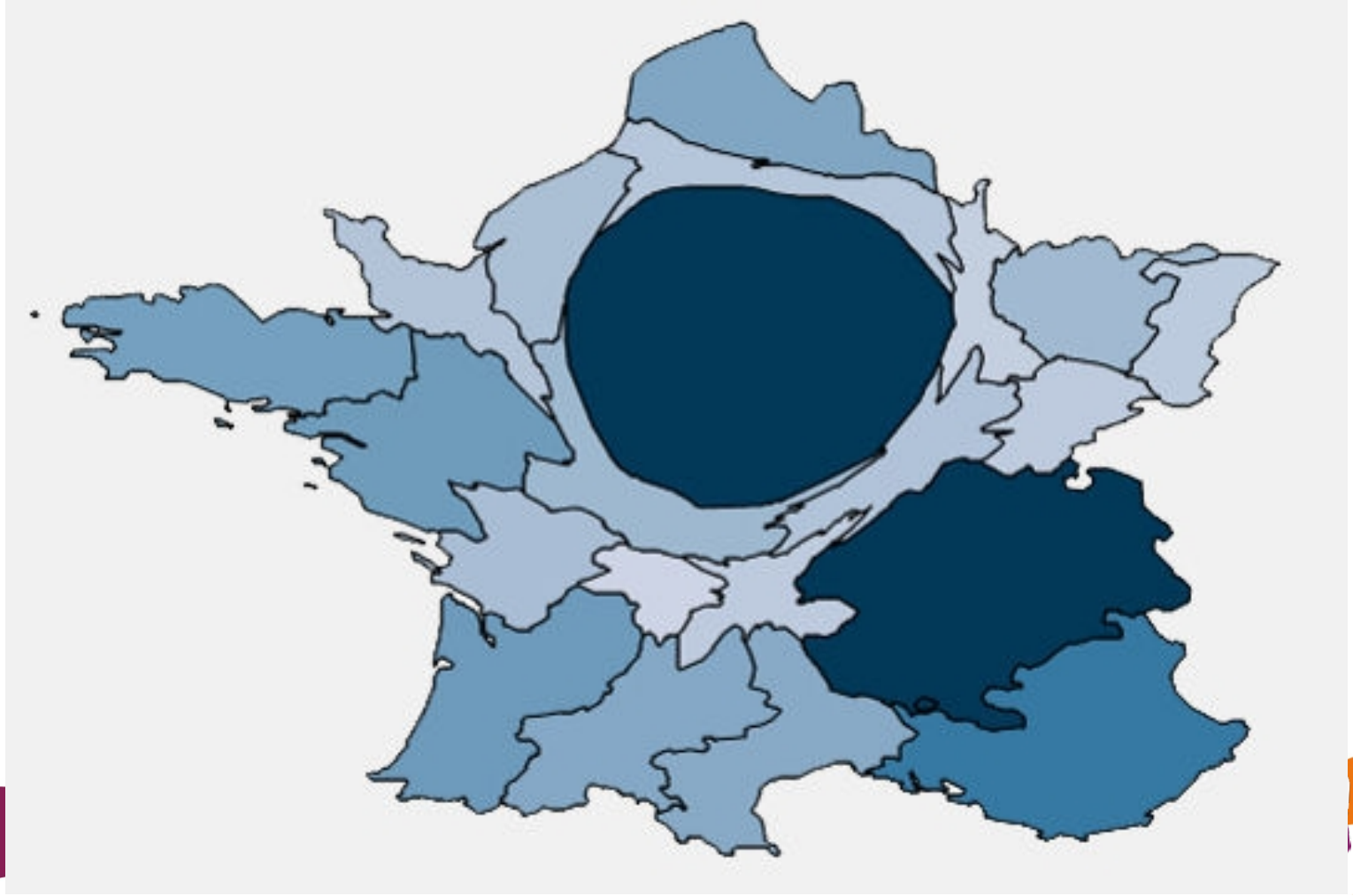
More OpenData: from Etalab (FR)

Organic agriculture per French region
(Le Journal Du Net, 07/09/2011)



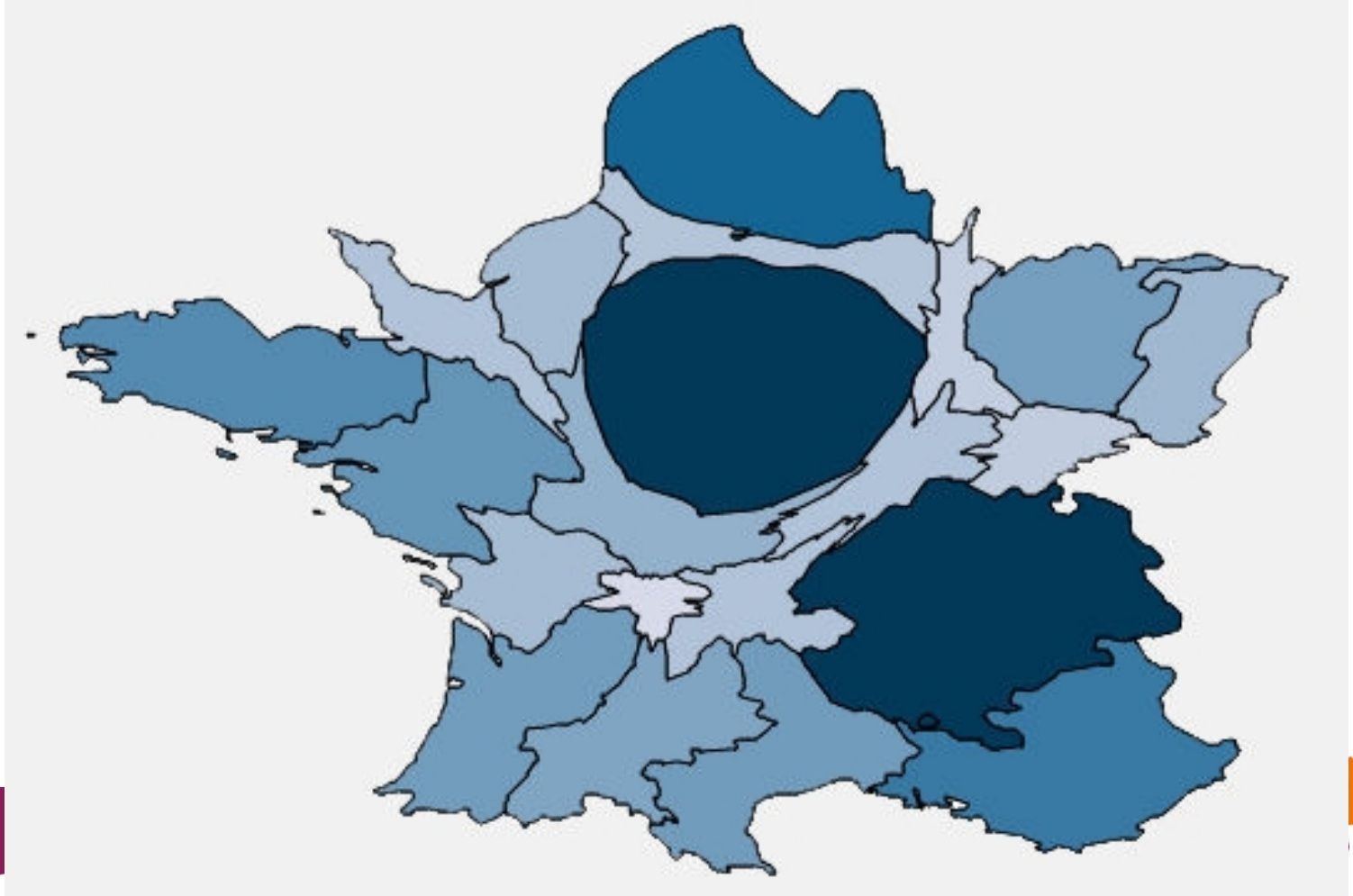
More OpenData: from Etalab (FR)

Cinemas/inhabitants per French region
(Le Journal Du Net, 07/09/2011)



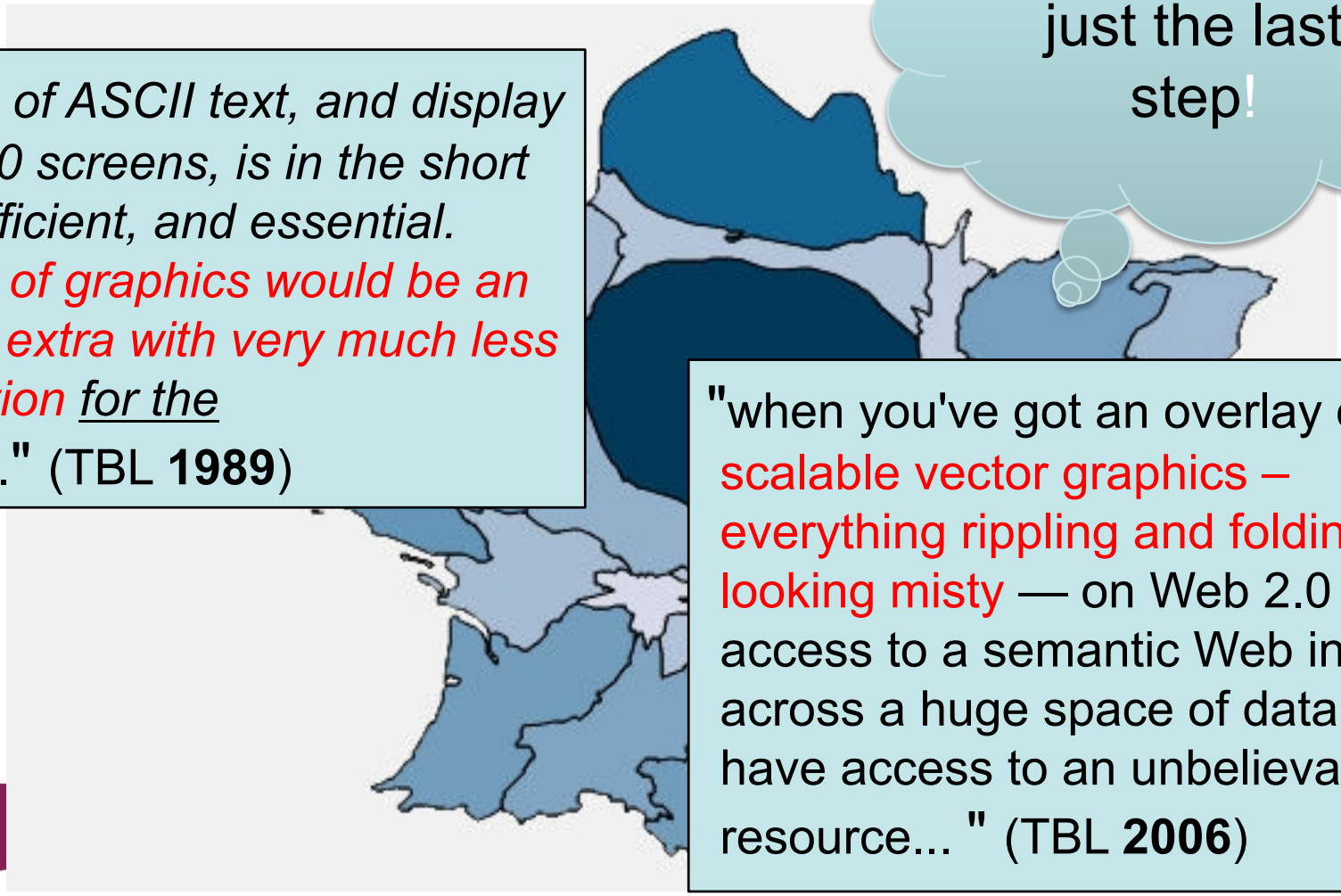
More OpenData: from Etalab (FR)

Boulangeries/inhabitants per French region
(Le Journal Du Net, 07/09/2011)



More OpenData: from Etalab (FR)

Boulangeries/inhabitants per Fr
(Le Journal Du Net, 07/0



Drawing it is just the last step!

"*Storage of ASCII text, and display on 24x80 screens, is in the short term sufficient, and essential. Addition of graphics would be an optional extra with very much less penetration for the moment.*" (TBL **1989**)

"when you've got an overlay of scalable vector graphics – everything rippling and folding and looking misty — on Web 2.0 and access to a semantic Web integrated across a huge space of data, you'll have access to an unbelievable data resource... " (TBL **2006**)

digiteo
Recherche en sciences & technologies de l'information

Some (more) scientific problems around Linked (Open) Data

FORUM 2011
4E ÉDITION

Problem: some Open Data comes in tables!

Estimated population per French region, January 2011

	2008 (p)			
	moins de 20 ans	**de 20 ans à 59 ans**	**60 ans ou plus**	**Total**
Alsace	451 588	1 018 528	367 384	1 837 500
Aquitaine	717 986	1 662 453	795 061	3 175 500
Auvergne	291 186	698 477	351 837	1 341 500
Bourgogne	375 451	840 149	420 400	1 636 000
Bretagne	766 518	1 624 204	750 278	3 141 000
Centre	609 250	1 316 698	609 052	2 535 000
Champagne-Ardenne	330 985	713 102	294 413	1 338 500
Corse	62 890	160 935	79 175	303 000
Franche-Comté	289 296	613 857	259 847	1 163 000
Île-de-France	3 027 497	6 639 779	2 005 224	11 672 500
Languedoc-Roussillon	604 809	1 326 046	656 645	2 587 500
Limousin	150 848	375 582	212 570	739 000
Lorraine	563 110	1 272 521	505 369	2 341 000
Midi-Pyrénées	649 999	1 493 182	694 319	2 837 500
Nord-Pas-de-Calais	1 090 023	2 167 695	764 282	4 022 000
Basse-Normandie	357 242	754 184	352 574	1 464 000
Haute-Normandie	472 337	968 365	378 798	1 819 500
Pays de la Loire	897 293	1 838 614	774 593	3 510 500
Picardie	500 963	1 023 654	378 883	1 903 500
Poitou-Charentes	395 199	894 239	460 062	1 749 500
Provence-Alpes-Côte d'Azur	1 149 753	2 520 542	1 230 205	4 900 500
Rhône-Alpes	1 560 992	3 257 459	1 294 549	6 113 000
France de province	**12 287 718**	**26 540 486**	**11 630 296**	**50 458 500**
France métropolitaine	**15 315 215**	**33 180 265**	**13 635 520**	**62 131 000**
Guadeloupe	122 398	208 383	71 719	402 500
Guyane	97 710	110 542	13 248	221 500
Martinique	113 181	209 972	76 347	399 500
Réunion	281 680	432 427	91 393	805 500
France métropolitaine et DOM	**15 930 184**	**34 141 589**	**13 888 227**	**63 960 000**

From tables to linked data

Population evolution in the area of Nord-Pas-de-Calais

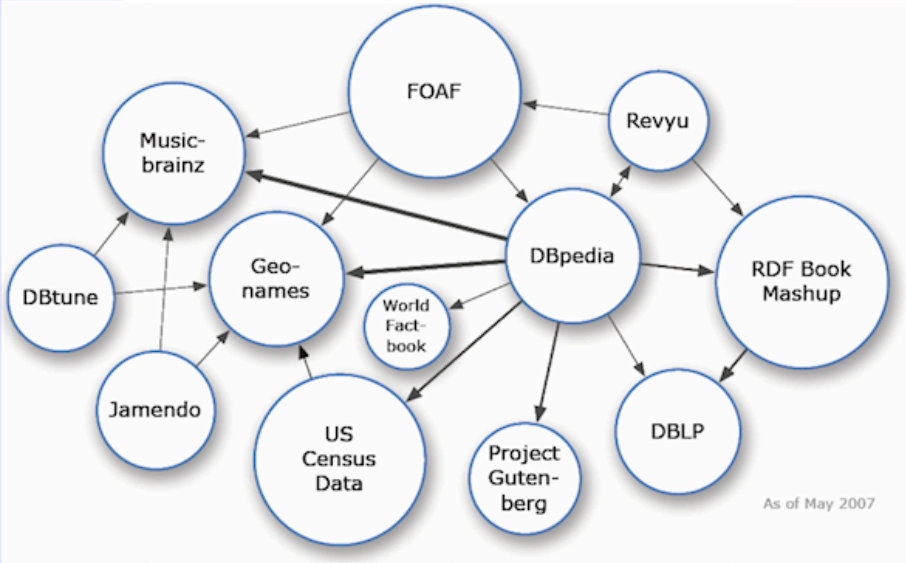
Nom de l'arrondissement	Population au 1er janvier 1999	Population au 1er janv 2008	Variation de population entre 1999 et 2008	Variation annuelle moyenne entre 1999 et 2008
Avesnes-sur-Helpe	238 557	234	– 4426	– 0,21
Cambrai	158 750	159	+ 812	+ 0,06
Douai	246 888	247	+ 738	+ 0,03
Dunkerque	379 602	375	– 3 982	– 0,12
Lille	1 181 724	1 198	+ 17 199	+ 0,16
Valenciennes				
Département du Nord	**2 554 449**	**2 564**	**+ 10 510**	**+ 0,05**
Arras	251 017	259	+ 8 729	+ 0,38
Béthune	279 775	283	+ 4 122	+ 0,16
Boulogne-sur-Mer	163 157	162	– 223	– 0,02
Calais	118 281	118	– 62	– 0,01
Lens	368 901	362	– 6 422	– 0,19
Montreuil	106 750	112	+ 5 862	+ 0,60
Saint-Omer	153 541	159	+ 6 103	+ 0,43
Département du Pas-de-Calais	**1 441 422**	**1 459**	**+ 18 109**	**+ 0,14**
Région Nord-Pas-de-Calais	**3 995 871**	**4 024**	**+ 28 619**	**+ 0,08**

From tables to linked data

Population evolution in the area of Nord-Pas-de-Calais

Nom de l'arrondissement	Population au 1er janvier 1999	Population au 1er janvier 2008	Variation de population entre 1999 et 2008	Variation annuelle moyenne entre 1999 et 2008
Avesnes-sur-Helpe			– 4426	– 0,21
Cambrai			+ 812	+ 0,06
Douai			+ 738	+ 0,03
Dunkerque			– 3 982	– 0,12
Lille			+ 17 199	+ 0,16
Valenciennes			+ 169	+ 0,01
Département du Nord			**+ 10 510**	**+ 0,05**
Arras			+ 8 729	+ 0,38
Béthune			+ 4 122	+ 0,16
Boulogne-sur-Mer			– 223	– 0,02
Calais			– 62	– 0,01
Lens			– 6 422	– 0,19
Montreuil			+ 5 862	+ 0,60
Saint-Omer	153 541	159 644	+ 6 103	+ 0,43
Département du Pas-de-Calais	**1 441 422**	**1 459 531**	**+ 18 109**	**+ 0,14**
Région Nord-Pas-de-Calais	**3 995 871**	**4 024 490**	**+ 28 619**	**+ 0,08**

How do we get here?

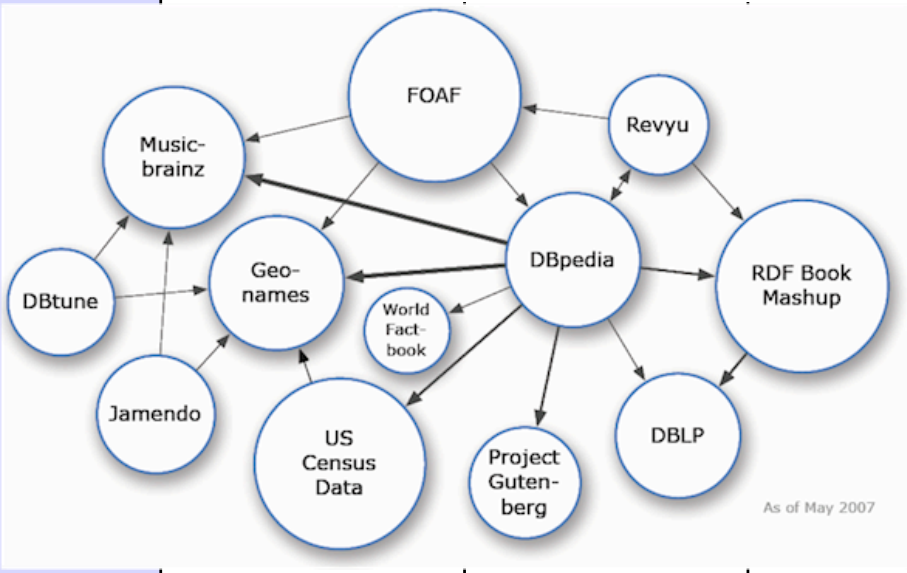


From tables to linked data

Population evolution in the area of Nord-Pas-de-Calais

Nom de l'arrondissement	Population au 1er janvier 1999	Population au 1er janvier 2008	Variation de population entre 1999 et 2008	Variation annuelle moyenne entre 1999 et 2008
Avesnes-sur-Helpe			– 4426	– 0,21
Cambrai			+ 812	+ 0,06
Douai			+ 738	+ 0,03
Dunkerque			– 3 982	– 0,12
Lille			+ 17 199	+ 0,16
Valenciennes			+ 169	+ 0,01
Département du Nord			**+ 10 510**	**+ 0,05**
Arras			+ 8 729	+ 0,38
Béthune			+ 4 122	+ 0,16
Boulogne-sur-Mer			– 223	– 0,02
Calais			– 62	– 0,01
Lens			– 6 422	– 0,19
Montreuil			+ 5 862	+ 0,60
Saint-Omer	153 541	159 644	+ 6 103	+ 0.43
Région Nord-Pas-de-Calais	**3 995 871**	**4 024 490**	**+ 28 619**	**+ 0,08**

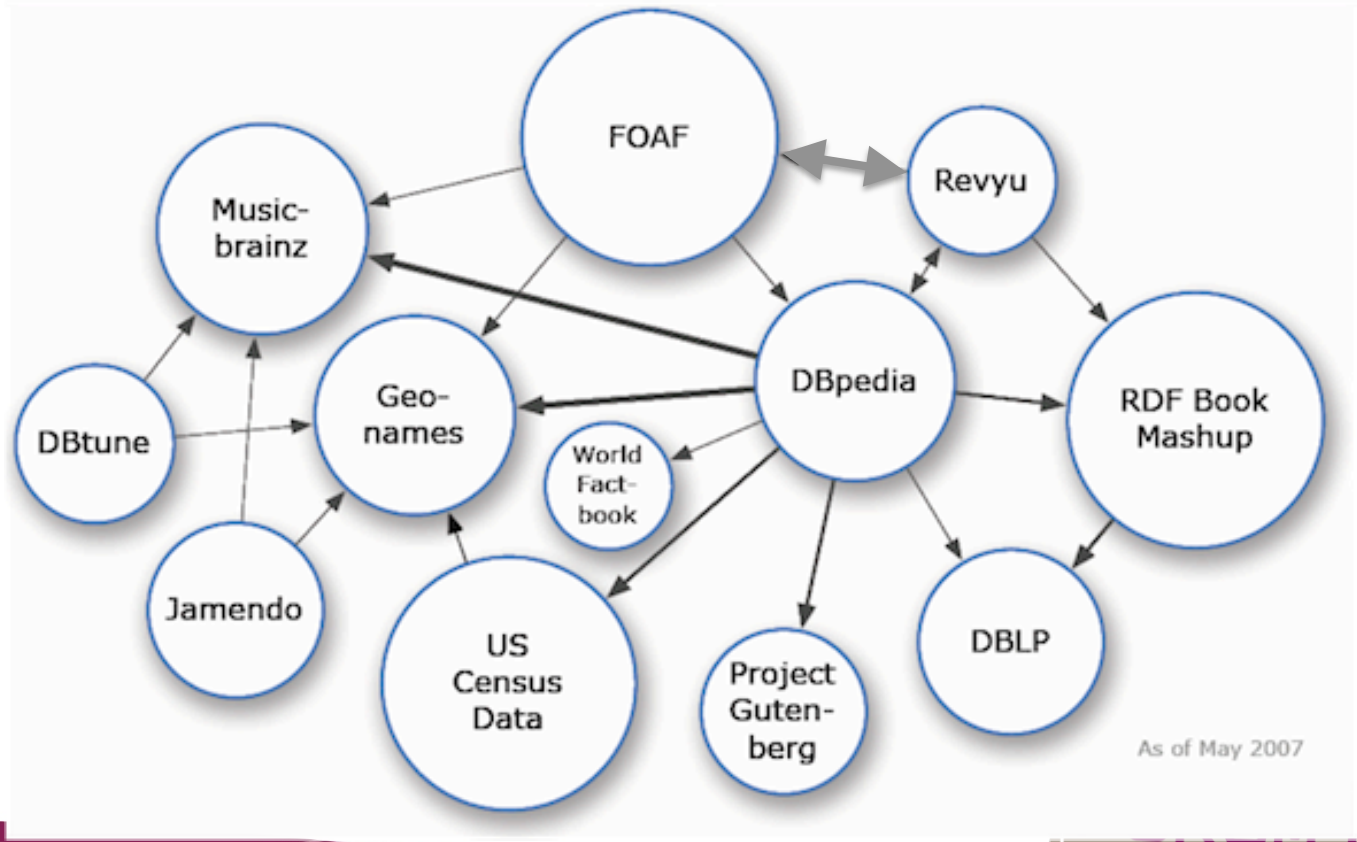
How do we get here?



Ongoing work in Leo, collaboration with DataPublica start-up

A different problem: RDF reconciliation

Build links between bubbles =
identify when the same entity appears in two data sets



Reference Reconciliation Problem

- Different identifiers refer to the same real world entity

SOURCE1	MuseumName	MuseumAddress	Located	inCountry
Museum11	“Madame Tussauds”	“Marylebone Road”	“London”	“England”
Museum12	“Science Museum”	“Exhibition Road”	“London”	“England”

SOURCE2	MuseumName	MuseumAddress	Located	inCountry
Museum21	“Madame Tussauds”	“Marylebone Road”	“London”	“UK”
Museum22	“British Museum”	“Great Russell Street”	“London”	“UK”

Reference reconciliation and key constraints

No knowledge about the properties ➔ give same importance to all

- Similarity(Museum11 , Museum21)=75%

Experts may specify key constraints

- Example: Key(MuseumName, MuseumAddress) →
  Similarity(Museum11, Museum21)=100%

Large volumes of data ➔ Keys harder to find; expert not always available or may be wrong…

Result: algorithm to automatically discover keys from data

– Complete and correct set of keys

Ongoing work within Leo

Many other groups worldwide! (data cleaning, entity resolution…)

Crucial to produce Linked Data

digiteo
Recherche en sciences & technologies de l'information

Wrap-up

FORUM 2011
4E ÉDITION

We forgot Web mining!

The Web is mined for:

- Data (extracting LOD)
- Complex information (who, what, when, why, … situations, relationships…)
- Knowledge / semantics / meaning (YAGO / ERC WebDam / Leo)
- Hidden structure

M. Vazirgiannis (Digiteo chair on Web mining)

DBWeb @ Telecom ParisTech

NLP teams at LIMSI and CEA (extraction of complex information from Web text)

Social Web analysis @ Alcatel Lucent…

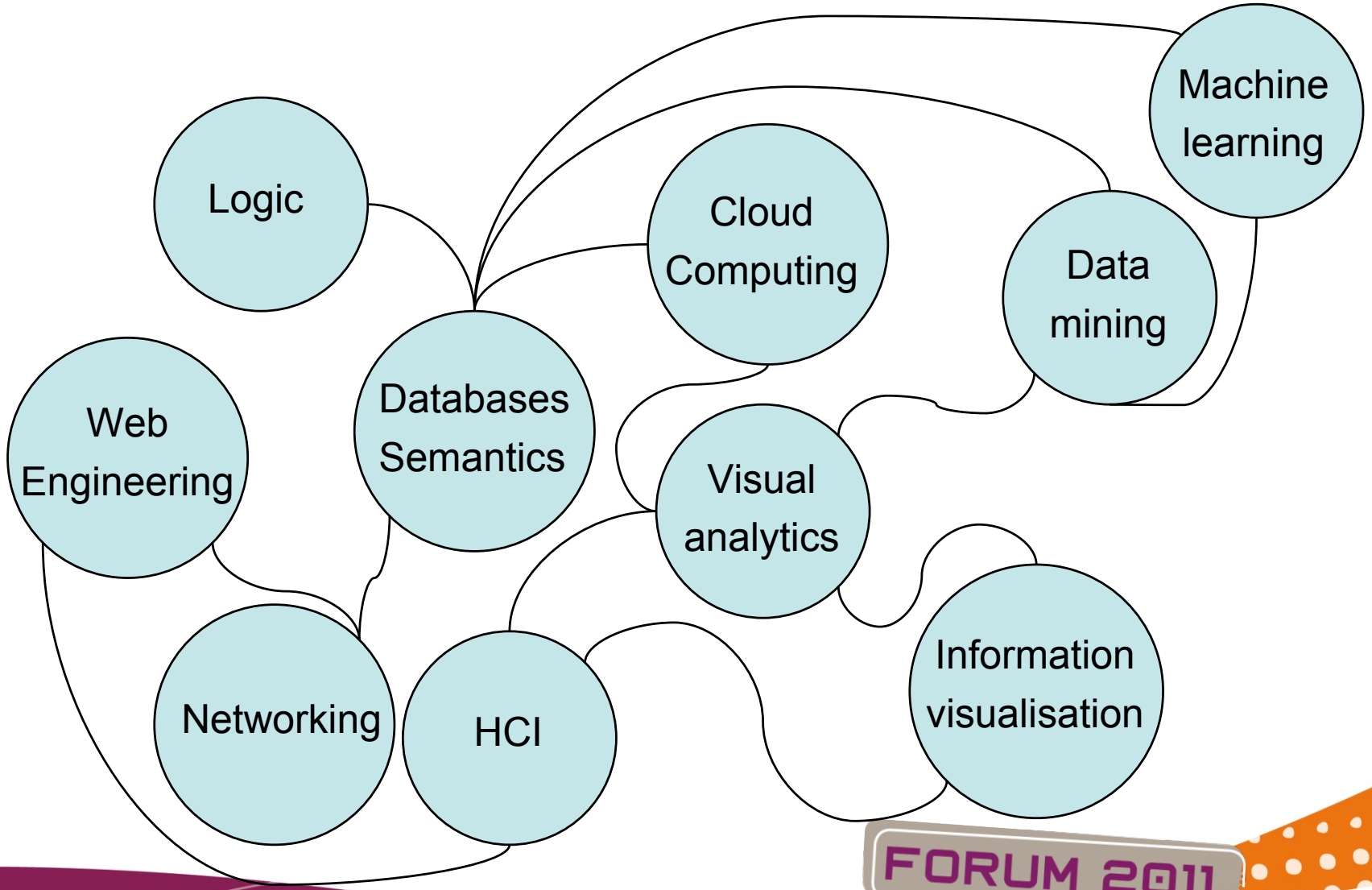
Web mining

The Web is mined for:

- Data (extracting LOD)
- Complex information (who, what, when, why, … situations, relationships…)
- Knowledge / semantics / meaning
- Hidden structure

"*An intriguing possibility, given a large hypertext database, is that it allows some degree of automatic analysis. It is possible to search, for example, for anomalies such as undocumented software or divisions which contain no people. It is also possible to look at the topology of an organisation or a project, and draw conclusions about how it should be managed, and how it could evolve. This is particularly useful when the database becomes very large, and groups of projects, for example, so interwoven as to make it difficult to see the wood for the trees.*"

Scientific domains for LOD and the Web



LOD extremely popular right now

In Databases, WWW, Web Engineering, Semantic Web venues

Connection increasingly being made with Big Data / Cloud Computing

- LOD reference reconciliation in a cloud environment

The "Universal Knowledge Base" is coming back. This isn't the CYC you used to know

Mining and extraction very important

- Still there after the last Facebook user quits…

Scalable (distributed) reasoning, maintenance of inferred knowledge?

Important to remember that **openness and platform-independence were essential to the Web** from the beginning.

Important to preserve.

Big picture (applications)

- Exploiting data:
  - Running marketplaces of specialized data, catering to specific business or personal needs.
- Making sense of data:
  - Web or social network mining for sentiment analysis, ads etc.
- Enriching data:
  - Augment the client's data with other public or proprietary information.
- Improving information systems:
  - Better classification / annotation of existing resources to enable finding, sharing, re-combining
- Improve the functioning of society at large:
  - Increase citizen awareness → better democracy
  - The Open-* movements have many interesting ideas. Also FING

digiteo
Recherche en sciences & technologies de l'information

Merci/questions?

FORUM 2011
4E ÉDITION