Le Web à l'heure du Linked Open Data: vers la base de données universelle

Ioana Manolescu-Goujot

Equipe OAK

Inria Saclay & LRI Université de Paris Sud-11

http://team.inria.fr/oak

INRIA Saclay– Île-de-France &
Université Paris Sud
LRI – UMR CNRS 8623

Large-scale complex data

Volume of digital data keeps increasing

- "Some areas of science are facing hundred- to thousand-fold increases in data volumes… compared to the volumes generated only a decade ago"

Bell, Hey and Szalay, *Beyond the Data Deluge*, Science, 2009

Web is the world's single biggest repository

- "Every two days we create as much information as we did up to 2003. Most of it is user-generated data"

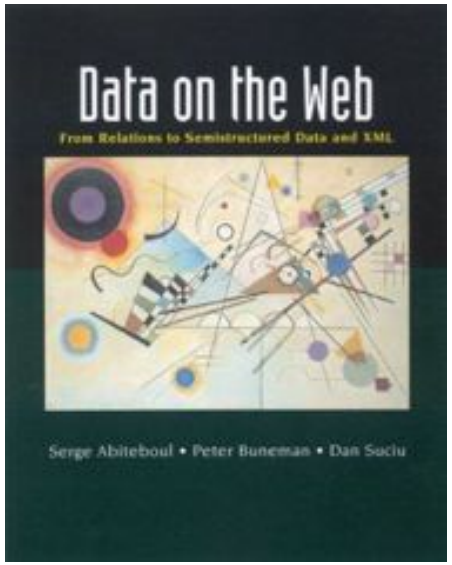
Eric Schmidt, former CEO of Google, April 2011

Increasing digitization of content leads to increased complexity in the data

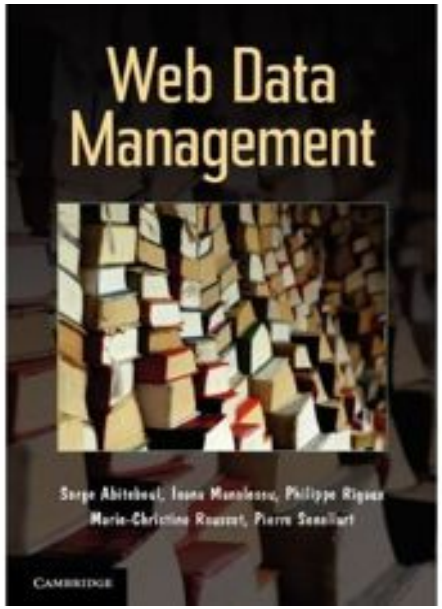
- Relational databases for transactions → cubes for analytics → OO for complex structure → XML for structured documents → RDF for metadata → schema transformation language for cube families…

OAK focuses on large-scale complex data

Plan



1. Short recall: data on the Web
2. The database approach: models and tools
3. Linked Open Data: towards an universal database on the Web
4. Selected related research topics (OAK and beyond)



1

Short recall: data on the Web

Data on the Web

Recall from earlier D'Alembert seminaries, notably A. Monnin (12/2011) and F. Gandon (today)

- Main concept in the architecture of the Web:
  URI (Universal Resource Identifier)
  - May identify a *digital resource (Web page, PDF, script…)* , an *institution*, a *human*, or a *concept* (e.g., "intellectual property as defined by EU legislation")
  - Not to be confused with an URL (Universal Resource Locator = "where to download from")
  - Allows distributed authorship of content (A. Monnin: "*everyone can say what they want about whatever they want*")
    - This raises many well-known data management problems (see next)
    - This is what the Semantic Web is about; also Linked Open Data (see next)

Claim: the original purpose of the Web was *a better data+knowledge base*

Short history of the Web

- 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)
  - Important and very often cited / referred to.

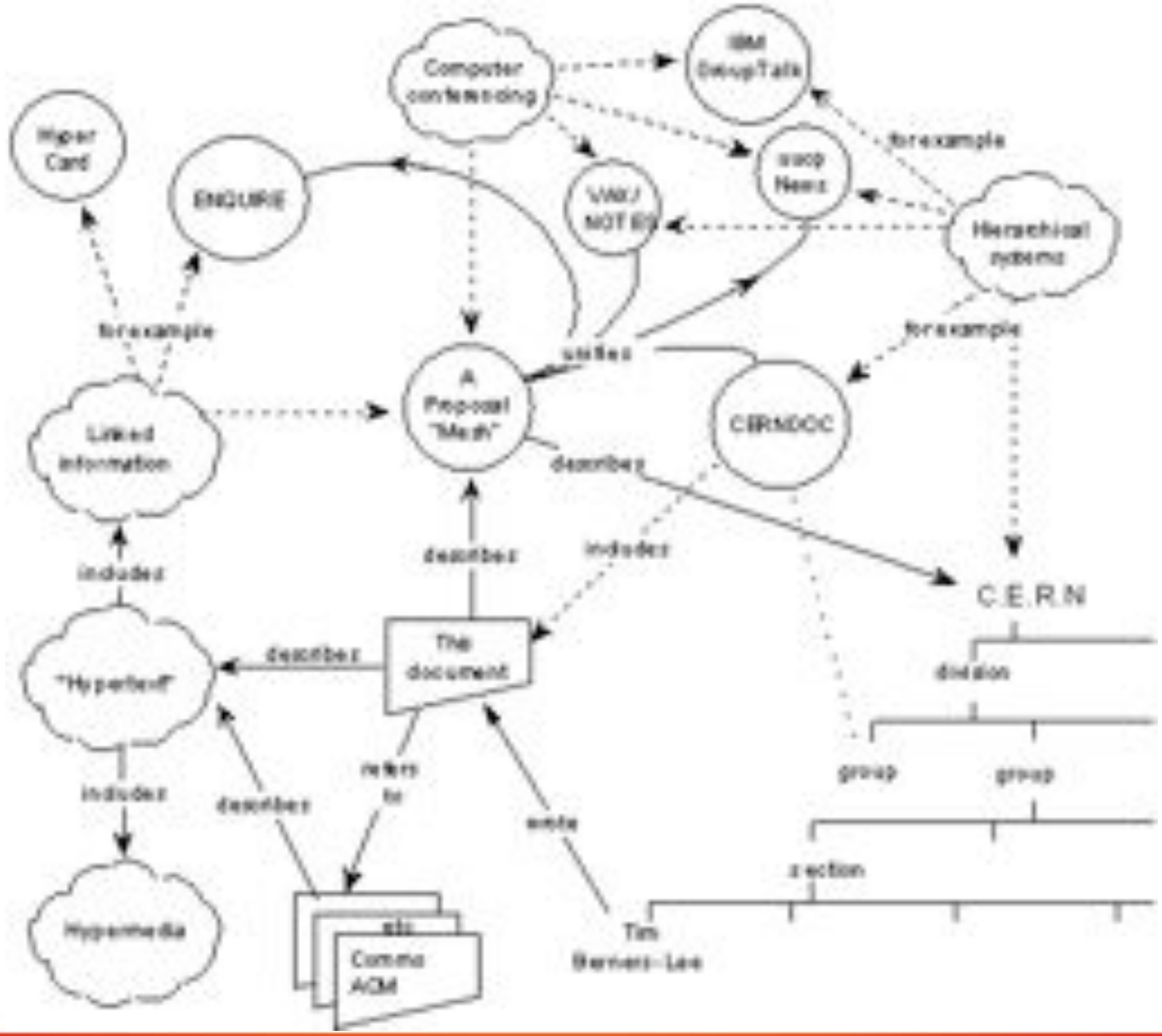
A short history of the Web

- 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)

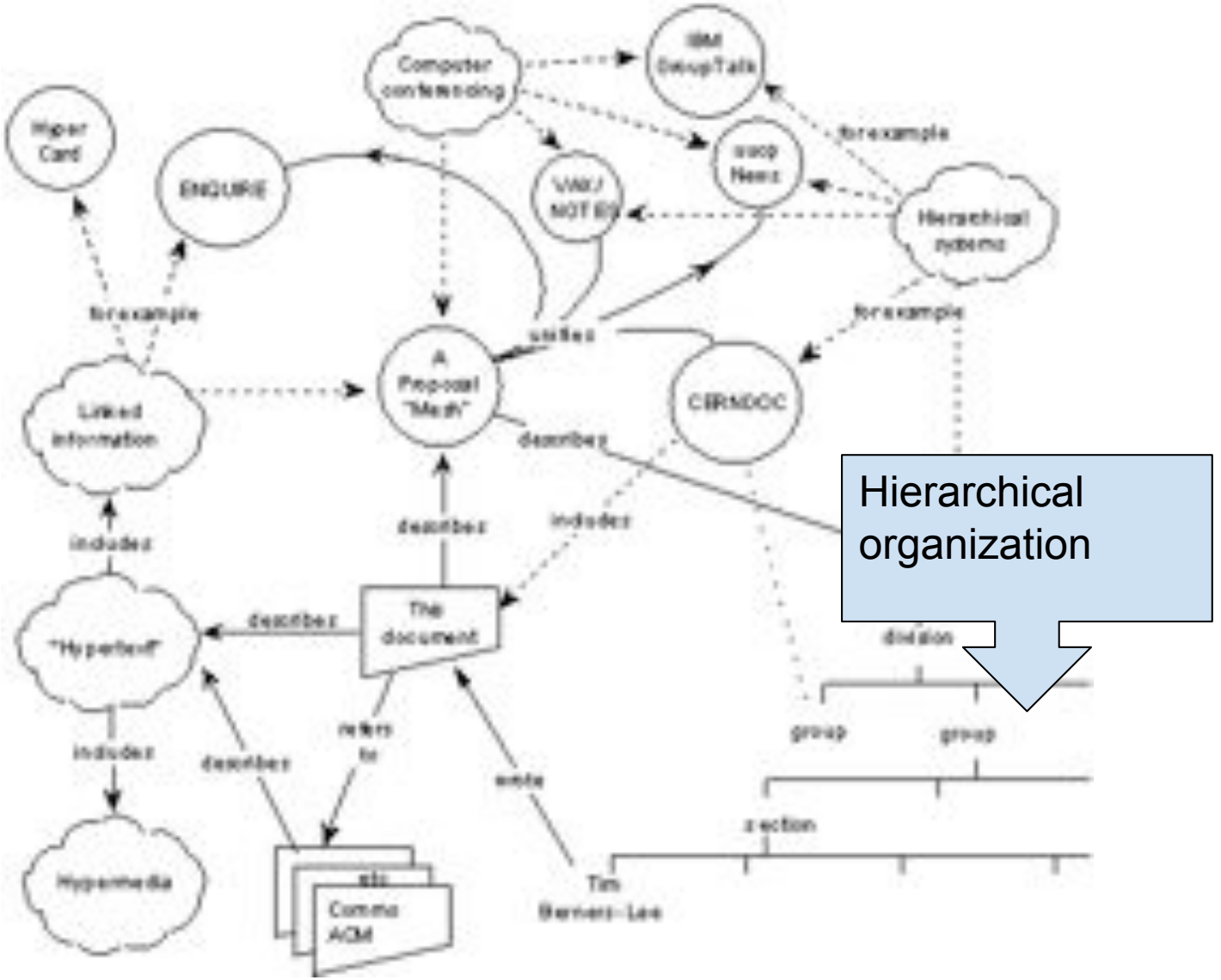
"*This proposal discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system. The sort of information we are discussing answers, for example, questions like:*

- *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)

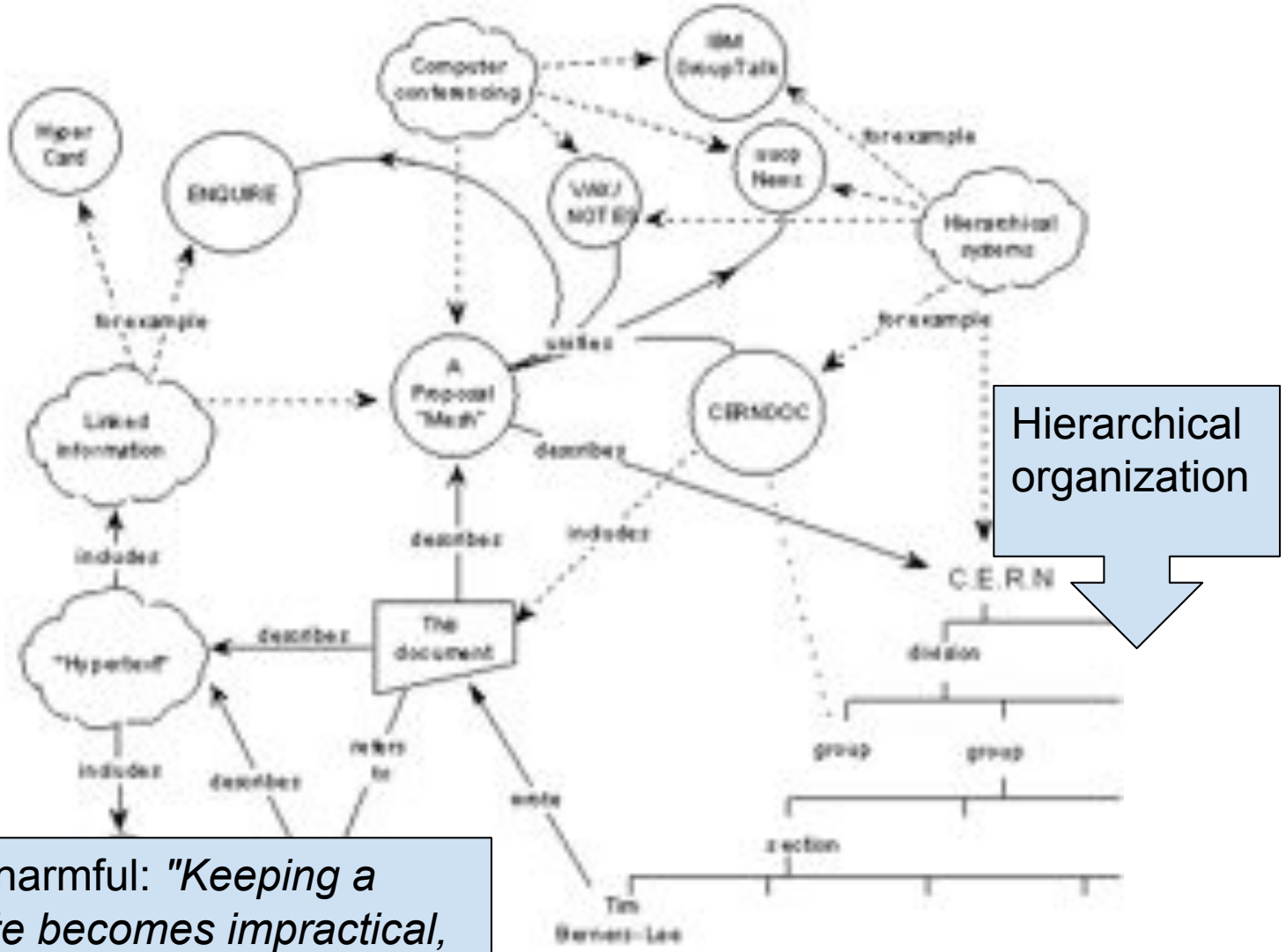


The Web (Tim Berners-Lee proposal, 1989)



Hierarchical organization

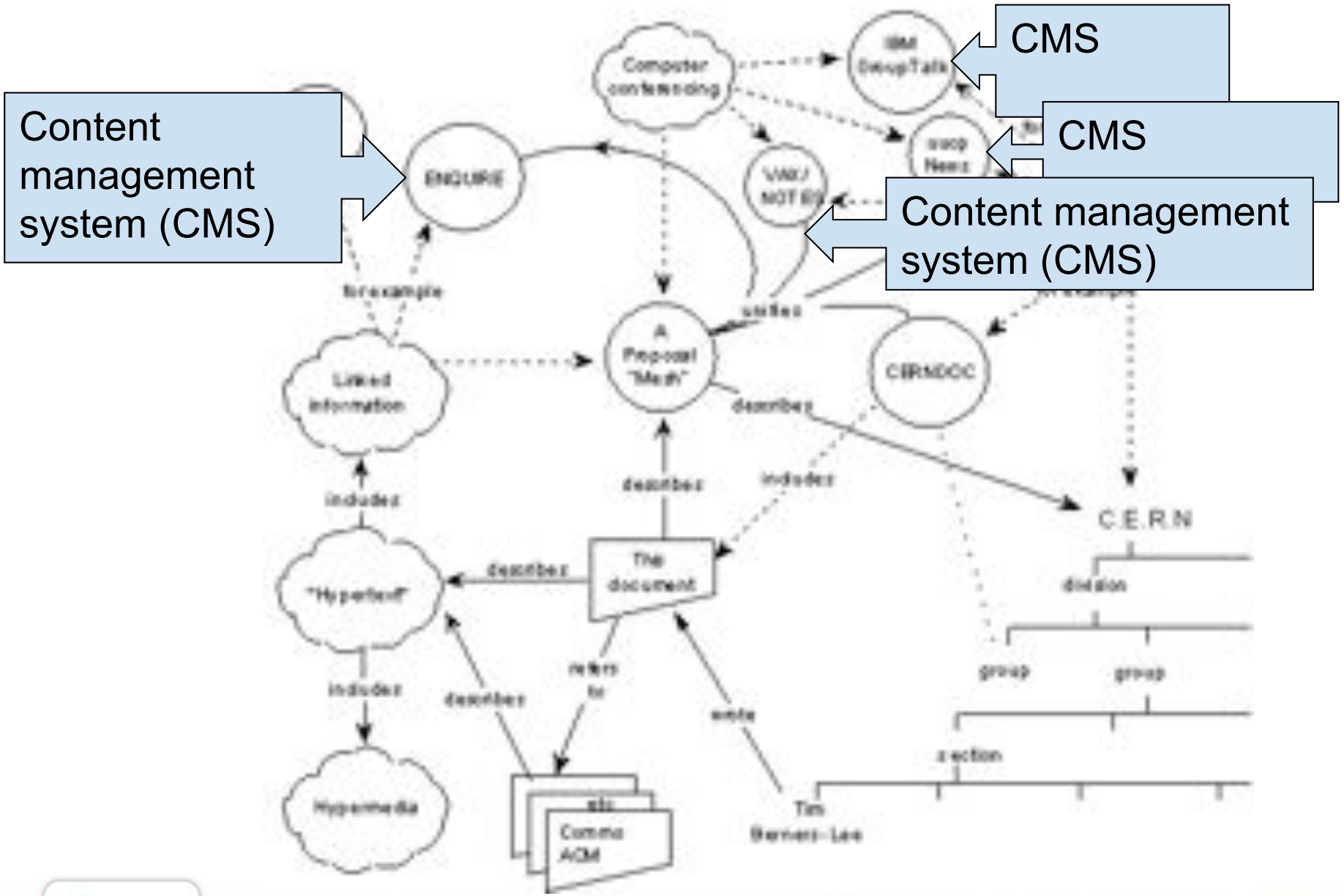
The Web (Tim Berners-Lee proposal, 1989)



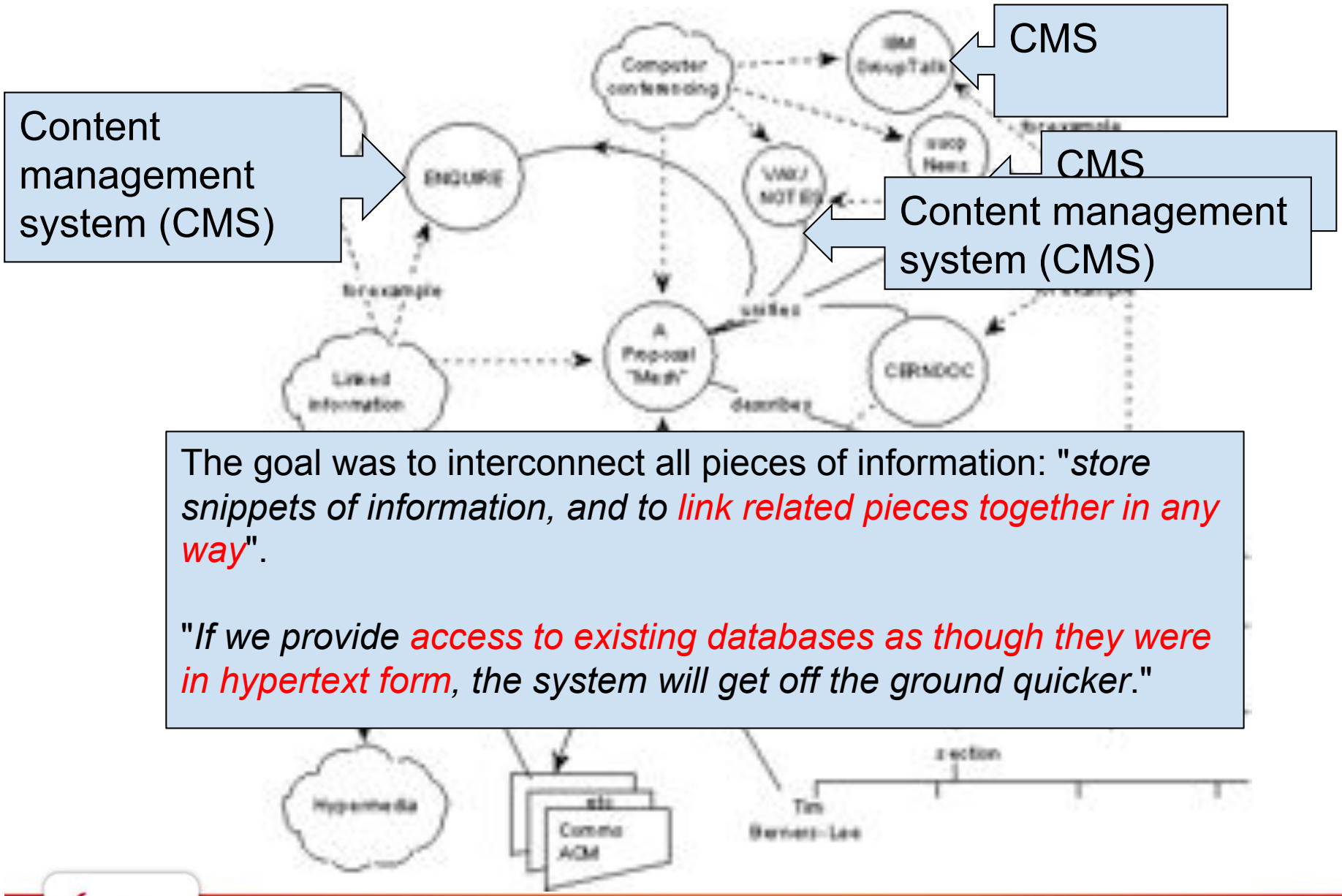
Hierarchical organization

…considered harmful: *"Keeping a book up to date becomes impractical, and the structure of the book needs to be constantly revised."*

The Web (Tim Berners-Lee proposal, 1989)



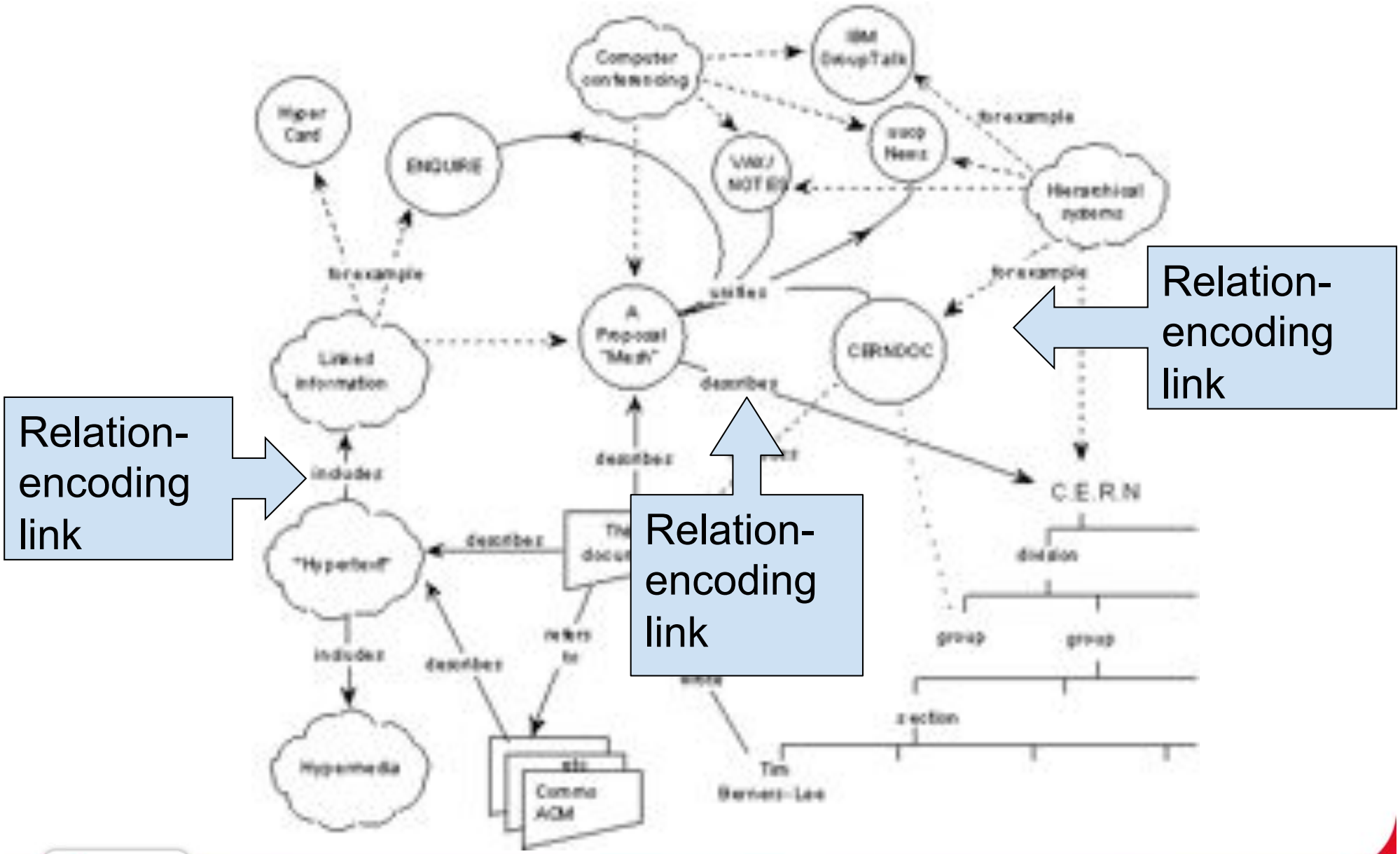
The Web (Tim Berners-Lee proposal, 1989)



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)

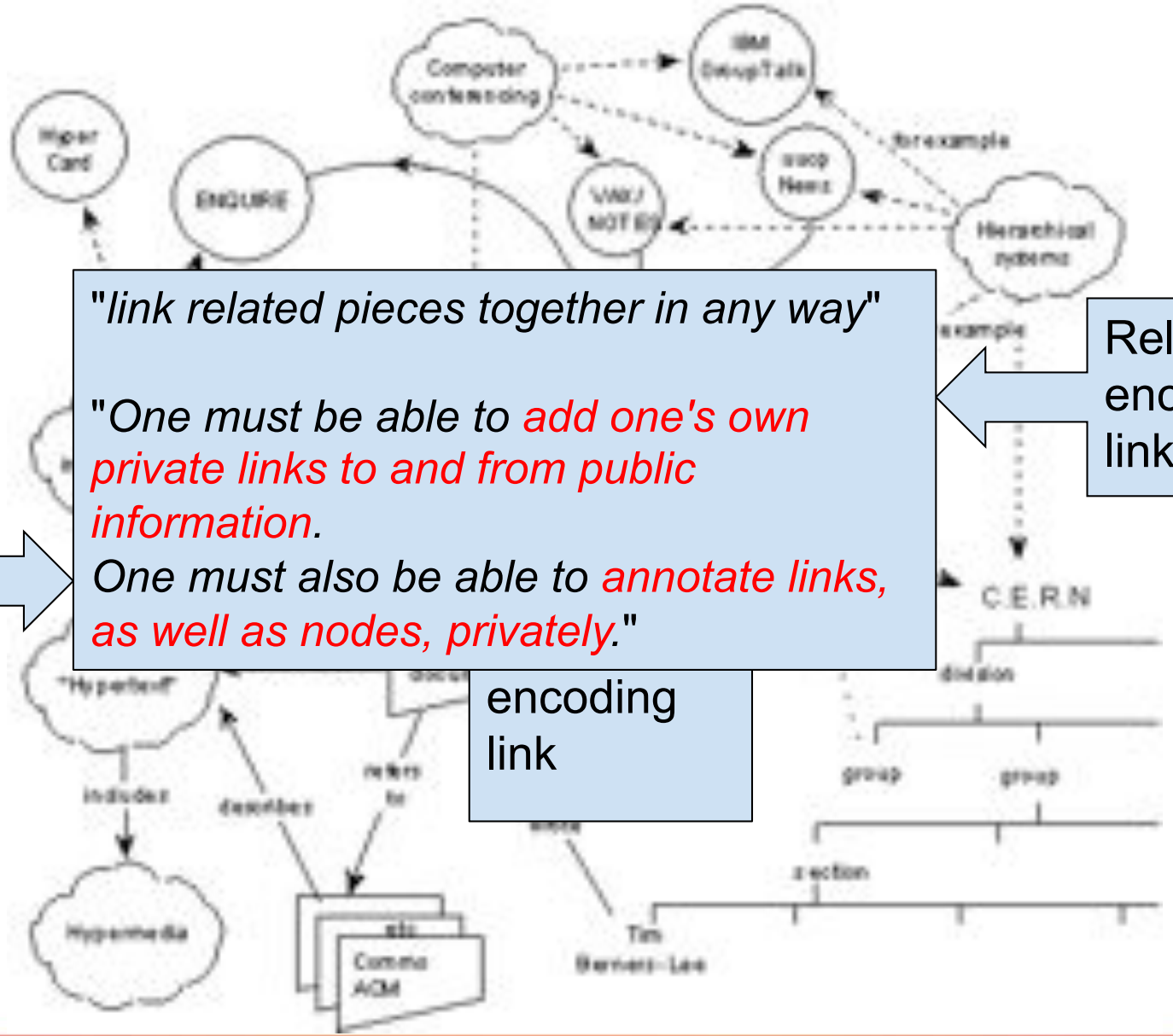


Relation-encoding link

Relation-encoding link

Relation-encoding link

The Web (Tim Berners-Lee proposal, 1989)



"*link related pieces together in any way*"

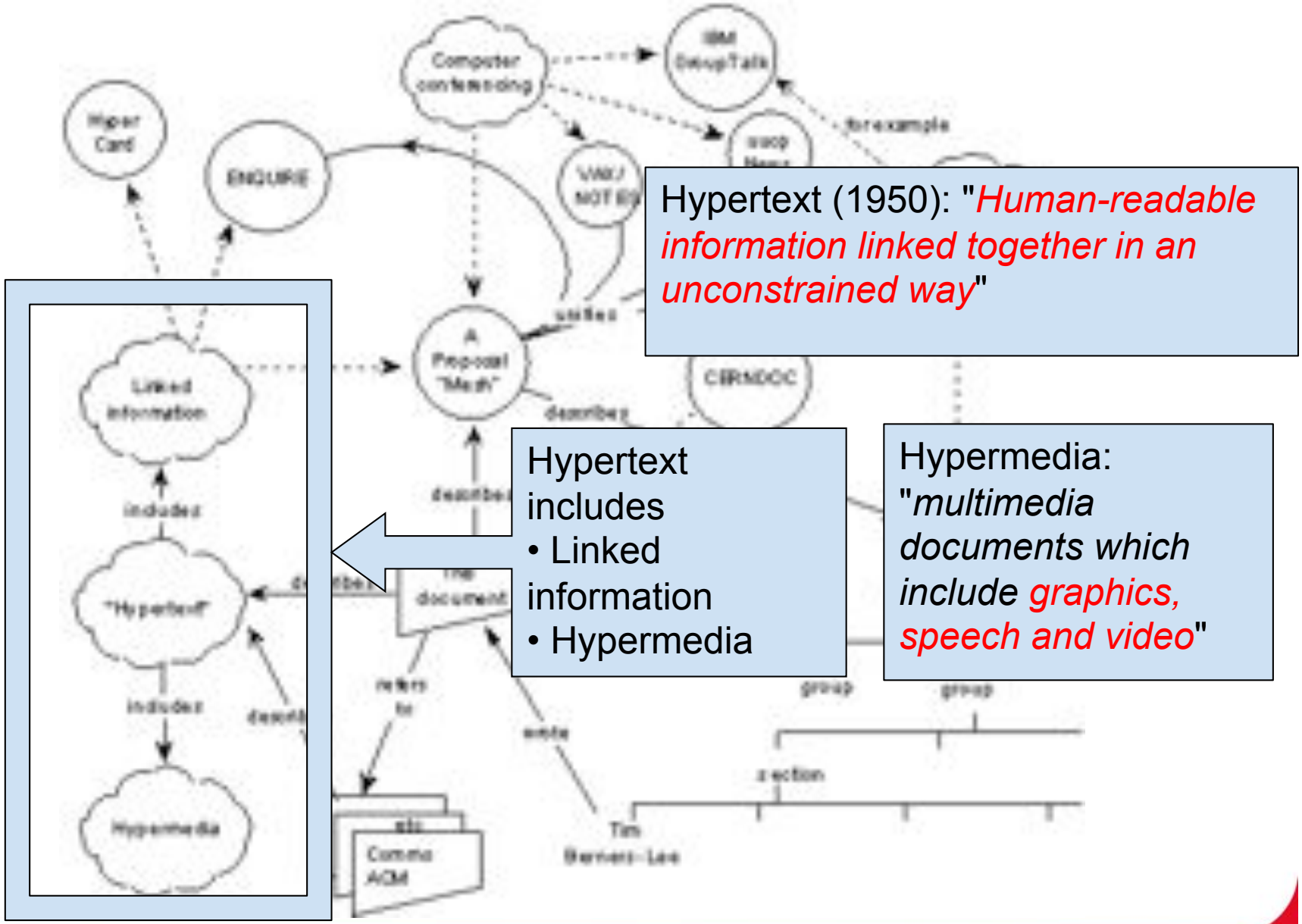
"*One must be able to add one's own private links to and from public information.*
One must also be able to annotate links, as well as nodes, privately."

Relation-encoding link

Relation-encoding link

encoding link

The Web (Tim Berners-Lee proposal, 1989)



What kind of data on the Web?

- 1989: Tim Berners-Lee proposal for an information system for the CERN
- 1991: HTTP, 1995: commercial Internet
- HTML content is served in answer to HTML requests.
  - Automated export from databases to HTML
- 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**

What kind of data on the Web?

- 1998: XML

  Tree-structured, "loose" format for complex data

  "Clean HTML": **separate content from presentation**

"*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***."

- Give the discussion an identifier, and point (refer) to it from all relevant contexts

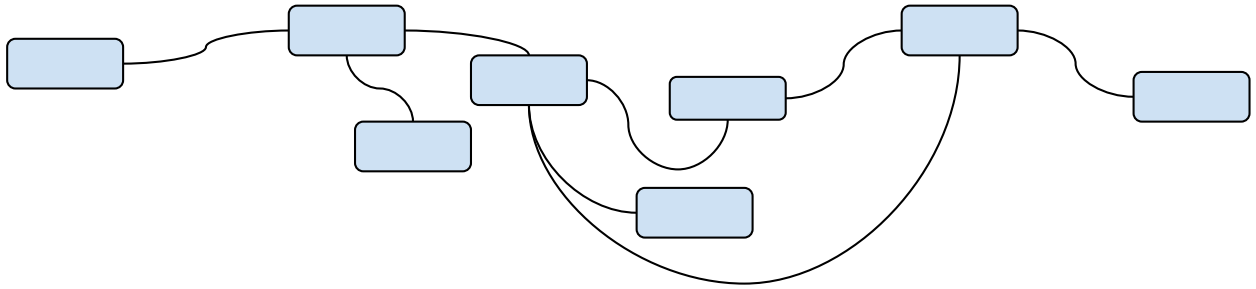
What kind of data on the Web?

- 1998: XML

  Tree-structured, "loose" format for complex data

  "Clean HTML": **separate content from presentation**

"*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***."



The system should be able to model the real world.
This is the premier ambition of databases and data management.

2

The database approach: models and tools

Crash course on data management

It all started with the banks!

Bank accounts pre-existed computers by a very long time

Here are your letters from the bank:

Stock market (around 1962 AC)

Seen from outside:

Stock market (around 1962 AC)

Seen from inside:

First goal of databases: model world data

Formally describe a set of data structures

- based on simple set or graph theory

The model should

- Be expressive: capture the characteristics of real world
  - A person's age is an integer; a person's name is a string…
  - A person has a *single* birth date
  - *Any* driver must be older than 16
  - Within a country, the zip code *determines* the city and viceversa.
- Be comprehensible (by the database administrator)
- Allow compact representation of data
  - An employee's address should be stored only once (in general)
  - This also simplifies updates (applied only in one place)

Modeling bank clients

Client

Nom: Julie

Adresse: 1, rue Dugommier

Ville: Paris

Age: 22

Client	Nom	Adresse	Ville	Age
	Julie	1 rue Dugommier	Paris	22

Modeling bank clients

Client

Nom: Marc

Adresse: 2, rue Archange

Ville: Orsay

Age: 25

Client	Nom	Adresse	Ville	Age
	Julie	1 rue Dugommier	Paris	22
	Marc	2 rue Archange	Orsay	25

Modeling bank clients

Client

Nom: Julie

Adresse: 1, rue Dugommier

Ville: Paris

Age: 22

Problem

Client				
	Julie	1 rue Dugommier	Paris	22
	Marc	2 rue Archange	Orsay	25
	Julie	1 rue Dugommier	Paris	22

Modeling bank clients

Client

Nom: Julie

Adresse: 1, rue Dugommier

Ville: Paris

Age: 22

Client	Nom	Adresse	Ville	Age
	Julie	1 rue Dugommier	Paris	22
	Marc	2 rue Archange	Orsay	25
	Julie	1 rue Dugommier	Paris	22

Modeling bank clients

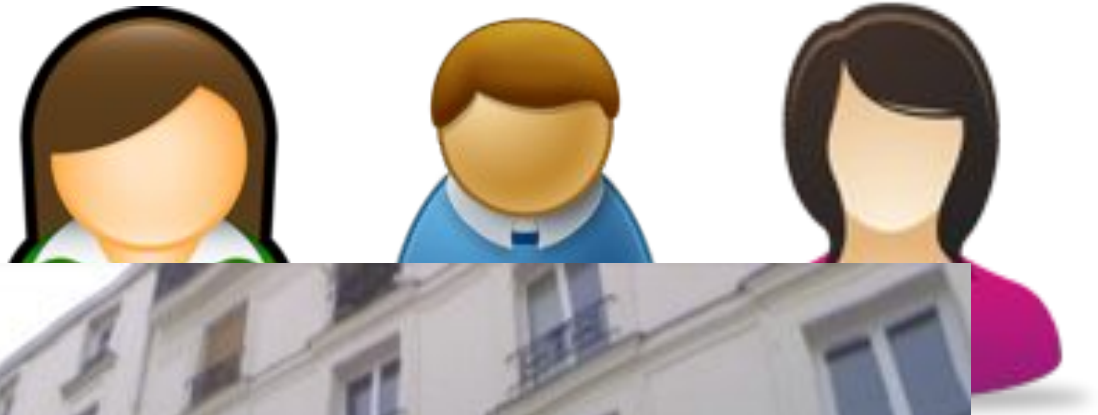
Client

Nom: Julie

Adresse: 1, rue Dugommier

Ville: P

Age: 22



Client

Modeling bank clients

La cliente

Nom: Julie

Adresse: 1, rue Dugommier

Ville: Paris

Age: 22

NumClient	Nom	Adresse	Ville	Age
1	Julie	1 rue Dugommier	Paris	22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

Modeling bank clients

La cliente
Nom: Julie
Adresse: 1, rue D
Ville: Paris
Age: 22

Primary key: knowing this value allows to exactly identify a record

NumClient	N			Age
1	Ju	Dugommier		22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

Modeling bank clients: primary key vs. URI

La cliente
Nom: Julie
Adresse: 1, rue D
Ville: Paris
Age: 22

Primary key:
allows to exactly identify a record

URI: allows to uniquely identify a resource

NumClient	N			Age
1	Ju	Dugommier		22
2	Marc	2 rue Archange	Orsay	25
3	Julie	1 rue Dugommier	Paris	22

Modeling the real world in a database

More goals

- Grouping together specific properties and active behavior
- Object-oriented databases (ODMG, Catell et al. Also O2)
  - class Person, class Car, …
- Modeling semantic relationships
  - a Driver is a Person

Data semantics!

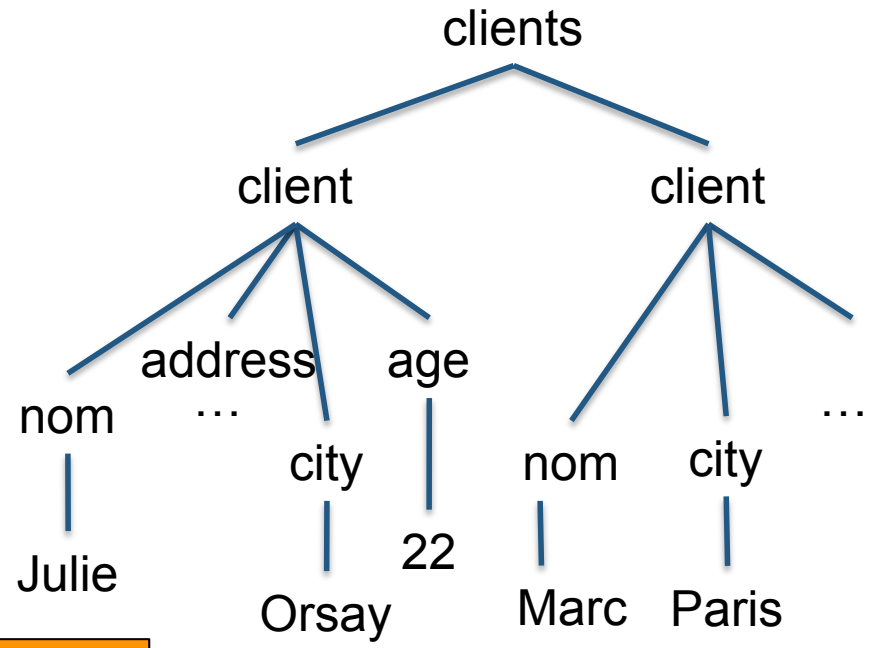
- Capturing time (R. Snodgrass, C. Jensen et al.; Baazizi et al.@TIME 2011)
  - The driver of this car *as of March 14, 2012* was Julie
  - The content of the page at URI1 as *of March 14, 2012, 3:05:06 PM EST* was…
- Capturing viewpoints or belief (e.g., Suciu et al., 2009)
  - Carole believes she has seen a bald eagle, and found a black feather
  - Bob does not believe that
  - Bob believes that Carole has found a purple-black feather
  - Alice believes Carole saw a crow

Modeling the real world in a database

Heterogeneous, varied structure: Object Exchange Model (OEM, 1996). Graph data with keys…

Self-describing data: 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

```
'/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>
```

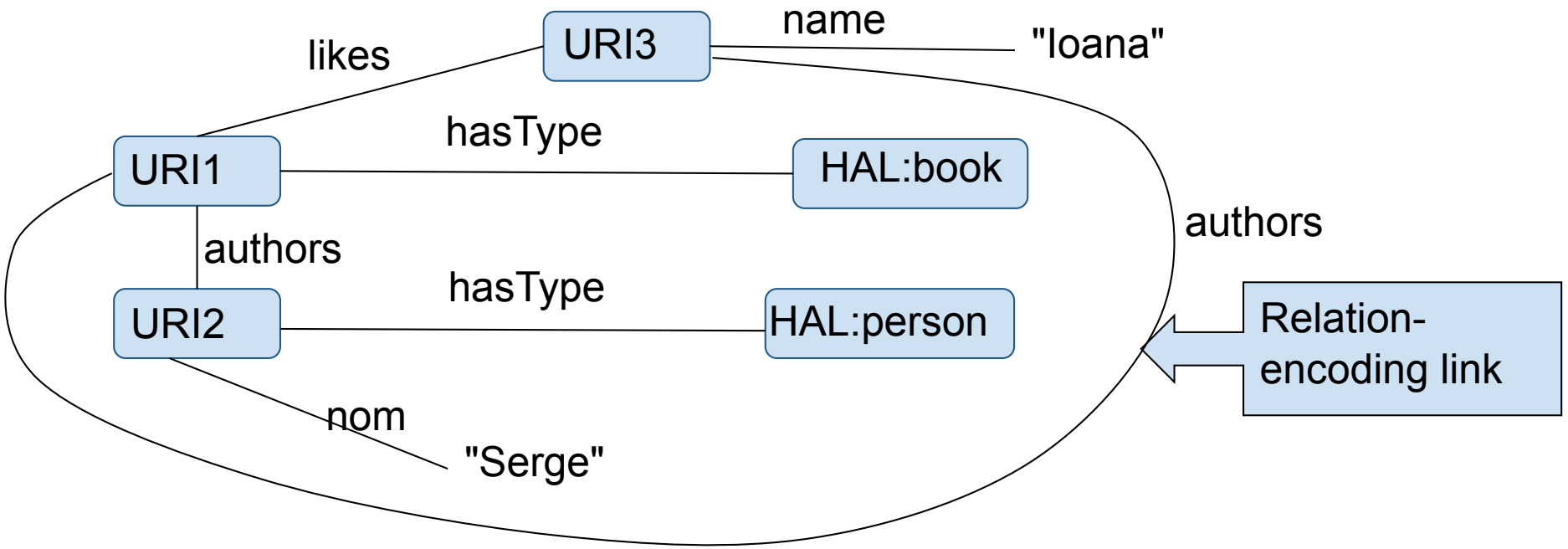
Modeling the real world in a database: RDF

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



Types enable reasoning according to the RDF semantics

Can we fully model the real world data?

U. Eco, Il Secondo Diario Minimo: "De l'Impossibilità da construire la carta dell'impero 1 à 1"

In quell'Impero, l'Arte della Cartografia raggiunse tale Perfezione che la mappa d'una sola Provincia occupava tutta una Città, e la mappa dell'impero, tutta una Provincia. Col tempo, codeste Mappe Smisurate non soddisfecero e i Collegi dei Cartografi eressero una Mappa dell'Impero, che uguagliava in grandezza l'Impero e coincideva puntualmente con esso.
Meno Dedite allo Studio della Cartografia, le Generazioni Successive compresero che quella vasta Mappa era Inutile e non senza Empietà la abbandonarono alle Inclemenze del Sole e degl'Inverni.

Umberto Eco vs. the URI of Umberto Eco

Can we fully model the real world data?

U. Eco, Il Secondo Diario Minimo: "De l'Impossibilità da construire la carta dell'impero 1 à 1"

In quell'Impero, l'Arte della Cartografia raggiunse tale Perfezione che la mappa d'una sola Provincia occupava tutta una Città, e la mappa dell'impero, tutta una Provincia. Col tempo, codeste Mappe Smisurate non soddisfecero e i Collegi dei Cartografi eressero una Mappa dell'Impero, che uguagliava in grandezza l'Impero e coincideva puntualmente con esso.
Meno Dedite allo Studio della Cartografia, le Generazioni Successive compresero che quella vasta Mappa era Inutile e non senza Empietà la abbandonarono alle Inclemenze del Sole e degl'Inverni.

J.L. Borges, "The other tiger"

To the tiger of symbols I hold opposed
The one that's real, the one whose blood runs hot […]
But by the act of giving it a name,
By trying to fix the limits of its world,
It becomes a fiction not a living beast,
Not a tiger out roaming the wilds of earth.

Second goal of databases: manipulate data

Create, Read, Update, Delete, Reason

In a declarative fashion

- The user specifies what data/processing she wants
  - Standard query/update languages: SQL, OQL, XQuery, SPARQL
- Not how to compute/evaluate it
- The database system translates the query into a lower-level evaluation plan
  - Query optimizer: millions of lines of code, crucial part of large commercial products (IBM DB2, Oracle, Microsoft SQLServer, SAP Hana…)
  - Optimization =
    - implementing efficient algorithms, materializing efficient data structures
    - exploring alternatives
    - estimating the cost of the alternatives

"*Relational databases are the cornerstone of modern civilization*"
Bruce Lindsay, IBM fellow
Distinguished Profiles in Databases,
ACM SIGMOD Record, 2003

Crucial requirement from databases: reliable transactions

1. When I buy something,
either the money moves from me to the buyer,
or both accounts are unchanged.

2. When I buy something, my balance will decrease accordingly

3. My husband and I should be able to shop at the same time.

4. The bank will not "forget" my balance (nor that of my loan…)

Atomicity

Consistence

Isolation

Durability

Crucial requirement from databases: reliable transactions

1. When I buy something,

either the money moves from me to the buyer,

Transaction theory (1970-…)
Jim Gray, David Lomet and others
Turing Award for Jim Gray in 1998

the same time.

4. The bank will not "forget" my balance (nor that of my loan…)

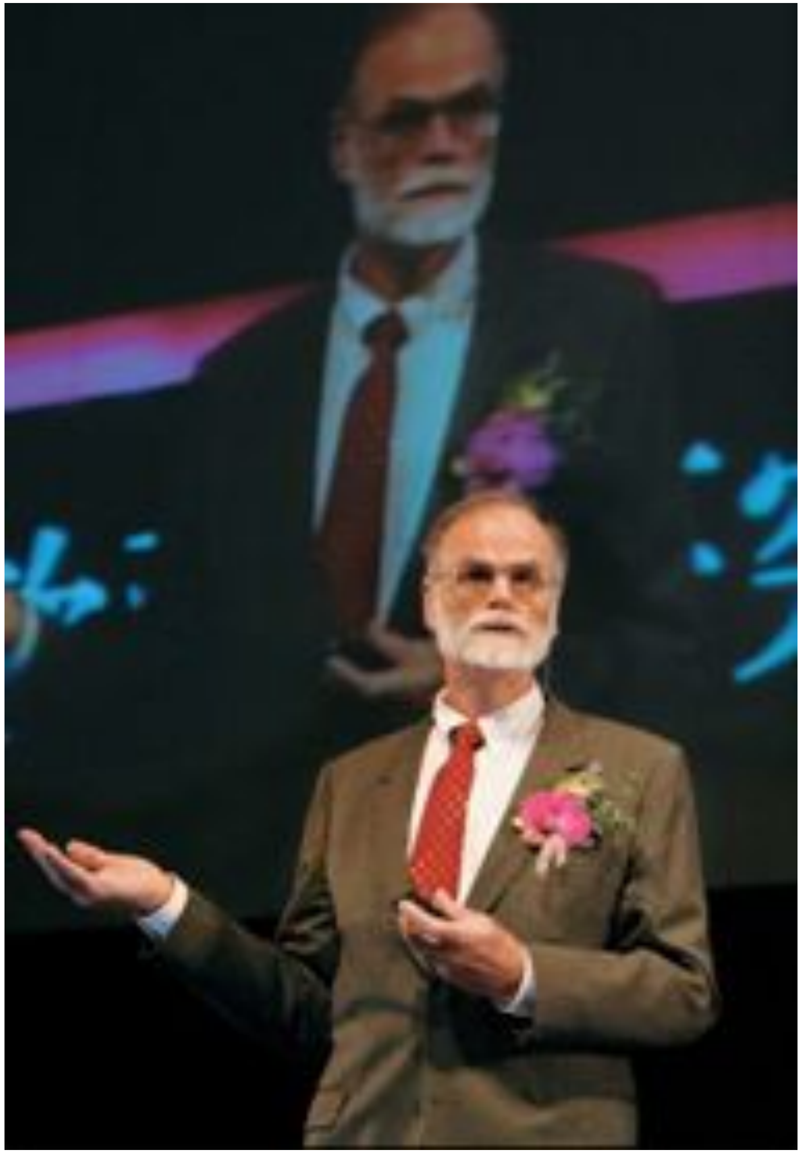
Atomicity

Consistence

Isolation

Durability

In memoriam: Jim Gray and the 4th paradigm of science



Jim Gray on eScience:
A Transformed Scientific Method

Data warehousing

Data warehouse = huge database where data is typically accumulating continuously

Analysis only make sense by (multi-dimensional) **aggregation**:

- Julie's expenses by year
- All clients' expenses by year and by city of residence
- All clients' expenses by year, city of residence, and age…

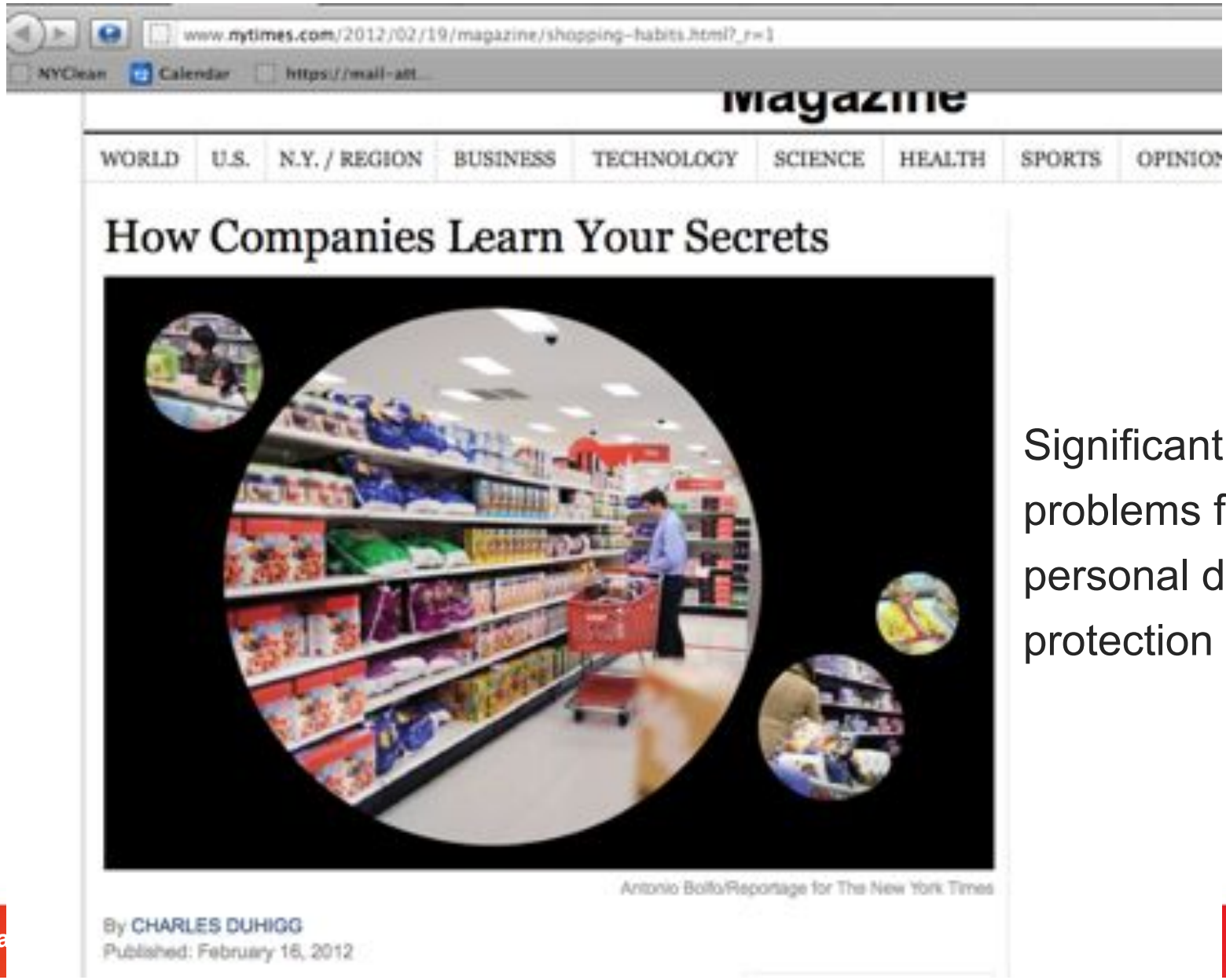
Biggest world's warehouses: sales data (Walmart, Target)

- A **cube** is an aggregation result by some set of dimensions
- A cube can be drilled down into (more detail)
- A cube can be rolled up (less detail, aggregation)

Warehouse tools (supporting cubes etc.) for relational databases are *crucial* tools for decision making (enterprise, but also administration, …)

Data mining

Large data analysis to identify trends and interesting associations



Significant problems for personal data protection

3

Linked Open Data

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"

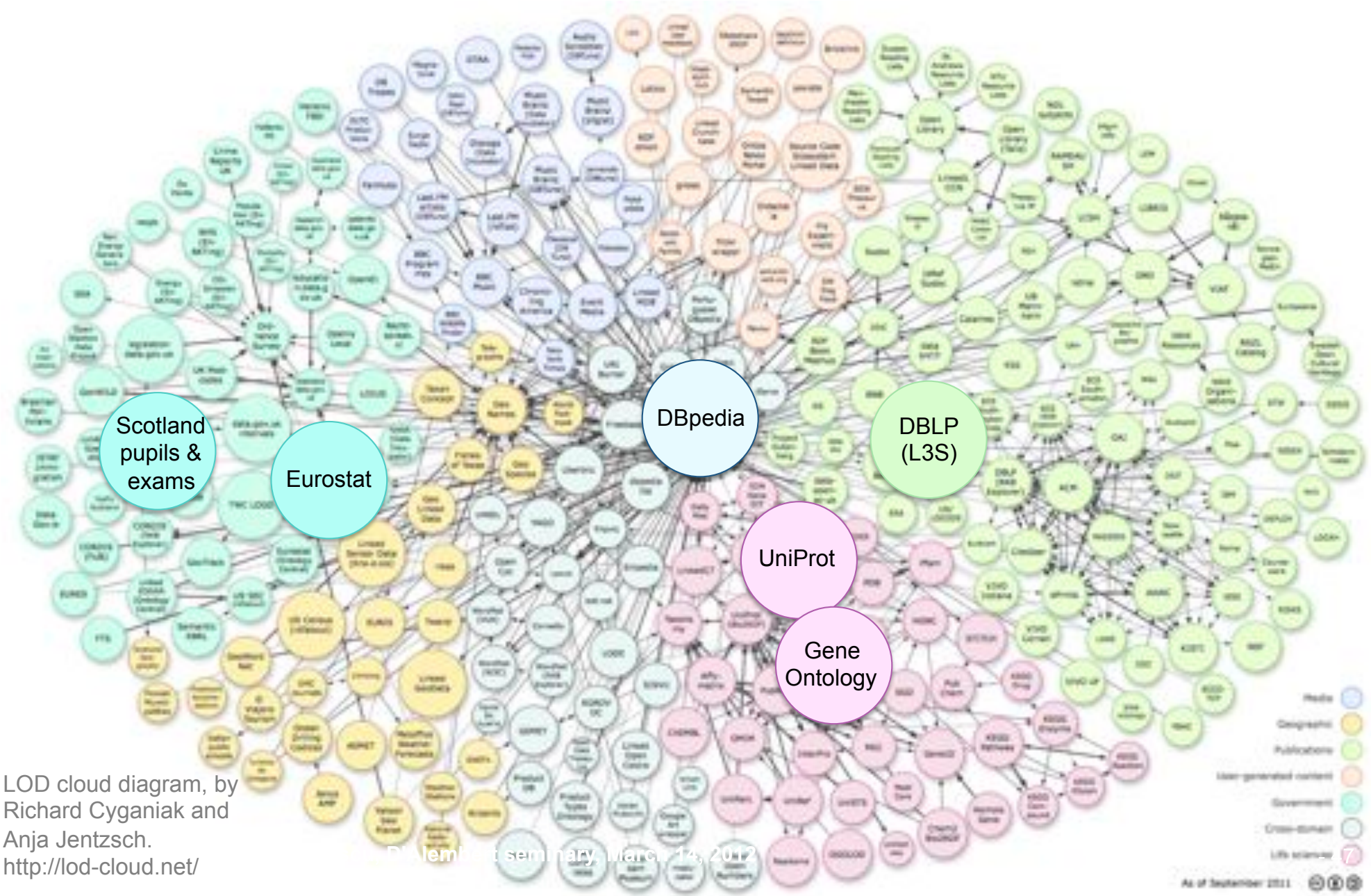
   - (Tim Berners-Lee) vision for the Web

2. **Open Data**:

   "**idea** 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"

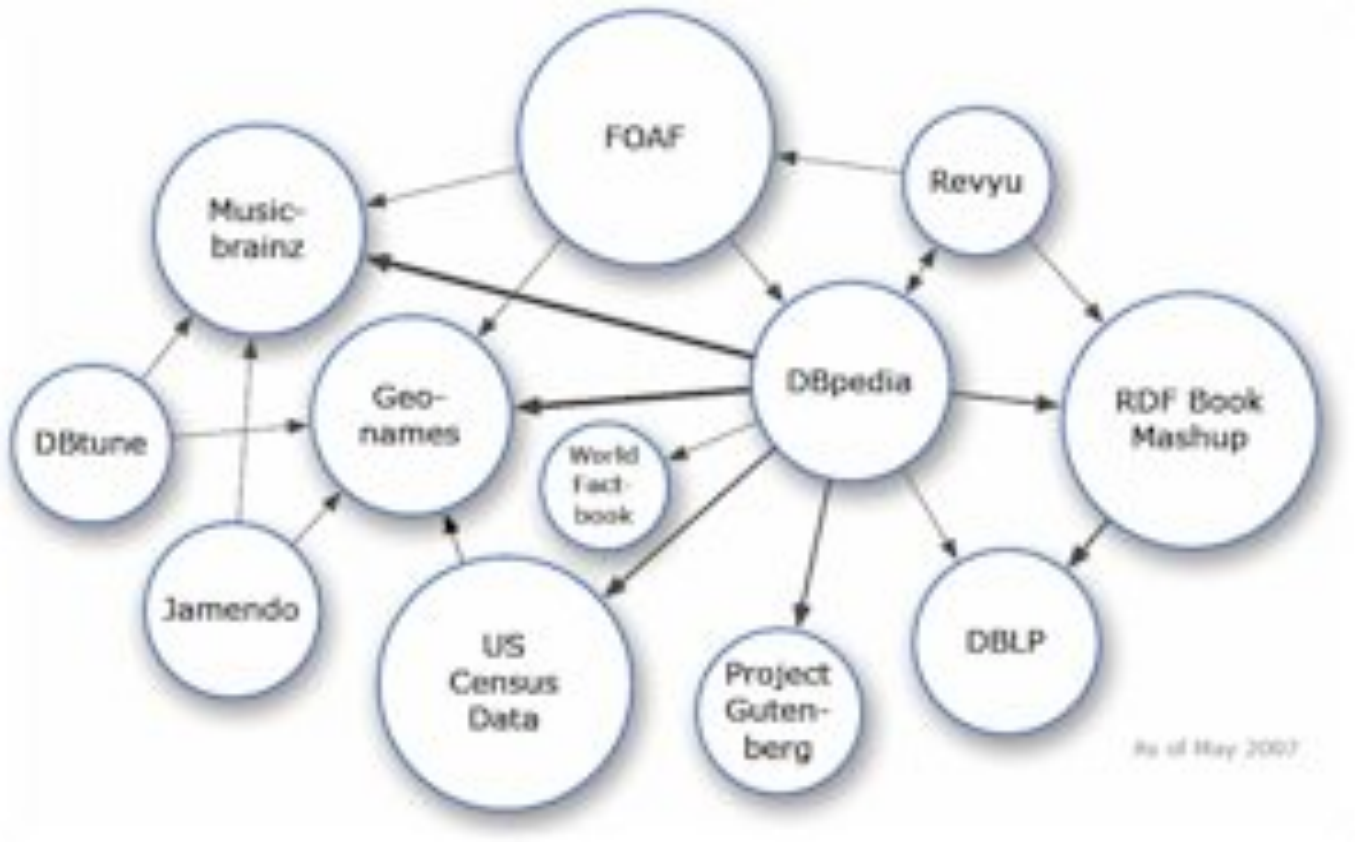
   - In principle, orthogonal to the Linked aspect
   - In practice, Linked is a technical mean toward Open

Linked Open Data Cloud



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

Linked Open Data Cloud (05/07)



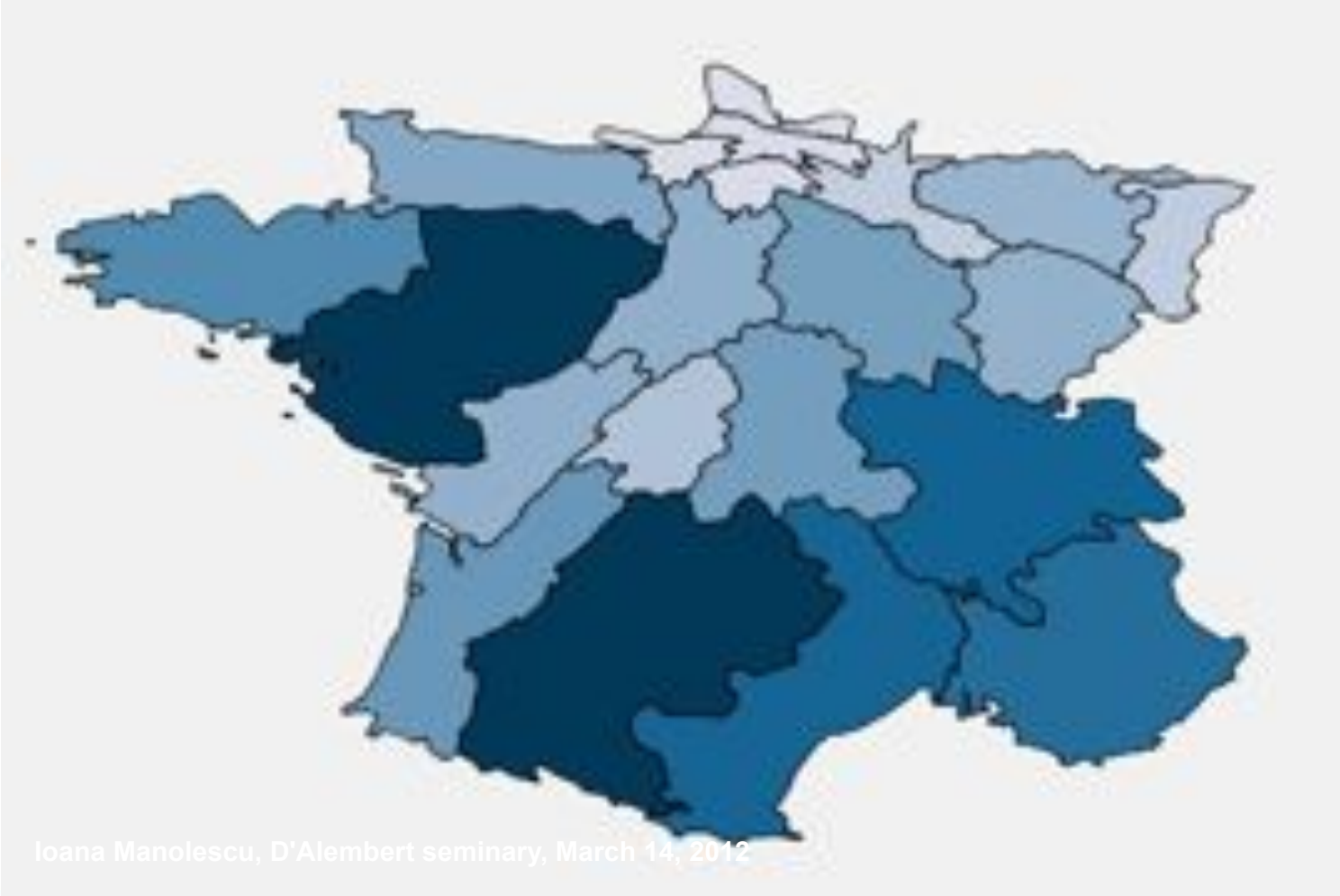
More Open Data: data.gov (US)

More OpenData: from Etalab (FR)

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

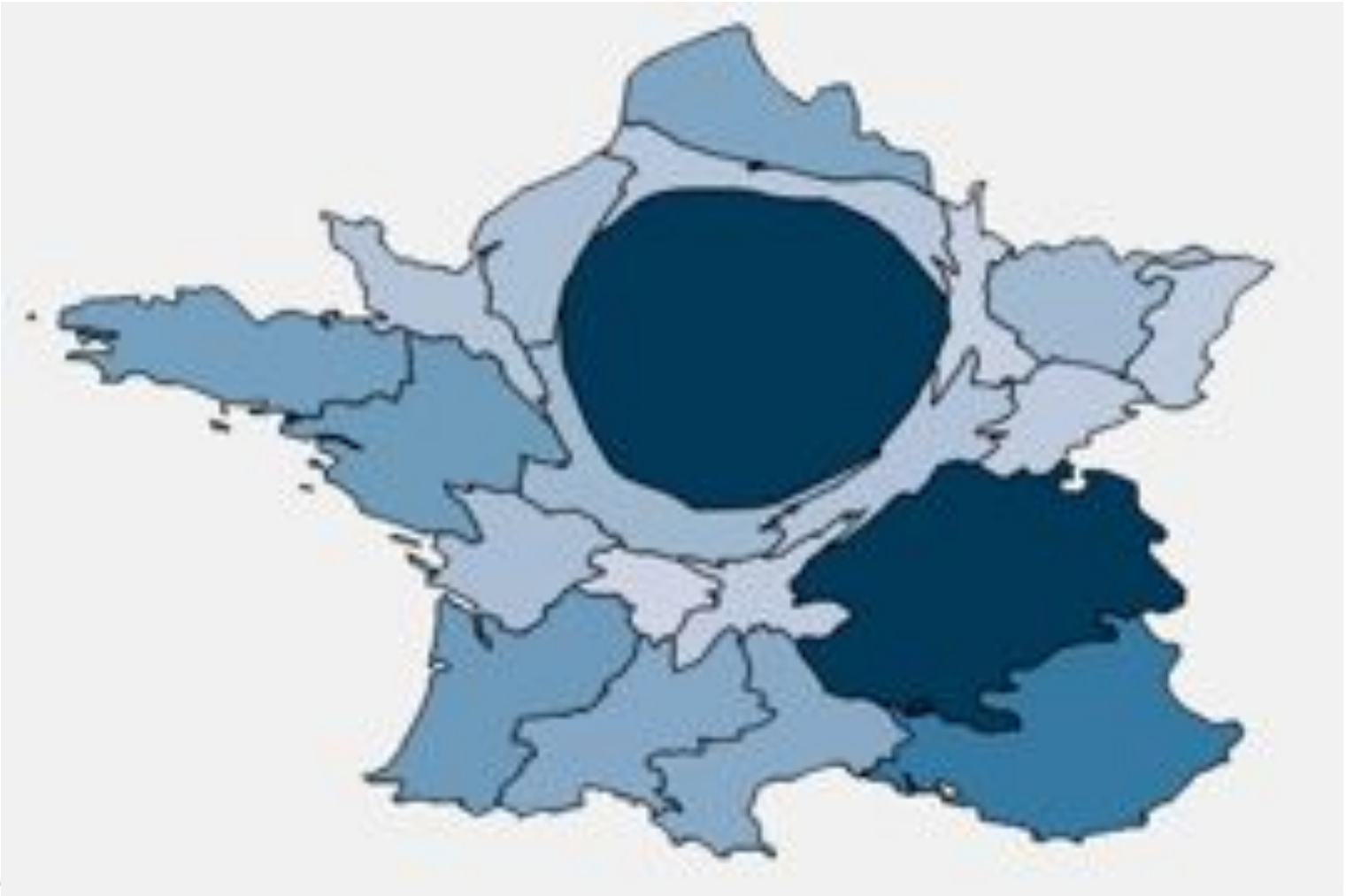
More OpenData: from Etalab (FR)

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



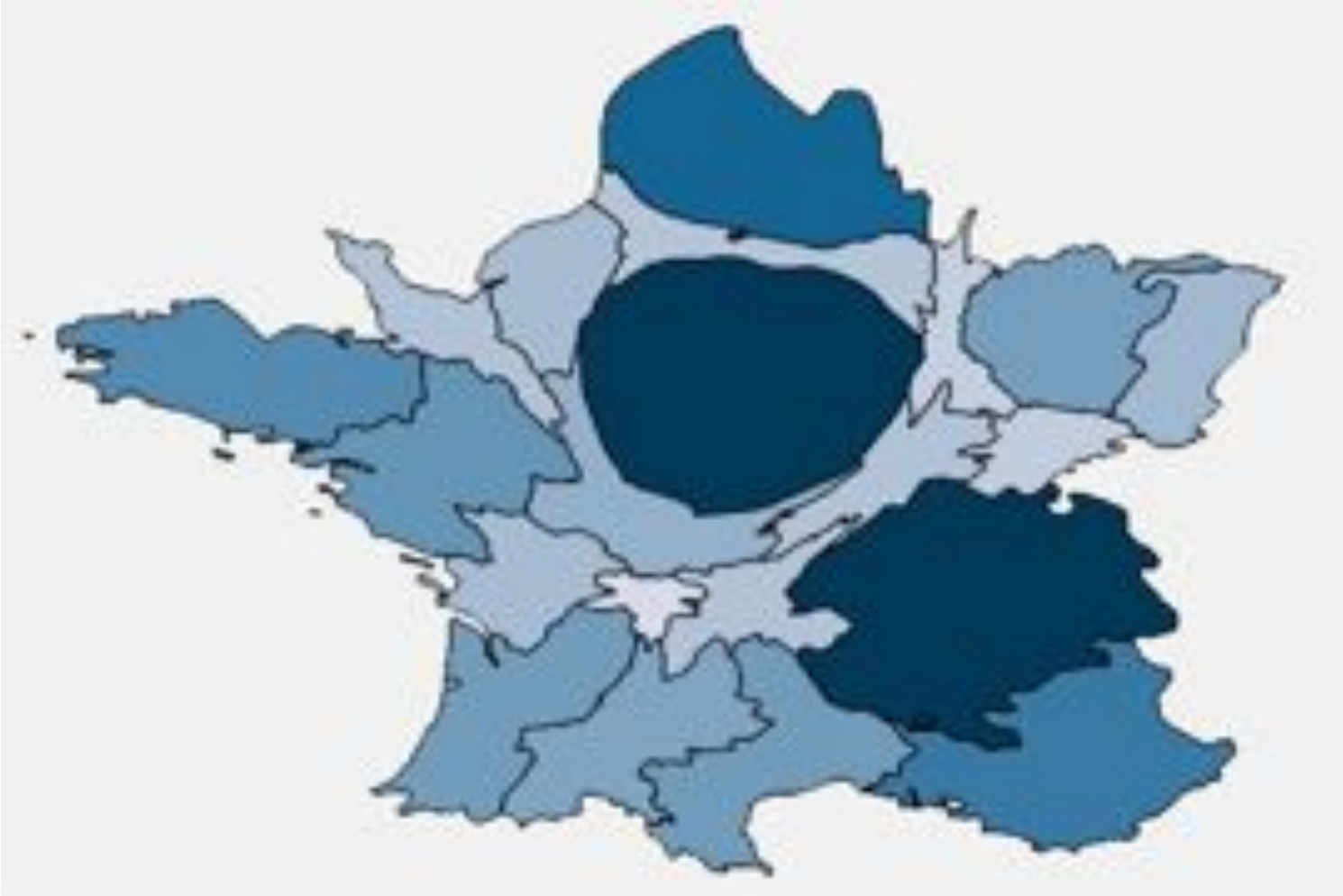
More OpenData: from Etalab (FR)

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)

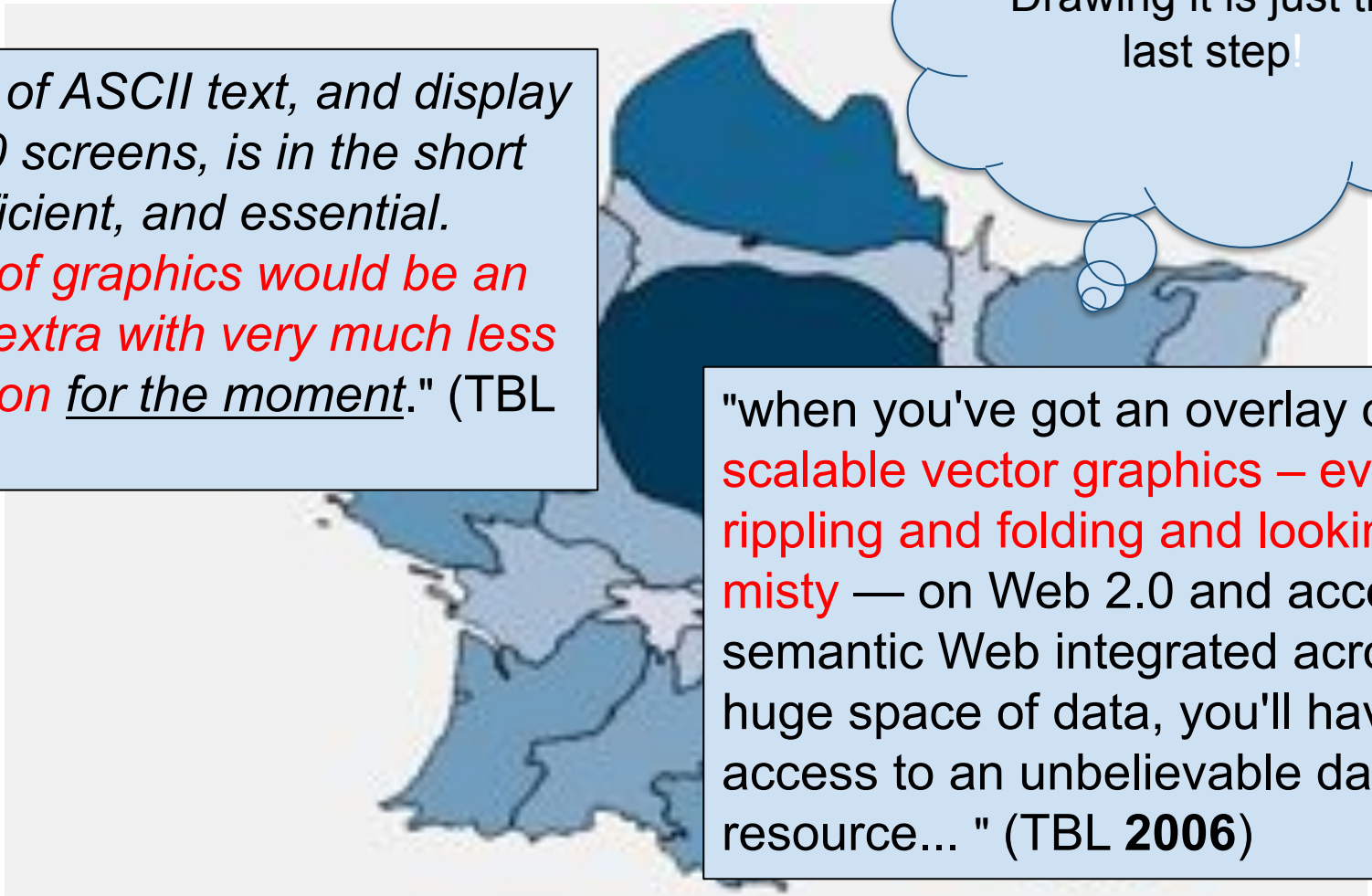


More OpenData: from Etalab (FR)

Boulangeries/inhabitants per French
(Le Journal Du Net, 07/09/20

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**)

4

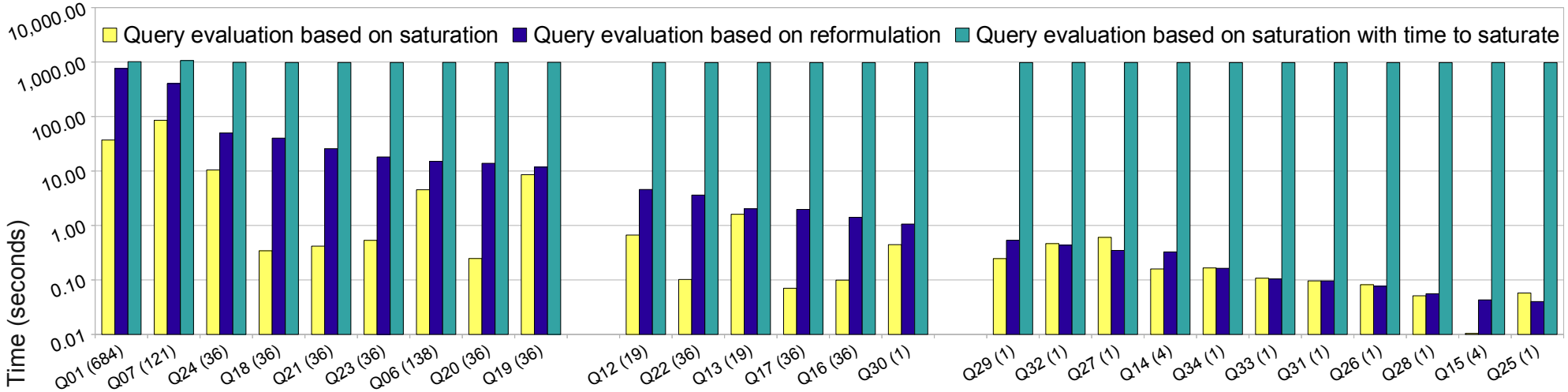
Selected research topics around Web/Linked Data

Ongoing work within OAK: efficient handling of implicit RDF data

Efficient query processing on very large volumes of RDF data.

How to **efficiently** handle implicit triples, due to RDF-S reasoning?

- Saturation: compute all such triples and store then in the database
- Query reformulation: keep the database unchanged, change the query



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)

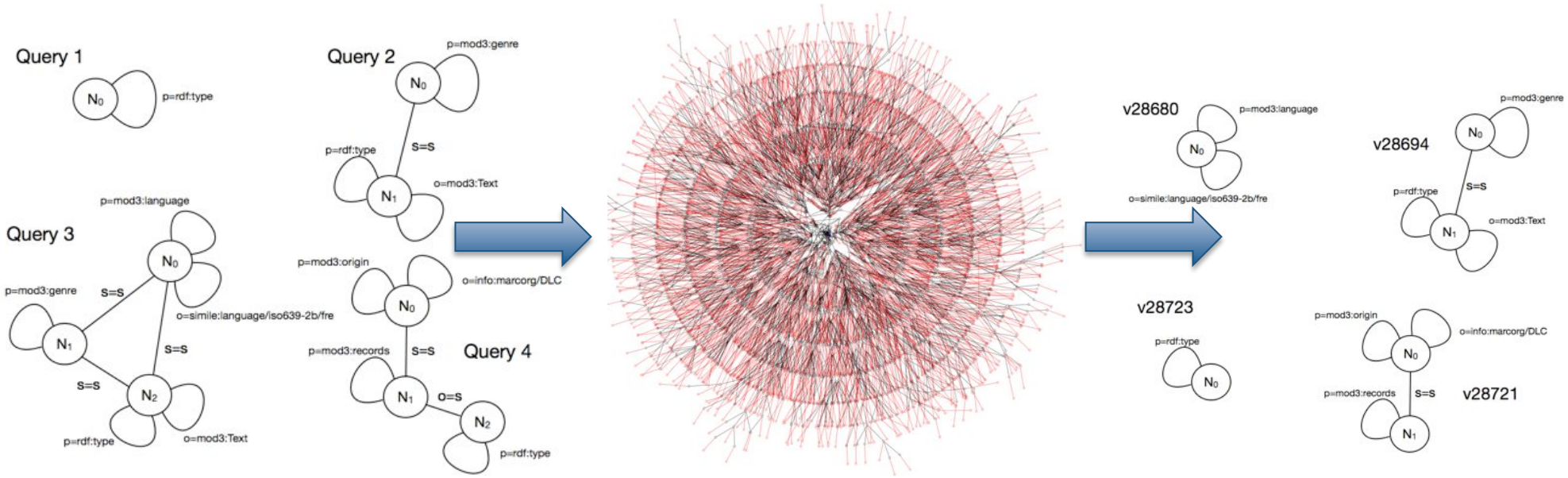
Difficulties: implicit RDF data, large queries

OAK paper @ VLDB 2012

Improving RDF query performance through materialized views

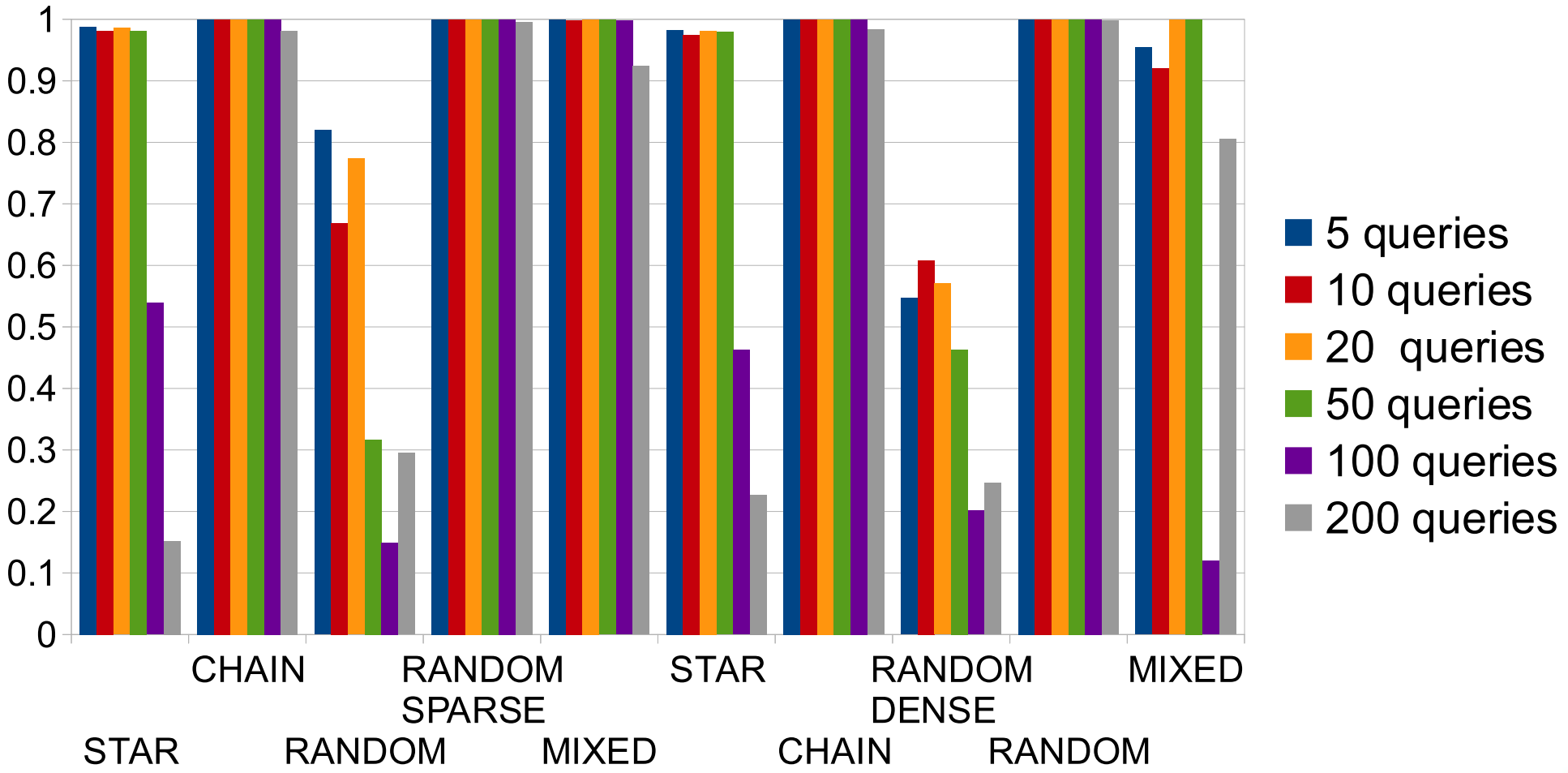
Input: RDF database D, RDF Schema S, workload $\{Q_1, Q_2, \ldots, Q_n\}$

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



View Selection in Semantic Web Databases (PVLDB 2011)

Cost reduction (%) achieved by our view search algorithm



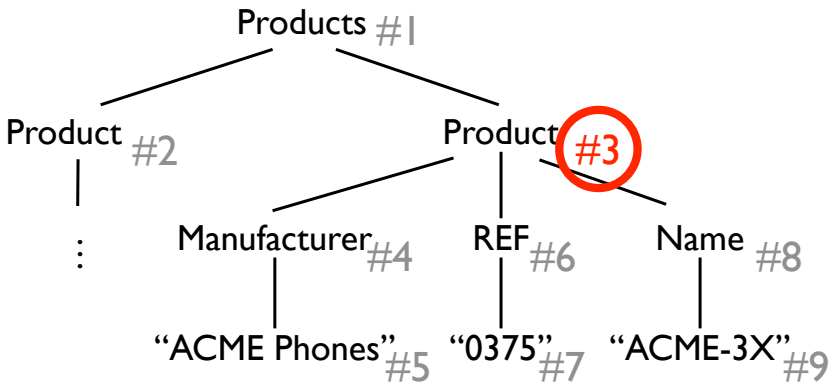
Ongoing work in OAK: documents with annotations

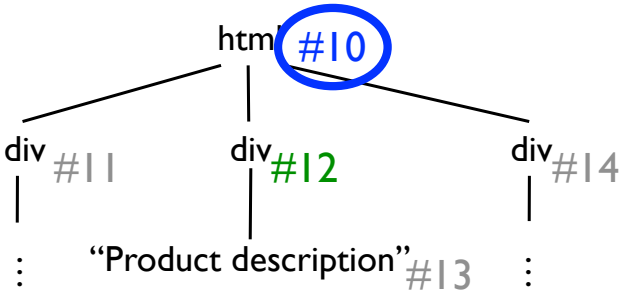
(⟨Alice⟩, ⟨likes⟩, #12),
(⟨Alice⟩, ⟨knows⟩, _:B),
(#3, rdf:type, ⟨MobilePhone⟩),
(#3, rdf:type, ⟨ElectronicDevice⟩)

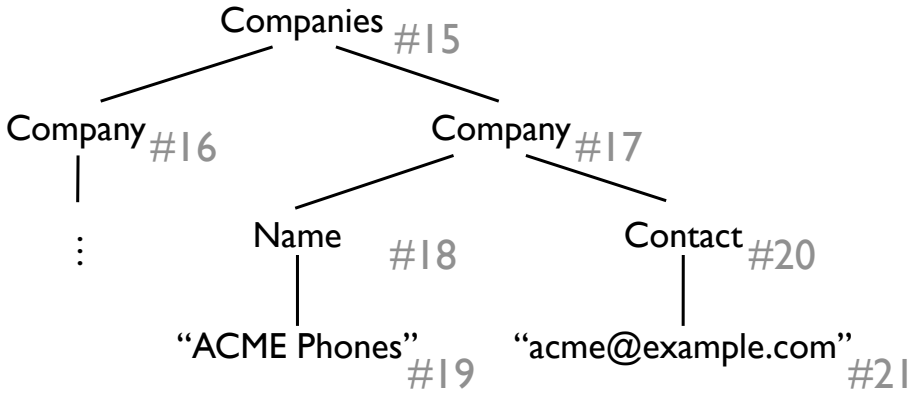
(#12, ⟨describes⟩, #3),
(_:B, ⟨email⟩, "bob@example.com"),
(⟨MobilePhone⟩, rdf:subClassOf, ⟨ElectronicDevice⟩),

(#10, ⟨date⟩, "23 May 2011"),

RDF sub-instance

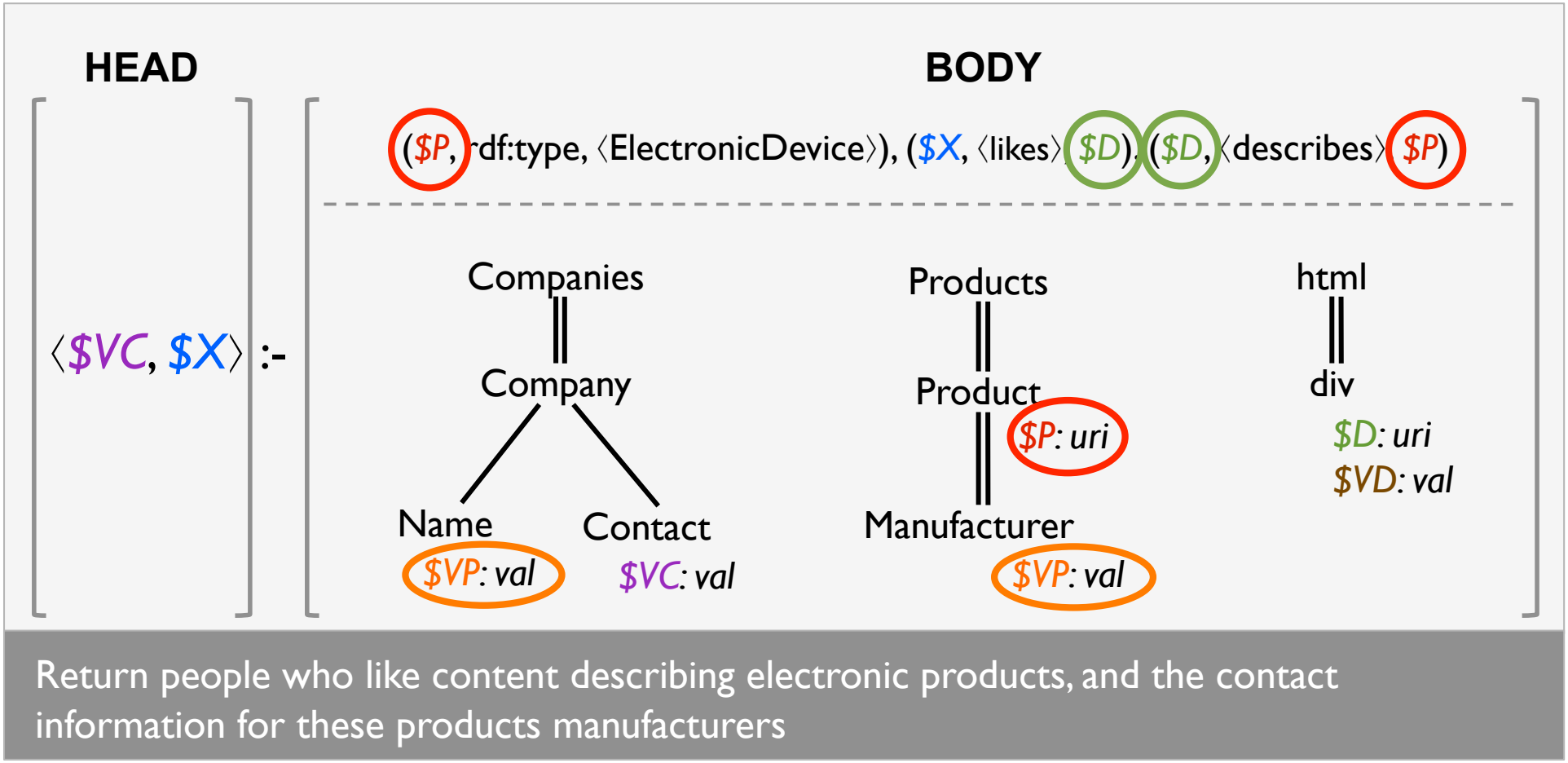




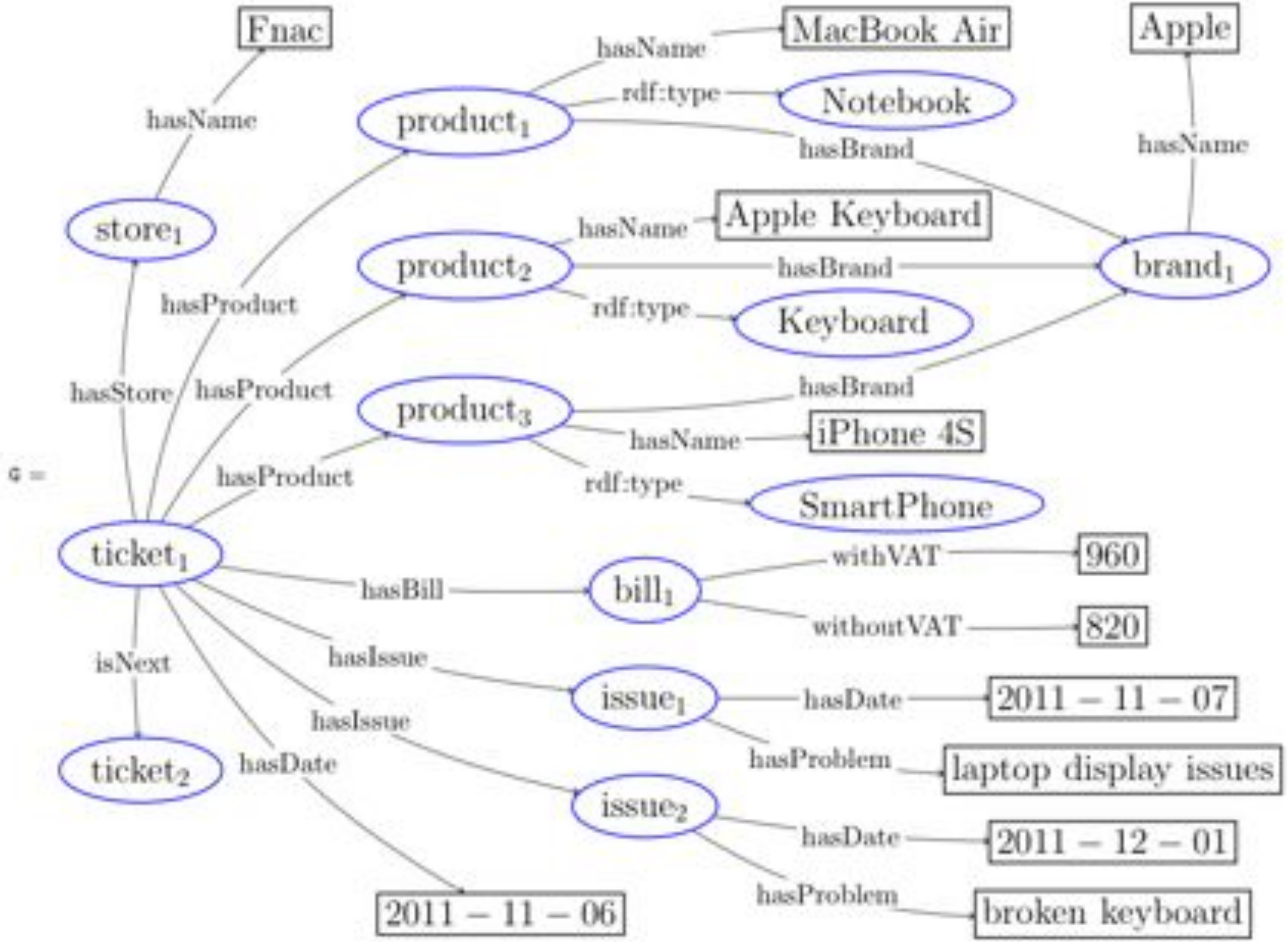


XML sub-instance

Ongoing work in OAK: querying documents with annotations



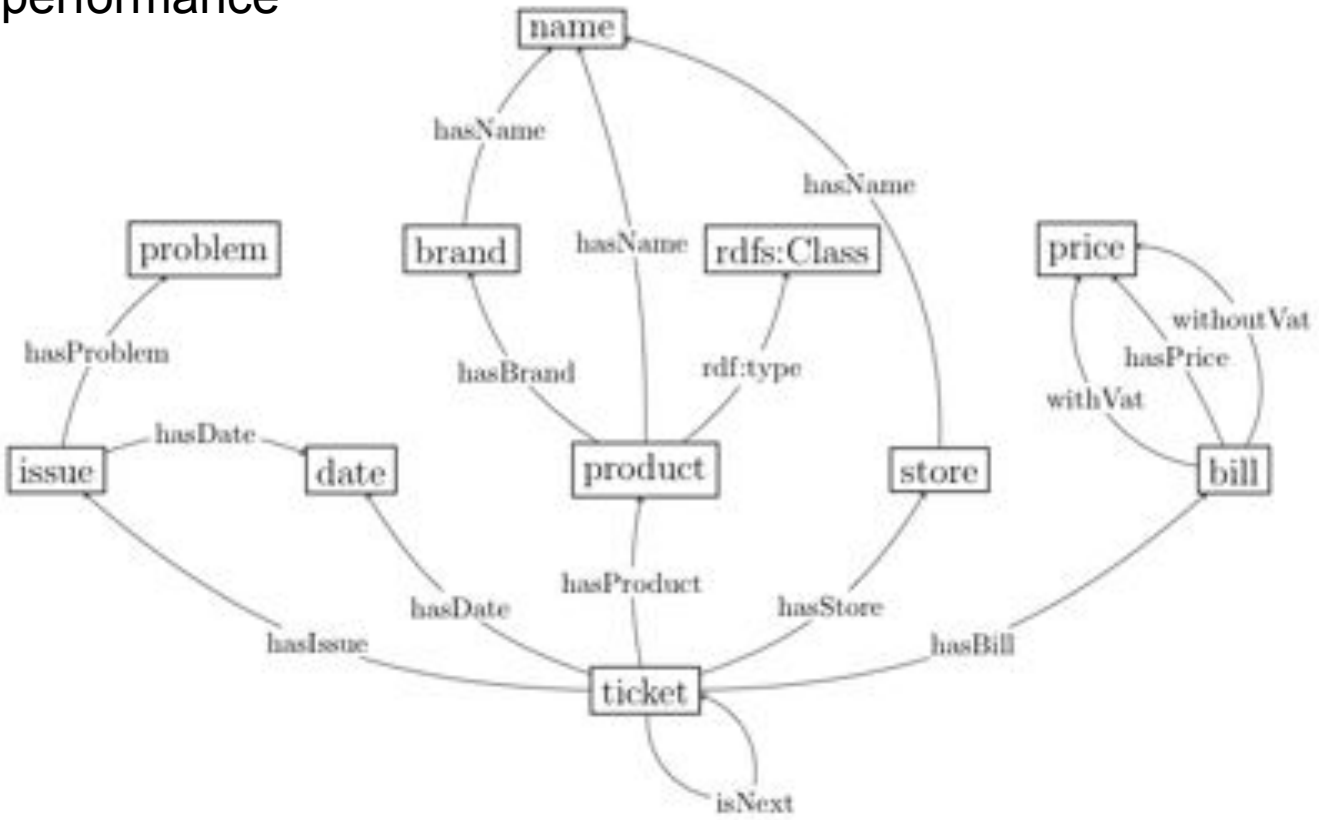
Return people who like content describing electronic products, and the contact information for these products manufacturers


G =
Fnac
hasName
store1
hasStore
ticket1
ticket2
isNext
hasProduct
hasProduct
hasProduct
product1
product2
product3
hasName
rdf:type
hasBrand
MacBook Air
Notebook
Apple Keyboard
Keyboard
iPhone 4S
SmartPhone
Apple
hasName
brand1
hasBrand
hasBill
bill1
withVAT
960
withoutVAT
820
hasIssue
hasIssue
issue1
issue2
hasDate
hasProblem
2011 – 11 – 07
laptop display issues
2011 – 12 – 01
broken keyboard
hasDate
2011 – 11 – 06

Ongoing work in OAK: warehousing RDF (and Linked) data

Two-step approach:

1. Model a cube based on a selected set of classes and properties of interest
2. Query the cube in the spirit of relational data cubes (roll-up, drill-down…)
   - With good performance



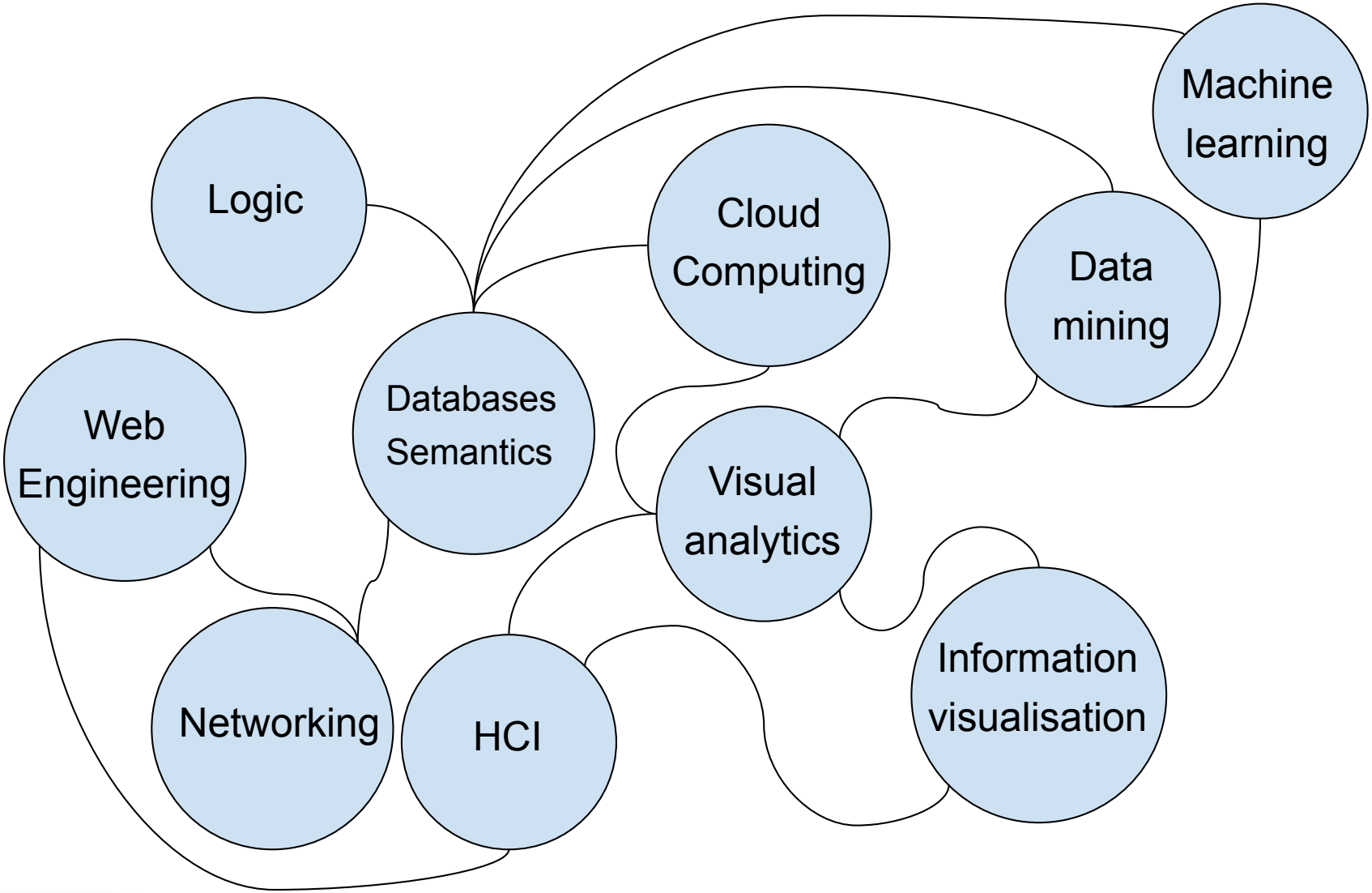
Web mining

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 We**b 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

merci / questions?