



## Dataförädling med semantiska tekniker

Olle Olsson

2

# ***ITARC 2010***

## ***Stockholm 100420***

***Olle Olsson***

***World Wide Web Consortium (W3C)***

***Swedish Institute of Computer Science (SICS)***

# Contents

Trends in information / data

Critical factors ... growing importance

Needs

Highlighting areas of needs

Technology

Enabling technology

Applications

Examples

# Contents:

***Trends in information / data***

Emerging needs

Technology

Applications

# Trends – Information/Data Volumes

Growing volumes

- Price/performance
- Networking
- Devices
- Storage, etc

Structured vs unstructured data!

Challenge: capacity and capability to handle huge data volumes

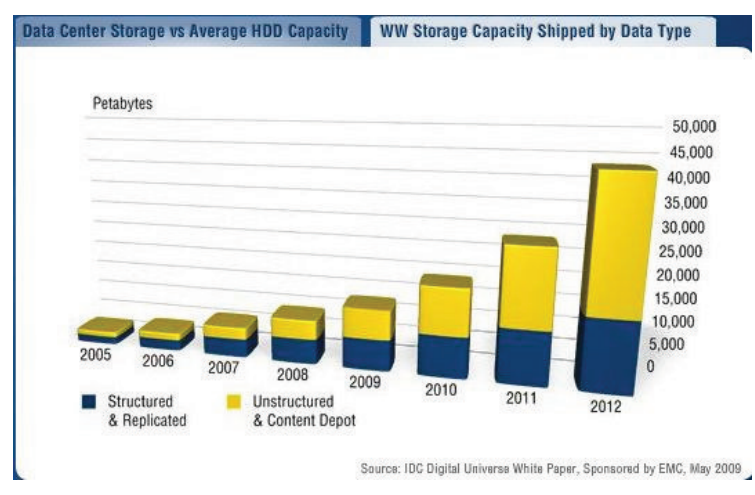


Illustration source: IDC/EMC "Digital Universe" (2009)



# Trends – Information/Data Volumes

Total volume vs number of units

Increased number of small data chunks

Challenge: keeping track of many small-size data chunks

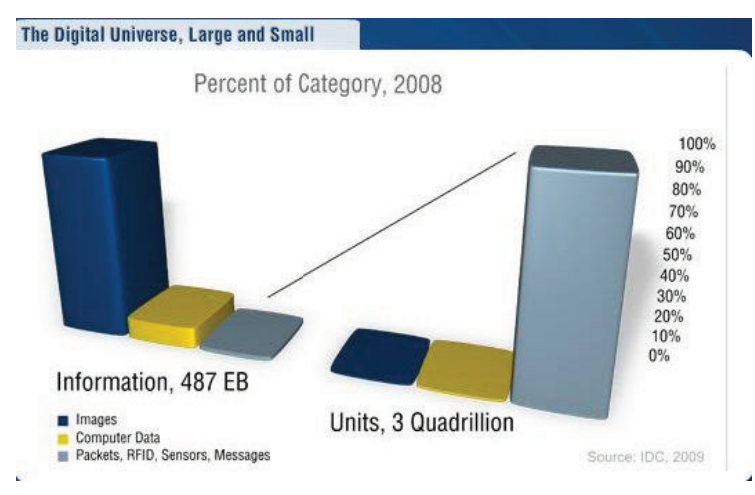


Illustration source: IDC/EMC "Digital Universe" (2009)

# Trends – Information/Data Sources

New types of devices arriving in accelerated pace

Sources of information / data

Challenge: Tracking data and dynamic sources

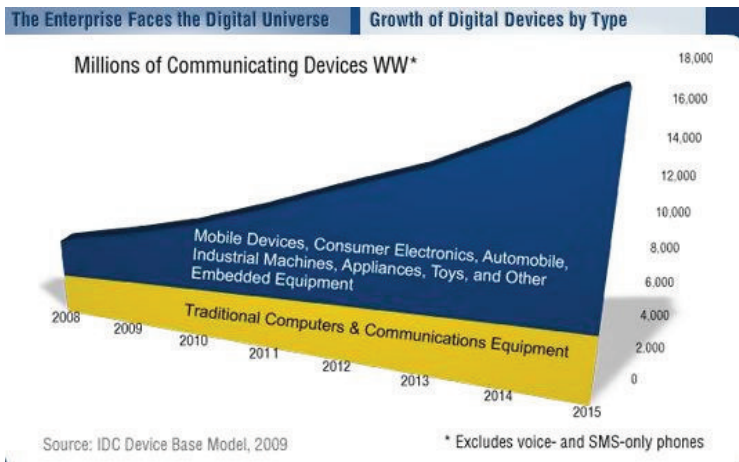


Illustration source: IDC/EMC "Digital Universe" (2009)

# Trends – Information/Data Sources

Data generating device examples

- Wireless sensor networks: multipurpose ...
- Mobile phones: GPS, temperature, ...

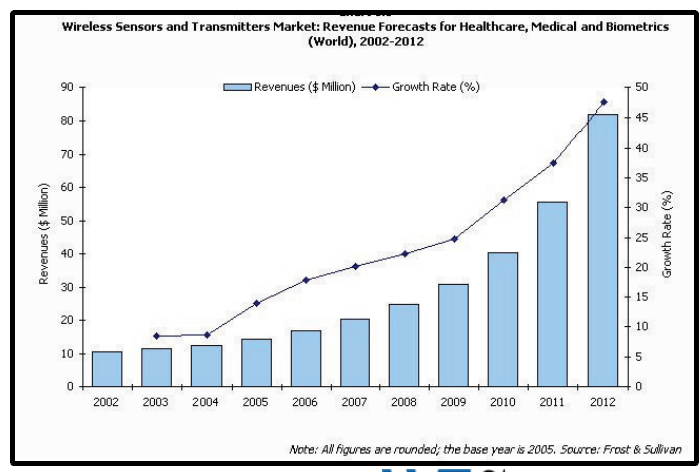
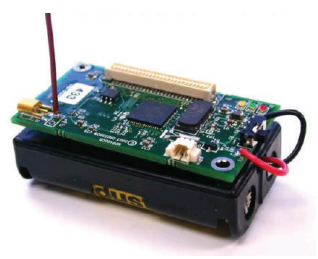


Illustration source: Frost&Sullivan (2009)

# Trends– Information/Data Communication

9

Internet paradigm evolving

- From *transmission centric*
- To *content centric*



Focus on data

- Content distribution
- (cf. the web!)

Challenge: how to package and describe data to profit from envisioned Internet functionality

Illustration source: UC San Diego (2009)

10

## Contents

Trends in information / data

**Emerging needs**

Technology

Applications

## ***Emerging needs – stream processing***

### Processing data on-the-fly

- Data/information generated continuously
- Streams of data
- Data stream processing

### Drivers, examples:

- Sensor networks – Internet-of-Things
- Messaging, blogging, micro-blogging

Challenge: how to process data efficiently/effectively

## ***Emerging needs – the web data space***

### Web context as picture of business context

- Interconnected sources of data
- Business interdependencies
  - “Data I need” vs. “Data I have & data you have”

### Drivers:

- growing volumes of available data
  - cf. “public sector information”
- data evolution
- Cost-efficiency!

Challenge: how to increase automation, data interoperability, adaptive processing



## Contents

Trends in information / data

Emerging needs

**Technology**

Applications

## **Technology – Semantic Web**

### Web technology

- Formats (XML, WS-\*, MathML, RDF, SVG, ...)
- Protocols (HTTP, SOAP, ...)
- Processing (XForms, DOM, Powder, Pipeline, ...)

### Semantic Web Technologies

- Rich representation:
  - RDF, RDFS, OWL, SKOS, ...
- Processing support:
  - OWL, RIF, SPARQL, ...

# The “*semantics*” in the Semantic Web

Semantic web is about what?

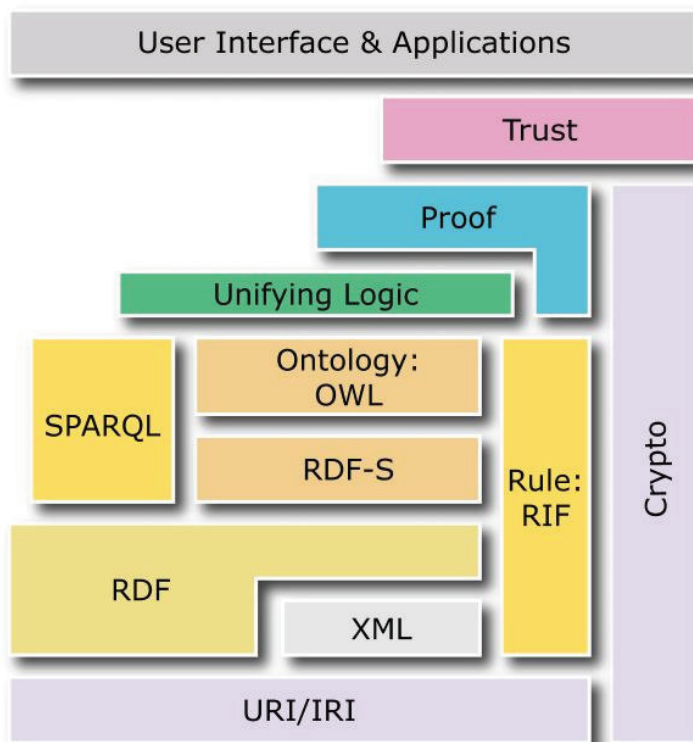
- about “meaning” and automation
  - “Meaning-based” automation

“Meaning” -- pragmatic approach in Semantic Web:

- a program “knows” what it *can* do with data
- self-describing data

No magic ... instead well-founded engineering

## Semantic web building blocks





# RDF – Resource Description Framework

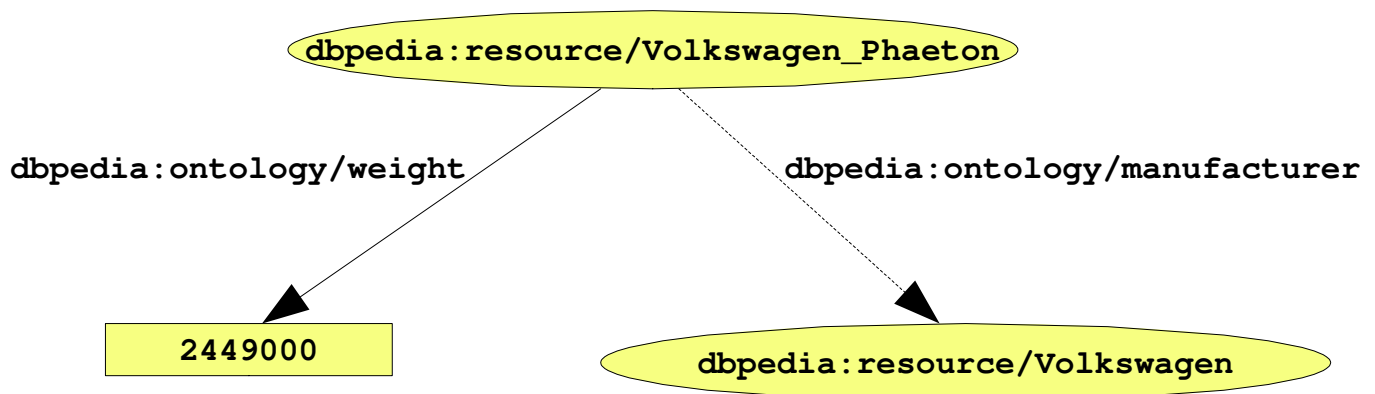
Basic data model – a “triple”

- triple (s, p, o) :
  - “s” -- subject
  - “p” -- predicate
  - “o” -- object
- Conceptually: “s” is related by “p” to “o”

RDF is a general model for such triples

- machine readable formats like RDF/XML, Turtle, n3, RXR

## RDF Example



```

dbpedia:resource/Volkswagen_Phaeton
dbpedia:ontology/weight
2449000
  
```

```

dbpedia:resource/Volkswagen_Phaeton
dbpedia:ontology/manufacturer
dbpedia:resource/Volkswagen
  
```

## OWL – Web Ontology Language

Define ontologies (conceptual model, ...) for data

Built on top of RDF

Basic components:

- *instance* – entity
- *class* – type
- *property* – relationship

Ontology enables:

- Checking consistency of instance graph
- Inferring implicit statements about instance graph

## SKOS – Simple Knowledge Organisation System

Define simple ontologies (conceptual model, ...)

Targeting traditional modelling approaches

- Taxonomies, classification schemes, thesauri, ...

Built on top of RDF

Compatible with modelling standards:

- NISO Z39.19 - 2005; ISO 5964:1985

# SPARQL – RDF Query Language

Retrieve RDF data from RDF data graphs

- RDF graph as answer to query

Syntax inspired by SQL

Example:

```

PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?manufacturer ?name ?car
WHERE {
  ?car skos:subject <http://dbpedia.org/resource/Category:Luxury_vehicles> .
  ?car foaf:name ?name .
  ?car dbo:manufacturer ?man .
  ?man foaf:name ?manufacturer
}
ORDER by ?manufacturer ?name

```

In progress, functionality for:

- update; subqueries; negation; etc

## Triple store

Database system for triples (RDF graphs)

- store, manage, query

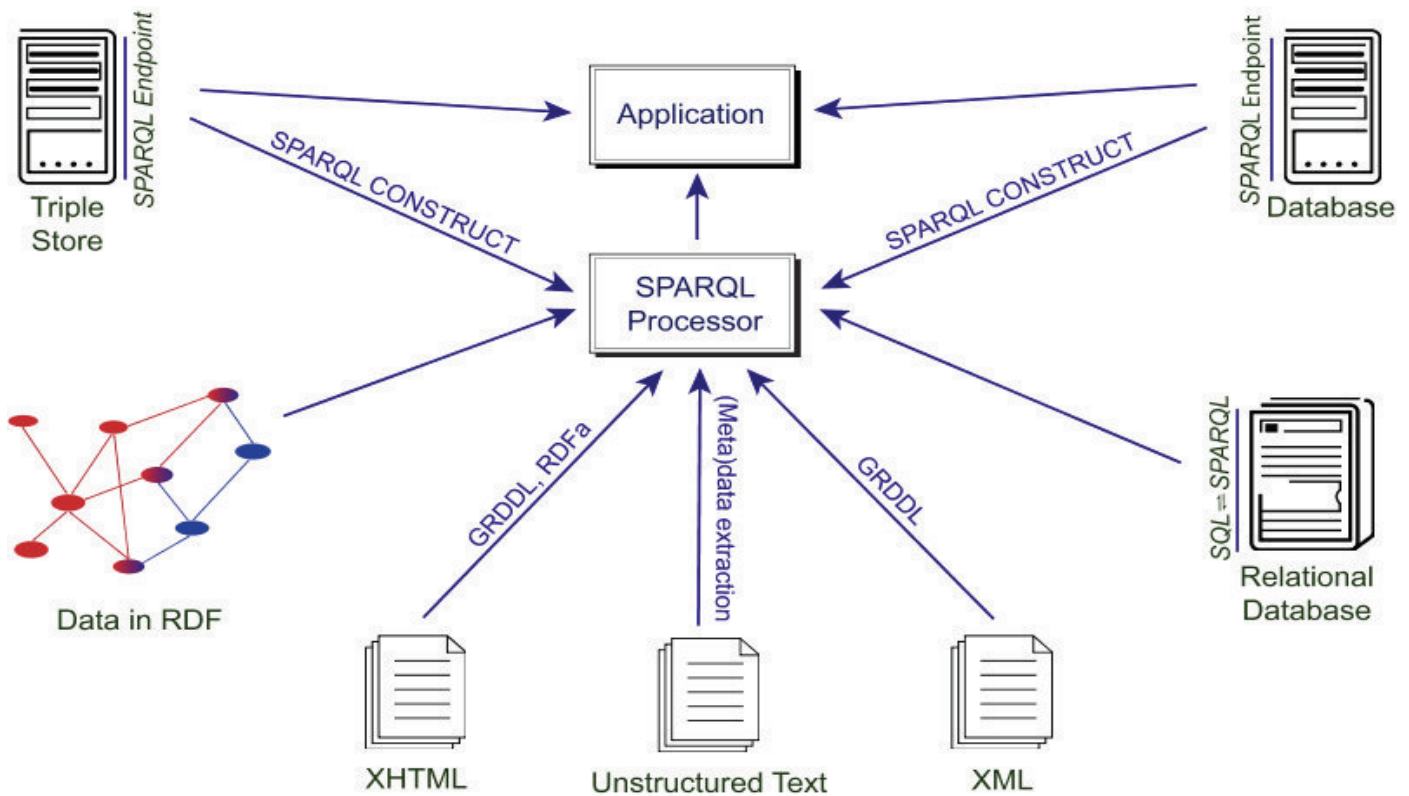
Implementation approaches

- Native triple store – RDF from ground up
- “triplified” RDB – basic storage in relational form

Interface:

- SPARQL

# Integration via SPARQL



## SemWeb technology implementations

Free / open source:

- Franz, Google, HP, Mozilla, ...

Commercial:

- IBM, Ontoprise, OpenLink, Oracle, Talis, ...

Still somewhat fragmented supplier space.

- “And the winner will be ...?”

But core standards are established!

## Contents

Trends in information / data

Emerging needs

Technology

***Applications***

## ***Applications of SemWeb***

### Aspects:

- Technology: availability; quality; cost; ...
- Methodology: scope; depth; ...
- Application areas: selection; scope; ...
- Technological environment: interoperability; ...
- Needs evolution: stable, isolated; dynamic, open; ...

### Ambition:

- Small, component-oriented, add-on, ...
- Large, all-embracing; ...

## *SemWeb technologies vs the Web*

Semantic web technologies :

- Support critical web requirements

Not necessarily used on the web

- Internal encapsulated component in some application

Web requirements positive effects on

- Interoperability; maintainability; evolution; ...

SemWeb acceptance analogous to XML, e.g.:

- Used in all contexts
- Standardised
- Uniform tool support
- Enables interoperability

## *Limited-scope application*

Adobe XMP

- Extensible Metadata Platform
- Copyright, Creator, Date, Location, ...

Aim:

- share metadata across applications, file formats, and devices

Metadata added by tools, e.g. Photoshop

- Formatted as RDF/XML



Advantage:

- Supports vendor-independent management of metadata



## Limited-scope application

### Dublin Core (DC)

- Document metadata (initial target: library metadata)
- “Dublin Core Metadata Element Set”
- Title, Creator, Date, Publisher, Language, ...

### Aim:

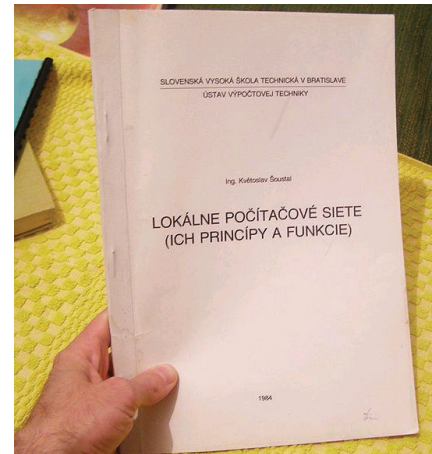
- Make existing DC metadata accessible as RDF

### Simple change of format (syntax)

- Semantics unchanged

### Advantage:

- Using standard technologies
- Simplifies data integration



## Limited-scope application

### Microsoft Interactive Media Manager (IMM)

- Extension to Sharepoint for media sector
- Metadata framework using RDF and OWL

### Aim:

- Support workflow for media assets

### Advantage:

- Metadata sharing between systems
- Simplifies data integration

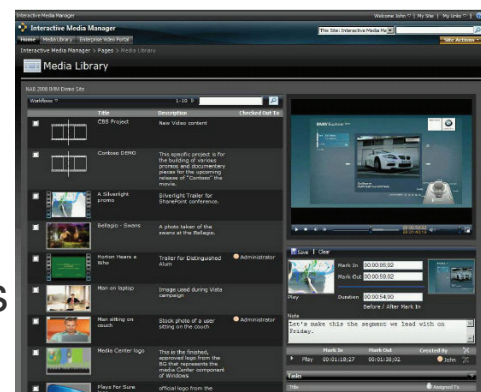


Illustration source: Microsoft

## Larger application

### Renault Automotive Repair

- Generate context-sensitive diagnosis/repair information
- Fetch general information from several sources

### Aim:

- Provide user with relevant information

### Source information

- Metamodel in OWL

### Advantage

- Sharable modeling.
- Unified view over distributed data.

Effectuez le test suivant :

Connecteur du calculateur de clim régulée - Contrôle visuel (1 mn)

Opérations préliminaires

- Autoradio "Haut de gamme" - Dépose (4 mn)
- Radio Navigation - Dépose (3 mn)
- Console centrale - Dépose (7 mn)
- Garniture de bas de planche de bord - Déclipper (1 mn)
- Garniture centrale - Dépose (3 mn)
- Tableau de commande de climatisation régulée - Dépose (7 mn)
- Connecteur du calculateur de clim régulée - Accès (0 mn)

Vérifier le branchement et l'état du connecteur

- Connecteur du calculateur de clim régulée - Rebrancher
- Connecteur du calculateur de clim régulée - Remplacement
- Autre...

Causes/Réparations possibles

- 13 % Câblage moteur de recyclage - calculateur de clim régulée - Remise en état
- 13 % Câblage moteur de recyclage - calculateur de clim régulée - Remplacement
- 13 % Calculateur de clim régulée - Remplacement
- 12 % Connecteur du calculateur de clim régulée - Rebrancher
- 13 % Connecteur du calculateur de clim régulée - Remplacement

Autoradio "Haut de gamme" - Dépose (4 mn)

Outillage spécialisé indispensable

Ms. 1544 Outil de dépose autoradio - Carminat Becker et Cabasse.

Autoradio chargeur CD

112232

112232

□ Déposer l'autoradio (1) à l'aide de l'outil (Ms. 1544)

□ Débrancher les connecteurs.

## Larger application

### NASA Expertise Finder - "POPS"

- Find expertise for task at hand
- 70,000 experts in organisation and contractor workforce

### Aim:

- Search tool against several databases

### Solution

- RDF-based integrating framework

### Advantage

- Consistent data model
- Loosely coupled infrastructure
- Share and reuse information

Space v.41 - Connected to "POPS on FatDuck" - Using Model "POPS on FatDuck Model" - Logged in as "Michael Grove"

File Options Bookmarks Advanced Help

NASA Center (15)

- ARC
- DTIC
- GRC
- SRPC
- HQ
- IVV
- JPL
- JSC
- Consolidation-A
- LARC
- MAF
- MSFC
- NSC
- JSC
- OSTC

Project (176)

- CCMC
- CMSI
- CLUSTER-B
- Carbon Cycle science team
- Cassini
- Center Investment Accounts
- Cloud-Aerosol Interact and Infrared PathM...
- Communications, Computing, Electron...
- Constellation-A
- Contract Management
- Corp Labor - Other HD Personnel Costs
- Corporate Labor
- Crew Exploration Vehicle
- (DAWN)
- (Data Acquisition) zifira

Competency (14)

- Acquisition and Contract Management
- Advanced Analysis and Design Method
- Advanced Mission Analysis
- Aerospace Systems Concept Developm...
- Air Traffic Systems
- Astronomy and Astrophysics
- Assesses
- Budgeting Management
- Business IT Systems
- Business Management
- Business Work & Team Management
- Communication Networks & Engineering
- Computer Systems and Engineering
- Configuration Management
- Construction Management

People (11)

- Bonita L. Hill
- Christopher J. Hill
- Jonathan S. Hill
- Kenneth W. Hill
- Michael E. Hill
- Ronald N. Hill
- Robert N. Hill
- Ryan L. Hill
- Steven D. Hill
- Timothy B. Hill
- William J. Hill

Information Panel

**Robert N. Hill**

fax: 301.286.1111

hasEmployer: NASA

phone: 301.286.1111

firstName: Robert

lastName: robert.n.hill@nasa.gov

mailingAddress: cmo-Robert N. Hill, 1400+Goodland Space Flight Center

hasRequirement: GSFC-581.0

roomNumber: 1111

Details Query Alternate Paths Web View

## Larger application

### BBC Music

- Provide rich information about broadcast music
- Not only broadcast data

### Aim:

- Offer users fresh data, enable navigation to non-BBC sites

### Solution

- RDF-based; uses other RDF sources

### Advantage

- Minimizes own data management
- New sources appear: easy to extend



## Linked Open Data Initiative

### Web of data:

- Many open datasets on the web
- Interoperable when accessible as RDF

### Examples:

- Wikipedia (“text”) ==> dbpedia (RDF);
- Scientific data sets (experimental data)
- Public sector information (geodata, census data, statistics, ...)

### Different aims and coverage

- But semantically interrelated
- Increasingly so over time!

## Linked Open Data Initiative

### Experimental initiative

- Explore opportunities
- Identify technology strength/weakness
- Synthesize best-practice guidance

### Stakeholders

- Data owners
- Data re-users

### Specific case, profit from experience

- Open public data
- PSI directive
- How to make data more easy to reuse?

## Linked Open Data Initiative

### Linked Open Data (LOD) objective:

“expose” open datasets via RDF

*set RDF links among the data items* from different datasets

a typical example is to set an owl:sameAs between two items in different datasets that refer to the same “thing”

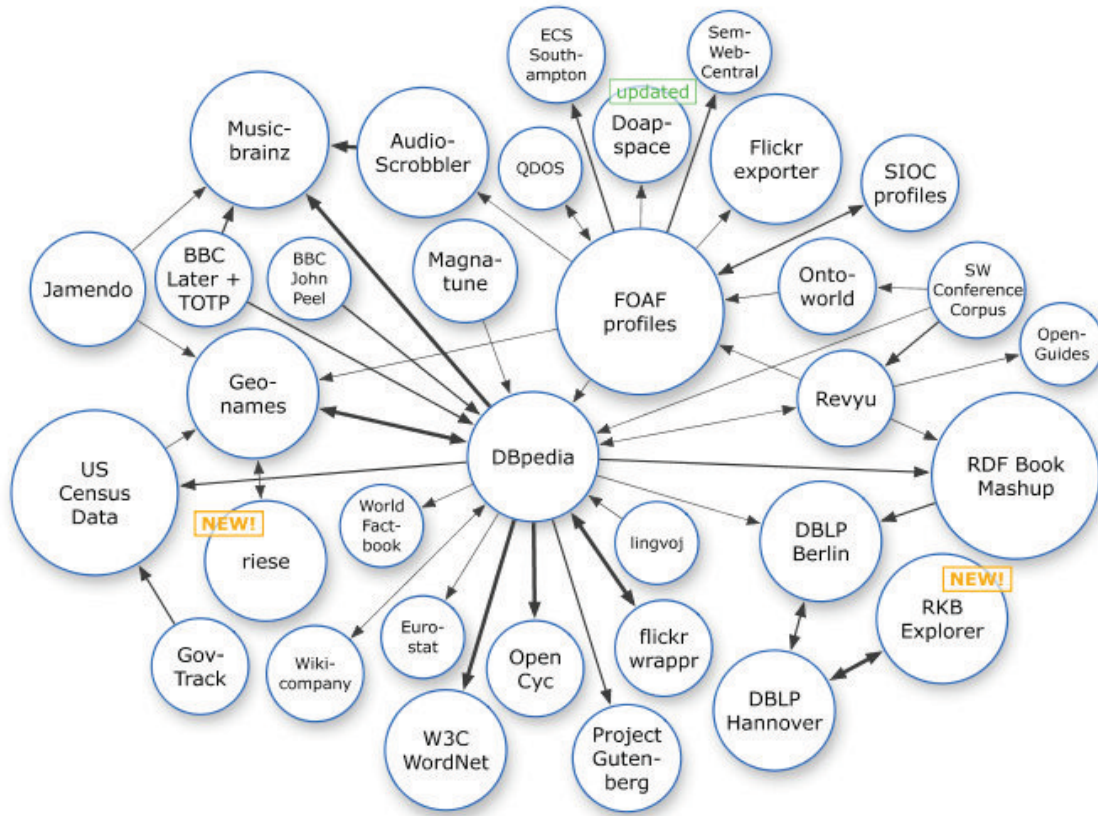
set up query endpoints (usually SPARQL)

Altogether billions of triples, millions of links...

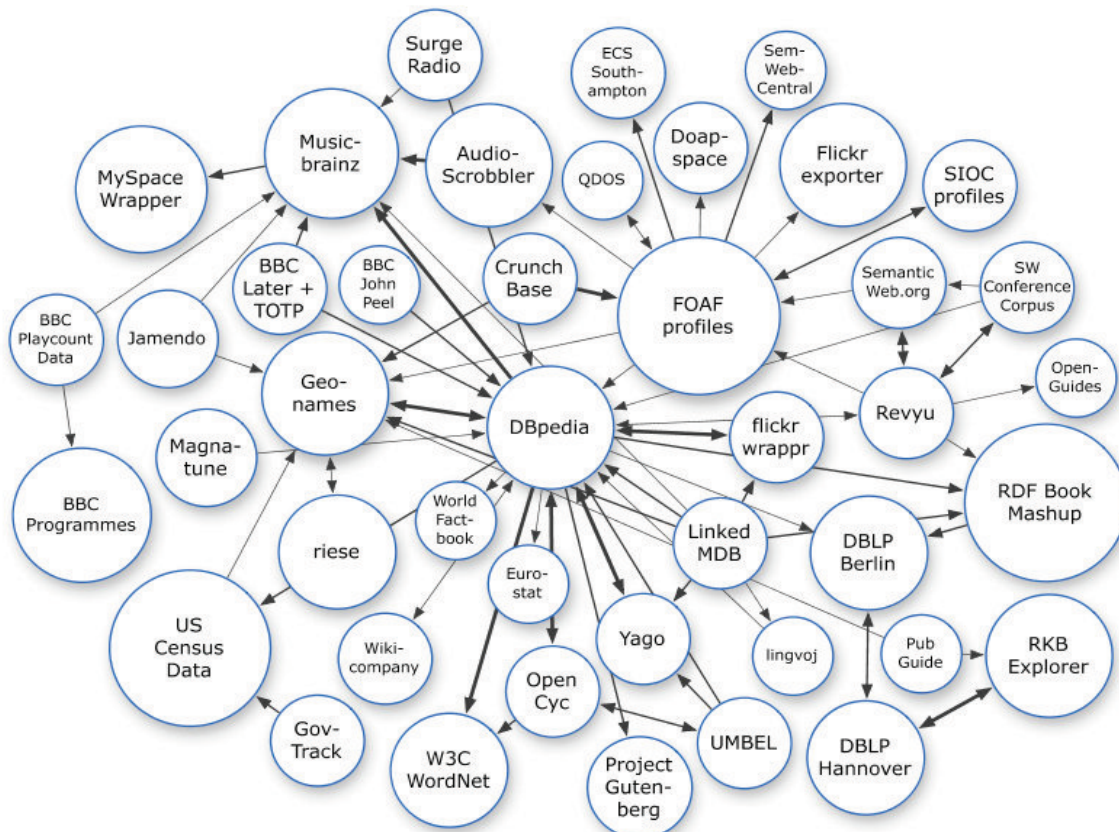
The “seed” for a general Web of Data



# The LOD "cloud", March 2008



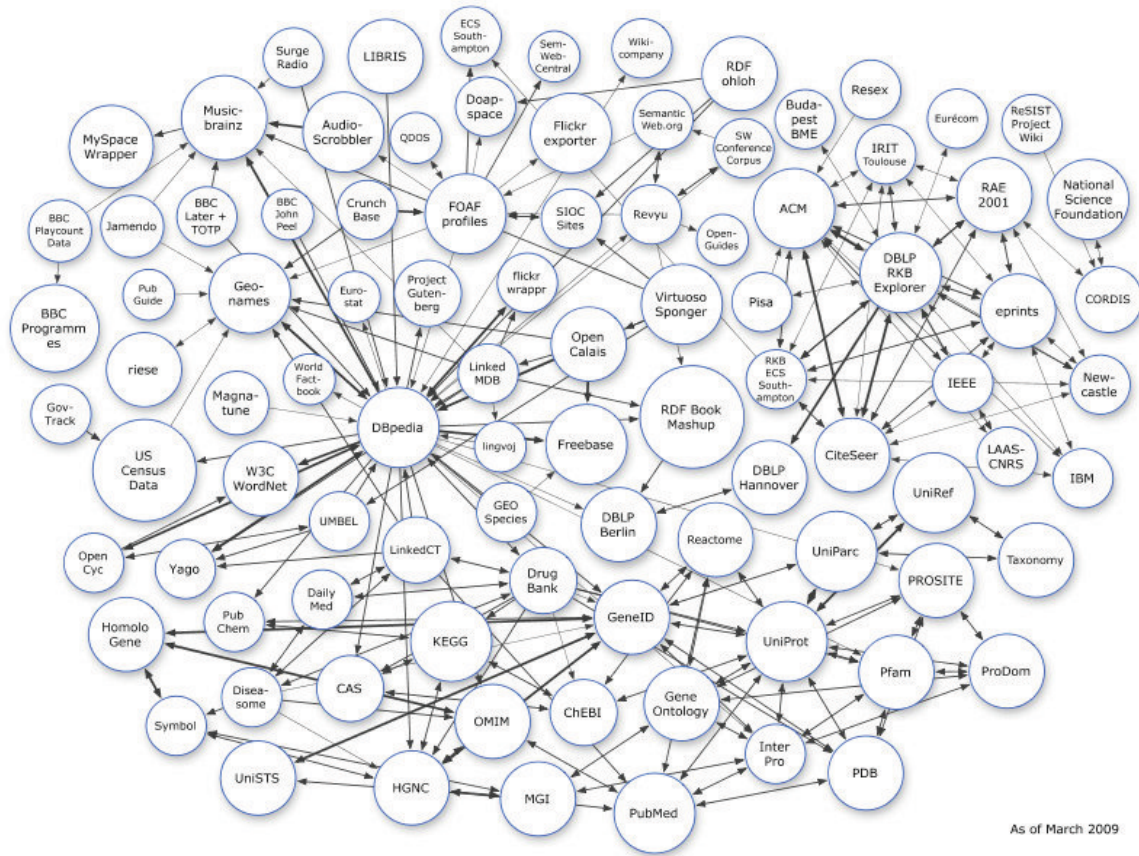
# The LOD "cloud", September 2008



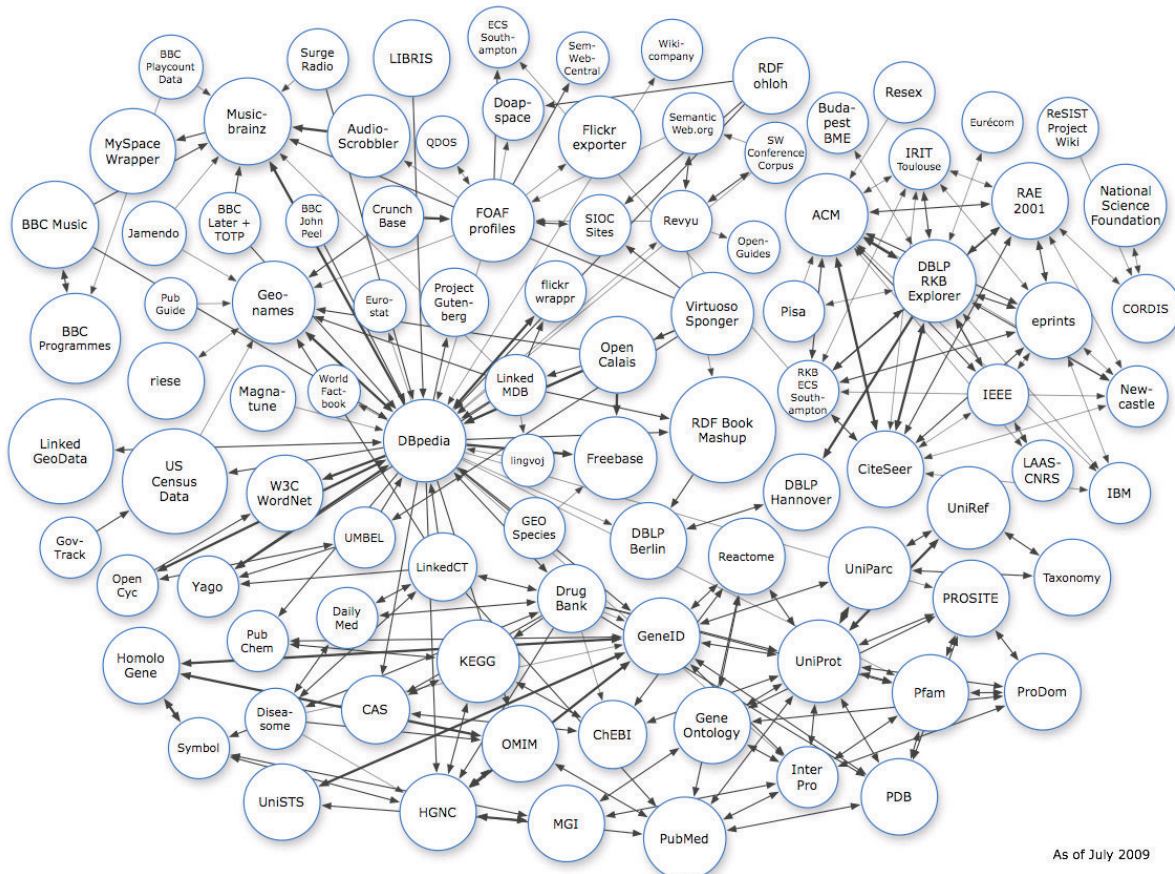
As of September 2008



# The LOD "cloud", March 2009



# The LOD "cloud", July 2009





# Applications using the cloud emerge

Bookmarking systems, exploration of social graphs, financial reporting

LOD nodes (eg, DBPedia) provide a set of referenceable URI-s for many things

Worth looking at the proceedings of the latest workshop, for example

April 2009, at WWW2009

<http://events.linkeddata.org/ldow2009>

## Semantic Sensor Web

Sensor networks deliver data

Semantic Sensor Web (OGC: Sensor Web Enablement)

- Deliver semantically represented data
- Control sensing & delivery via semantic specifications
- Discovery, access, tasking, alerts

Towards a semantic  
Internet-of-Things

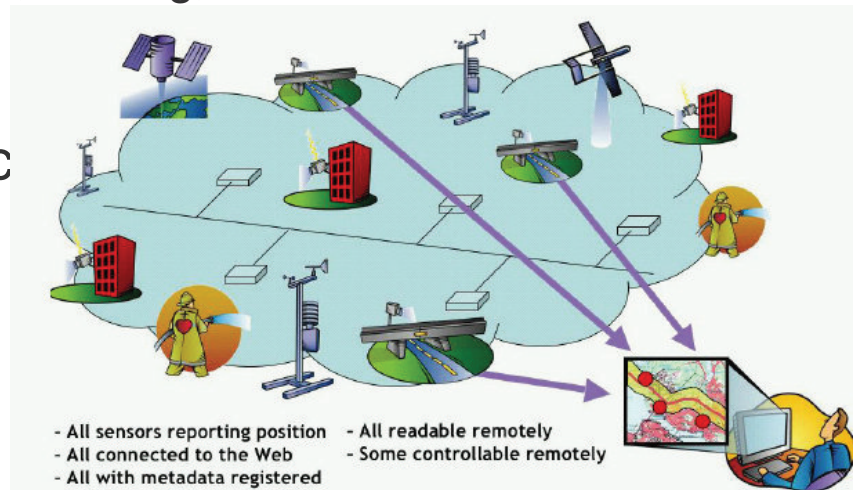


Illustration source: OGC (2007)

## *Reflections and Conclusions*

## *Semantic techniques*

### Toolset

- Basic standardised technologies
- Proven through use
- Emerging extensions and additions (needs driven)

### Applications

- Many products and applications in use

### Competence

- Awareness and experience growing

### Drivers

- More demanding business needs
- Public sector initiatives

## *Growth of data / information*

### Increased volumes

- Moore's law helps!

### Increased diversity (more types of data)

- Semantic techniques definitely appropriate

### Increased complexity (more interconnections)

- Semantic techniques are enabling

### Increased rate of change

- (the tough challenge!)

*Thank you for your attention*