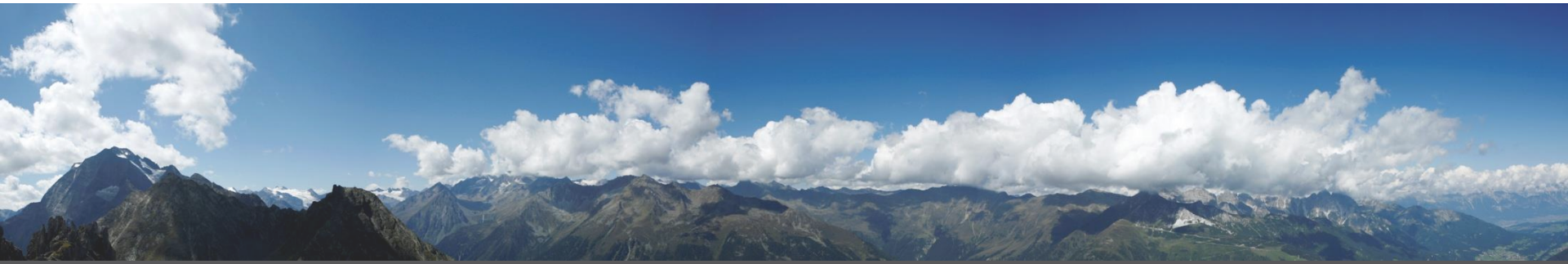
STI · INNSBRUCK

Steps towards a Data Value Chain

Dieter Fensel, University of Innsbruck

Salzburg, June 2013



Contents

STI · INNSBRUCK

1. Big Data
2. Public Open Data
3. Linked (Open) Data
4. Data Economy

STI · INNSBRUCK

BIG Data

What is Big Data?

STI · INNSBRUCK

- Every day, we create 2.5 quintillion* bytes of data — so much that 90% of the data in the world today has been created in the last two years alone.
- These data come from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.
- These data are **big data.****

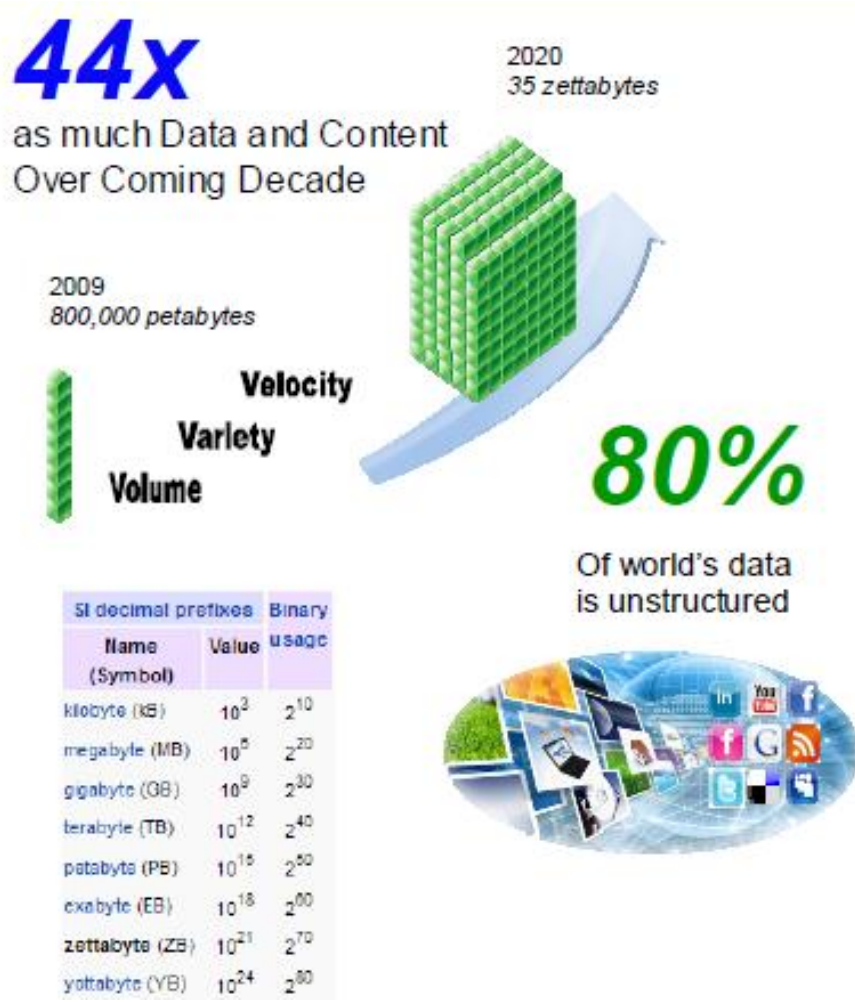
10^{30}

* 10
** http://www-01.ibm.com/software/data/bigdata

What is Big Data?

- **"Big data"** is a loosely-defined term
- used to describe data sets so large and complex that they become awkward to work with using on-hand database management tools.
  - White, Tom. Hadoop: The Definitive Guide. 2009. 1st Edition. O'Reilly Media. Pg 3.
  - MIKE2.0, Big Data Definition http://mike2.openmethodology.org/wiki/Big_Data_Definition

Infromation Explosion in data and real world events (IBM)



SI decimal prefixes		Binary usage
Name (Symbol)	Value	
kilobyte (kB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Big Data Application Areas

STI · INNSBRUCK

Big Data @ Work

Organizations in all industries are under increasing pressure to capitalize on data

Healthcare

The average amount of data per hospital will increase from 167TB to 665TB in 2015, driven by the enormous growth of medical images and electronic medical records.[1]

With Big Data

Medical professionals can improve patient care and reduce costs by extracting relevant clinical information from vast amounts of data to better understand the past and predict future outcomes.

Customer Service

Today, 86% of consumers quit doing business with a company because of a bad customer experience, up from 59% four years ago.[2]

With Big Data

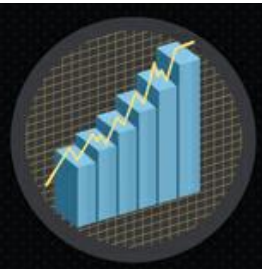
Service representatives can use data to gain a more holistic view of their customers, understanding their likes and dislikes in real-time in order to resolve a problem or capitalize on happy clients faster.

Insurance

Insurance companies and government agencies each gather fraud data related to their own individual missions. But the kind, quality and volume of data compiled varies widely.[3]

With Big Data

An insurance or citizen services provider can apply advanced analytics to data and detect fraud quickly, before funds are paid out.



Financial Services

Wall Street alone delivers 5 new research documents every minute. Dow Jones publishes upwards of 19,000 news items per day.[4]

With Big Data

Financial services professionals can better understand market changes through improved business insight from data, helping to anticipate performance gaps and more accurately assess investment alternatives.

Retail

$165 billion in total sales are missed each year because retailers don't have the right products in stock to meet customer demand.[5]

With Big Data

Retailers can better understand their customers by analyzing sales trends and incorporating more accurate forecasting, ultimately increasing customer loyalty and revenue.

Communications

5 billion global subscribers in the telco industry are demanding unique and personalized offerings that match their individual lifestyles.[6]

With Big Data

Communications providers can use data to create a more personalized customer experience and avoid losing customers to competitors.

*Information gathered by IBM

1 North American Health Care Provider Information Market Size & Forecast, Enterprise Strategy Group
2 Customer Experience Impact Report, Harris Interactive
3 Coalition Against Insurance Fraud
4 The Future of News in Trading, Dow Jones, June 2011
5 IHL Group, Dec 2011
6 IBM Global CFO Study 2010

Picture taken from http://www-01.ibm.com/software/data/bigdata/industry.html

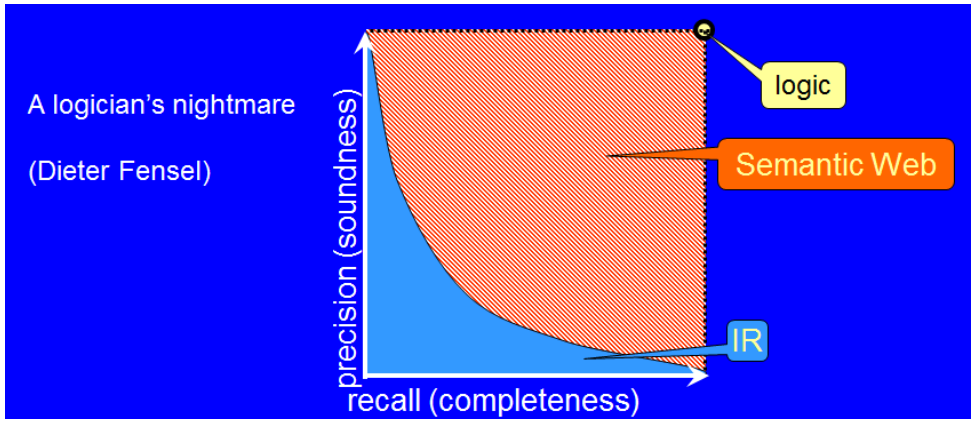
Use case : Climate Research

- Eiscat and Eiscat 3D are multimillion reserch projects doing environmental research as well as evaluation of the built infrastructures.
  - Observation of climate: sun, troposphere, etc.
  - Simulations, e.g. Creation of artificial Nothern light
  - Run by European Incoherent Scatter Association

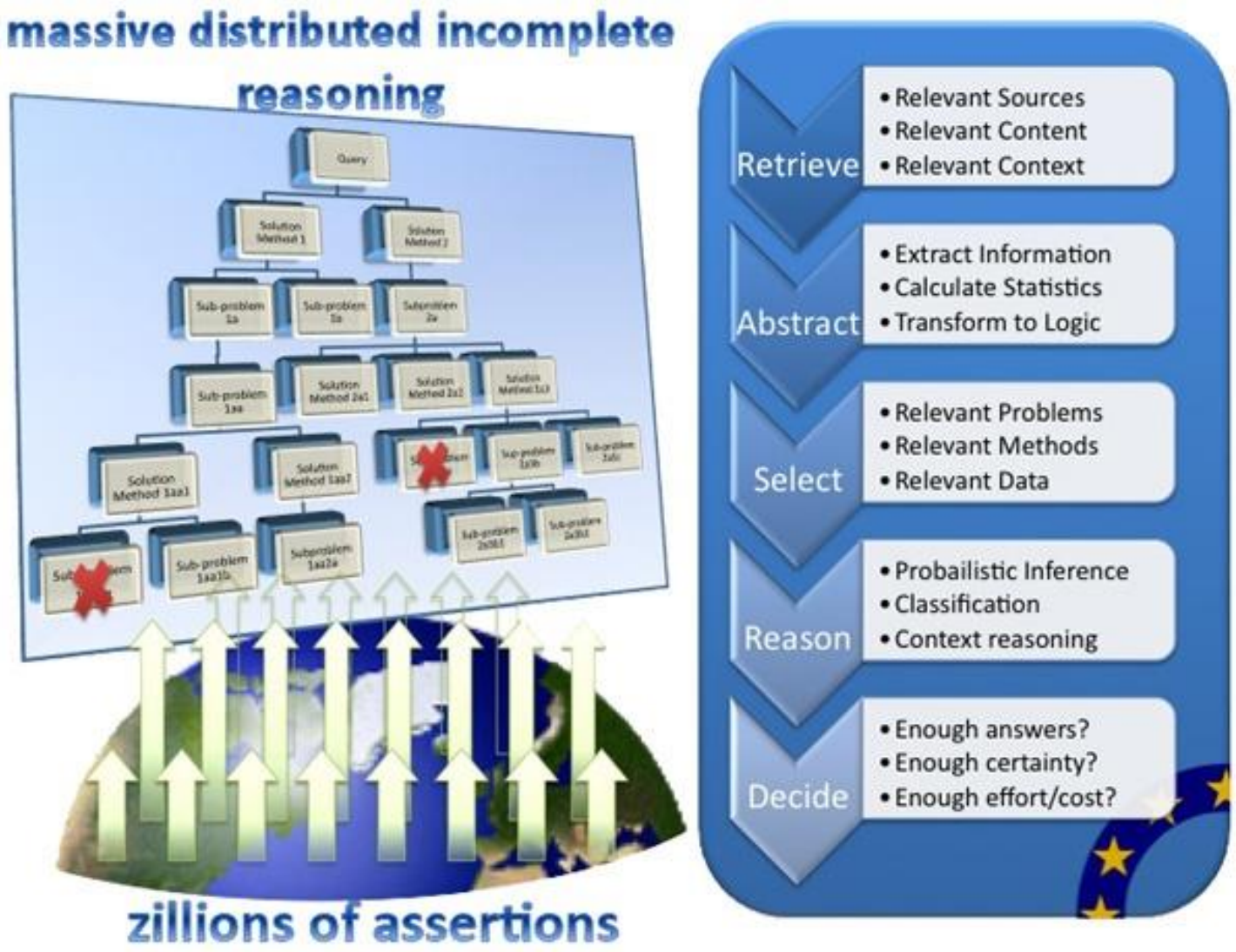
- 1,5 Petabytes of data are generated daily (1,5 Million Gigabytes).
  - Processing of this data would require 1K petaFLOPS performance
  - Or 1 billion Euro electricity costs p.a.

Large Scale Reasoning

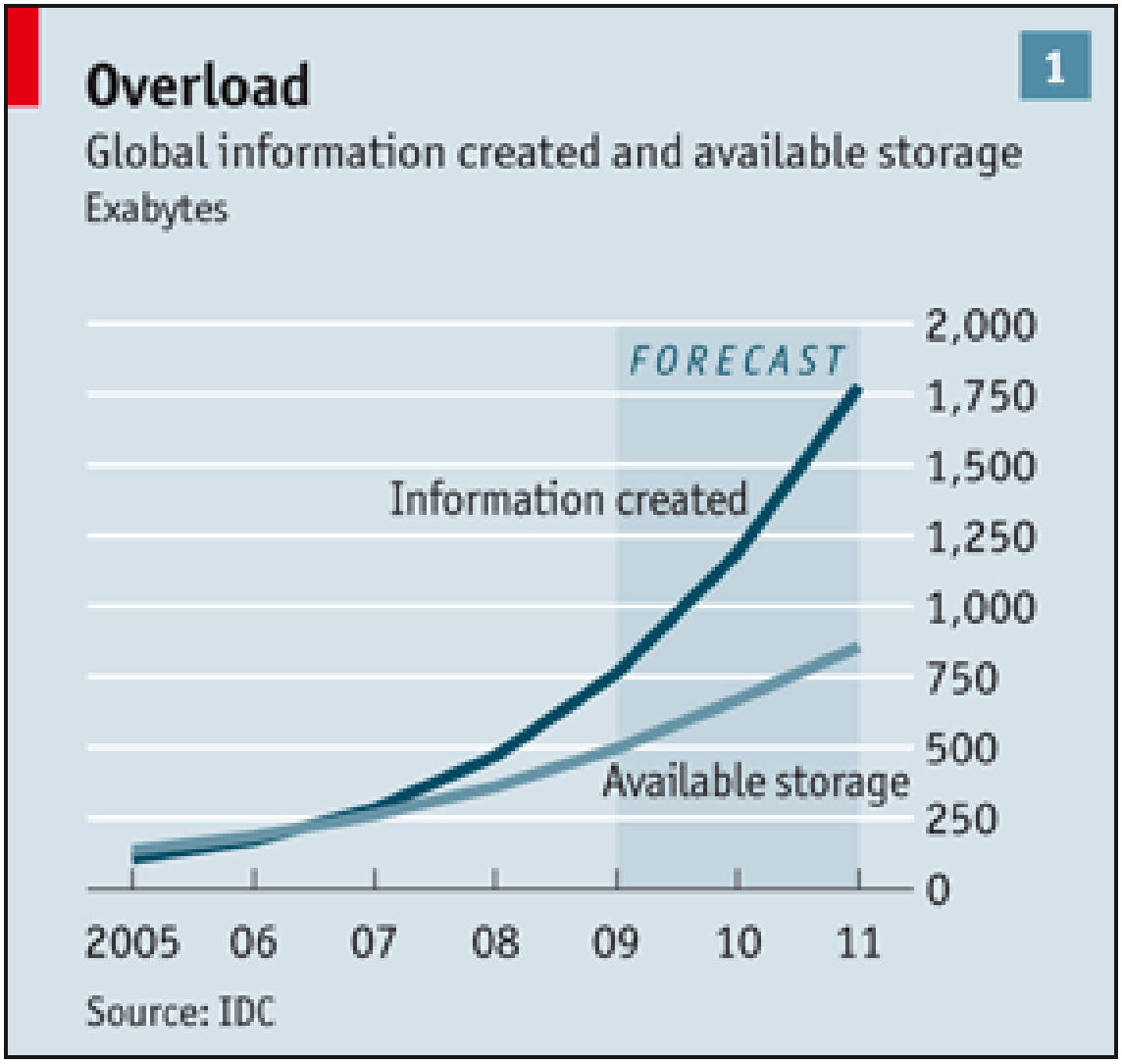
- Performing deductive inference with a given set of axioms at the Web scale is practically impossible
  - Too manyRDF triples to process
  - Too much processing power is needed
  - Too much time is needed
- LarKC aimed at contributing to an ‘infinitely scalable’ Semantic Web reasoning platform by
  - Giving up on 100% correctness and completeness (trading quality for size)
  - Include heuristic search and logic reasoning into a new process
  - Massive parallelization (cluster computing)



Large Scale Reasoning



Volumes of Data Exceed the Availale Storage Volume Globally

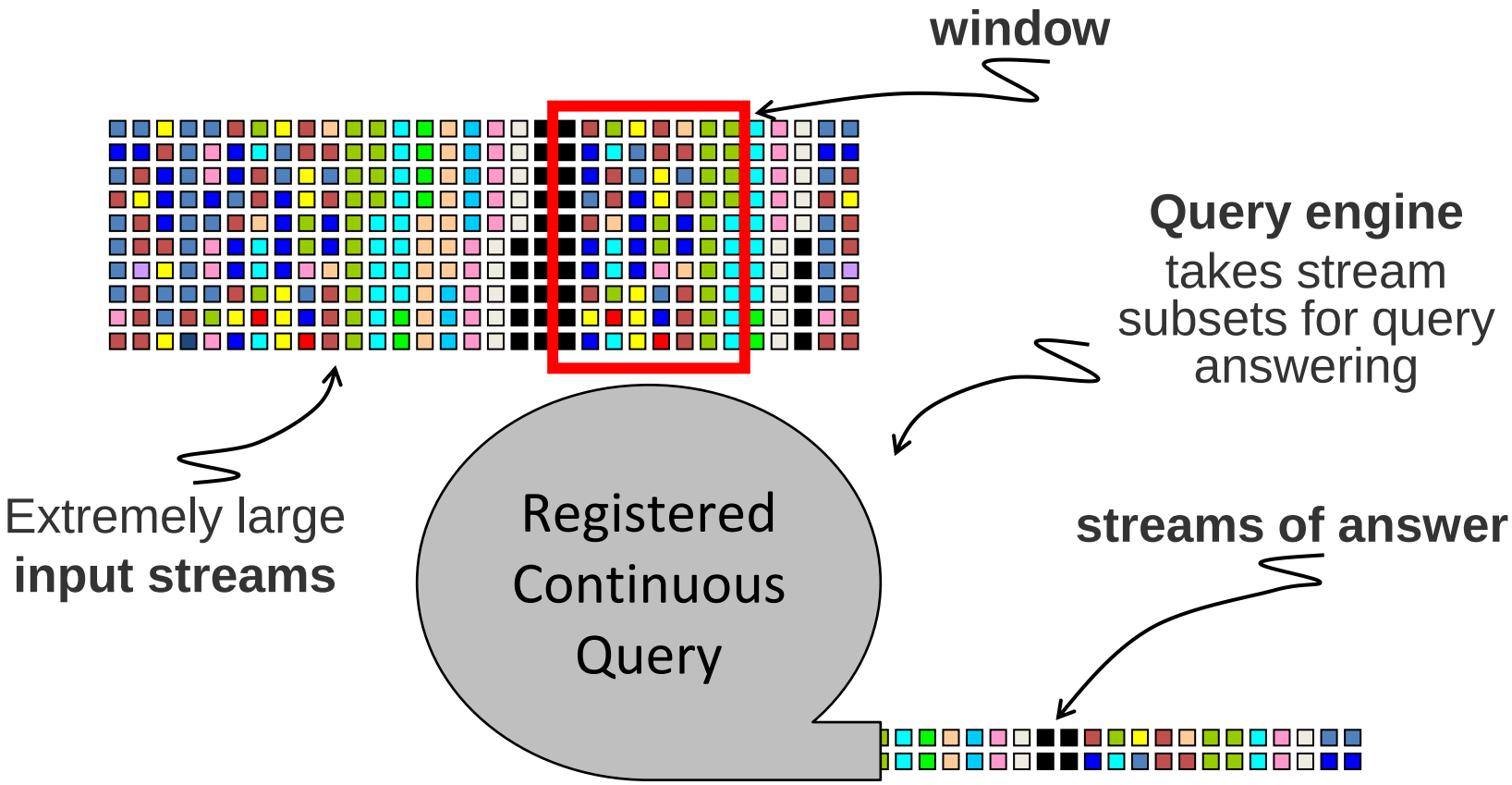


There is a need to throw the data away due to the limited storage space.

- Before throwing the data away some processing can be done at run-time
  - Processing streams of data as they happen
- Embracing the streaming model
  - Data is seen as a constant flow (sequence) of transient elements
  - Fits naturally with many application domains (sensors, social media, etc.)
- “Big data” is bringing an inherent set of complexities
  - Data structures exceeding the available memory
  - Approximate/incomplete results are taken as granted
    - Always look at the latest part of a dataset

- Logical reasoning in real time on multiple, heterogeneous, gigantic and inevitably noisy data streams in order to support the decision process…

-- S. Ceri, E. Della Valle, F. van Harmelen and H. Stuckenschmidt, 2010



Picture taken from Emanuele Della Valle "Challenges, Approaches, and Solutions in Stream Reasoning", Semantic Days 2012

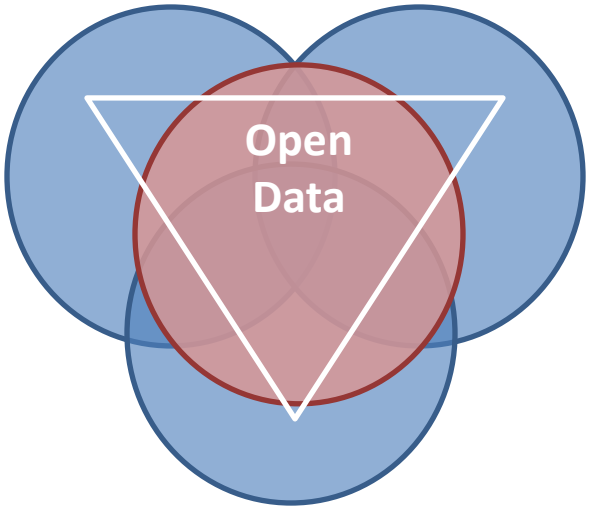
Conclusions

STI · INNSBRUCK

- Big Data describes datasets so large and complex that they become awkward to work with using on-hand database management tools.
- Big data application domains are diverse
  - Embracing a big data processing strategy can have a significant impact
- Tacking the issues of big data processing requires to loose the requirements on completeness and precision
- Big data on Web scale suffers from an inherent heterogeneity and different levels of expressiveness
- **Complexity is more than just size!**

STI · INNSBRUCK

Public Open Data



Public Open Data: What is Open Data?

STI · INNSBRUCK

Definitions:

- Open data is non-personally identifiable data produced in the course of an organisation's ordinary business, which has been released under an unrestricted licence (like the Open Government Licence).

- Open public data is underpinned by the philosophy that data generated or collected by organisations in the public sector should belong to the taxpayers, wherever financially feasible and where releasing it won't violate any laws or rights to privacy (either for citizens or government staff).

[linkedgov project]
http://linkedgov.org

Public Open Data: What is Open Data?

STI · INNSBRUCK

Definitions:

The idea behind open data is that information held by government should be freely available to use and re-mix by the public. It's a movement to make non-personal data:

- open so that it can be turned into useful applications
- support transparency and accountability
- make sharing data between public sector partners more efficient.

The Government is committed to making much more public data openly available. On 22 March 2010, the Prime Minister announced that the Government was going to:

"...use digital technology to open up data with the aim of providing every citizen in Britain with true ownership and accountability over the services they demand from government."

http://www.idea.gov.uk/

Public Open Data: Features of Open Data

STI · INNSBRUCK

Open Data principles [1]:

1. **completeness** – all data that can be open (w.r.t. privacy and security) should be open

2. **primary source** – all open data should be gathered at their source in raw format

3. **temporal closeness** – all open data should be up-to-date

4. **easy access** – all open data should be easily accessible

5. **machine readability** – all open data should be structured for machine processing

[1] Source [Kaltenböck M., Thurner T., (Hg.): Open Government Data Weißbuch, 2011]

Public Open Data: Features of Open Data

STI · INNSBRUCK

Open Data principles [1]:

6. **non-discriminating** – all open data should be accessible for everyone

7. **open standards** – all open data should use open standards

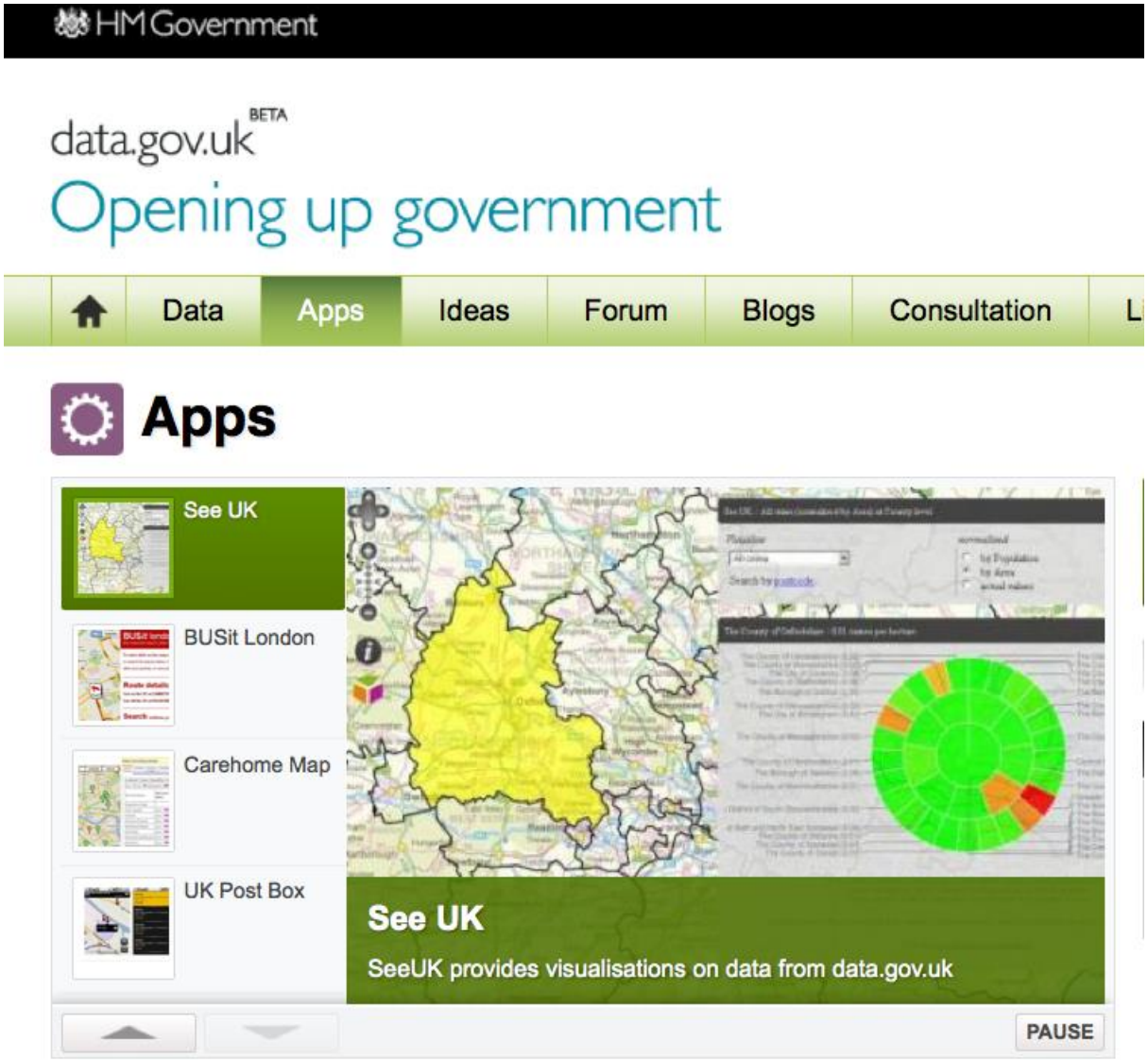
8. **liberal licensing** – all open data should use a liberal licensing without huge obligations for potential users

9. **durability** – all open data should be available on a long term basis

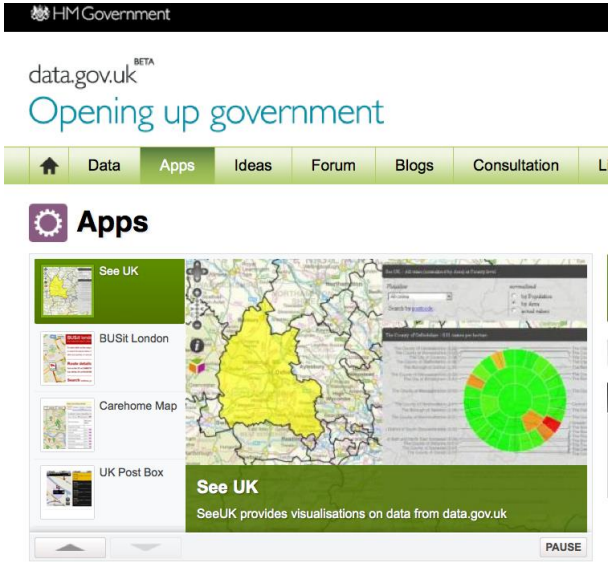
10. **non-discriminating usage costs** – some open data might involve usage costs. These should be kept as low as possible.

[1] Source [Kaltenböck M., Thurner T., (Hg.): Open Government Data Weißbuch, 2011]

Public Open Data: And it works (SeeUK)



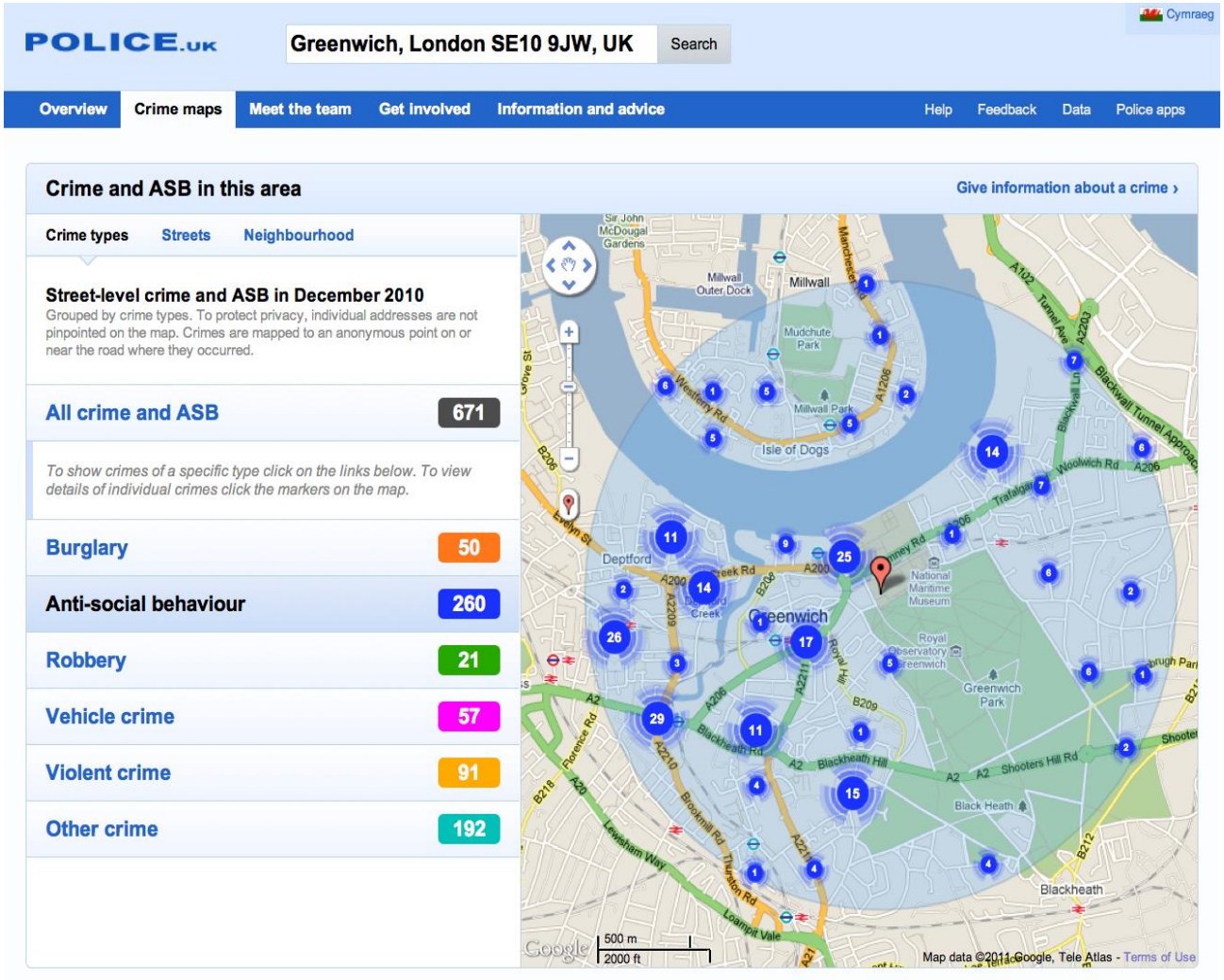
Public Open Data: And it works (SeeUK)



- See UK uses data that have been sourced from data.gov.uk and processed into Linked Data .
- All the datasets are enriched and cross-linked to additional sources.
- The visualisation provides a view centred on a chosen region of the specified size, and most noticeably gives a "pie-chart" that shows the viewer how that region compares with similar regions around it.

Public Open Data: And it works (police.uk)

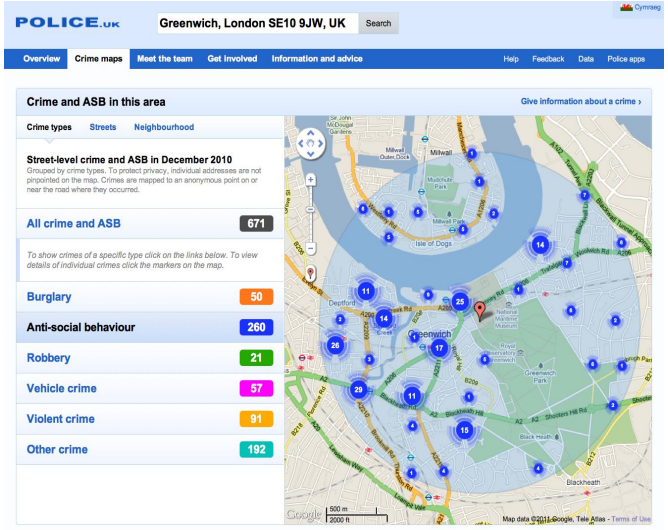
STI · INNSBRUCK



Public Open Data: And it works (police.uk)

- Different apps such as „Vehicle Crime & Road Accident Map“, „Crime Sounds“ and „UK Crimeview“ are provided

- The user can get a quick idea about different areas of cities and towns and their crime statistics.



Public Open Data

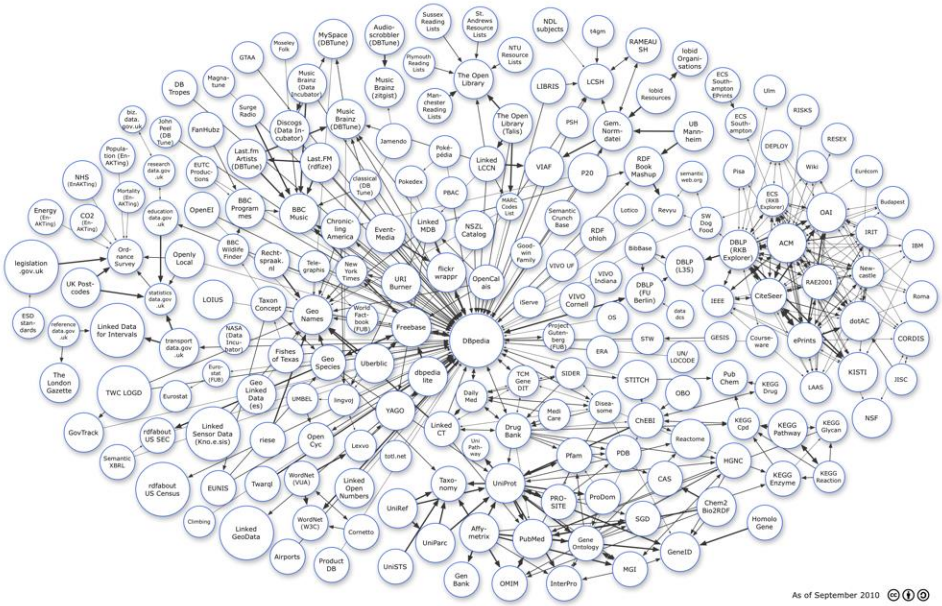
STI · INNSBRUCK

- **Openess**:
  Open Data is about changing behaviour
- **Heterogenity**:
  Different vocabularies are used
- **Interlinkage**:
  Need to link these data sets to prevent data silos

- → **Linked** Open Data

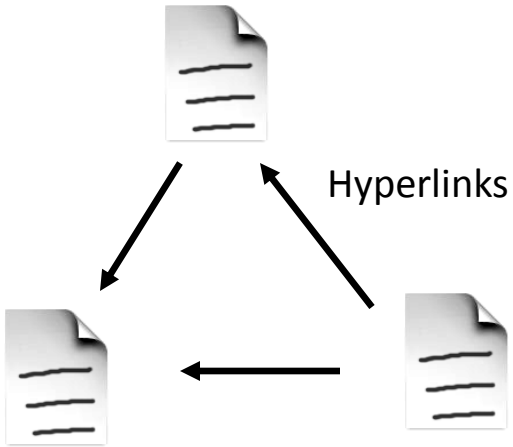
STI · INNSBRUCK

Linked Open Data



Motivation: From a Web of Documents to a Web of Data

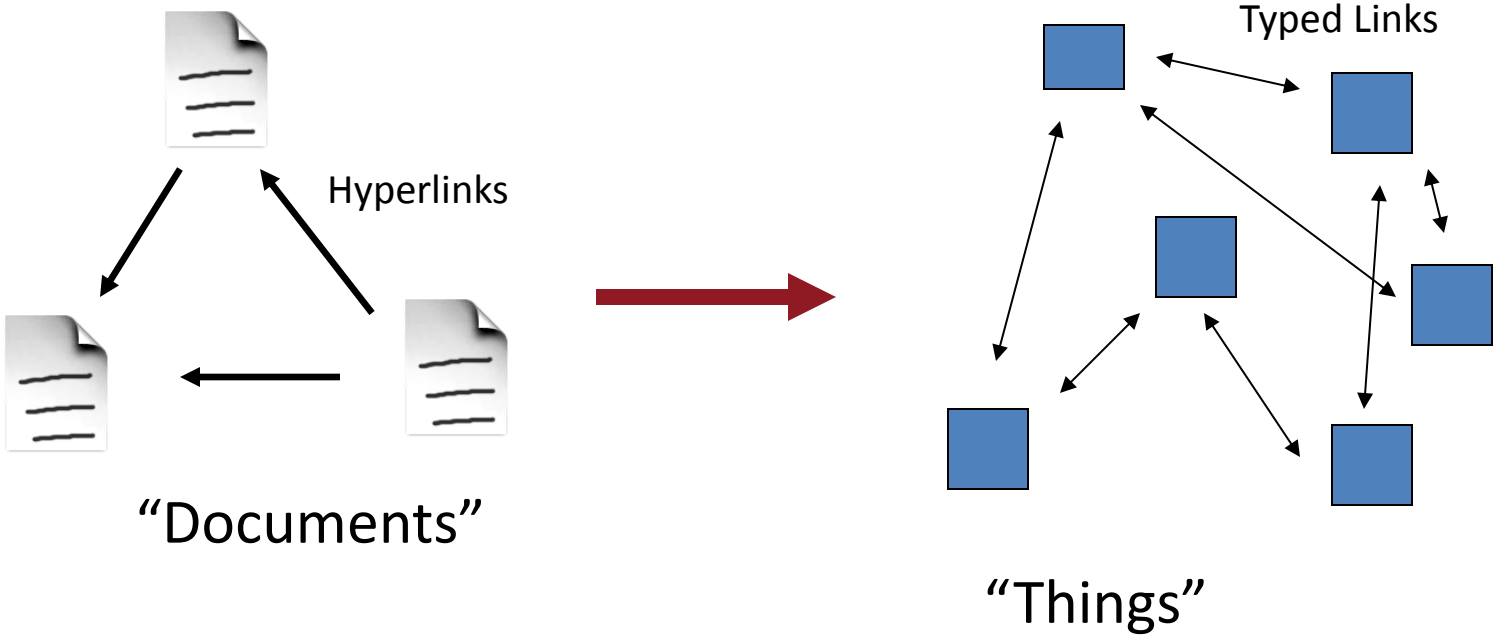
- Web of Documents



- Fundamental elements:
  1. Names (URIs)
  2. Documents (Resources) described by HTML, XML, etc.
  3. Interactions via HTTP
  4. (Hyper)Links between documents or anchors in these documents
- Shortcomings:
  - Untyped links
  - Web search engines fail on complex queries

Motivation: From a Web of Documents to a Web of Data

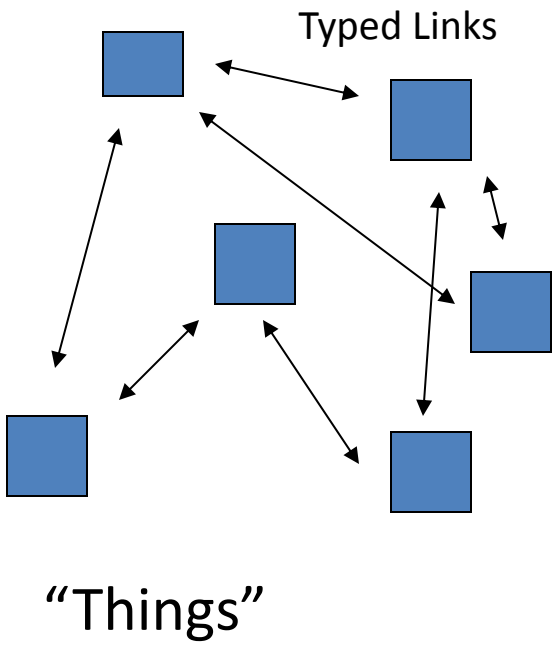
- Web of Documents
- Web of Data



Motivation: From a Web of Documents to a Web of Data

- Characteristics:
  - Links between arbitrary things (e.g., persons, locations, events, buildings)
  - Structure of data on Web pages is made explicit
  - Things described on Web pages are named and get URIs
  - Links between things are made explicit and are typed

- Web of Data



Google Knowledge Graph

- *"A huge knowledge graph of interconnected entities and their attributes"*.

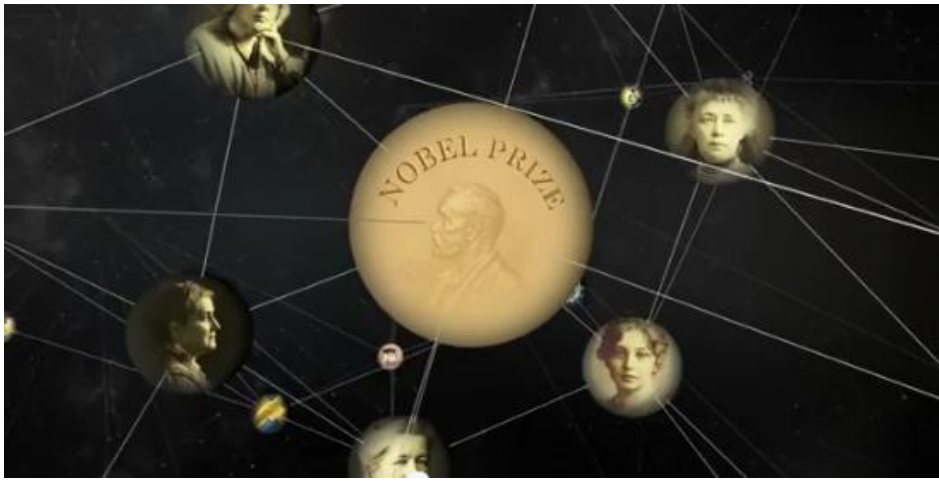
  Amit Singhal, Senior Vice President at Google

- *"A knowledge based used by Google to enhance its search engine's results with semantic-search information gathered from a wide variety of sources"*

  http://en.wikipedia.org/wiki/Knowledge_Graph

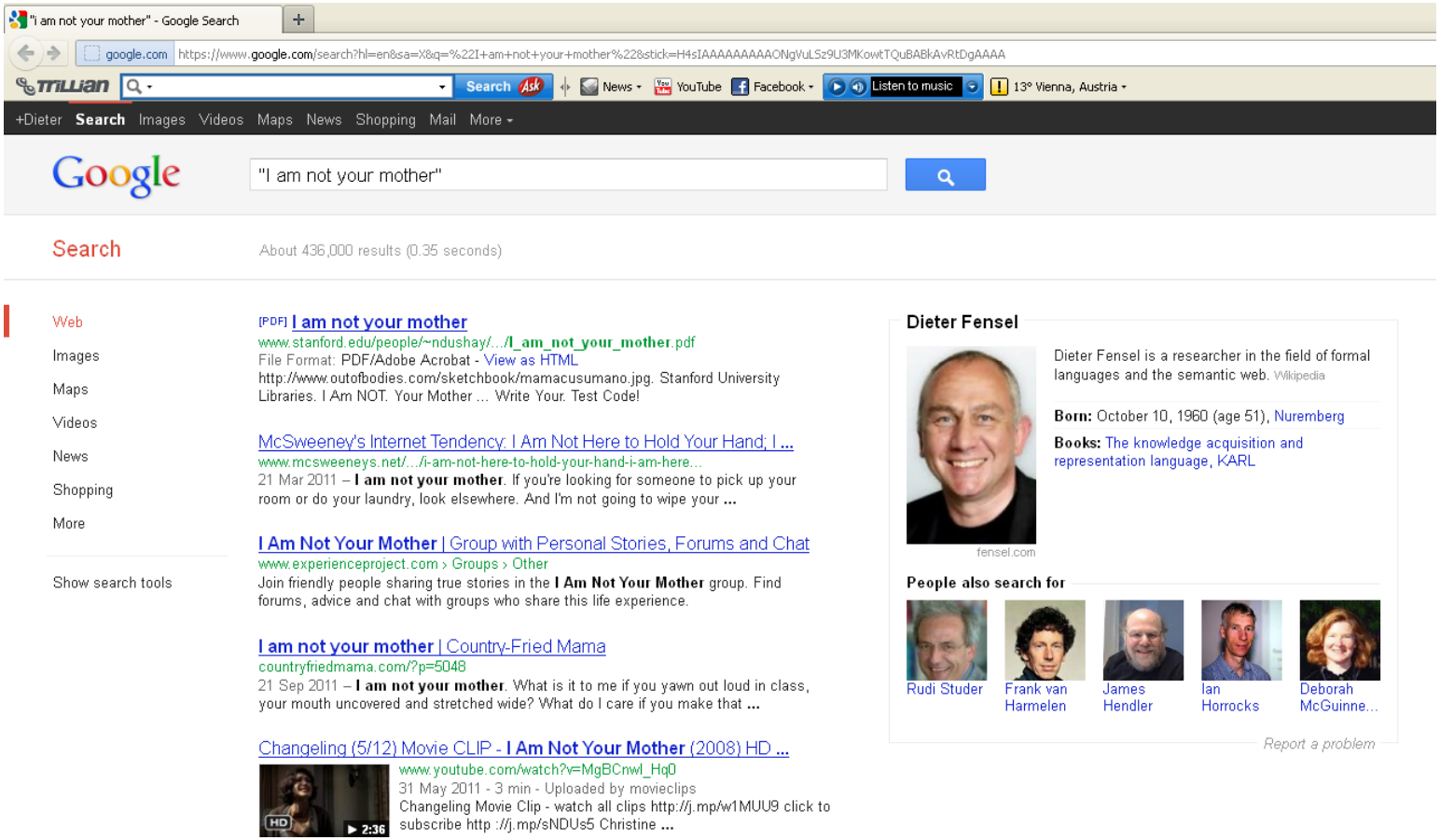
- Based on information derived from many sources including *Freebase*, *CIA World Factbook*, *Wikipedia*

- Contains about 3.5 billion facts about 500 million objects



Google Knowledge Graph

https://www.google.com/search?hl=en&sa=X&q=%22I+am+not+your+mother%22&stick=H4sIAAAAAAAAAONgVuLSz9U3MKowtTQuBABkAvRtDgAAAA



Linked Data – a definition and principles

- **Linked Data** is about the use of Semantic Web technologies to publish structured data on the Web and set links between data sources.

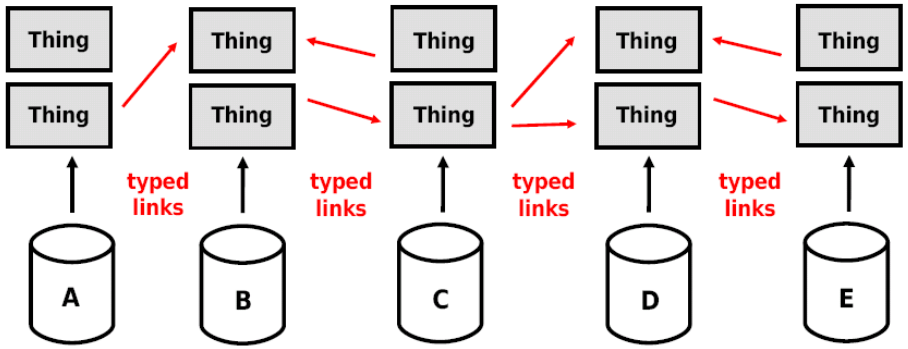


Figure from C. Bizer

LOD Cloud May 2007

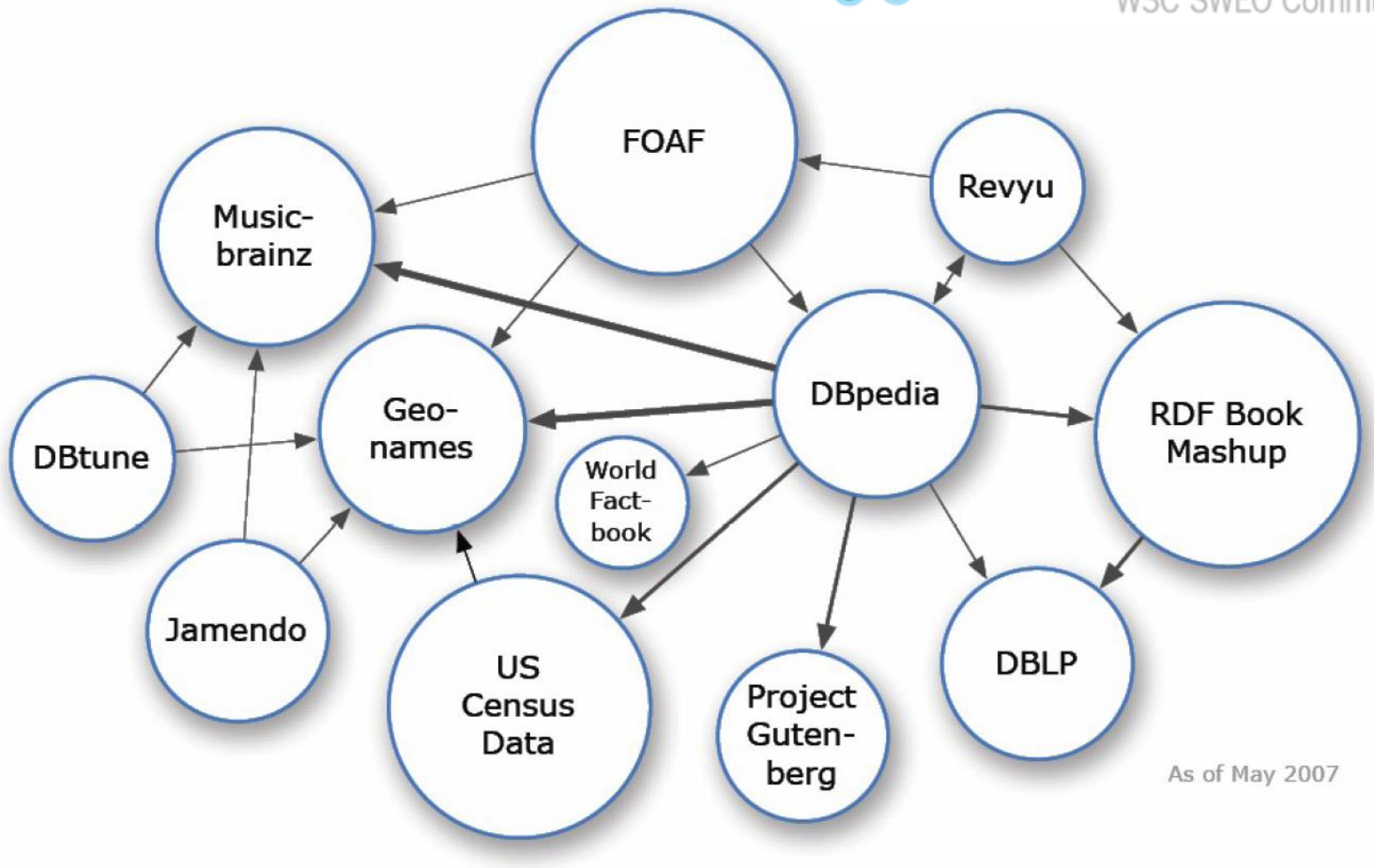


Figure from http://linkeddata.org/

LOD Cloud May 2007

STI · INNSBRUCK

Basics:

The Linked Open Data cloud is an interconnected set of datasets all of which were published and interlinked following the Linked Data principles.

Facts:

- Focal points:
  - DBPedia: RDFized vesion of Wikipiedia; many ingoing and outgoing links
  - Music-related datasets
- Big datasets include FOAF, US Census data
- Size approx. 1 billion triples, 250k links

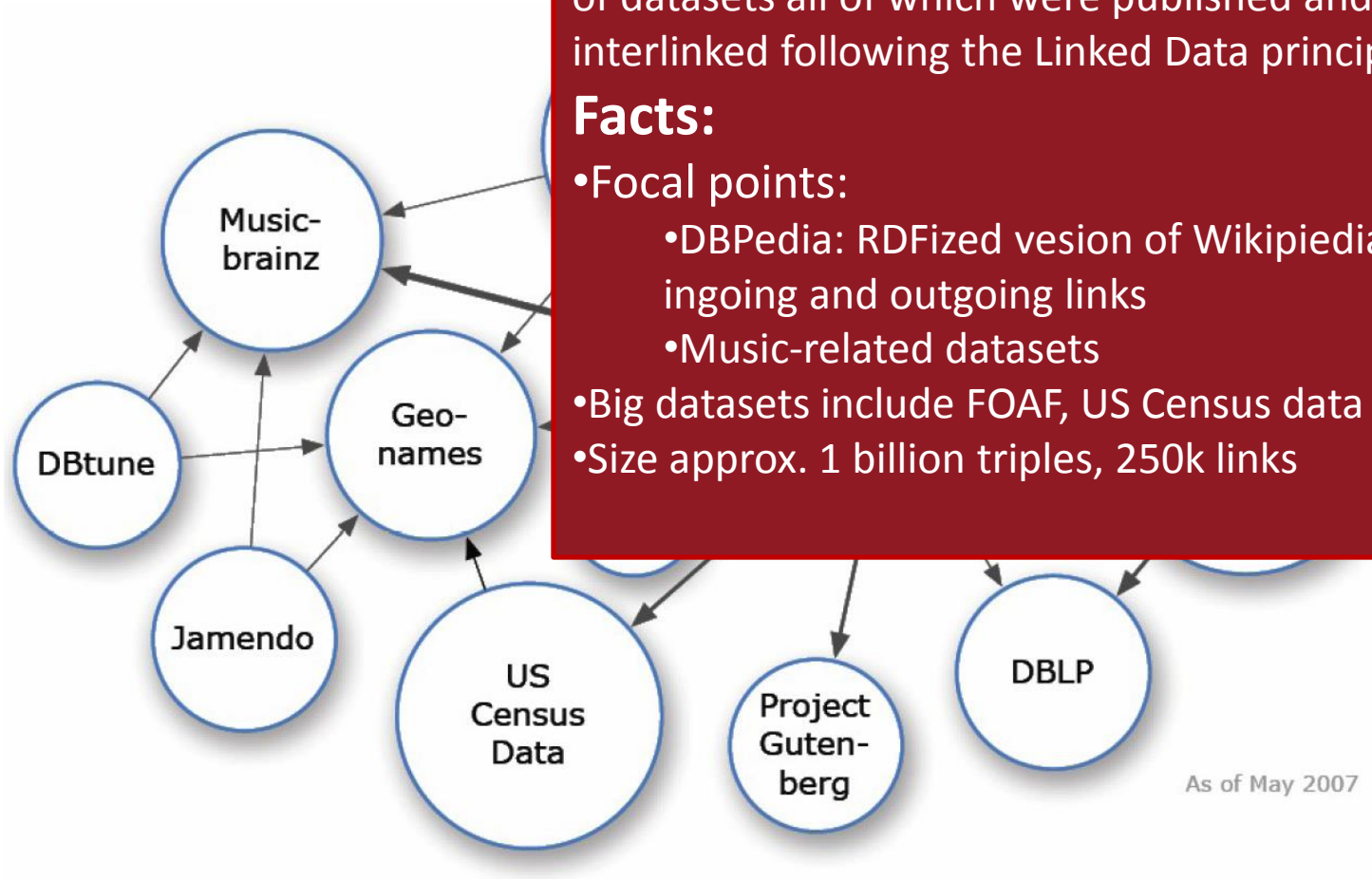


Figure from http://linkeddata.org/

LOD Cloud March 2009

STI · INNSBRUCK

LINKINGOPENDATA
W3C SWEO Community Project

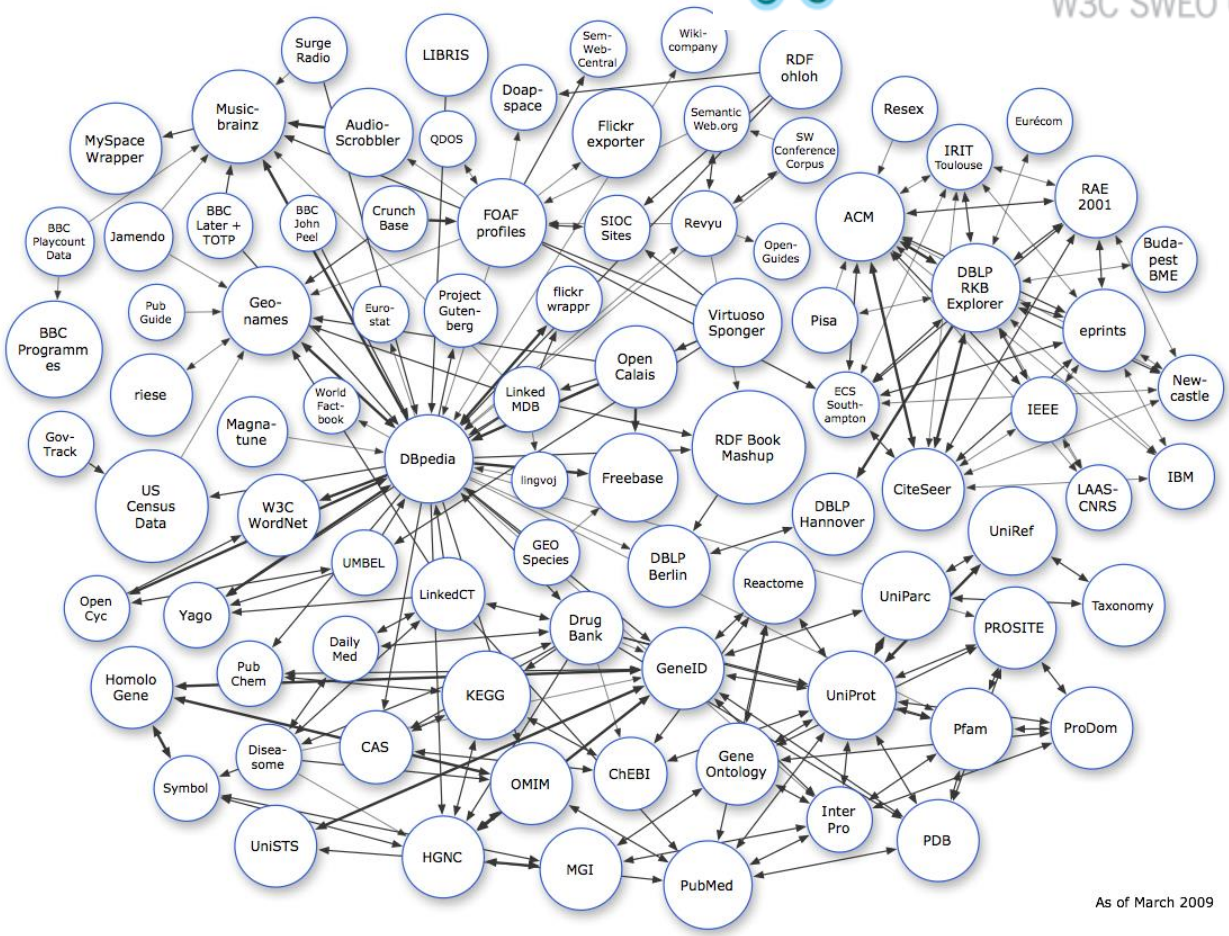


Figure from http://linkeddata.org/

LOD Cloud September 2011

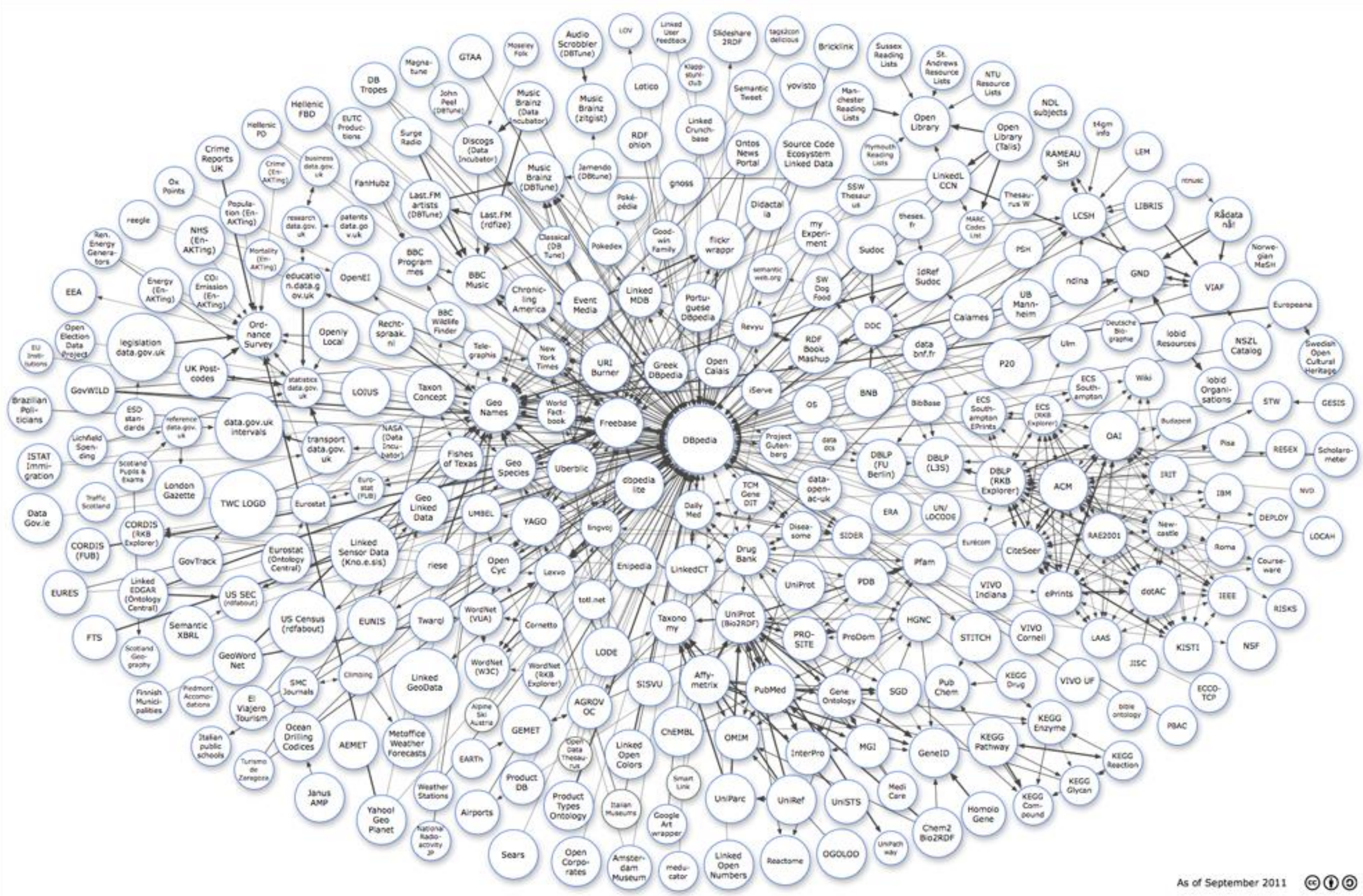
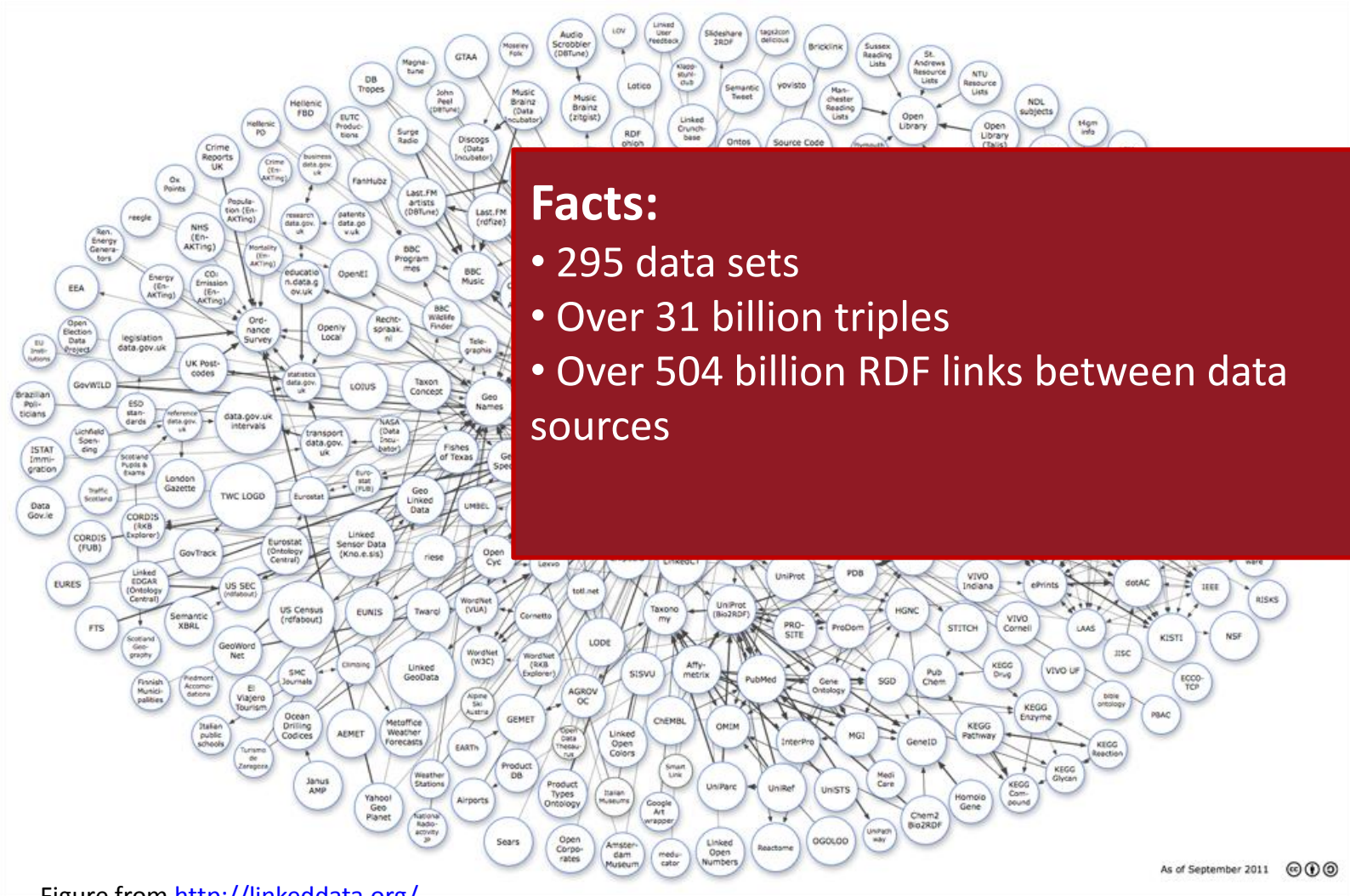


Figure from http://linkeddata.org/

LOD Cloud September 2011

STI · INNSBRUCK



Facts:

- 295 data sets
- Over 31 billion triples
- Over 504 billion RDF links between data sources

Figure from http://linkeddata.org/

Linked Open Data – silver bullet for data integration

STI · INNSBRUCK

- Linked Open Data can be seen as a global data integration platform
  - Heterogeneous data items from different data sets are linked to each other following the Linked Data principles
  - Widely deployed vocabularies (e.g. FOAF) provide the predicates to specify links between data items

- Data integration with LOD requires:
  1. Access to Linked Data
     - HTTP, SPARQL endpoints, RDF dumps
     - Crawling and caching
  2. Normalize vocabularies – data sets that overlap in content use different vocabularies
     - Use schema mapping techniques based on rules (e.g. RIF, SWRL) or query languages (e.g. SPARQL Construct, etc.)
  3. Resolve identifies – data sets that overlap in content use different URIs for the same real world entities
     - Use manual merging or approaches such as SILK (part of Linked Data Integration Framework) or LIMES
  4. Filter data
     - Use SIVE ((part of Linked Data Integration Framework)

See: http://www4.wiwiss.fu-berlin.de/bizer/ldif/

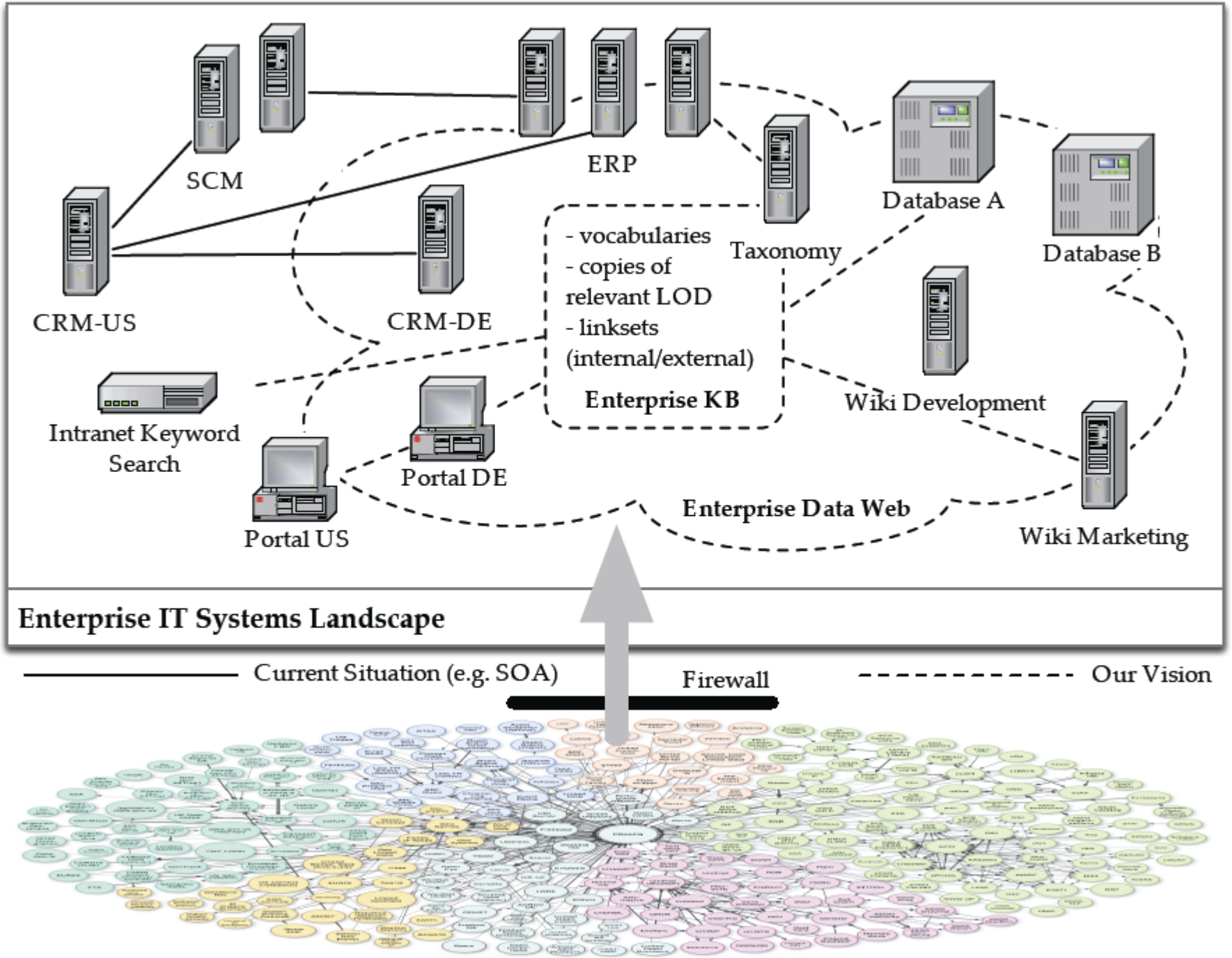
Example - Mashup: DBPedia Mobile

- Geospatial entry point into the Web of Data.
- It exploits information coming from DBpedia, Revyu and Flickr data.
- It provides a way to explore maps of cities and gives pointers to more information which can be explored

Pictures from DBPedia Mobile

Try yourself: http://wiki.dbpedia.org/DBpediaMobile

From Intranet to Enterprise Data Web around a knowledge hub



STI · INNSBRUCK

Data Economy

"Your data is worth more if you give it away."

Commission Vice President

Neelie Kroes

What is Data Economy?

STI · INNSBRUCK

- Non tangible assets (i.e. data) play a significant role in the creation of economic value
- Data is nowadays more important than, for example, search or advertisement
- The value of the data, its potential to be used to create new products and services, is more important than the data itself

Why a Data Economy?

STI · INNSBRUCK

- Total market for public sector information **€ 28 billion in 2008** for the EU27
- Annual growth of 7% leading to around **€ 32 billion in 2010**
- Estimated **€40 billion annual** boost for the European economy.
- The total direct and indirect economic gains across the whole EU27 economy would be in the order of **€ 140 billion annually**.

See: Review of recent studies on PSI re-use and related market developments, G. Vickery, August 2011.

Why a Data Economy?

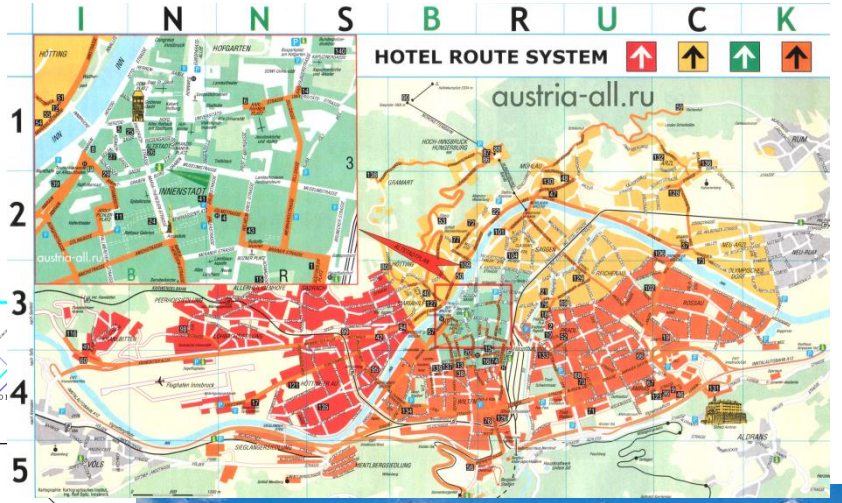
- **New businesses can be built on the back of these data**: Data are an essential raw material for a wide range of new information products and services which build on new possibilities to analyse and visualise data from different sources. Facilitating re-use of these raw data will create jobs and thus stimulate growth.
- **More Transparency**: Open data is a powerful tool to increase the transparency of public administration, improving the visibility of previously inaccessible information, informing citizens and business about policies, public spending and outcomes.
- **Evidence-based policy making and administrative efficiency**: The availability of solid EU-wide public data will lead to better evidence-based policy making at all levels of government, resulting in better public services and more efficient public spending.

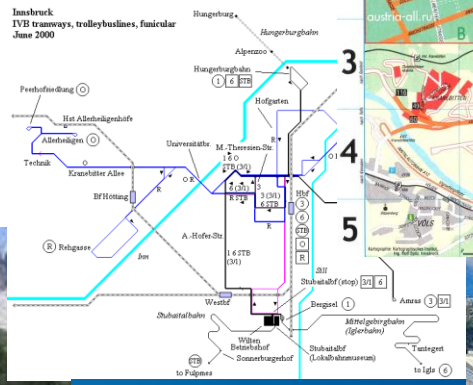
See: http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO11/891&format=HTML&aged=0&language=EN&guiLanguage=en

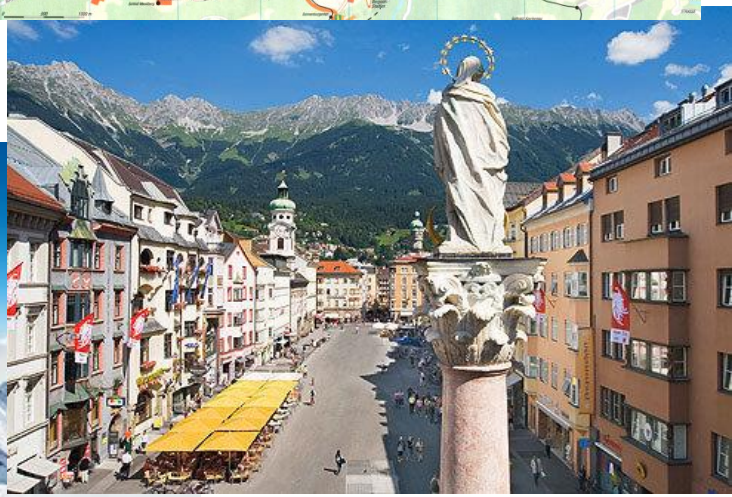
Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

- Use LOD to integrate and lookup data about
  - places and routes
  - time-tables for public transport
  - hiking trails
  - ski slopes
  - points-of-interest



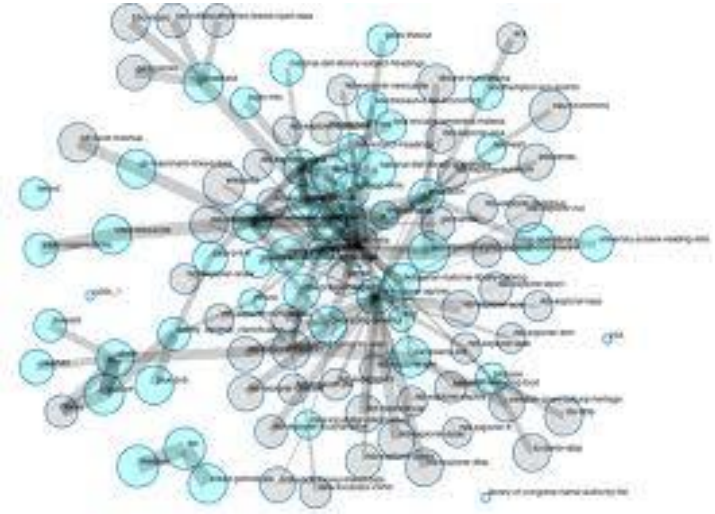




Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

LOD data sets



- Open Streetmap
- Google Places
- Databases of government
  - TIRIS
  - DVT
- Tourism & Ticketing association
- IVB (busses and trams)
- OEBB (trains)
- Ärztekammer
- Supermarket chains: listing of products
- Hofer and similar: weekly offers
- ASFINAG: Traffic/Congestion data
- Herold (yellow pages)
- City archive
- Museums/Zoo
- News sources like TT (Tyrol's major daily newspaper)
- Statistik Austria
- Innsbruck Airport (travel times, airline schedules)
- ZAMG (Weather)
- University of Innsbruck (Curricula, student statistics, study possibilities)
- IKB (electricity, water consumption)
- Entertainment facilities (Stadtcafe, Cinema...)
- Special offers (Groupon)

Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

- Data and services from destination sites integrated for recommendation and booking of
  - Hotels
  - Restaurants
  - Cultural and entertainment events
  - Sightseeing
  - Shops

Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

- **Web scraping** integration
  - Create wrappers for current web sites and extract data automatically
  - Many Web scraping tools available on the market

Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

- Integration into a comprehensive map of ***multi-channel communication***, ***seekda booking engine, Linked Open Data*** and ***on the fly service integration as you pay*** to generate added value for businesses as well as customers

- Combination of multi channel communication and yield management
  - Semantic Communication Engine Innsbruck (SCEI)
  - seekda booking solutions
- enriched with Linked (Open) Data
  - Machine understandable interlinked data
  - Bike and hiking trails, sight information, etc.
- and on the fly service integration as you pay
  - Solutions for ad-hoc service integration for touristic destination sites
  - Bike rental, ski passes, etc.
  - Services are quickly integrated through scraping (and later through legal frameworks and backend integration in case of business volumes)

Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

- Based on **Open Street Map**

Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

- Based on **Open Street Map**
- Increase on-line visibility for hotels and destinations via **multi-channel communication – SCEI**

Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

- Based on **Open Street Map**
- Increase on-line visibility for hotels and destinations via **multi-channel communication – SCEI**
- Hotels, ski passes, etc. are directly bookable – **seekda engine**

Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

- Based on **Open Street Map**
- Increase on-line visibility for hotels and destinations via **multi-channel communication – SCEI**
- Hotels, ski passes, etc. are directly bookable – **seekda engine**
- **LOD** to integrate and lookup data about hiking trails, ski slopes, etc.

Combining Open Data and Services – Tourist Map Austria

STI · INNSBRUCK

- Based on **Open Street Map**
- Increase on-line visibility for hotels and destinations via **multi-channel communication – SCEI**
- Hotels, ski passes, etc. are directly bookable – **seekda engine**
- **LOD** to integrate and lookup data about hiking trails, ski slopes, etc.
- On the fly service integration as you pay

“There's No Money in Linked (Open) Data”

http://knoesis.wright.edu/faculty/pascal/pub/nomoneylod.pdf

- It turns out that using LOD datasets in realistic settings is not always easy.
  - Surprisingly, in many cases the underlying issues are not technical but legal barriers erected by the LD data publishers.
  - Generally, mostly non-technical but socio-economical barriers hamper the reuse of date (do patents and IPR protections hamper or facilitate knowledge reuse?).
  - Business intelligence
  - Dynamic Data
  - On the fly generation of data