

STI International Off-Site  
Costa Adeje, Tenerife, May 30th, 2008

## Quality-Driven Information Filtering in the Context of Web-Based Information Systems

Chris Bizer, Freie Universität Berlin



## Hello

- **Chris Bizer**
- **Junior-Professor at Freie Universität Berlin**
- **Projects:**
  - **RAP - RDF API for PHP (together with Universität Leipzig)**
  - **D2RQ und D2R Server (together with HP Labs)**
  - **Named Graphs and NG4J (together with HP Labs)**
  - **Fresnel Display Vocabulary (together with MIT and INRIA)**
  - **DBpedia (together with Universität Leipzig and OpenLink)**
  - **Linking Open Data (community project sponsored by W3C)**

## Overview

1. **Information Quality and the Web**
  - Information Quality
  - Information Quality Assessment
2. **Representing Quality-Related Meta-Information**
  - The Named Graphs Data Model
3. **Expressing Information Filtering Policies**
  - The WIQA-PL Policy Language
4. **Use Cases**
  - What might this be good for?

Christian Bizer: Information Quality Assessment (5/29/2008)

## Problem Statement

Information providers have

- different levels of knowledge
- different views of the world
- different intentions

Therefore, provided information will be

- wrong
- biased
- outdated
- Incomplete
- inconsistent

Christian Bizer: Information Quality Assessment (5/29/2008)

## Information Quality Assessment in the Offline World

In our everyday life, we might

- accept information from a friend on restaurants, but distrust him on computers.
- regard scientific papers only as relevant, if they have been published within specific journals.
- or believe foreign news only when they are reported by several independent sources.

Christian Bizer: Information Quality Assessment (5/29/2008)

## Goal

**Empower the users of Web-based systems to employ a similar wide range of information quality assessment policies as they are using in the offline world.**

Christian Bizer: Information Quality Assessment (5/29/2008)

## Information Quality

### “fitness for use”

1. task-dependent
2. subjective

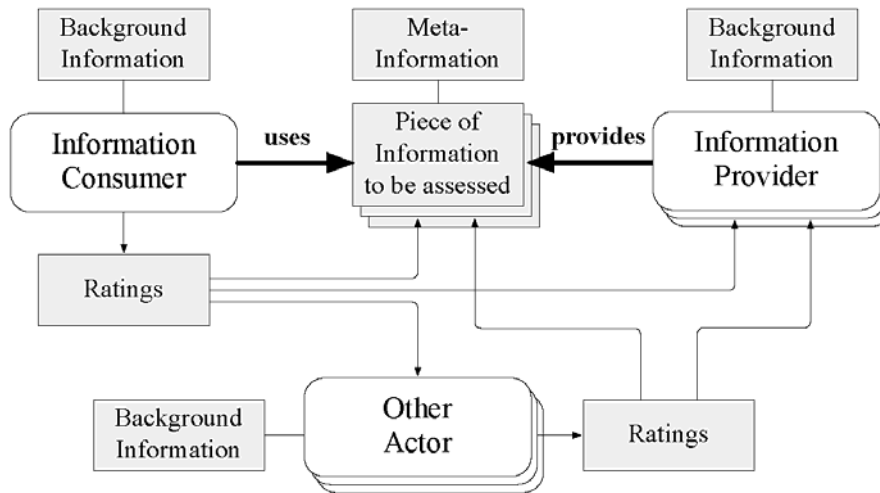
Christian Bizer: Information Quality Assessment (5/29/2008)

## Information Quality Dimensions

<i>Category</i>	<i>Dimension</i>	<i>Wang</i> [WS96]	<i>Redman</i> [Red96]	<i>Jarke</i> [JV97]
Intrinsic Dimensions	Accuracy	✓	✓	✓
	Consistency			✓
	Objectivity	✓		
	Timeliness	✓	✓	✓
Contextual Dimensions	Believability	✓		✓
	Completeness	✓	✓	✓
	Understandability	✓		
	Relevancy	✓	✓	✓
	Reputation	✓		
	Verifiability		✓	
Representational Dimensions	Amount of Data	✓	✓	
	Interpretability	✓	✓	✓
	Rep. Conciseness	✓	✓	
	Rep. Consistency	✓	✓	✓
Accessibility Dimensions	Availability	✓	✓	✓
	Response Time			
	Security	✓		

Christian Bizer: Information Quality Assessment (5/29/2008)

## Quality Indicators



Christian Bizer: Information Quality Assessment (5/29/2008)

## Information Quality Assessment Metrics

### ■ Content-Based Metrics

- use information to be assessed itself as quality indicator.
- Examples: statistical outlier detection methods, text analysis methods, domain specific rules.

### ■ Context-Based Metrics

- employ meta-information about the information content and the circumstances in which information was created as quality indicator.
- Example: "Disbelieve everything a vendor says about its competitor."

### ■ Rating-Based Metrics

- rely on explicit ratings about information itself, information sources, or information providers.
- Various authors have proposed different scoring functions.

Christian Bizer: Information Quality Assessment (5/29/2008)

## Factors restricting the Choice of Assessment Metrics

1. Availability of Quality Indicators
2. Quality of Quality Indicators
3. Understandability
4. Subjective Preferences

Christian Bizer: Information Quality Assessment (5/29/2008)

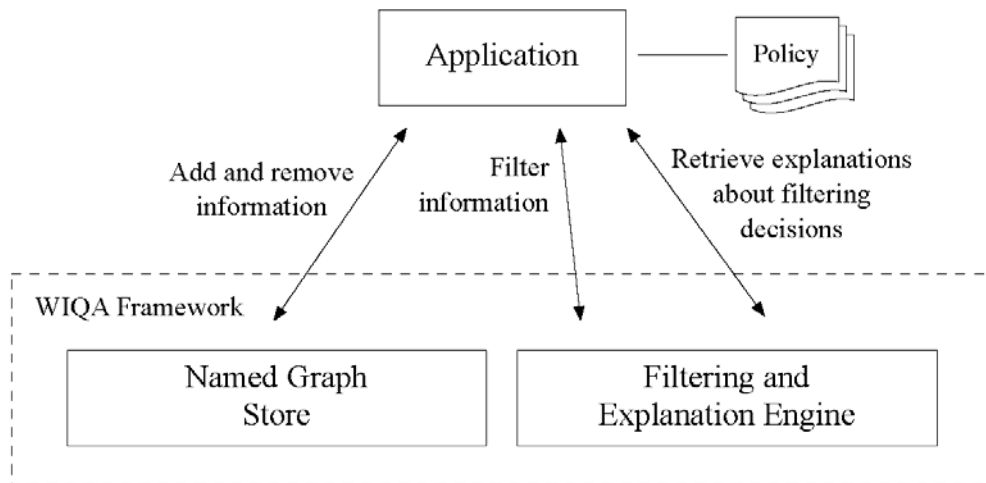
## The WIQA - Information Quality Assessment Framework

**WIQA is designed to fulfill the following requirements:**

1. Flexible representation of information together with quality-related meta-information
2. Enable users to employ different information filtering policies
3. Ability to generate explanations about the filtering process

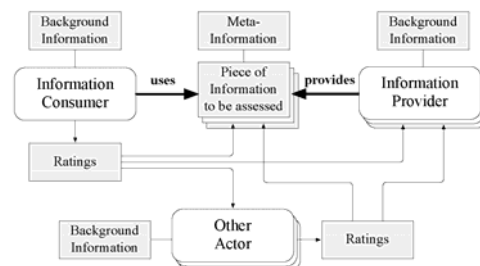
Christian Bizer: Information Quality Assessment (5/29/2008)

## The WIQA - Information Quality Assessment Framework



Christian Bizer: Information Quality Assessment (5/29/2008)

## 1. Representation of Quality-Related Meta-Information



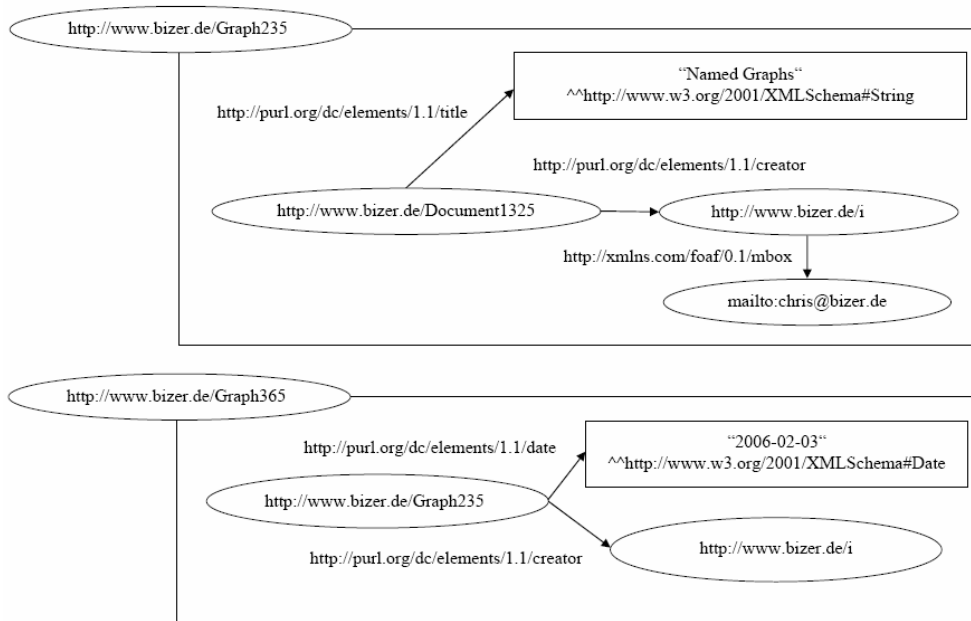
### ■ RDF Reification mechanism is not suitable for the representation of complex meta-information

- triple bloat
- does not work together with SPARQL
- reification of reified statements?

### ■ Therefore, RDF is extended to the Named Graphs Data Model

Christian Bizer: Information Quality Assessment (5/29/2008)

## The Named Graphs Data Model



Christian Bizer: Information Quality Assessment (5/29/2008)

## The TriG Syntax

```

fd:GraphFromIntel {
  <http://www.intel.com/c>
    rdf:type fin:Corporation ;
    fin:country iso:US ;
    foaf:homepage <http://www.intel.com> .
}

fd:GraphFromYahooFinance {
  <urn:x-ISIN:US4581401001>
    rdf:type fin:Share ;
    fin:emitter <http://www.intel.com/c> .
}

```

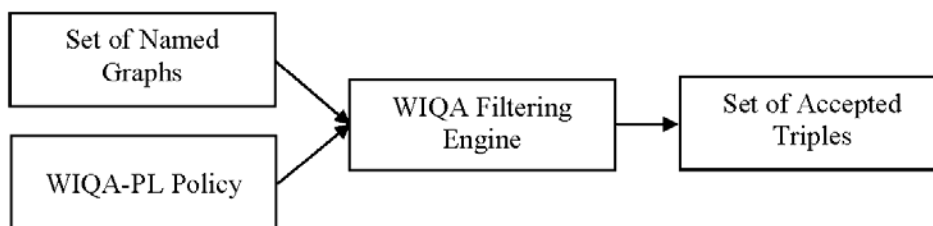
Christian Bizer: Information Quality Assessment (5/29/2008)

## The Semantic Web Publishing Vocabulary (SWP)

```
fd:GraphFromAggregator {  
  fd:GraphFromIntel  
    swp:assertedBy fd:IntelWarrant .  
  fd:IntelWarrant  
    swp:authority <http://www.intel.com/c> ;  
    dc:date "2007-10-21"^^xsd:date .  
  fd:GraphFromYahooFinance  
    swp:assertedBy fd:YFWarrant .  
  fd:YFWarrant  
    swp:authority <http://www.yahoo.com/c> ;  
    dc:date "2007-11-20"^^xsd:date .  
}
```

Christian Bizer: Information Quality Assessment (5/29/2008)

## 2. Expressing Information Filtering Policies



### ■ WIQA-PL policies are expressed as a combination of

- Graph patterns
- Filter clauses
- Extension function calls

Christian Bizer: Information Quality Assessment (5/29/2008)

## WIQA-PL Referring Variables

Variable	Description
?SUBJ	Reference to the subject of a triple.
?PRED	Reference to the predicate of a triple.
?OBJ	Reference to the object of a triple.
?GPAPH	Reference to the graph containing a triple.

Christian Bizer: Information Quality Assessment (5/29/2008)

## WIQA-PL Example Policy

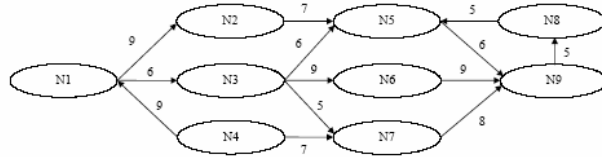
```
NAME "Information from German analysts"
DESCRIPTION "Use only information which has been
            asserted by German analysts."
PATTERN
{
  GRAPH fd:GraphFromAggregator
  { ?GRAPH swp:assertedBy ?warrant .
    ?warrant swp:authority ?authority . }

  GRAPH fd:BackgroundInformation
  { ?authority rdf:type fin:Analyst .
    ?authority fin:country iso:DE . }
}
```

Christian Bizer: Information Quality Assessment (5/29/2008)

## The TidalTrust Extension Function

$$t_{is} = \frac{\sum_{j \in \text{suc}(i) \mid t_{ij} \geq \text{max}} t_{ij} t_{js}}{\sum_{j \in \text{suc}(i) \mid t_{ij} \geq \text{max}} t_{ij}}$$



NAME "TidalTrust rating above 5"

DESCRIPTION "Only accept information from analysts with a Tidal Trust rating above 5."

```
PATTERN {
  GRAPH fd:GraphFromAggregator {
    ?GRAPH swp:assertedBy ?warrant .
    ?warrant swp:authority ?authority .
    FILTER (wiqa:TidalTrust(?USER, ?authority) > 5)
  }
}
```

Christian Bizer: Information Quality Assessment (5/29/2008)

## 3. Explaining Filtering Decisions

User's final decision whether to trust or distrust assessment results depends on her understanding of the assessment process.

The WIQA framework combines two explanation generation mechanisms:

- A template mechanism to explain why constraints that are expressed as graph patterns are satisfied.
- Custom explanations generated by the extensions functions.

Christian Bizer: Information Quality Assessment (5/29/2008)

## Explanation Patterns

```
NAME "Asserted by analysts with at least 3 positive
ratings."
PATTERNS {
  GRAPH fd:GraphFromAggregator
    { ?GRAPH swp:assertedBy ?warrant .
      ?warrant swp:authority ?auth .
      EXPL "it was asserted by " ?auth " and " . }
  GRAPH ANY
    { ?rater fin:positiveRating ?auth .
      FILTER (wiqa:count(?rater) > 2) .
      EXPL ?auth "has received positive
        ratings from" . }
  GRAPH fd:BackgroundInformation
    { ?rater fin:affiliation ?company .
      EXPL ?rater "who works for" ?company . }
}
```

Christian Bizer: Information Quality Assessment (5/29/2008)

## Example Explanation

### The triple:

- Siemens AG has positive analyst report: "As Siemens agrees partnership with Novell unit SUSE ..."

### fulfills the policy:

- Accept only information that has been asserted by people who have received at least 3 positive ratings.

### because:

- it was asserted by Peter Smith and
- Peter Smith has received positive ratings from
  - Mark Scott who works for Siemens.
  - David Brown who works for Intel.
  - John Maynard who works for Financial Times.

Christian Bizer: Information Quality Assessment (5/29/2008)

# The WIQA Browser

The screenshot shows the WIQA Browser interface in Mozilla Firefox. The main content area displays search results for 'Deutsche Bank' and 'Google Inc.'. The interface includes several key components:

- Filter Criteria:** Located at the top left, showing the current filter 'is a: Corporation'.
- Search Panel:** A search bar at the top right with the placeholder text 'Type here to search'.
- Value Selection Panel:** A dropdown menu on the right side showing 'country' with options for 'Germany (2)' and 'United States (2)'.
- Property Selection Panel:** A panel on the right side showing a list of properties: 'is a', 'name', 'country', 'has homepage', and 'news'.
- Information Item:** A callout box pointing to the 'Deutsche Bank' search result.

The search results for 'Deutsche Bank' include properties like 'country', 'has homepage', 'is a', 'name', and 'news'. The 'news' property shows a snippet of a news article about Deutsche Bank's investigation.

# Applying Filtering Policies

The screenshot shows the WIQA Browser interface in Mozilla Firefox. The main content area displays search results for 'Siemens'. The interface includes several key components:

- Filter Criteria:** Located at the top left, showing the current filter 'is a: Share'.
- Policy Selection Panel:** A panel on the right side showing a list of filtering policies: 'Information from German analysts', 'Information from positively rated information providers', 'New information from highly rated analysts', 'Only German or English information', 'Accept only information from Deutsche Bank', 'More positive Ratings', 'Trust rating above 5', 'Asserted by two different analysts', 'Asserted by analysts with at least 3 positive ratings', and 'Accept everything'.
- Oh, yeah? Button:** A callout box pointing to a 'Share' button next to a search result.

The search results for 'Siemens' include properties like 'is a', 'name', 'discussion forum posting', 'emitted by', 'positive analyst report', and 'negative analyst report'. The 'positive analyst report' property shows a snippet of a news article about Siemens' partnership with Novell.

# Retrieving Explanations

The screenshot shows the WIQA Browser interface. On the left, a list of items is displayed, including a 'positive analyst report' and a 'negative analyst report'. The 'positive analyst report' is selected, and its explanation is shown in a separate window titled 'EXPLANATION'. The explanation details the triple, the policy it fulfills, and the reasons for it.

**EXPLANATION**  
WIQA Browser

**The Triple:**  
Siemens Share positive analyst report Siemens agree partnership with Novell unit SUSE. Siemens Business Services (SBS), the IT services arm of German technology conglomerate Siemens <SIEGn.DE>, said on Tuesday it had agreed a partnership deal with Novell (nasdaq:NOVL - news - people) newly acquired unit SUSE Linux. Linux software is open-source, meaning it can be freely copied and modified, unlike proprietary software such as Microsoft (nasdaq:MSFT - news - people) Windows. In the past months clients have been asking more and more for open-source platforms, SBS said in a statement which said SUSE would have premier partner status. SBS is one of Europe top 10 information technology service providers. Linux, once the exclusive province of a few dedicated enthusiasts, is now seen as the only serious rival to Windows and is supported by U.S. giant International Business Machines (nyse:IBM - news - people), among others. Its advocates, who include big businesses and government departments, argue it is cheaper, simpler and more secure than Windows.

**fulfils the policy:**  
Use only information which has been asserted by German analysts.

**because:**

- it is stated in the document **Information from Peter Smith**, which is asserted by the German analyst **Peter Smith**.

Close

# Explanation for a Policy using the TidalTrust Metric

The screenshot shows the 'EXPLANATION' window for a policy. It details the triple, the policy, and the reasons for it, including a detailed calculation of the Tidal Trust rating.

**EXPLANATION**  
WIQA Browser

**The Triple:**  
Intel Share discussion forum posting As we have already seen in in the past, investing into this company is no good idea.

**fulfils the policy:**  
Only accept information from information providers with a Tidal Trust rating above 5.

**because:**

- it was asserted by **Mark Scott**. The WIQA extension function 'Tidal Trust' inferred a rating of 6.7 from **Chris Bizer** for **Mark Scott**. (Detail number 1)

**Details:**

**Detail number 1**

- The inferred rating arises from the following calculation:
  - The shortest path between **Chris Bizer** and **Mark Scott** has length 3. There are 4 different paths with that length. (Detail number 2)
  - The maximum strength of the paths is 6.0. Therefore, ratings below 6.0 are ignored.
  - The calculation yielded a result of 6.7. (Detail number 3)

**Detail number 2**

- Paths from the source to the sink:
  - Chris Bizer** -6.0-> **Anne Richards** -6.0-> **John Gevner** -9.0-> **Mark Scott** (Strength of the path: 6.0)
  - Chris Bizer** -9.0-> **Siddhartha Katakli** -7.0-> **Mary Louis** -6.0-> **Mark Scott** (Strength of the path: 6.0)
  - Chris Bizer** -6.0-> **Anne Richards** -6.0-> **Mary Louis** -6.0-> **Mark Scott** (Strength of the path: 6.0)
  - Chris Bizer** -6.0-> **Anne Richards** -6.0-> **Ivan Versivic** -8.0-> **Mark Scott** (Strength of the path: 5.0)

**Detail number 3**

- Chris Bizer** -6.7-> **Mark Scott** was calculated from these ratings:
  - Siddhartha Katakli** -6.0-> **Mark Scott**, weighted with **Chris Bizer's** rating of 9.0 for **Siddhartha Katakli**
  - Anne Richards** -7.0-> **Mark Scott**, weighted with **Chris Bizer's** rating of 6.0 for **Anne Richards**
- Mark Scott** was calculated from these ratings:
  - Mary Louis** -6.0-> **Mark Scott**, weighted with **Anne Richards's** rating of 6.0 for **Mary Louis**
  - John Gevner** -9.0-> **Mark Scott**, weighted with **Anne Richards's** rating of 9.0 for **John Gevner**
  - Ivan Versivic** -8.0-> **Mark Scott**, weighted with **Anne Richards's** rating of 5.0 for **Ivan Versivic**
- Siddhartha Katakli** -6.0-> **Mark Scott** was calculated from these ratings:
  - Mary Louis** -6.0-> **Mark Scott**, weighted with **Siddhartha Katakli's** rating of 7.0 for **Mary Louis**
- Ivan Versivic** -6.0-> **Mark Scott** is a direct rating.
- John Gevner** -9.0-> **Mark Scott** is a direct rating.
- Mary Louis** -6.0-> **Mark Scott** is a direct rating.

Close

## Use Cases for the WIQA Framework

- **Financial Information Portals**
- **News Portals**
- **Electronic Markets**
- **Online Communities like MySpace or Facebook**
- **Knowledge Management Systems**
- **Web Search Engines**

Christian Bizer: Information Quality Assessment (5/29/2008)

## My preferred Use Case

**The Web of Linked Data (aka Semantic Web)**

Christian Bizer: Information Quality Assessment (5/29/2008)

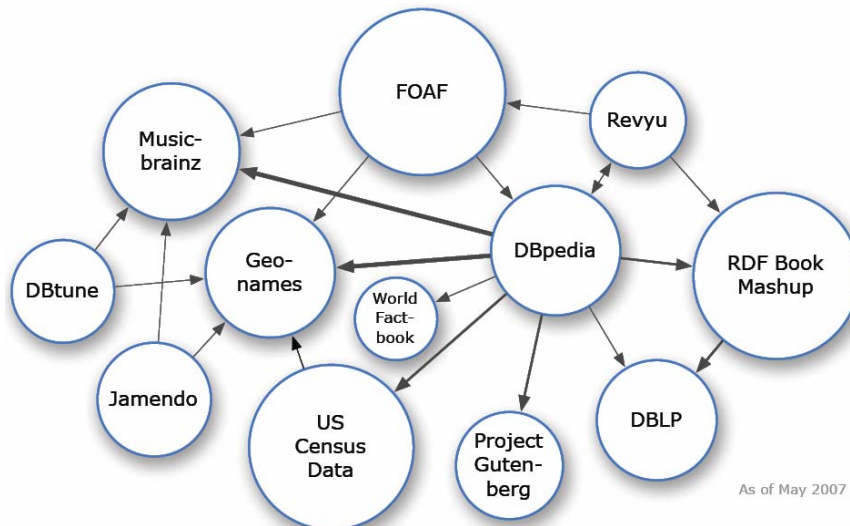


**Community effort stated in 2007 to**

- publish existing open license datasets as Linked Data on the Web
- interlink things between different data sources

Christian Bizer: Information Quality Assessment (5/29/2008)

**LOD Datasets on the Web: May 2007**

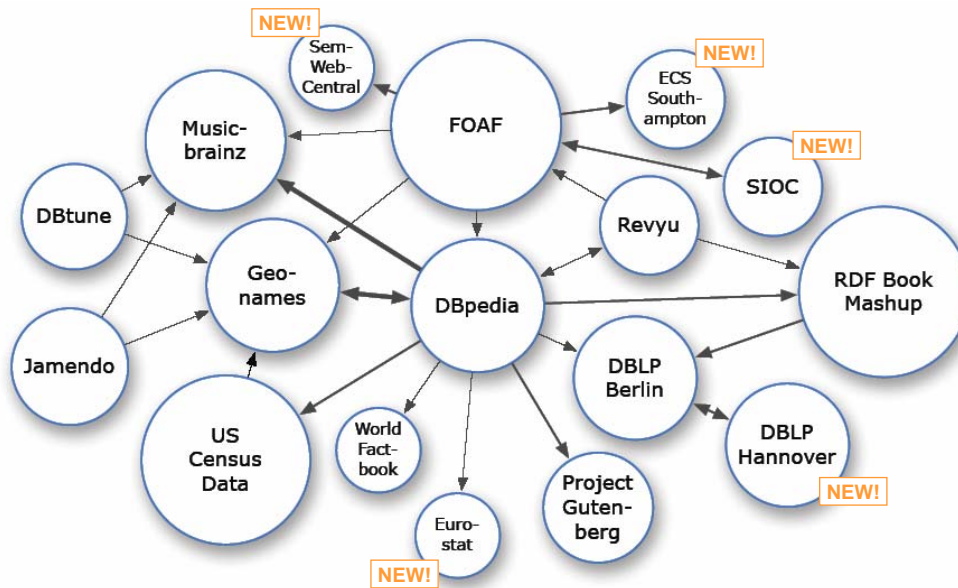


As of May 2007

- Over 500 million RDF triples
- Around 120,000 RDF links between data sources

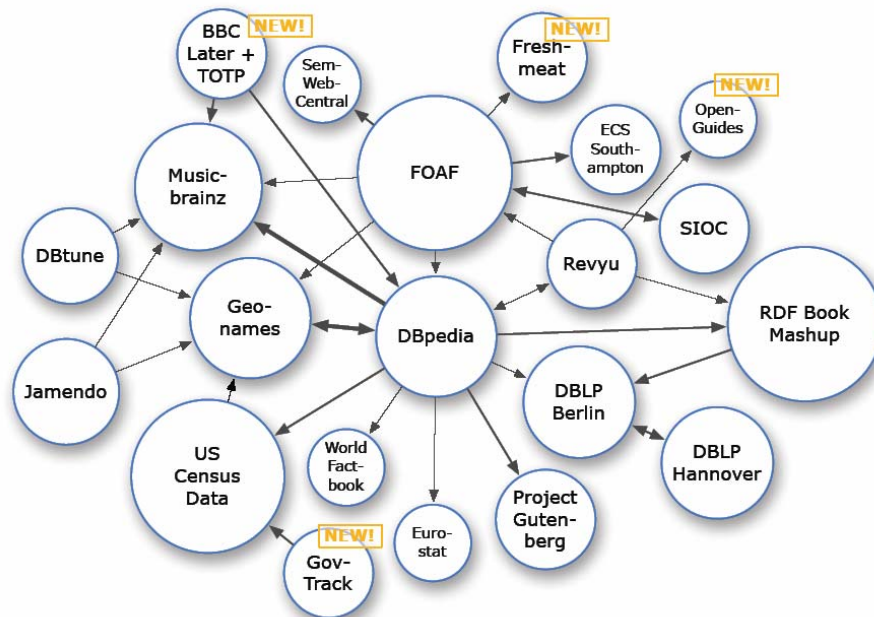
Christian Bizer: Information Quality Assessment (5/29/2008)

## LOD Datasets on the Web: July 2007



Christian Bizer: Information Quality Assessment (5/29/2008)

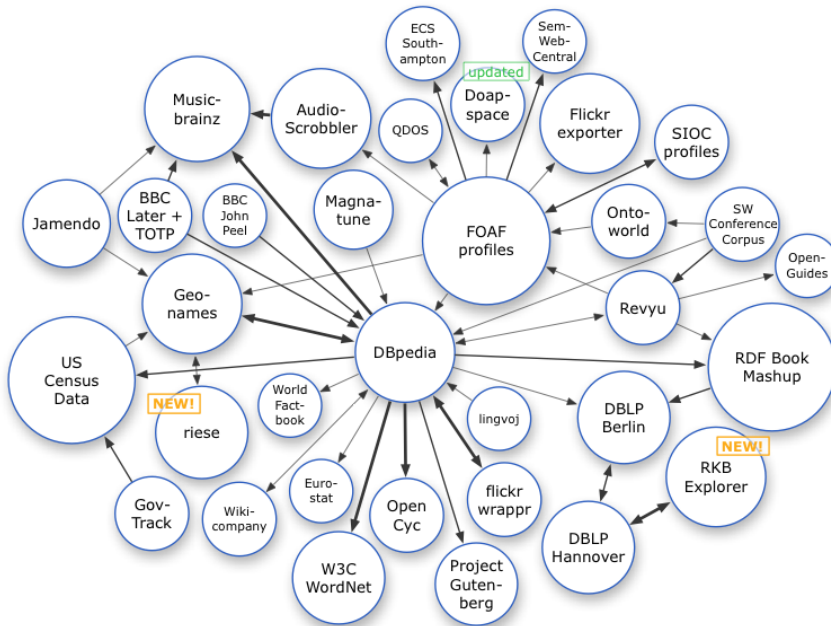
## LOD Datasets on the Web: August 2007



Christian Bizer: Information Quality Assessment (5/29/2008)



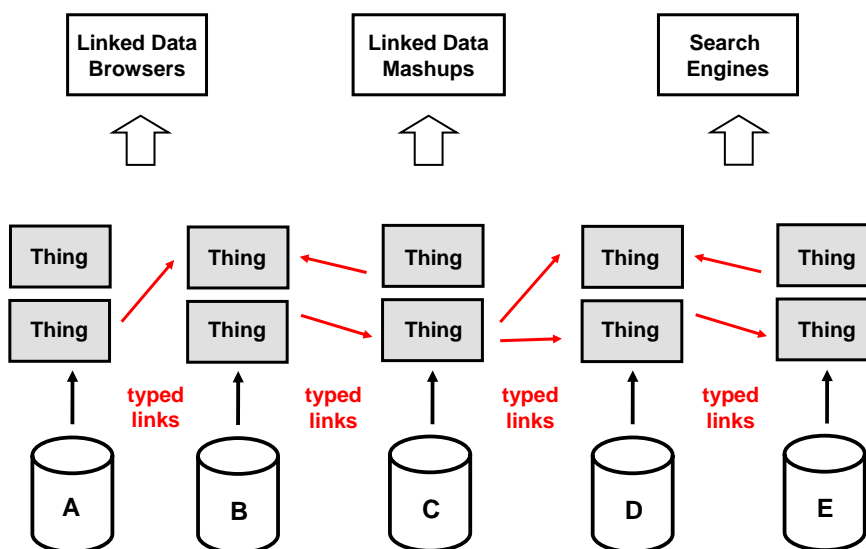
## LOD Datasets on the Web: April 2008




More than 2 billion RDF triples interlinked by 3 million RDF links.

Christian Bizer: Information Quality Assessment (5/29/2008)

## Sem-Web Apps that need Info-Quality Assessment





Christian Bizer: Information Quality Assessment (5/29/2008)


http://dbpedia.org/resource/Beijing\_Capital\_International\_Airport Open 

## Beijing Capital International Airport

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>


- **Building** 
- **Feature** 

**comment**


- Der Flughafen Peking ist einer der wichtigsten Flughäfen Chinas und liegt im Nordosten Pekings. Er liegt 20 Kilometer nördlich vom Stadtzentrum, im Gebiet Shunyi. Genauer ist aber der Flughafen innerhalb des Verwaltungsgebiets des Stadtbezirks Chaoyang. 
- L'aéroport international de Pékin (code AITA : PEK ; code OACI : ZBAA) est le principal aéroport de Pékin, capitale de la République populaire de Chine. Il est situé à une vingtaine de kilomètres au nord-est du centre de la ville, dans le district de

⋮

**is sameAs of**

- [Beijing Capital International Airport](#) 


**depiction**



⋮

## Linked Data Mashups like DBpedia Mobile

- **Geospatial entry point into the Web of Data**
- **Uses DBpedia, Revyu and Flickr**



Christian Bizer: Information Quality Assessment (5/29/2008)

## Semantic Web Search Engines like Falcons

The screenshot displays the Falcons Semantic Web search engine interface. The main window shows search results for the query 'Beijing'. The results are categorized by type: Capital City, City, Document, Group, Institution, Landmark, Location, Noun Synset, Ontology, Organization, Person, Publication, Subject, and System. The search results list several entries for 'Beijing', including its types and labels in various languages (e.g., 北京, Peking, Пекин). A detailed summary of the resource 'http://dbpedia.org/resource/Beijing' is shown in a separate window. This summary includes a table with columns for General, GEO, P, FODF, SKOS, and Authorized Sources. The summary also provides a comment in multiple languages, including Chinese, Japanese, and English, describing Beijing as the capital of the People's Republic of China (PRC).

## Take-away Message

**The (public) Semantic Web will be a huge mess!**

**The next generation of Semantic Web applications should take this more into account as applications do today.**

# Thanks!

## References

- **PhD Thesis**  
Christian Bizer: Quality-Driven Information Filtering.  
VDM Verlag Dr. Müller, ISBN 978-3-8364-2232-1, 2008
- **WIQA Framework**  
<http://www.wiwiss.fu-berlin.de/suhl/bizer/wiqa/>
- **WIQA Browser**  
<http://www.wiwiss.fu-berlin.de/suhl/bizer/wiqa/browser>

Christian Bizer: Information Quality Assessment (5/29/2008)