

XBRL, RDF and the Semantic Web

Companies House
Cardiff, 22 July 2009

Dave Raggett dsr@w3.org

W3C Fellow – Financial data & Semantic Web
Member – XBRL International Technical Standards Board

Agenda

- ▶ Why does XBRL matter
 - Core value proposition
 - Boosting innovation through better investment
 - Application to corporate and government data
- ▶ Semantic Web and linked open data
 - Basic principles
 - Relationship to database technologies
 - Linked open data
 - Support from PM and Sir Tim Berners-Lee
- ▶ How XBRL could feed an ecosystem of value added services
 - RSS feeds and web services
 - Rendering XBRL data
 - Search engines and financial data
 - Mapping XBRL to RDF and OWL
 - Hard and soft facts, people and computers
 - Towards a pilot project

Why does XBRL matter?

What is XBRL?

- ▶ Financial reports contain numeric and textual facts
- ▶ XBRL tags these facts with the reporting concept
- ▶ These concepts are defined by reference to generally accepted accounting principles

Demo of XBRL viewer, showing XBRL info as pop-ups

<http://apps.xbrlspy.org/test/index.php>

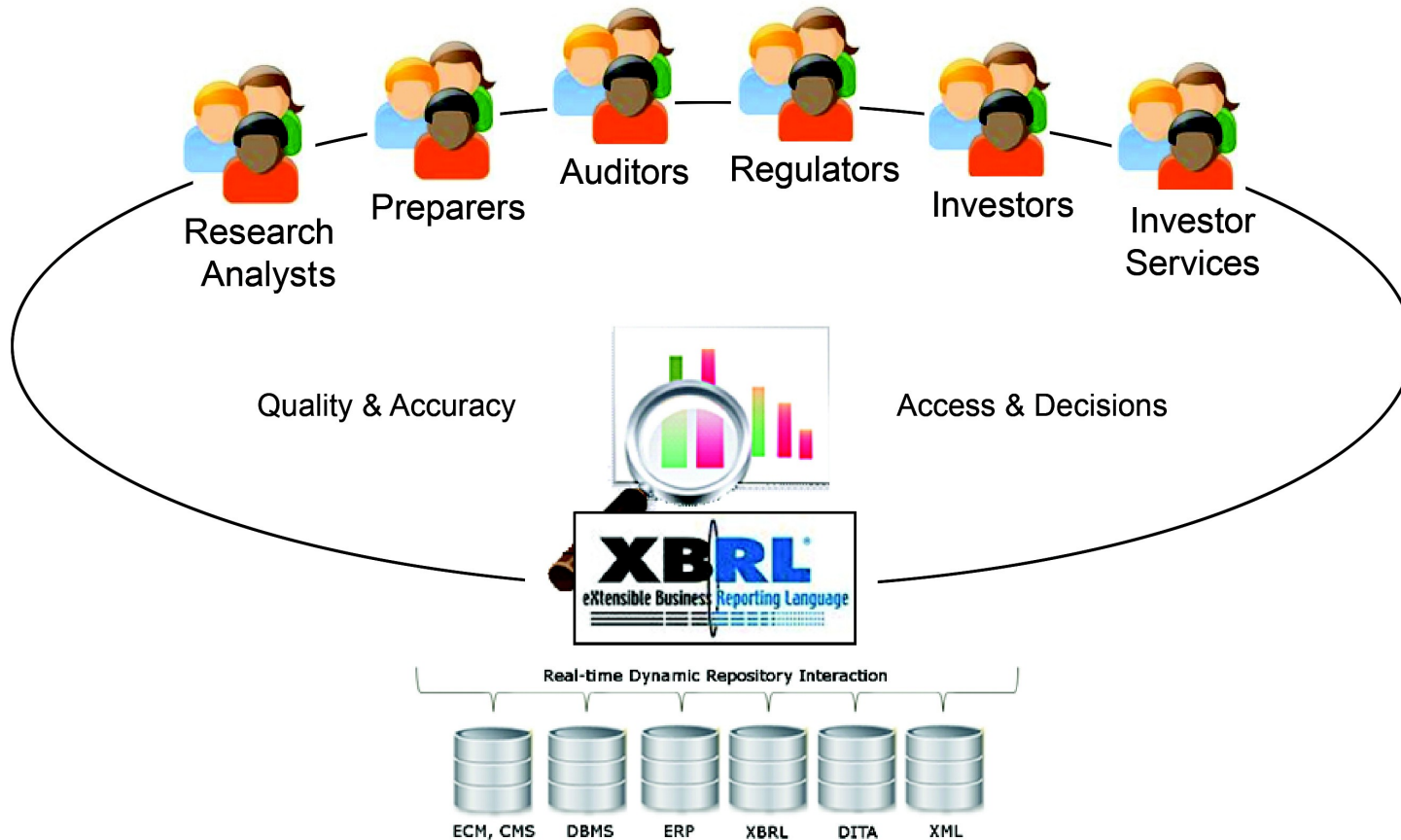
Value Proposition for XBRL

- ▶ Reducing costs and errors for dealing with financial information
 - No more error prone manual re-keying of data
 - Democratizing access to financial data
- ▶ Wooing Investors
 - Managers who run businesses for themselves rather than putting shareholders first
 - But businesses are competing for investor £££
- ▶ Investors are willing to pay for good analysis
- ▶ Governments Worldwide are requiring better reporting through use of XBRL

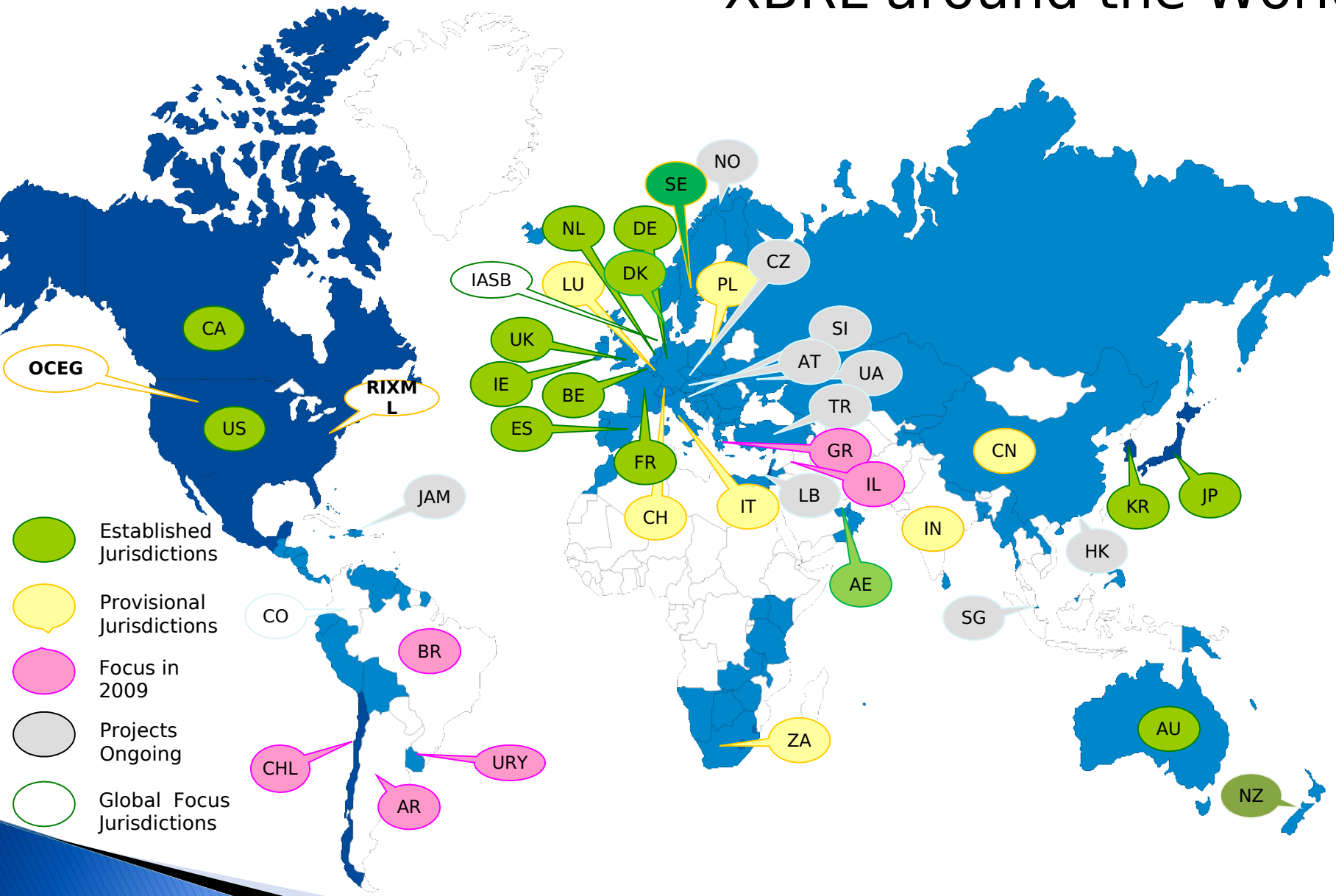
Who Benefits from XBRL

- ▶ Four categories of users:
 - business information preparers,
 - Intermediaries in the preparation and distribution process,
 - users of this information and
 - the vendors who supply software and services to one or more of these three types of user..
- ▶ A major goal of XBRL is to improve the business report product.
- ▶ It facilitates current practice; it does not change or set new accounting or other business domain standards.

Who Benefits from XBRL?



XBRL around the World



Transparency in Government

- ▶ US Transparency in Government Act 2008
 - Making information about Congress and the executive branch publically available online
- ▶ Obama memo on Transparency and Open Government
 - Government should be transparent, participatory and collaborative
- ▶ Sunlight is the best disinfectant
 - Transparency has the potential to reduce waste of taxpayers' money and provide more effective government

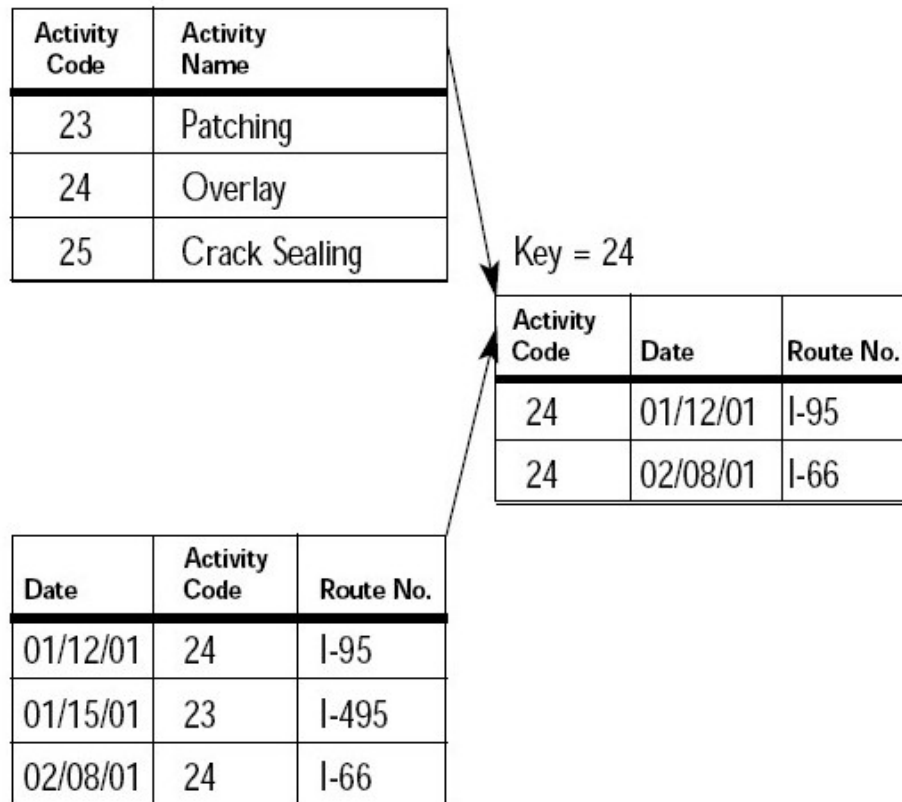
UK focus on government transparency

- ▶ Generally seen as a good idea
 - Temporarily sidetracked by MP allowances
 - Put information into public domain and encourage innovation
 - Encourage use of shared vocabularies and data formats, especially as a long term goal
 - Charging for access along with restrictive licenses will put a cold blanket on innovation
- ▶ <http://innovate.direct.gov.uk>
 - Innovation around open use of government data
- ▶ But how to balance costs and benefits?

The Semantic Web and Linked Open Data

Relational Data Model

- ▶ Based upon tables



Relational Data Model

- ▶ Uses shared values to link rows in different tables
- ▶ The naming scheme is essentially local
- ▶ The data model is not an integral part of the database
 - Tends to fall into disrepair as database changes in response to new needs
 - This makes it hard to combine data from different databases
- ▶ No standard way to access database as a web service

Evolution: databases \Rightarrow Semantic Web

- ▶ Extension of database technologies to deal with information on a Web scale
 - Combining information across many servers
 - Globalization of knowledge representation in the same way that the Web did for hypertext
- ▶ Relationships as the building block
 - Subject \rightarrow Relationship \rightarrow Object
- ▶ Where each of these are named with URIs
 - Universal resource identifiers, e.g. HTTP addresses

OWL Ontologies

- ▶ Used to describe data models for the Semantic Web
- ▶ Rich knowledge representation
 - X is a subclass of Y
 - P is an instance of class Q
 - A is a named part of B
 - Plus data types such as numbers, dates, and strings
- ▶ SPARQL query language
 - Applies to triple stores
 - Analogous to SQL for relational databases

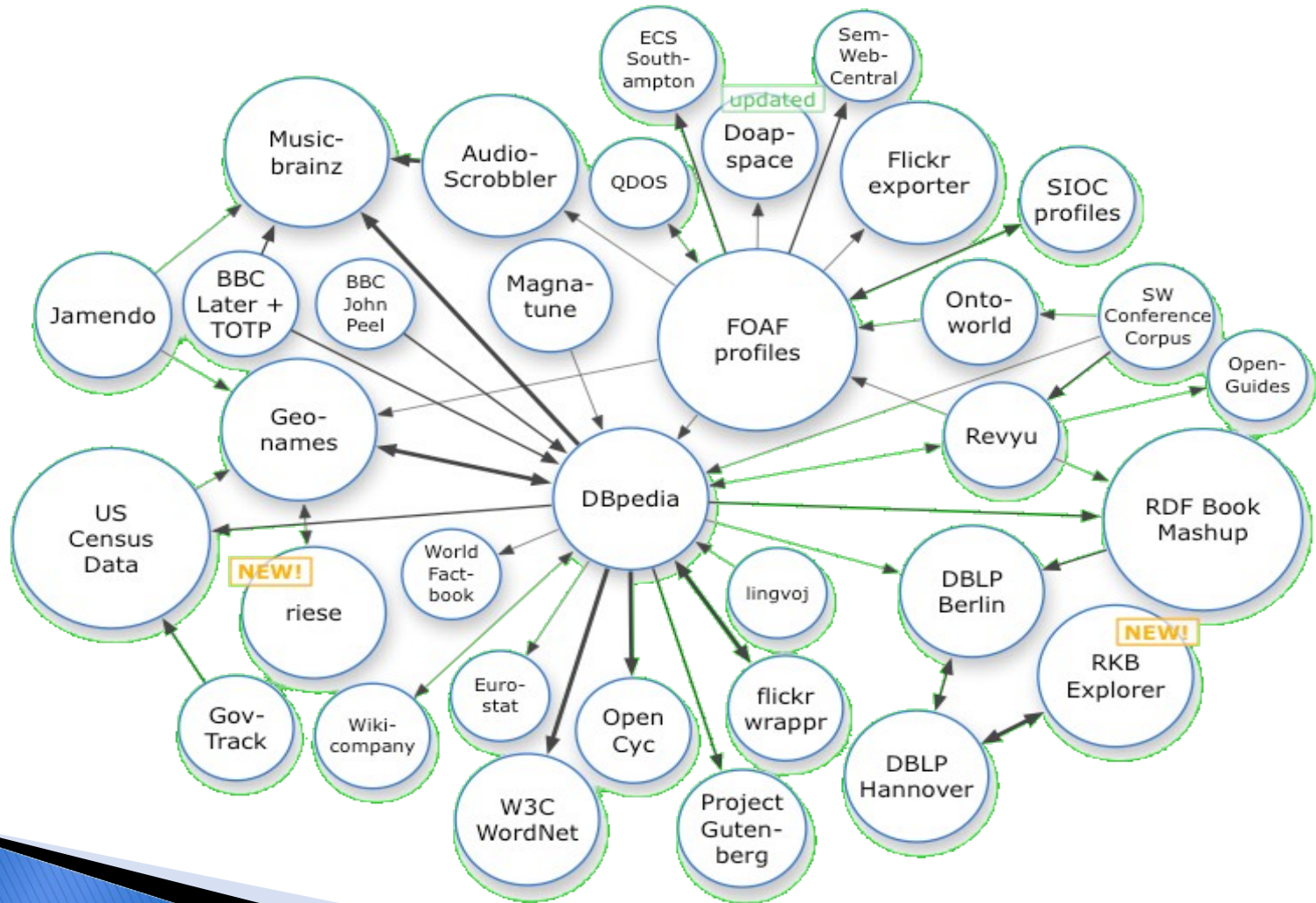
Rule languages

- ▶ Logic based rules
 - If *condition/event* then *action*
- ▶ Support richer kinds of reasoning than is possible with SPARQL and OWL ontology
- ▶ Can be used for
 - Access control
 - Integrity constraints
 - Other kinds of business logic
 - Analogous to XBRL formula
- ▶ Easier to inspect and maintain than procedural code, e.g. Java

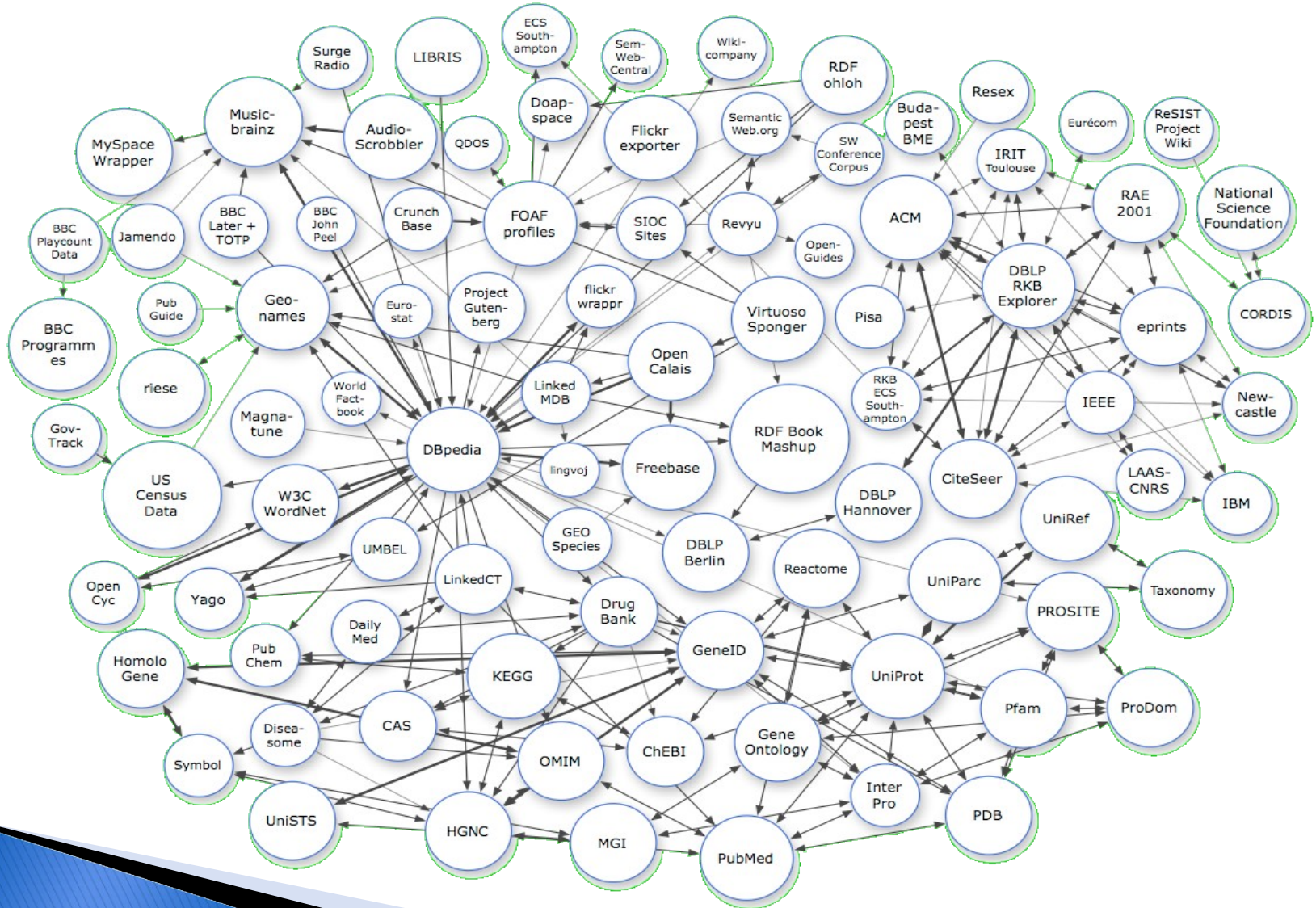
Linked Open Data

- ▶ Using the Web to connect related data that wasn't previously linked
- ▶ Allowing people and machines to explore and make use of this data
- ▶ Using standards that encourages re-use
 - HTTP URIs as names for things
 - SPARQL for querying data
 - Including links to other URIs so that you can discover more things

Linked Open Data - 2008-03-31



Linked Open Data - 2009-03-27



Gordon Brown on UK Initiative

10 June 2009

"So that government information is accessible and useful for the widest possible group of people, I have asked Sir Tim Berners-Lee who led the creation of the world wide web, to help us drive the opening up of access to Government data in the web over the coming month."



Inventor of the World Wide Web

<http://www.guardian.co.uk/technology/2009/jun/10/berners-lee-downing-street-web-open>

Sir Tim Berners-Lee



- ▶ From BBC News interview, 12 June 2009
 - “Growing public demand for access to government data”
 - “This is our data, this is our taxpayer's money which has created this data so it should be available for us to see!”
- ▶ This should include both
 - Data created by the government, and
 - Public data collected by the government

How XBRL could feed an ecosystem of value added services

View from America

- ▶ SEC voluntary filing program for XBRL
 - Allowed businesses to learn by experience
 - Filings made public on SEC website
 - SEC too learned about what rules to apply before accepting filings
- ▶ Free access via RSS, HTTP and FTP
 - <http://www.sec.gov/Archives/edgar/xbrlrss.xml>
 - <http://www.sec.gov/Archives/edgar/data/.../ndaq-20081231.xml>
 - <ftp://ftp.sec.gov/>
- ▶ XBRL viewer for voluntary filings
 - <http://viewerprototype1.com/viewer>
- ▶ Online validator for pre-filing checks

Purpose of XBRL

- ▶ XBRL provides users with a standard format in which to **prepare reports** that can subsequently be presented in a variety of ways.
- ▶ XBRL provides users with a standard format in which information can **be exchanged** between different software applications.
- ▶ XBRL permits the automated, efficient and reliable **extraction of information** by software applications.
- ▶ XBRL facilitates the **automated comparison** of financial and other business information, accounting policies, notes to financial statements between companies, and other items about which users may wish make comparisons that today are performed manually.

Drawbacks with XBRL

- ▶ Expensive to process with XML tools
 - XSLT is bad at dealing with Xlink
- ▶ Bad practices
 - Embedding untagged data as HTML
 - Missing roles e.g. periodEnd
 - Misuse of XBRL tuples
- ▶ Missing certain kinds of knowledge
 - Supports ordering within tables
 - But not across tables
 - Good for numeric facts
 - But poor for non-numeric information such as a prospectus for a mutual fund

Feeding the Semantic Web

Typical XBRL filing consists of

- ▶ Instance file with reported facts
 - Numeric and textual facts
 - Dates and Periods
 - Currencies
 - Footnotes
 - Reporting dimensions
- ▶ Schema for the instance file
 - Definition of markup elements for facts
 - References to reporting taxonomy
- ▶ Taxonomy extensions
 - Labels, additional concepts and relationships
- ▶ Around 10 to 50 Mb including the taxonomy!

Why translate XBRL?

- ▶ Very expensive to process 10-50Mb of XML per filing on each query
 - Memory and CPU intensive, about 1 sec per 10 Mb
- ▶ Better to pre-process filings into a persistent format designed to match needs of queries
 - Current tools use proprietary relational model
- ▶ RDF and OWL as natural target formats
 - Mature standards
 - Mashing financial and other kinds of data
 - Web APIs and standards would enable an ecosystem of value adding players

The Semantic Web as a Layer Cake

```
rdfproc store query sparql - "PREFIX rdf:  
<http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX usfr-pte:  
<http://www.xbrl.org/us/fr/common/pte/2005-02-28> PREFIX xl:  
<http://www.xbrl.org/2003/XLink> PREFIX xbrli:  
<http://www.xbrl.org/2003/instance> SELECT ?v WHERE {?f rdf:type  
usfr-pte:NetIncome . ?f xl:type xbrli:fact . ?f rdf:value ?v}"
```

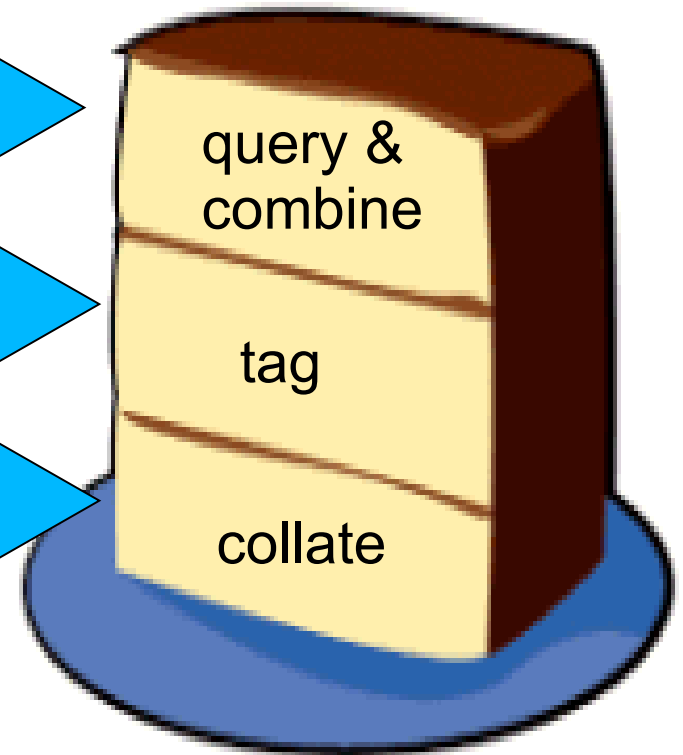
RDF

```
<import namespace="http://xbrl.us/us-gaap-all/2008-03-31"  
schemaLocation="http://xbrl.us/us-gaap/1.0/elts/us-gaap-all-2008-03-  
31.xsd"/>
```

XBRL



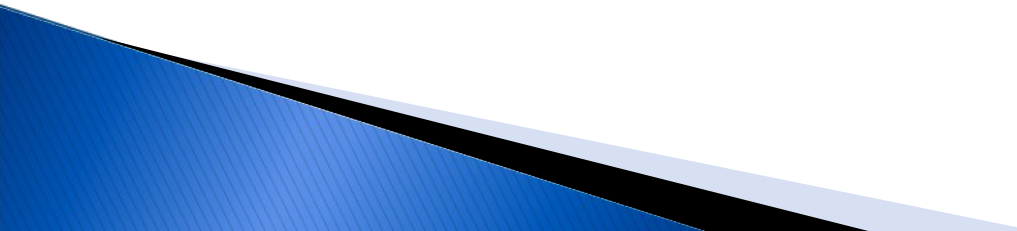
Raw Data



XBRL and OWL

- ▶ XBRL Taxonomy loosely equates to OWL ontology
 - ▶ But note XBRL's taxonomy overrides
- ▶ Automated mapping is mostly feasible
 - ▶ As demonstrated by Rhizomik XSD2OWL
- ▶ XBRL's formal semantics are weak
- ▶ XBRL versioning standard will describe differences between different versions of the same taxonomy, e.g. US GAAP 2008, 2009
 - ▶ Unaware of work on mapping this into OWL
- ▶ Reasoning across different taxonomies remains a major challenge
 - ▶ e.g. US GAAP vs IFRS

Web APIs for Financial Data

- ▶ Support for an ecosystem of value-adding players
 - ▶ First stage is data aggregators who pull XBRL from SEC and other sources and expose it as triples
 - ▶ Access to raw triples via SPARQL queries
 - ▶ Consumer uses scripts to add value
 - ▶ High-level APIs for common queries, where the results are provided as charts or tables
 - ▶ For embedding in web pages
 - ▶ Yet higher-level APIs for financial analytics that combine data from multiple filings
 - ▶ Complicated by variations across ontologies
- 

Smart Search Engines

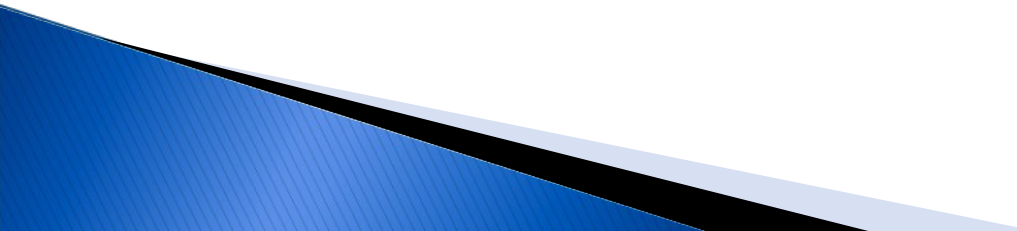
- ▶ Imagine search engines that provide selected financial highlights for each company that matches the search criteria you just entered
 - ▶ With salient numbers and charts
- ▶ The search results tailor the data provided according to your interests
 - ▶ Based upon analysis of the search criteria and other information gleaned from previous searches
 - ▶ Subject to your privacy preferences, of course! **
- ▶ Interactive data you can drill down on

Web Scale Queries

- ▶ SPARQL & RDF offer generality but sacrifice speed for complex queries
- ▶ For predetermined models and queries a persistent object store can allow queries to execute at native C or Java speeds
 - ▶ Sub-second response times
- ▶ Use of cloud computing solutions for web scale performance
 - ▶ Executing a query across thousands of servers
 - ▶ Exploiting really large data sets
 - ▶ Changing the kinds of questions we can ask
- ▶ Dependent on ecosystem of players
 - ▶ Not a single algorithm unlike text-based search

Soft and Hard data

*Combining the strengths of
people and computers*



Open Calais



Buttons: Show RDF, Entry Page

Navigation: +, -, left arrow, right arrow

Topics:

Environment	100%
-------------	------

Social Tags:

Environment	☆☆☆
-------------	-----

Entities:

- Company**
 - British Broadcasting Corporation (C:1 R:18%)
- Country**
 - United Kingdom (C:3 R:35%)
- Organization**
 - Environment Agency (C:2 R:51%)
- Person**
 - Liz Parks (C:4 R:80%)
- Position**
 - Head (C:1 R:33%)

Events & Facts:

- Generic Relations**
- Person Communication**
- Quotation**

The UK is working with Brazilian authorities to return more than 1,400 tonnes of toxic waste to Britain, the Environment Agency has said. Head of waste Liz Parks said plans were being made to bring back the rubbish, but it could take a number of weeks. An inquiry into how the waste, including syringes, condoms and bags of blood, was sent to three Brazilian ports has been launched by the UK. The Environment Agency says those responsible could face prosecution. Ms Parks told the BBC's News hour she understood the waste, found in about 90 shipping containers, was currently being held by the Brazilian authorities. "They haven't yet released it, as far as I'm aware. But arrangements are being made for that to happen. And it will take a number of weeks for the waste to be returned," she said.

<http://viewer.opencalais.com/>

Understanding and Valuing Businesses

- ▶ The numbers in financial reports are only one source of information
 - The notes to the financial statements
 - News stories about the business, sector and economy
 - Global influences e.g. on currencies
- ▶ Much of this extra information is soft!
 - Can't be extracted by a computer as it relies on human judgement
- ▶ This is where human generated content comes in
 - Allowing people to contribute their analysis as part of an ecosystem

Where next?

Next Steps

W3C/XBRL International Workshop

- ▶ 5-6 October, Washington DC, hosted by the FDIC
 - Focus on use cases and challenges for realizing an ecosystem of services
 - See <http://www.w3.org/2009/03/xbri/cfp.html>
 - Anticipate follow up workshops in Europe and Asia
- ▶ Expected to influence further standards work in both W3C and XBRL International

Opportunities for the UK?

- ▶ Done right, Web-based access to financial and business data will boost investment in UK companies
- ▶ This will be a learning curve for all involved
- ▶ Opportunity for a Pilot project
 - Exploring practical issues for exposing public data collected by Companies House and HMRC
 - Export data as XBRL and as RDF/OWL
 - Collaboration with UK-based partners
- ▶ What's the best way to approach this?

Thanks for listening!

Questions?