



**PhD Thesis
Progress and future work**

Querying and optimising access to RDF datasets

Carlos Buil Aranda

Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
cbuil@fi.upm.es
27th May 2010

- Introduction
- Hypothesis and objectives
- PhD Progress
 - First Year
 - Second Year
 - Third Year

- **Introduction**
- Hypothesis and objectives
- PhD Progress
 - First Year
 - Second Year
 - Third Year

- Everyday more RDF data is published on the Web

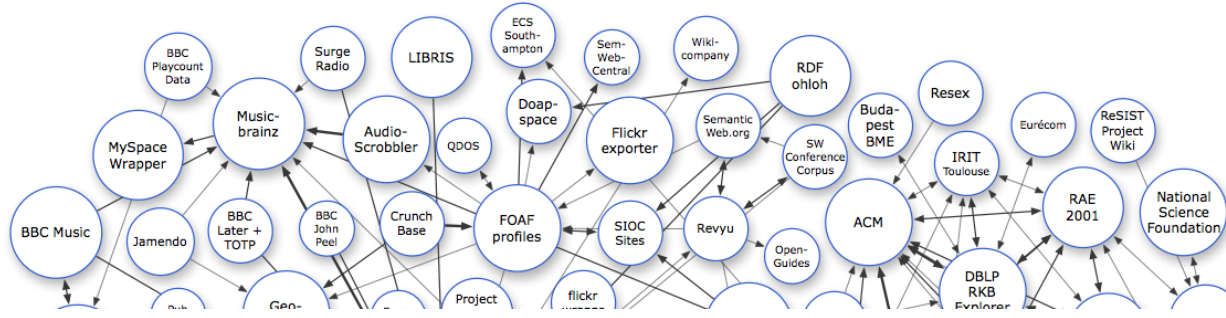
- Mostly via SPARQL endpoints

- For

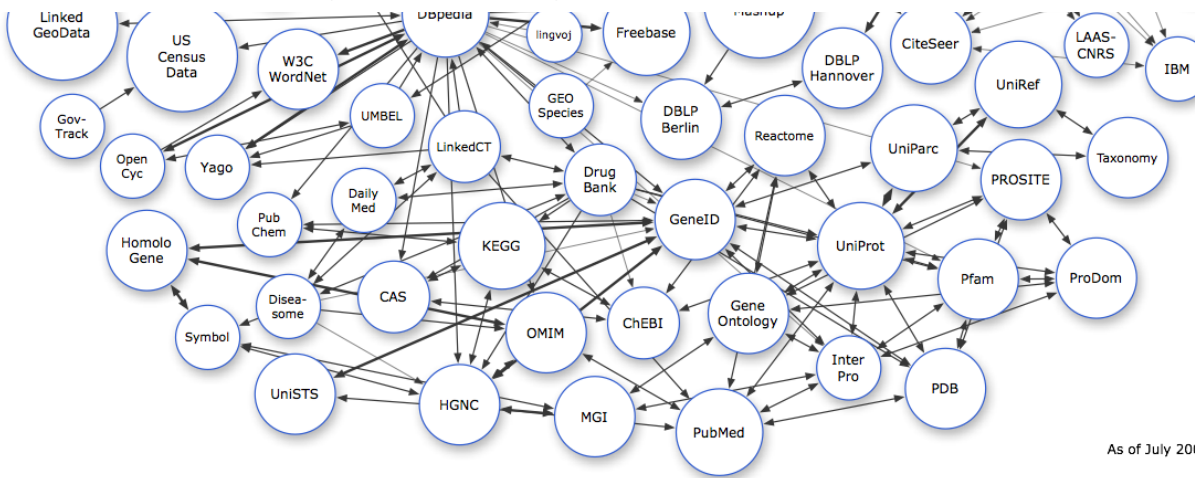
- Cl

- htt

- htt



They are not fully connected, i.e. To mash up the data it is necessary to query each of the RDF datasets



As of July 2009

- Currently there are only a few approaches to federate SPARQL queries
 - DARQ (Cyganiak 2006)
 - SemWIQ (Langegger, iiWAS2008)
 - Networked Graphs (Staab WWW2008)
 - Executing SPARQL Queries over the Web of Linked Data (Hartig, Bizer, Freytag, ISWC09)
- Problems from the previous approaches
 - They implement a basic system for optimising SPARQL queries
 - They do not take into account neither blank nodes nor complexity
 - They do not implement many optimisation methods

- Can RDB2RDF Tools Feasibly Expose Large Science Archives for Data Integration?, ESWC 2009
 - No, among other reasons there is no research in SPARQL Query Optimisation (authors dixit)
- Need of query optimisation
 - A Flexible Architecture for Virtual Information Integration based on Semantic Web Concepts, A Langeegger, PhD Thesis (beginning 2010)
 - Foundations of SPARQL Query Optimization, M Schmidt, M Meier, G Lausen (Tech. Report 2008)

- Introduction
- **Hypothesis and objectives**
- PhD Progress
 - First Year
 - Second Year
 - Third Year

Assumptions

- There is a large set of distributed RDF stores with a large amount of triples (billions of triples)
- RDF data is linked, or has links between them
- A SPARQL query can be translated into a SQL query, without losing expressivity

Restriction

- Users have a minimal knowledge of the vocabulary used in those RDF datasets

Hypotheses to test

- It is possible to use existing SQL distributed query processing techniques to process distributed SPARQL queries

Objectives

- Design, implement and evaluate a Distributed Query Processor to federate SPARQL queries
- Create new SPARQL optimisation techniques
- Validate how existing SQL optimisation techniques perform in a distributed SPARQL query environment

- Introduction
- Hypothesis and objectives
- **PhD Progress**
 - **First Year**
 - **Second Year**
 - **Third Year**

- First Year
 - Read, read, read
 - About Grid computing
 - About Description Logics
 - About Computational Complexity
 - Attendance to ISGC 2008 as student
 - Attendance to OGF 25
 - Which changed a little bit my research topic
 - Paper accepted in Web Semantics workshop in conjunction with DEXA2009: Robust service-based semantic querying to distributed heterogeneous databases, Buil-Aranda C, Corcho O, Krause A
 - Suficiencia investigadora (beginning of second year)

- Second year
 - Extended an existing SQL Distributed Query Processing system
 - SPARQL-DQP implementation
 - Extends a well know DQP system: OGSA-DQP
 - Research visit: EPCC and NeSC in Edinburgh (July to end September)
 - Submitted paper to ER2010 conferences, waiting for an answer (an others)
 - Many contacts with international institutions, to collaborate in some way
 - Carleton Univ. (bio2rdf, Michel Dumontier, thanks to Alex), EPCC & NeSC (Uni Edinburgh), PUC (Chile, M Arenas and C Gutierrez), UCL (London), Uni Vienna, AIST (Japan)
 - Collaboration with DAIS-WG

- Third year (since April)
 - Improvement of the SPARQL-DQP software
 - Research visit to Universidad Católica de Chile (Claudio Gutierrez and Marcelo Arenas)
 - Goals: Merge both theoretical and practical approaches to SPARQL query processing
 - Treat blank nodes problem?
 - Paper to ESWC: outcome from the research visit
 - Paper to ISWC? Journal of Web Semantics?
 - Write the PhD thesis
 - If there is time research visit to?
 - Vienna (with A Langegger)?
 - U. South California (with JL Ambite)?



**PhD Thesis
Progress and future work**

Querying and optimising access to RDF datasets

Carlos Buil Aranda

Ontology Engineering Group
Facultad de Informática
Universidad Politécnica de Madrid
cbuil@fi.upm.es
27th May 2010