Open Genomic Data Web

Dr Jun Zhao
Image Bioinformatics Research Group
Department of Zoology
University of Oxford
6 August 2009
GMOD Meeting Europe
Web of Data: “may more accurately be described as a web of things in the world, described by data on the Web.”
Linked Data Design Issues

Linked Data

1. Use URIs as names for things

2. Use HTTP URIs so that people can look up those names.

3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)

4. Include links to other URIs, so that they can discover more things.
Resource Description Framework (RDF)
Resource Description Framework (RDF)
SPARQL queries

SPARQL

PREFIX chado: <http://purl.org/net/chado/schema>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xs: <http://www.w3.org/2001/XMLSchema#>
SELECT ?flybaseID
WHERE {
  ?feature rdf:type chado:Feature ;
  chado:name "schuy"^^xs:string ;
  chado:uniquename ?flybaseID .
}

SQL

SELECT ?feature.uniquename AS flybaseID
FROM feature
WHERE feature.name = "schuy"
SPARQL protocol

HTTP GET
GET /query/flybase?query=[URL encoded query] HTTP/1.1
Host: openflydata.org
Accept: application/sparql-results+json

HTTP POST
POST /query/flybase HTTP/1.1
Host: openflydata.org
Accept: application/sparql-results+json
Content-Type: application/x-www-form-urlencoded
Content-Length: 456
query=[URL encoded query]
open interoperable
Two Exemplar Applications

- OpenFlyData.org
- Connect TCM with Western Medicine
OpenFlyData: mRNA gene expression study

- Microarray analysis
  - How much of a given transcript (mRNA) is present in a sample
  - In a quantitative way
  - Lack of spatial information

- RNA *in situ* hybridization
  - Reveal both spatial and temporal aspects of gene expression during the development
  - But not quantitative
Barriers for accessing these data

- Data are scattered at different web sites
- Searches have to be repeated, different search interfaces, different use of terminology
- Limited (if any) programmatic access to data ... hard work to answer questions that span data sources
OpenFlyData.org demonstration

- Three gene express cross-database search applications
  - Search by gene, gene expression mashup: [go]
  - Search gene expression by gene batch [go]
  - Search gene expression by tissue expression profile [go]
System architecture

Client side

Web browser
FlyUI application
FlyUI widget

HTTP
SPARQL endpoint

Server side

SPARQL server (SPARQLite, Tomcat, Apache)
RDF cache (Jena TDB)

FlyBase
BDGP
FlyTED
FlyAtlas
AffyMetrix
Creating RDF from data sources

- **D2RQ mapping**
  - FlyBase and BDGP, native relational databases
  - Conservative mapping, with minimum interpretation

- **OAI2SPARQL**
  - Harvesting N3 RDF metadata via the OAI-PMH protocol, built-in support by Eprints
  - Further from ESWC2008 paper

- **Custom Python program**
  - FlyAtlas
  - Generating N3 from spreadsheet table
Performance

- **Loading:** Our datasets ~175 million triples
  - Jena / TDB gives much better load performance (~15-30K tps), on 64 bit system with Amazon EBS storage (~3hrs)

- **Querying:**
  - Good enough for real time user interaction, e.g., <1s for single gene search, 1-4s for multigene search (unions)
  - No significant slowdown when scale from 10m to 175m triples

- **Text matching and case insensitive search**
  - Problems with using SPARQL regex filter, the only mechanism for case-insensitive search in SPARQL
  - Pre-generated lower-case gene names and loaded into the FlyBase RDF DB
  - Tried with OpenLink Virtuoso, still ~10 seconds for a case-insensitive search
TCM-LODD: Background

- Connect the knowledge about alternative medicine and western drugs
- Demonstrate the value of Linked Data
- Demonstrate a novel technique for creating interlinks between datasets on a large scale
- A joint effort of the BioRDF and LODD (Linked Open Drug Data) task forces of the World Wide Web Consortium (W3C) Health Care Life Science Interest Group
Demo

- Search for herbs associated with a particular disease … [go]
Benefits of SW technologies

- RDF provides a uniform and flexible data model
  - RDF dump is cheaper and quicker
  - Maintaining a separate SPARQL endpoint for each data source makes it easier than a data warehouse approach for handling data updates
- RDF facilitates data re-use and re-purposing
- SPARQL raises the point of departure for an application
  - Expressive, open-ended query protocol
  - Support for unanticipated queries
 Costs & Risks

- Mapping data to RDF requires expertise and experience
- Expressive query protocol is a double-edged sword
- Performance is good for some queries, not for others...
Web creator job 'beyond politics'

Sir Tim Berners-Lee has told the BBC that the job he has been given by Gordon Brown is an important one that goes beyond party politics.

The inventor of the world wide web has been asked by the prime minister to help open up access to government data.

"I think there's a public demand for transparency. This is way beyond party politics and beyond global borders," Sir Tim said.

“So that government information is accessible and useful for the widest possible group of people, I [Gordon Brown] 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.”
Further information

- About Linked Data
  - http://linkeddata.org/
  - http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
- About the projects
  - http://www.flyweb.info
  - http://esw.w3.org/topic/HCLSIG/AlternativeMedicineUseCase/
Acknowledgements

- Alistair Miles, Graham Klyne and David Shotton
- Anja Jentzsch, Matthias Samwald, Kei-Hoi Cheung and others from W3C HCLSIG
- Dr Helen White-Cooper and her research group
- BBSRC for funding building the FlyTED database
- BDGP, FlyAtlas and FlyBase for making the data available
- JISC, for funding the FlyWeb project
- The Jena team, esp. Andy Seaborne
- OpenLink Virtuoso
Thank you!

jun.zhao @ zoo.ox.ac.uk