

Open Data, Linked Data

Janet McKnight
Oxford University Computing Services

July 2010



Our panellists

- 1 Graham Klyne (Zoology)
- 2 Sally Rumsey (Bodleian Library)
- 3 Neil Jefferies (Bodleian Library)
- 4 Howard Noble (OUCS)



What is open data?

'A piece of knowledge is open if you are free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and share-alike.'

[From <http://www.opendefinition.org/>]



What is linked data?

'The semantic web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related data.'

[Tim Berners-Lee, from
<http://www.w3.org/DesignIssues/LinkedData.html>]



Four rules of linked data

- Use URIs for things
- Use HTTP URIs so that people can look up those names
- Return data using standards (RDF)
- Link to other things!



RDF is (stock description)

"The RDF data model is similar to classic conceptual modeling approaches such as Entity-Relationship or Class diagrams, as it is based upon the idea of making statements about resources (in particular Web resources) in the form of subject-predicate-object expressions. These expressions are known as triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object."

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

All true, but probably unhelpful, so let's try another approach...

“Connolly's Bane”

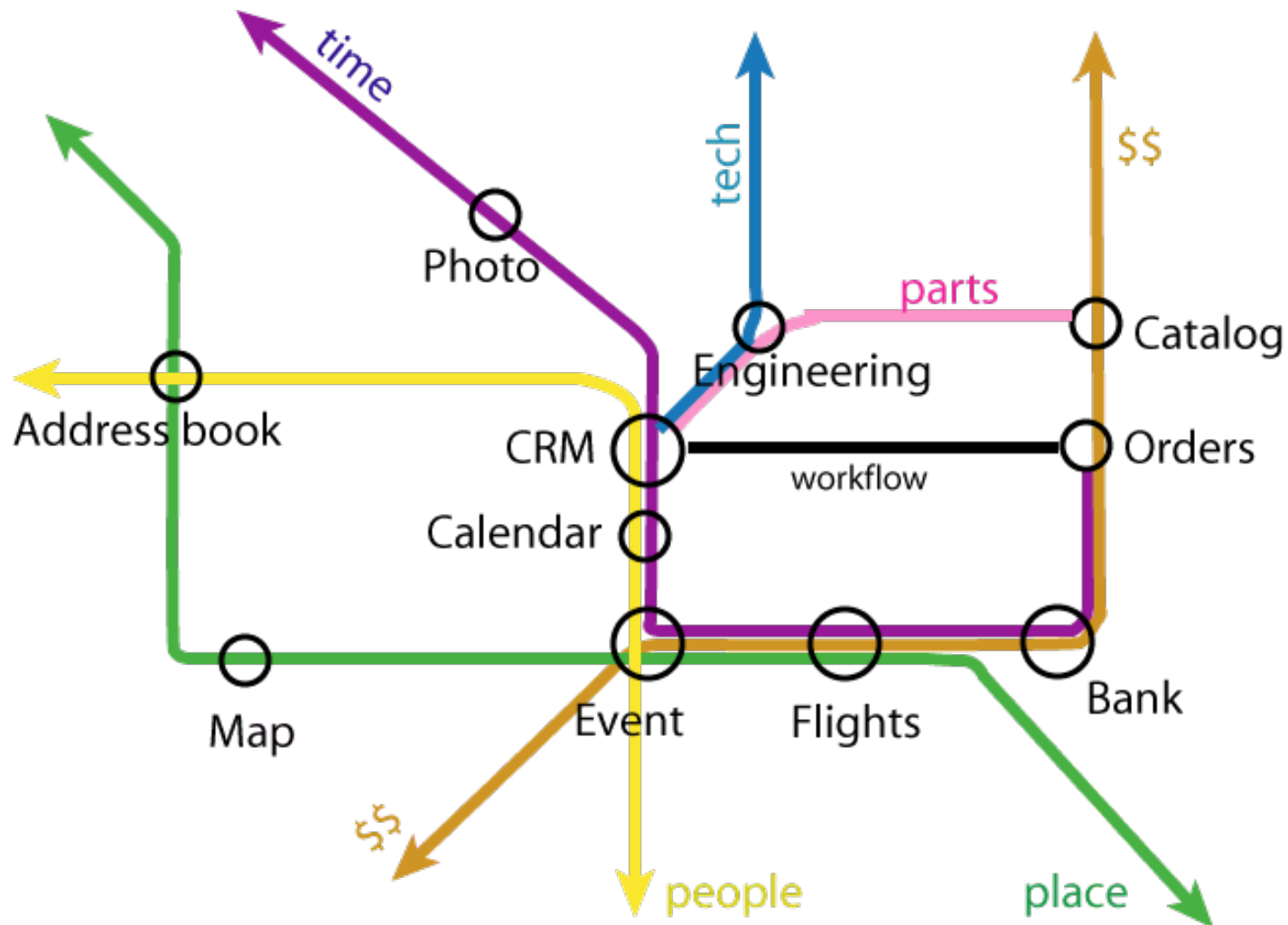
"The bane of my existence is doing things that I know the computer could do for me"

– Dan Connolly, <http://www.nature.com/nature/webmatters/xml/xml.html>

Many of these things involve combining information from diverse, independently-developed applications

“Semantic subway”

Combining diverse information



Tim Berners-Lee: <http://www.w3.org/2006/Talks/0314-ox-tbl/>

RDF is

"RDF is a general method to decompose **any type of knowledge** into **small pieces**, with some **rules about the semantics**, or meaning, of those pieces. The point is to have a method so **simple** that it can **express any fact**, and yet so **structured** that **computer applications can do useful things** with it."

<http://rdfabout.com/quickintro.xpd>

- a universal data format
- schema-free
- open-world
- uniformly structured
- uses global identifier namespace
- semantically formalized

CLAROS

Classical art data fusion

- Maps data from diverse independent data sources
- Search/browse over merged data

Universal – target for relational, XML, and other data

Schema-free – enhancing data structures incrementally

Open world – new data sources

Uniform – merging via RDF

Global names – URIs for objects

Semantics – inference engine

The screenshot shows the CLAROS website interface. At the top, there is a search bar and navigation tabs for Home, CLAROS, Pottery, Gems, Sculpture, Iconography, Antiquaria, Dictionary, and Tools. The main content area is titled 'Timeline for calyx krater' and features a bar chart showing the number of occurrences in each period. Below the chart is a map of the region with a callout for CAPUA, indicating 15 occurrences. To the right, there is a 'Summary results' section listing museum holdings, including the Ashmolean Museum in Oxford, the National Museum in Stockholm, and the British Museum in London. The bottom of the page includes a footer with copyright information and a last updated date of 13 May, 2008.

<http://www.clarosnet.org>

ADMIRAL

Small-scale research data curation

- Research data curation
- “Curation by addition”

Start with what is available, figure out what we've got, enhance incrementally until ready for publication

universal – can represent arbitrary (meta)data

scheme-free – no foreknowledge of structures needed

open-world – can add new data, missing isn't broken

uniform structure – merge (meta)data from diverse sources

global names – link to public datasets; open publication

- <http://imageweb.zoo.ox.ac.uk/wiki/index.php/ADMIRAL>

Use of linked data within ORA and the research registry

Three things you should know

- What?
- Why?
- Plans

Sally Rumsey, ORA Service & Development Manager

sally.rumsey@bodleian.ox.ac.uk





Oxford University Research Archive (ORA) contains research publications and other research output produced by members of the University of Oxford. Content includes copies of journal articles, conference papers, theses and other types of research publications. The full text of many of these items is freely available to be used in accordance with copyright and end-user permissions.

Oxford University Research Archive is a growing repository of Oxford research publications and is therefore not a complete record of the research output from the university. Members of the University of Oxford may deposit items in Oxford University Research Archive. For more details contact us.

Newest Additions: (follow via [Twitter](#))

Yeast forms dominate fungal diversity in the deep oceans - Journal Article	David Bass, Alexis Howe, Nick Brown, Hannah Barton, Maria De... et al	Biological sciences, Plant Sciences	2007	Peer Reviewed
Migration and stopover in a small pelagic seabird, the Manx shearwater Puffinus puffinus: insights from machine learning - Journal Article	T. Gullott, J. Meade, J. Willis, R. A. Phillips, D. Boyle, ... et al	Biological sciences, Engineering & allied sciences	2009	Peer Reviewed
The universality and demarcation of lexical categories cross-linguistically - Oxford Thesis	Lindsay A. Morcom	Linguistics, Indigenous peoples	2009	
The fantasy of reunion: the rise and fall of the Association for the Promotion of the Unity of Christendom - Journal Article	Mark D. Chapman	Theology and Religion, Christianity and Christian spiritual ...	2007	Peer Reviewed
Does God know what it is like to be me? - Journal Article	William J. Mander	Philosophy, Theology and Religion, Philosophy, psychology and ...	2002	Peer Reviewed
Omniscience and pantheism - Journal Article	William J. Mander	Philosophy, Philosophy, psychology and sociology of religion	2000	Peer Reviewed
The place of the officer-offender relationship in assisting offenders to desist from crime - Journal Article	Ros Burnett, Fergus McNeil	Law, Criminology	2005	Peer Reviewed
Experiencing modernization: frontline probation perspectives on the transition to a National Offender Management Service - Journal Article	Deen Robinson, Ros Burnett	Law, Criminology	2007	Peer Reviewed
Paid annual leave and the long-term sick: third time lucky for the United Kingdom? - Journal Article	Alan Bogg	Law	2007	Peer Reviewed
Employment Relations Act 2004: another false dawn for collectivism? - Journal Article	Alan L. Bogg	Law	2005	Peer Reviewed

<http://databank.ouls.ox.ac.uk/>



Oxford University Library Services' Databank

Oxford DataBank is a small collection of non-bibliographic works created by Oxford researchers. The purpose of DataBank is to provide a pilot version of a robust and efficient system for the safe storage of and open access to (where appropriate) Oxford research data. Content could be described as 'data produced as a result of academic research.' Items might comprise files such as spreadsheets, databases, audio files and images (still and moving). Numerical or other data might be raw data or the data may have been manipulated or processed in some way.

DataBank is not intended to store large-scale data sets such as grid data or other vast data sets. Neither is it intended to replace national, subject or other established data collections. It is envisaged that the future role of DataBank will be to provide a secure and accessible store of small to medium-sized files which have no other option for similar safe storage. The role of DataBank will be clarified as Oxford University works towards a solution for the storage of and access to its research data. Please note DataBank is a pilot system and still under development.

Users should adhere to the terms and conditions stated for each item.

Datasets

Latest Datasets:

Title	Created by	Date	Rights
Dataset: Tick1 audio corpus	Greg Kochanski	2010-03-04T13:58:46.334Z	Unestablished
Robert Darnton at Oxford	Robert Darnton	2009-06-19T14:34:54.235Z	Copyright 2009

Copyright resides with the noted authors/creators of the works held here, unless otherwise noted.

Vocabulary

@ OULS - a space for vocabularies.

Introduction

The following vocabularies are currently hosted here:

[Academic Research Project Funding Ontology \(ARPEO\)](#)

A vocabulary for describing the structure and relationships between organisations and individuals involved in providing funding or receiving funds for academic research or projects.

HTML produced from the toolchain at vocab.org - a copy of which can be downloaded (with example rdf from the funding vocab) [here](#).

<http://vocab.ox.ac.uk>

<http://163.1.127.171/index>

Research and Researchers Blue Pages



The Blue Pages makes it easier than ever before to find information about research and researchers at Oxford.



Search

Search the Blue Pages

Exact phrase Any of these words

Or browse:

People

Principal investigators, research staff...

Research Projects

Research projects, clinical trials, surveys...

Funders

Research councils, sponsors...

Academic Units

Divisions, departments, colleges...

Why?

Semantic web used

- De facto standard
- Reflect complex structure
- Easy data sharing & re-use

Problems

- Aggregation
- Sharing
- Not a database
- Creating linked data version

Plans

- Future developments
- Timing



UNIVERSITY OF
OXFORD

Google Libraries Project Bodleian Update

Neil Jefferies
R&D Project Manager
Systems & eResearch Service (SERS)
Bodleian Libraries, Oxford University

Source Material

- Over 400K digitized books, approx 300 pp
 - After dedup, pruning of collection
- JPEG2000 images + raw OCR text + “metadata”
- Volume packages generated on-demand
- 50-100TB of raw data
- Currently held on Google servers

Ingest Process

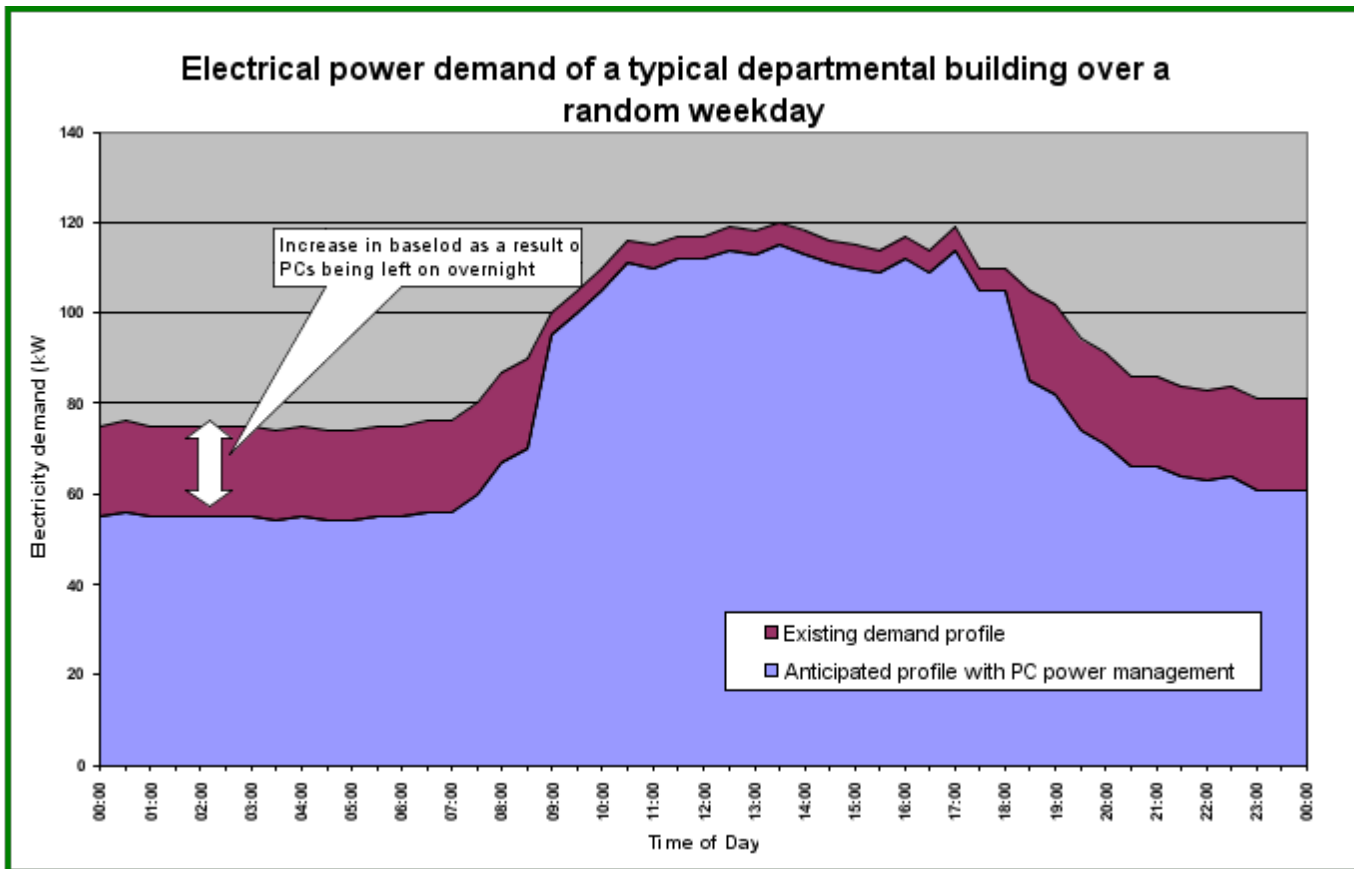
- Download volume packages
 - Bandwidth limited, checksum to validate overall package
- Unpack and validate
 - Check against manifest, zero length files, missing pages etc.
 - Manual sampling/QA of content
- Cross-reference with OLIS for bib metadata
- Generate dissemination formats
 - J2K → JPEG → PDF (Raw book vs image+OCR)
- Stored as objects with RDF descriptions

Dissemination

- Human-readable formats
 - Browse, faceted search (SOLR), page-turner, e-Citation
- Machine-readable formats
 - RDF/Linked data, OAI-PMH/ORE, RSS, Atom
 - Links to author/place thesauri such as CERL
- Added value services
 - Open Annotation (www.openannotation.org)
 - Annotation, translation, transcription – as RDF objects
 - PDF's including annotations?
 - Print-on-demand?

ICTF 2010

Open data



Three *policies*:

1. Always on: $110/1000 \times (24 \times 365) \times 16,000 \times \text{£}0.12 = \text{£}1,850,112$
2. Power down after work: $110/1000 \times (10 \times 223) \times 16,000 \times \text{£}0.12 = \text{£}470,976$
3. Hope for the best = £?

The social side of green IT

Consider a world where:

- 20% of people have changed their lives to reduce their energy consumption significantly (eco-warriors)
- 30% of people consume as much electricity as they can afford (sceptics and deniers)
- 50% of people consume as they see the people around them consuming (migrating birds)

Question: What effect do the eco-warriors have on total energy consumption?

- A. Total consumption will decrease
- B. Total consumption will stay the same
- C. Total consumption will increase

Social Capital + Open Data



Social Capital + Open Data



Questions?
Over to you...



Thank you!

- opendata@maillist.ox.ac.uk

- janet.mcknight@oucs.ox.ac.uk
- graham.klyne@zoo.ox.ac.uk
- sally.rumsey@bodleian.ox.ac.uk
- neil.jefferies@bodleian.ox.ac.uk
- howard.noble@oucs.ox.ac.uk

