Doing data in the social sciences and humanities: links to and from published work

Peter Burnhill

Director, EDINA national academic data centre,
University of Edinburgh, Scotland UK

Beyond Books: What STM & Social Science publishing should learn from each other
Marriott Hotel/Kensington, London, 22 April 2010

EDINA®

Overview

1. A bit about EDINA
2. Research data & research publications
3. All that is digital are not data
4. Autobiography as brief commentary on data facilities
   + Digital library, Information Science & the two traditions
5. Citation and linking
   - with switch and intro Linked Data
     * Semantic Web anyone?

If there is time

6. Suggestions about who should / could do what
   - Researchers/Authors, Editors, Publishers
   - Universities, Data centres, Libraries, Curators

EDINA®

EDINA®

providing resources for staff and students in higher and further education in the UK and beyond

Services Select Go Projects Select Go

→ Reading & Reference

Article References

- → CAB Abstracts
- → Land Life Leisure

Journal Catalogues

- → SUNCAT (UK) 🔓
- → SALSER (Scottish) 🔓

E-books

- → The Statistical Accounts of Scotland 🔓

Deposit Academic Papers

- → the Depot

→ Maps & Data

Geo-data Portal

- → Go-Geo! 🔓

Maps & Datasets

- → Digimap - Ordnance Survey
- → Geology Digimap
- → Historic Digimap
- → Marine Digimap
- → UKBORDERS

Agricultural Census Data

- → agcensus

Web Services

- → Unlock

Deposit Data

- → ShareGeo facility

→ Multimedia & Education

Film, Images & Sound

- → Education Image Gallery
- → Film & Sound Online
- → NewsFilm Online

Learning Materials

- → Jorum User

Deposit Learning Materials

- → Jorum Contributor

News, Training & Events

EDINA gets social
13 April 2010

Digimap and OS OpenData
5 April 2010

AddressingHistory tool combines historical data and maps
1 April 2010

all news

Quarterly Newsletter

Follow us

Ways to contribute online ▶

Ensuring continuous access ▶

Our projects and middleware ▶

For library and support staff ▶

🔓 = *open to all*

What services can I use?

EDINA is a JISC National Data Centre based at the University of Edinburgh | Accessibility | Acknowledgements

JISC

Reading and Reference Room

In mid-90s, our strategy was based on **hosting key A&I databases** (Art Abstracts, BIOSIS, Compendex, Inspec etc)

- but market changed: commercial rush for retail frontage

Since 2002 we have been re-making our future with:

- SUNCAT, UK national union catalogue of serials
- **National OpenURL Router, as registry of OpenURL resolvers in use**
- Technical (metadata) Operator for UK Access Management Federation
  - Investigated Shibboleth for JISC and developed SDSS pilot
- Digital preservation as part of **'continuity of access'**
  - CLOCKSS Access Host for orphaned content; Edinburgh University as Archive Node
  - Technical support for UK LOCKSS Alliance cooperative
  - Piloting an e-journals preservation registry, with ISSN-IC **[PEPRS]**
  - Working with JISC Collections for assured access to back issues of e-journals
- supporting JISC with e-learning …
  - (with Mimas) developing and managing Jorum, **repository** of learning materials

having already **diversified with GeoSpatial and Multimedia resources**

EDINA®

2. research data & research publications

A. "We Need Publishing Standards for Datasets and Data Tables", OECD Publishing White Paper, OECD Publishing. T. Green (2009)

A. Nature Editorial: Data's shameful neglect, Nature, 461, p.145. (2009, September)

"Three major responsibilities are covered:

1. **preservation** of the original data on which the paper is based,
2. **verification** that the figures and conclusions accurately reflect the data collected and that manipulations to images are in accordance with Nature journal guidelines, and
3. **minimisation of obstacles to sharing** materials, data & algorithms through appropriate planning."

Researchers' viewpoint: a cultural shift?

"You are not finished until you have done the research, published the results, *and* published the data, receiving formal credit for everything."

Mark A. Parsons (2006)
International Polar Year

Preserve or Perish

"A scholar's positive contribution is measured by the sum of the original data that he contributes. Hypotheses come and go but data remain."

in *Advice to a Young Investigator* (1897) Santiago Ramón y Cajal
(Nobel Prize winner, 1906)

EDINA®

3. All that is digital are not data (& vice versa)

a) Data derive importance from their evidential value
   - the empirical base for (scholarly) statement & decision-making
   - Provenance (where data comes from) is very important

b) Differences in ways that disciplines in Humanities & Social Sciences assess scholarship and evidence
   - In what they regard as data, as value for their subject
     * Arts: performance
     * Humanities: long view (including history/philosophy of science)
     * Social Sciences: Big Societal Challenges; flirt with policy
   - mix of approach to phenomenology, inc document tradition

c) Data represented (encoded) as numbers or words
   - often derived from observation (with issues of ontology!)
   - or as pictures or sounds (not encoded - pre-data?)
   - or algorithmic models (as with physical & life sciences)

Our shared task:

To ensure ease & continuing access to record of scholarship

- research publications and research data

Consider at least three types of (research) data:

A. Supplementary data

- multimedia files: part of the published article that presents research argument and conclusions
  - more than linear text, limited tabular and graphical display
  - enhances user experience with various multimedia objects

B. Research dataset(s) upon which conclusions based

- check analysis of those data to support statements made

C. Database(s) from which datasets were assembled

- for reproducibility (exposure to refutation) and new work via alternative analysis and updates to the database(s)

EDINA®

4. autobiography as commentary: data facilities

1. **Scottish Education Data Archive,** late 1970s – mid '80s
   - Survey statistician: for school leaver, YTS & 16-19 cohort surveys

2. **Edinburgh University Data Library,** mid- 1980s & on
   - Manager: set-up and development

3. **ESRC Regional Research Laboratory** for **Scotland** 1986/90
   - Co-director: early days of Geographical Information Systems (GIS)

4. **EDINA national data centre,** mid-1990s to present
   - Director: set-up and continuous development

5. **Digital Curation Centre,** 2004 & 2005
   - Interim Director: set-up & 'data curation' & 'digital preservation'

Began as a data manufacturer

- **Scottish Education Data Archive,** late 1970s – mid '80s
  - Survey statistician: for school leaver, YTS & 16-19 cohort surveys
    * Database of derived data made available online, used for Government statistics
    * Successive survey data -> trend datasets, changing classifiers (eg Social Class)

comment

- This was based in a research centre at University of Edinburgh
  - Prototypical of what is now widespread, in universities &research institutes
  - The data, **curated as databases**: the **working capital** for research group
  - There was access by others, but as 'privileged access' [join our gang]
  - There is always/often threat to continuity because of funding

Became a data broker

- **Edinburgh University Data Library,** mid- 1980s & on
  - Manager: set-up and development
    * A library of datasets and analysis software
      - social surveys (Govt & academic), economic series + Population & Agricultural Censuses
    * Providing **ease of access to data held elsewhere**
      - eg UK Data Archive; Oxford Text Archive

Comment

- Focus on data for the social sciences, public health and rural studies
- Demand-driven, for ***secondary data analysis***
    * Could not generate the data they needed to address their questions
    * Could not command the resources (funding/expertise)
      - few research groups and Government could get funding to manufacture original data

Became a data broker

- **Edinburgh University Data Library,** mid- ‘80s & on
  - * A library of datasets and analysis software
  - * Providing **ease of access to data held elsewhere**

Comment

- IASSIST International Association for Social Science Information Service & Technology
  - * annual conference; www.iassistdata.org; Past President, 1997/200

- Words, as text full of meaning, came into view via the Text Encoding Initiative (TEI)
  - – a document markup language, SGM ISO 8879:1986 SGML
    - * precursor to HTML, DTD and XML

- EUDL plays lead role in DISC-UK, a group of data libraries in UK universities
  - * Datashare project to support institutional responsibilities for data
    - – *alongside* Institutional Repositories

EDINA®

DISC-UK DataShare Project Edinburgh, LSE, Oxford, Southampton

to formal institutional arrangement

from informal storage and sharing

Quality assured metadata; guidance available for depositors; suitably anonymised/consent for sharing obtained from subjects; thorough documentation about data creation and methodology included; permanent IDs; formats validated and suitable for distribution; migration-based preservation commitment	Network of distributed repositories: subject and/or institutionally based
Data files with minimal documentation (e.g. readme file describing each data file) downloadable from Internet	Zip and ship Open access
Metadata record of dataset on website or in repository; possibly with embargo and contact information to request access	Search and discovery enabled; restricted access
Networked drive, available to research group, version control	Email dissemination by request
Password protected, networked drive (backup procedures)	Privileged access
Personal hard drives, un-networked	Simple data storage

Robin Rice, September 2007

DataShare exemplars aiming here

Institutional Repositories

Typical status quo

Robin Rice, Data Librarian, University of Edinburgh

a move into interesting spaces

- **ESRC Regional Research Laboratory** for **Scotland** 1986/90
  - Co-director: early days of Geographical Information Systems (GIS)
    - Integrating ‘large-scale’ data, mainly geographic or geo-spatial

Comment on the now:

- Recurrent focus on the geo-spatial
  - Resurgence of interest, launch of EDINA Digimap in 2000
  - MultiMap, StreetMap, GoogleMap; location-based services
  - Geo-tagging, mobile phones, cameras, social websites
  - EU INSPIRE directive: all public bodies, including universities
- Part of overall strategic purpose
  - to build the academic spatial data infrastructure
    - “over 75% of all research resources are geo-spatial” anon.
  - to enhance discoverability of online resources
  - to provide context for the analysis of data
    - geo-parsing (to extract place names from documents)
    - geo-tagging (to ensure names have geo-feet)

❖ *Unlock the place in your online resource!*

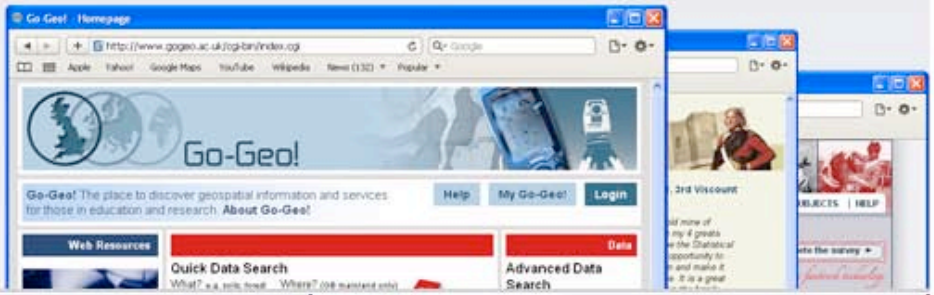
EDINA®

Unlock

Questions? Contact Us

home | places | geocodes | text/places | register

Unlock the potential in your data with our simple web services



places

Find the locations of place names

Find >>

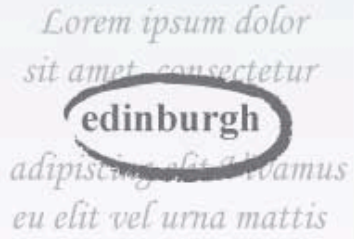
geocodes

Convert postcodes, grid references, etc to coordinates

Convert >>

text/ places

Extract placenames from text and find their locations



Extract >>

EDINA

JISC

Move into national data services & data curation

- **EDINA national data centre,** mid-1990s to present
  - Director: set-up and continuous development
    - **online access** to wide range of A&I/bibliographic, multimedia & OS mapping data
    - **national repositories of digital content**: Jorum learning materials & ShareGeo

Comment on the now:

- **Digital Curation Centre,** 2004 & 2005; now in its Phase 3
  - Interim Director: set-up/strategy for 'data curation' & 'digital preservation'
    - even wider range of databases (e-science), held by others
- Growth of data-driven science
  - importance of the data curator for managed open databases
- Growth of institutional and subject repositories
  - mostly research papers but increasingly research data
    - DataShare (Edinburgh, LSE, Oxford, Southampton)

Re-stating our shared task:

To ensure ease & continuing access to record of scholarship

- research publications and research data

Consider at least three types of (research) data:

A. Supplementary data
   - multimedia files: part of the published article that presents research argument and conclusions
     * more than linear text, limited tabular and graphical display
     * enhances user experience with various multimedia objects

B. Research dataset(s) upon which conclusions based
   - check analysis of those data to support statements made

C. Database(s) from which datasets were assembled
   - for reproducibility (exposure to refutation) and new work via alternative analysis and updates to the database(s)

EDINA®

5. Citation, then linking

- Citation of database(s) (Type C data)
  - for reproducibility (exposure to refutation)
  - to prompt new work via alternative analysis and updates to the database(s)
  - to credit those who curate the data needed for scholarship

- Citation of the datasets used (Type B data)
  - verification of analysis, that the figures and conclusions accurately reflect those data

Plus hyperlink *to* the dataset *from* the published article

… and back again *from* the dataset *to* the published article

\+ Links to presentations, blogs, websites, funders etc related to the same research activity and same researcher(s) (Type D data?)

EDINA®

Standards to cite data (A long running saga)

There is no universal standard for citing data and computer files, but …

- Dodd, Sue. (**1979**) "Bibliographic references for numeric social science data files: Suggested guidelines." *Journal of the American Society for Information Science*, 30 (2), 77-82.
  - ISO 690: **1987** Bibliographic references - Content, form and structure
- Dodd, Sue. (**1990**) "Bibliographic References for Computer Files in the Social Science: A Discussion Paper." Chapel Hill, NC: Institute for Research in Social Science, University of North Carolina. presented to IASSIST 1990 Poughkeepsie, N.Y. http://www.people.virginia.edu/~pm9k/info/compRef.html
  - ISO 690-2: **1997** Bibliographic references, Part 2: Electronic documents
- Schneider, Jeri. (**2006**) "Why we need a data citation standard: Lessons learned from compiling ICPSR's Bibliography of Data-Related Literature." *ICPSR Bulletin*, 26 (2), 9-12. http://www.icpsr.umich.edu/org/publications/bulletin/spr06.pdf

EDINA®

Obtaining the citation at source

1. CIESIN

   "Most of our datasets and products contain a suggested citation on the Web site as to where the data was obtained"

   "Whenever possible, we urge you to cite the use of data and web resources in the reference section"

   – http://sedac.ciesin.columbia.edu/citations/

2. How to Cite Statistics Canada Products:

   "This guide has been developed for authors, editors, researchers, academics, students, librarians and data librarians.

   "It describes, in three steps, how to build your reference when citing Statistics Canada products"

   – http://www.statcan.gc.ca/pub/12-591-x/12-591-x2006001-eng.htm

Get it from those who make the data available: the data publishers

cf Cataloguing in Publication!

EDINA®

Data registration, citation & identifier initiatives

- DataCite: an international consortium
  - easier access to scientific research data on the Internet
  - increase acceptance of research data as legitimate, citable contributions to the scientific record, and
  - support data archiving that permits results to be verified and re-purposed for future study.
    * http://www.datacite.org
- ANDS: Australian National Data Service
  - Identify My Data service
  - to persistently identify your data
    * http://ands.org.au/services/identify-my-data.html
- Identifiers for authors/creators
  - Open Researcher and Contributor ID (ORCID)
  - NAMES, EU Interparty, ISNI, VIAF

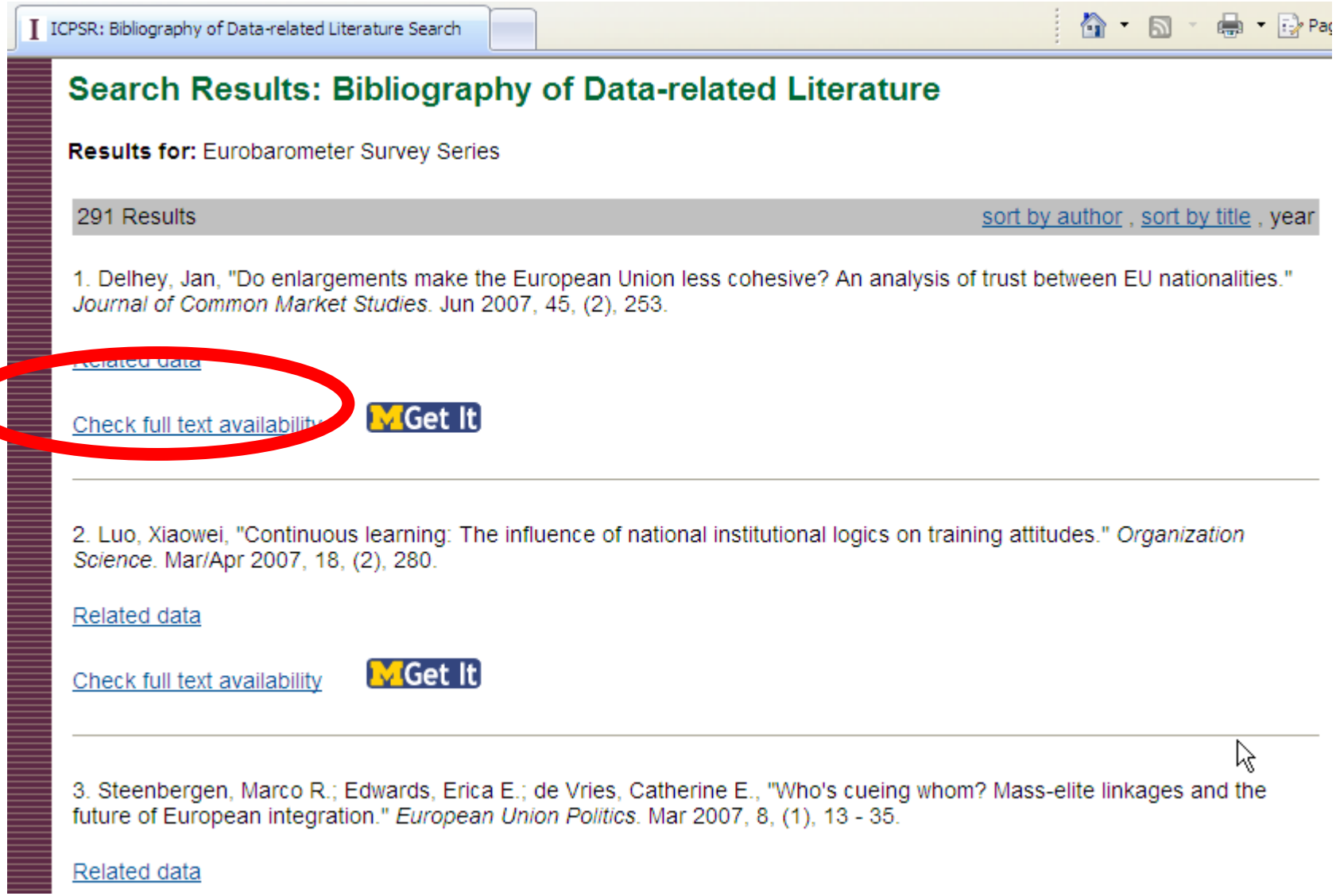
EDINA®

Examples of hand-crafted, hard-coded linking

hyperlink from the published article *back to* the dataset

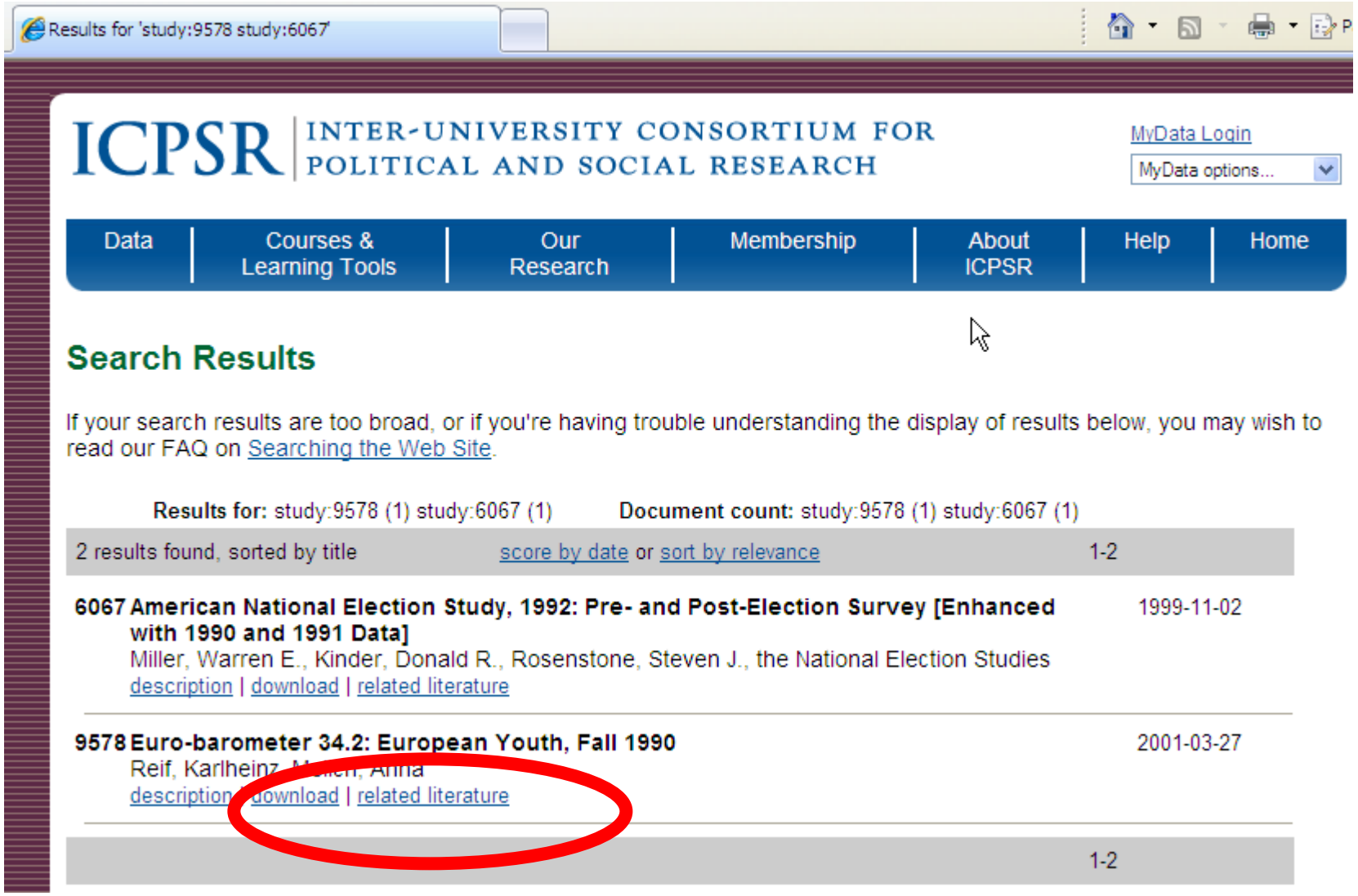
... and *forward* from the dataset to the published article

search on bibliography and hyperlink to data



Myron Gutmann, Inter-university Consortium for Political & Social Research

From data to (subsequent & known) published literature



Works with supplemental files

from “Dissertations, Data Sets and ProQuest UMI”, Austin McLean, IASSIST, May 2008

14. **Actin organization and activity in lily pollen tube growth: A role forpH**
by *Lovy-Wheeler, Alenka*, Ph.D., **University of Massachusetts Amherst**, 2006, 142 pages; AAT 3216956
Abstract | 24 Page Preview | Full Text - PDF (8 MB) | Supplemental files | Order a copy

15. **An ethnographic account of language documentation among the Kurripako of Venezuela**
by *Granadillo, Tania*, Ph.D., **The University of Arizona**, 2006, 235 pages; AAT 3207639
Abstract | 24 Page Preview | Full Text - PDF (2 MB) | Supplemental files | Order a copy

16. **"Book": A graphic score for any musicians or non-musicians**
by *Redman, G. William*, Ph.D., **State University of New York at Buffalo**, 2006, 106 pages; AAT 3213624
Abstract | 24 Page Preview | Full Text - PDF (48 MB) | Supplemental files | Order a copy

17. **Education under enemy occupation: Experiences of selected Dutch students during World War II**
by *King, Rebecca Celia Maria*, Ed.D., **Boise State University**, 2006, 218 pages; AAT DP15570
Abstract | 24 Page Preview | Full Text - PDF (814 K) | Supplemental files | Order a copy

18. **ESL learners' perceptions of American dialects**
by *Damann, Melissa M.*, M.A., **The University of North Carolina at Chapel Hill**, 2006, 73 pages; AAT 1435003
Abstract | 24 Page Preview | Full Text - PDF (657 K) | Supplemental files | Order a copy

27. **Speech enhancement using transient speech components**
by *Tantibundhit, Charturong (Paul)*, Ph.D., **University of Pittsburgh**, 2006, 180 pages; AAT 3224052
Abstract | 24 Page Preview | Full Text - PDF (7 MB) | Supplemental files | Order a copy

28. **The calabozo: Virtual reconstruction of a prison cell based on personal accounts**
by *Aroztegui Massera, Carmen*, Ph.D., **Texas A&M University**, 2006, 240 pages; AAT 3219138
Abstract | 24 Page Preview | Full Text - PDF (9 MB) | Supplemental files | Order a copy

29. **The Chevalier de Saint-George: His violin style and eighteenth-century musical aesthetics**
by *Zinck, Bernard F.*, D.M.A., **Temple University**, 2006, 207 pages; AAT 3211909
Abstract | 24 Page Preview | Full Text - PDF (12 MB) | Supplemental files | Order a copy

What about supplementary data (Type A data)?

ProQuest

Basic | Advanced | Browse | My Research 0 marked items

Databases selected: Dissertations & Theses

Document View « Back to Results

Print | Email | Copy link | Cite this | Mark Document Translate

The stylistic diversity of the concert saxophone

by *Rubinoff, Daniel I.*, M.A., **York University (Canada)**, 2007, 84 pages; AAT MR32037

Abstract (Summary)

This thesis examines the sonic parameters and musical versatility of the concert saxophone. Invented in 1840, the instrument failed to become a regular member of the symphony orche and is thus underrepresented in classical music. This researcher argues that the saxophone's unique sonic design makes it an effective contemporary instrument in a wide variety of ger Specifically, the techniques of subtone, harmonics, and false fingerings are examined from both a performance and compositional perspective. Additionally, the instrument's resemblanc the human voice is documented.

An examination of five original saxophone compositions highlights the instrument's flexibility as a solo instrument or as a member of an ensemble. This work adds to the number of origi compositions for the saxophone and explores the reasons behind the instrument's success in contemporary music.

Supplemental files

Some files may require a special program or browser plug-in. More information

- Track_1.mp3 (4MB)
- Track_2.mp3 (3MB)
- Track_3.mp3 (6MB)
- Track_4.mp3 (4MB)
- Track_5.mp3 (4MB)
- Track_6.mp3 (5MB)
- Track_7.mp3 (5MB)
- Track_8.mp3 (3MB)

Indexing (document details)

School:	York University (Canada)
School Location:	Canada
Keyword(s):	with Original composition

UMI
Dissertation Publishing

How supplemental files appear

ProQuest

Basic | Advanced | Browse | My Research 0 marked items

Databases selected: Dissertations & Theses

Document View « Back to Results

Print | Email | Copy link | Cite this | Mark Document

Concert dance and social/political discourse in the arts

by Snodgrass, Jeanne L., M.F.A., **The University of New Mexico**, 2007, 50 pages; AAT 1444784

Abstract (Summary)

This paper looks at concert dance, dance presented in a traditional theatre venue, that chooses to present works that deal with social and/or political issues. This exploration i serve as a broad-spectrum analysis of all dance engaging with politics or as a comparison of political versus non-political dance. Rather I have chosen to look at a selection of modern dance artists that have successfully broached social and political issues in their work in an artistically critical manner. This paper examines the roles of form, content the creation of these works and the ways in which the methodology employed by these artists may be used as a guideline for other dance choreographers wishing to work in t

Specifically, this paper explores these issues by: looking at the debate over the value of politicizing art; framing the historical context through an examination of the New Dance 1930s and Judson Dance Theater of the 1960s; contemplating contemporary artists Bill T. Jones and Liz Lerman; and looking at my own work, an evening of three dances deal social/political themes, in terms of choreographic intention, methodology and audience reaction.

I believe that the presentation of topic driven works can serve to reinvigorate both contemporary dance artists and audiences and it is my hope that this exploration will encoura choreographers and performers to see concert dance as a viable venue for engaging an audience in social and political discourse."

"This dissertation is a compound document (contains both a paper copy and a CD as part of the dissertation). The CD requires the following system requirements: Windows M RealPlayer.

Supplemental files

Some files may require a special program or browser plug-in. More information

11.zip (1572MB zipped)
archive for component

Indexing (document details)

Information Science has had (other) ideas ...

1. World Wide Web
   - intended for resource sharing by/for a science community
     - **took off in wider world in way that we all know**
   - Putting the Web to work: for our related business / industry
     * **'appropriate copy problem'** for digital library / publishing
       - **OpenURL**
         » linking between the A&I/reference world and online source(s) of the full text of the (digital) article

2. Re-working the Web: adding new weft and weave
   - The social networking (web 2.0) thing
     * user generated content, tagging and collaborative spaces
   - **The semantic web (web 3.0) thing** – machine as user

Emergence of Digital Library: Information Science

Michael Buckland, Presidential Address, American Society for Information Science, JASIS's 50th (1998)

- 2 traditions/mentalities co-exist in Information Science
  1. **Document tradition**: signifying record-ness
  2. **Computational tradition**: various uses of formal techniques

     * non-convergent mentalities working to build the 'digital library'

     a) modernisation of library services

     b) infrastructure to access complex databases

Link remains the key verb

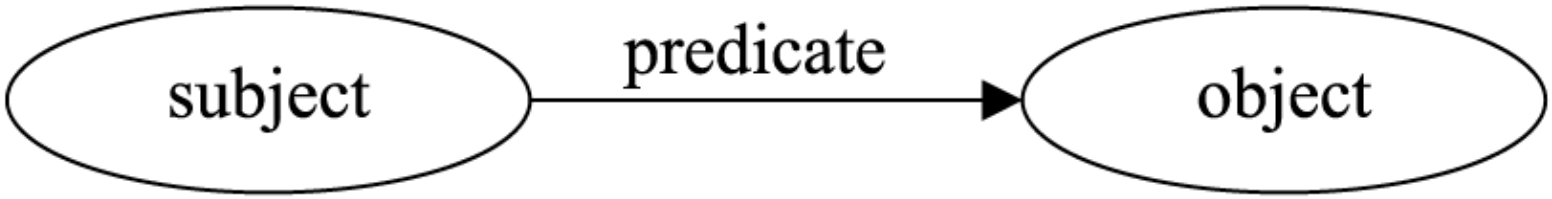
But need to shift attention from

- ***Linking resolver*** (unidirectional)
  - From metadata reference to full text of article
    * SICI-Citation | Z39.50
    * DOI | OpenURL | http

to

- ***Linked Data*** (relational, bi-directional)
  - Between resources in the weave of the Web
    * Using URIs as names for things
      - Not just URLs (the addresses on the web) but the URIs
    * Using RDF/XML to define the relationships between the resources
      - RDF triples: subject / relationship / object

Resource Description Framework (RDF)



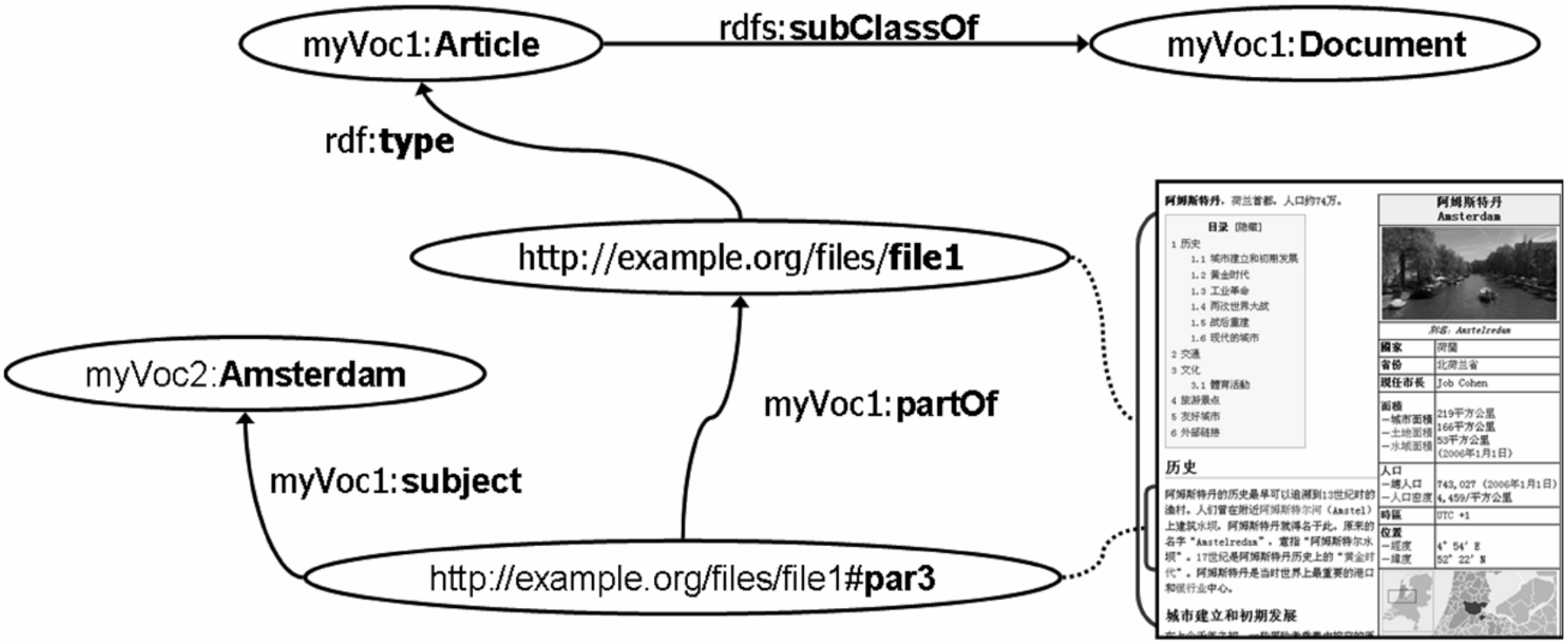
Source: W3C (2004)

Resource Description Framework (RDF), and URIs

- framework for representing information in Web; identifiers
  - http://www.w3.org/TR/rdf-concepts/
  - http://www.w3.org/TR/rdf-primer/

RDF graph: Article & Supplementary Data

http://www.emeraldinsight.com/fig/0350570303002.png



1. Build and publish as metadata in XML format to be found on the web
2. Publishing text and data/multimedia content in XML will delight researchers
   - Researchers want to access ‘article as data’, via computational algorithm

EDINA®

Linked Data ...

A note from Tim Berners Lee now in circulation proposes 4 steps:

1. Use URIs as names for things
2. Use http URIs so that people [& computers?] 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.

- may become the principles/rules/definition of 'Linked Data'

OpenURL to OAI-ORE

Note that the man who gave you OpenURL …

- "Reference Linking in a Hybrid Library Environment. (Part 1: Frameworks for Linking)", **Herbert Van de Sompel** and Patrick Hochstenbach

  D-Lib Magazine ISSN 1082-9873 Volume 5 Issue 4 April 1999

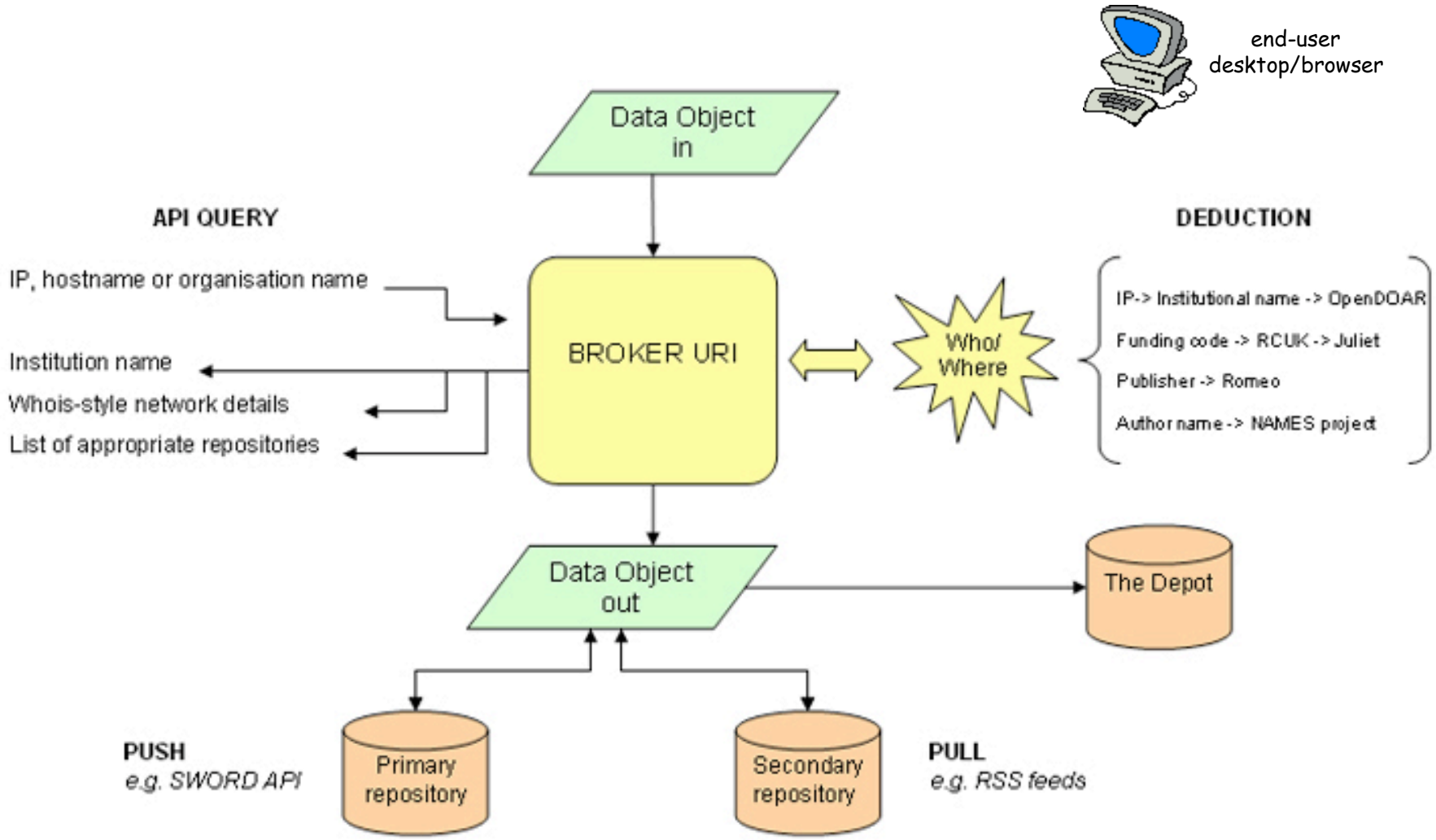
… is now into Linked Data

- "Adding eScience Assets to the Data Web", **Herbert Van de Sompel**, Carl Lagoze, Michael L. Nelson, Simeon Warner, Robert Sanderson, Pete Johnston

  Proceedings of Linked Data on the Web (LDOW2009) Workshop, [v1] Thu, 11 Jun 2009 15:33:37 GMT
  http://arxiv.org/abs/0906.2135v1

EDINA®

‘Repository Junction’



A broker to discover nodes for deposit

- for long-term stewardship and added services
- for others to re-analyse for (secondary) research purposes

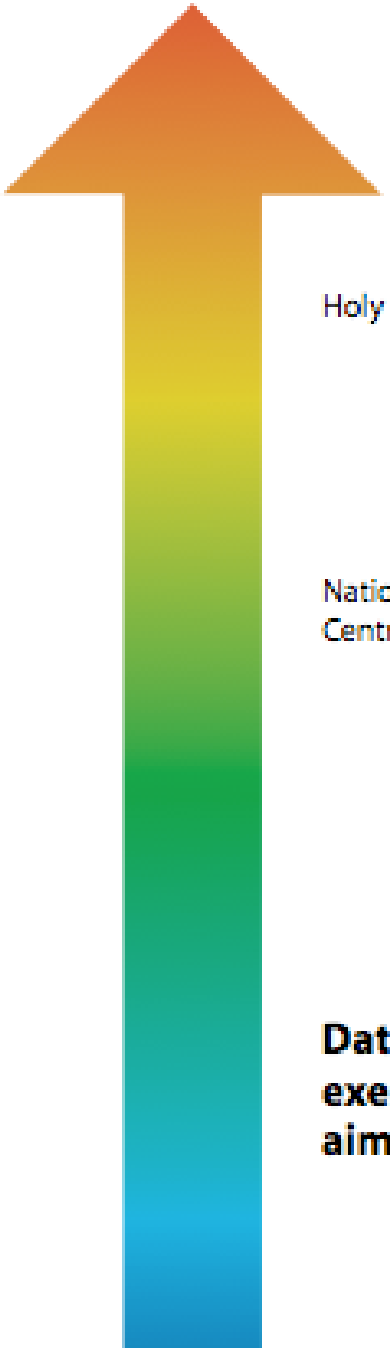
DataShare2

DISC-UK DataShare: Data Sharing Continuum

to formal publishing into (linked) data infrastructure

Distributed high performance computing; analysis tools applied to data over secure international network; M2M interfaces	Data Grid
Peer review of datasets; seamless link to publications; role-based layers of access; data overlay journals	Data publishing
Graphs, charts, maps configurable online	Data visualisation
"Actionable" marked up dataset installed in a data browser tool subsetting capability	Data manipulation online
Original format plus XML markup of data or XML database; open standards used appropriate to domain; metadata or setup files may be bundled with dataset for importing elsewhere	Data enhanced for re-use
Quality assured metadata; guidance available for depositors; suitably anonymised/consent for sharing obtained from subjects; thorough documentation about data creation and methodology included; permanent IDs; formats validated and suitable for distribution; migration-based preservation commitment	Network of distributed repositories: subject and/or institutionally based

from formal institutional arrangement



Holy grail

National Data Centres/Archives

DataShare exemplars aiming here

Time for me to stop ...

Hoping that I have left some space/place for questions

- Thank you

Acknowledgements

All at EDINA & to Robin Rice, Stuart Macdonald and Nicola Osborne in particular

p.burnhill@ed.ac.uk
http://edina.ac.uk
Tel.: +44 (0)131 650 3302
Fax: +44 (0)131 650 3308

EDINA®