Digital Libraries and Open Access: Interoperability Strategies

Open Repositories and Interoperability Challenges in UK

Peter Burnhill

Director, EDINA National Data Centre,
University of Edinburgh, Scotland UK

DL.org Workshop, The British Academy, London UK, 4 February 2011

EDINA

+ expertise, in areas such as geo-enabling, access management, etc

research, learning & teaching in UK universities & colleges

↑

acting as platform for network-level services
& helping to build the JISC Integrated Information Environment

National Data Centres
EDINA MIMAS

↑

JISC Collections — JISC — JISC Sub-Committees

↑

UK funding councils
hefce Department of Education www.deni.gov.uk hefcw Scottish Funding Council Research Councils UK

autobiography as commentary on data facilities

Began as a data manufacturer

- **Scottish Education Data Archive,** late 1970s – mid '80s
  - Survey statistician: for school leaver, YTS & 16-19 cohort surveys
  - **curated as databases**: the **working capital** for research group + 'guests'

Became a data broker

- **Edinburgh University Data Library,** mid- 1980s & on
  - Providing **library and ease of access to data held elsewhere**
  - Connected to **IASSIST**, international group for data librarians/archivists

Learning about interesting spaces [time/place referencing]

- **ESRC Regional Research Laboratory** for **Scotland** 1986/90
  - Co-director: early days of Geographical Information Systems (GIS)
    - * Inter-Agency Committee on Global Environmental Change Data Task Force

Moved into national data services; learnt more about data curation

- **EDINA national data centre,** mid-1990s to present
  - Director: set-up and continuing development of access/deposit services
- **Digital Curation Centre,** 2004 & 2005
  - Director: set-up & digital curation = 'data curation' & 'digital preservation'

Spot the repository!

EDINA

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)

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

document tradition & computation tradition

Emergence of Digital Library: Information Science

"considerable simplification, … helpful to think … of two traditions, or mentalities, even cultures, co-exist in area of Information Science

1. "Approaches based on a concern with documents, with signifying records: archives, bibliography, documentation, librarianship, records management, and the like

1. "approaches based on uses for formal techniques, whether mechanical (such as punch cards and data-processing equipment) or mathematical (as in algorithmic procedures)."

Michael Buckland (UC Berkeley), Presidential Address, American Society for Information Science, JASIS's 50th (1998)
http://people.ischool.berkeley.edu/~buckland/asis62.html

Semantics of 'Open Repositories & Interoperability'

- R is for Repository
  - "university-based institutional repository is a set of services … for the management and dissemination of digital materials created by the institution & its community members. … organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access ..." (C. Lynch, 2003)
  - Digital repository differs from other digital collections in that:
    * "content is deposited, whether by content creator, owner or third party
    * "architecture manages content as well as metadata
    * "repository offers a minimum set of basic services
    * "must be sustainable & trusted, well-supported & well-managed"

    Digital Repositories Review (R.Heery and S.Anderson, 2005)
- O is for Open
  - OA (for publications) not the only 'open' policy:
    * OER: Open Educational Resources
      - Open means 'not closed': making teaching & learning materials visible
      - Open CourseWare – often as open stack of webpages
    * Open Data
      - Datasets tradition (IASSIST); 'open/privilege access to databases; open data.gov
    * Open Source Software
      - OSS has its own way of doing things

'Open Repositories & Interoperability' [in the UK]

Heretical and Haphazard Thinking About The Brief …

- Are Repositories the (only) way to support an Open Agenda?
  and
- Is Open really what Repositories are for?
  Or
  Is this usage just intended to help us avoid issues of IP and access management?

And should the focus be on:

- Interoperability between Repositories?
  or
- Interoperability of Repositories with the wider Internet?

EDINA 9

Interoperability Strategies; Interoperability Challenges

- Whose strategy, and towards what purpose?
  - 'within & for the research & education sector'? Or beyond?
  - for the institution, the UK, EU, global anybody; for the researcher?
  - for the machine as user ["Provider/Consumer"]
- Interoperability as technical [& semantic] means to support interworking by persons or systems has challenges
  - policy/technology/infrastructure/management/metadata
  - Internet engineering & semantic web
    - Repositories and Linked Data

Beyond PUT, KEEP and GET of the singular Repository

- Connecting Repositories with
  - NOTIFY & EXCHANGE; TRANSFER of objects or metadata only
    - So that the content of the object (or its metadata) may be re-used
- Linking & Identifiers Really Matter!
  - and Registries have a key role as authorities and cross-walks

EDINA 10

Maybe can we agree our shared & central task...

- to ensure ease & continuity of access to (online/digital) scholarly resources
  - for researchers, students and their teachers, *now and into the future*

My perspective

- In a University-based organisation (EDINA) that is a provider of content services & infrastructure services within national (UK) policy framework
  - number of content services based on use of repository software
    - Eprints, DSpace, IntraLibrary
  - number of infrastructure activities
    - OpenDepot.org, OA Repository Junction; OpenURL Router
  - member of SONEX and indirectly of COAR and UK-CORR
- Later focus on on repository-related progress in the UK; where is the value, how this is assessed/expressed?
  - Size of investment in recent times; cost-effectiveness and 'impact'
  - Effort at institutional & inter/national level and the 'shared services' agenda?

EDINA 11

Nostalgia for Days of Plenty as we worry about the future

JISC as well-funded agent for change

- JISC Repositories & Preservation Programme - April 2006; March 2009
  "£14m investment in H.E. repository and digital content infrastructure"
- A drive to assist institutions, including JISC RepositoryNet?
  - Repository Support Project; Repository Research Project
  - Intute Repository Search; 'interim repository' | the Depot | OpenDepot.org
- Services/Tools like OpenDOAR; Romeo; SWORD; OARJ
- Check the JISC website
  - http://www.jisc.ac.uk/whatwedo/programmes/inf11/sue2.aspx
  - under the heading of 'key digital repository activities' are 21 funding programmes and 226 funded projects.

& then there were many meetings, including a new 'regular' street event:

- RepoFringe2010: Repository Fringe 2/3 September, Edinburgh

EDINA 12

SONEX

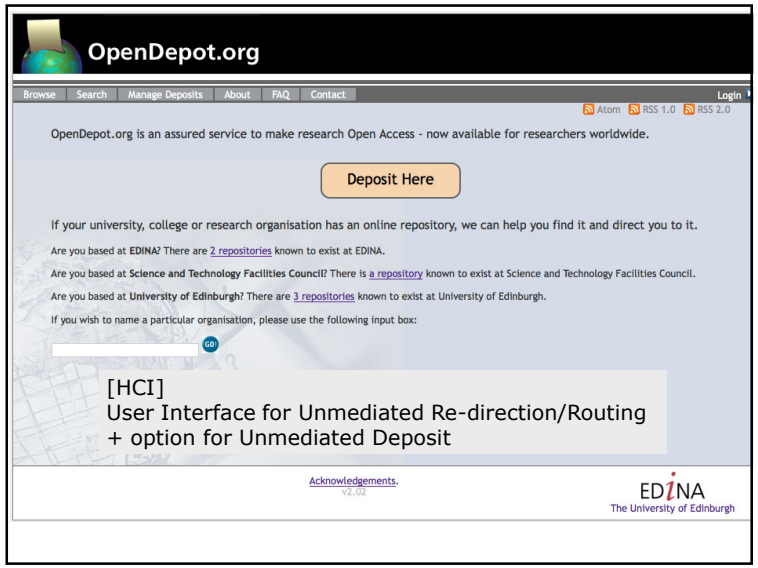
- **Began with focus on 'deposit opportunities'**
  - categorisation of repository types into which authors deposit
- **Began with research paper use case:** multi-author & multi-institutional
- **Looking at onward interoperability** (SWORD) not just technical interoperability but workflow

Day Job: as provider of services/tools & user of software

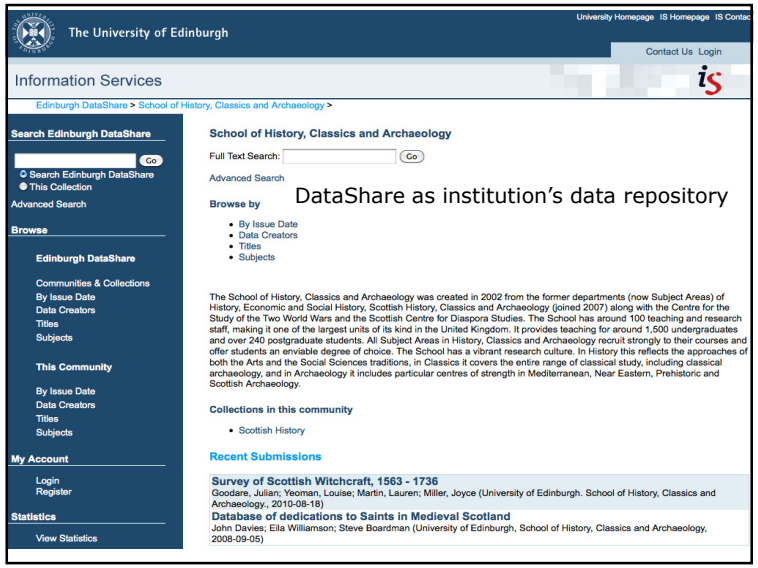
- EDINA and Edinburgh University Data Library run repositories, with and without JISC
  - **Jorum**: for learning materials [with Mimas]
    - OER and turnstile (UK); using DSpace
  - **OpenDepot** (the Depot): for research papers
    - OA (world); using Eprints
  - **ShareGeo**: for geo-spatial data
    - Open Data and turnstile (UK); using DSpace
  - **DataShare**: for research data (institutional, U of Ed)
    - Open Data; using DSpace
  - **OA Repository Junction** as shared service tool
    - using own code and Eprints as an 'escrow' repository during the transfer process.
- All seek to be 'standards-based' , reducing need to be mediated

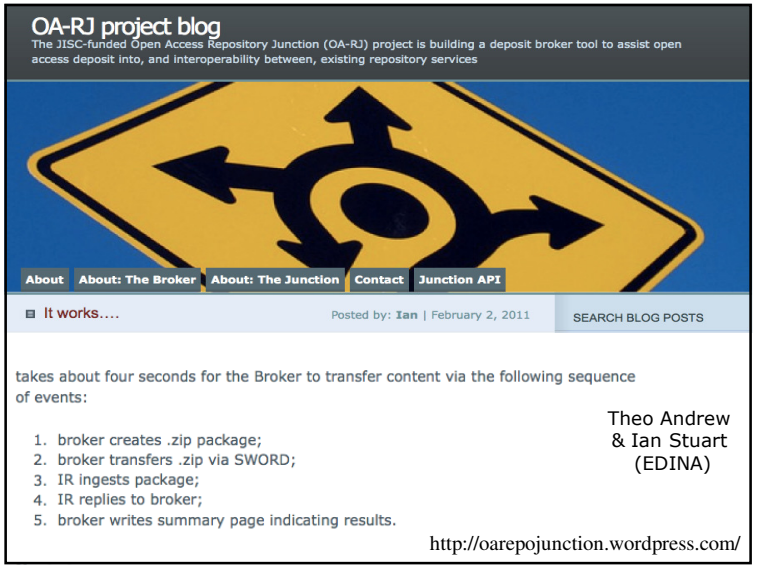
DSpace as open repository software

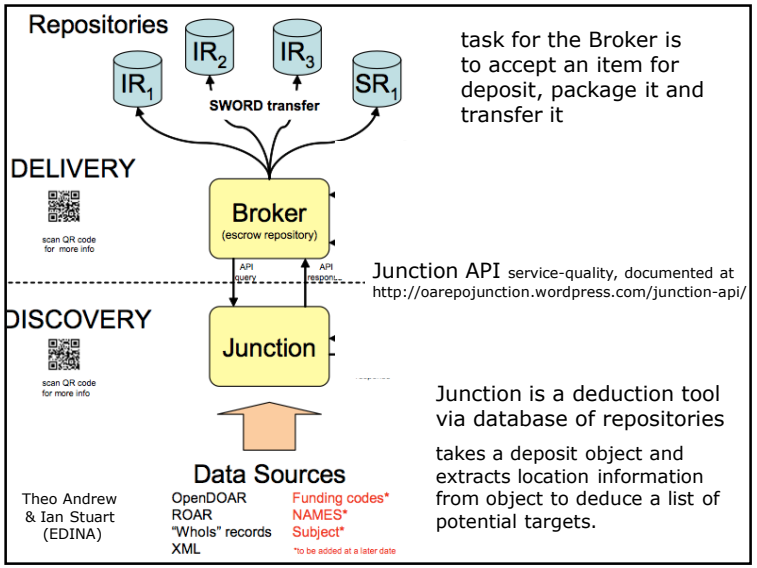
- Open Architecture: supports flexibility
- Active international development community
- Developer Community provides software add-on and pool of experience
  e.g. creating de-referenceable URIs (RDF/XML) for all in metadata store, to support both registry function and also engage with semantic web
- Implications for Jorum: self-deposit of learning materials
  - Particular requirements, as per Jorum Roadmap, based upon:
    - mediated ingest for multiple items (RSS for metadata; OAI-PMH for objects in IntraLibrary)
    - unmediated deposit (Selenium & SWORD/OARJ) in test-bed development
- cost of ownership, as per JISC OSS Watch:
  "*.. important that when procuring open source software solutions you also plan to properly resource collaboration work ..*"
  - need to embed developments back into DSpace codebase:
    - enables others to use, maintain and develop the new features
    - ensures that your extensions are in main code
    - increases the return on investment by making them available to all

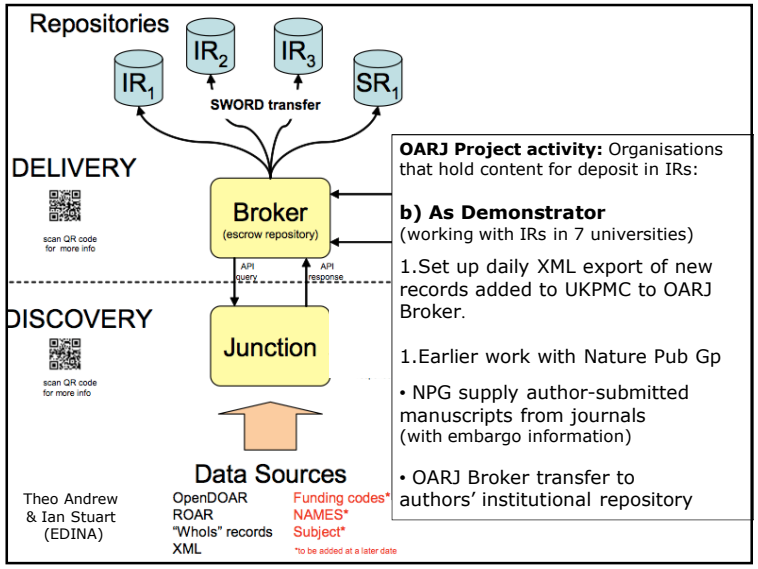

OpenDepot.org
Browse
Search
Manage Deposits
About
FAQ
Contact
Login
Atom
RSS 1.0
RSS 2.0
OpenDepot.org is an assured service to make research Open Access - now available for researchers worldwide.
Deposit Here
If your university, college or research organisation has an online repository, we can help you find it and direct you to it.
Are you based at EDINA? There are 2 repositories known to exist at EDINA.
Are you based at Science and Technology Facilities Council? There is a repository known to exist at Science and Technology Facilities Council.
Are you based at University of Edinburgh? There are 3 repositories known to exist at University of Edinburgh.
If you wish to name a particular organisation, please use the following input box:
GO!
[HCI]
User Interface for Unmediated Re-direction/Routing
+ option for Unmediated Deposit
Acknowledgements.
v2.02
EDINA
The University of Edinburgh

ShareGeo Open
EDINA
Home
Help | About | Contact Us | Login
Search ShareGeo
Search
Geo Search
Advanced Geo Search
Browse
All of ShareGeo
By Issue Date
By Depositor
Titles
Subjects
My Account
Login
Register
Contribute
The UK open source for your free geodata
Find data
Find and re-use geospatial datasets
Search for geospatial datasets by defining a search box on a map. Click on an items title to view its metadata and download the dataset.
Share your data
Share your geospatial datasets
Contribute geospatial datasets to share with others. A wide range of geospatial data formats are accepted.
ArcGIS Depositor Plug-in
Deposit data directly into ShareGeo from ArcGIS.
Download the new ArcGIS depositor tool now [.msi]
Download use guide [.pdf]
RSS Feed
Latest Items
UK Universities and Colleges (01 Feb 2011)
World Suicide Rates (31 Jan 2011)
World Political Boundaries - Detailed (19 Jan 2011)
3D Point data around Kingston University (19 Jan 2011)
Funded by
JISC
©University of Edinburgh
Accessibility
Terms of Use
Take Down Policy
DSpace


The University of Edinburgh
University Homepage
IS Homepage
IS Contac
Contact Us
Login
Information Services
Edinburgh DataShare > School of History, Classics and Archaeology >
Search Edinburgh DataShare
Search Edinburgh DataShare
This Collection
Advanced Search
Browse
Edinburgh DataShare
Communities & Collections
By Issue Date
Data Creators
Titles
Subjects
This Community
By Issue Date
Data Creators
Titles
Subjects
My Account
Login
Register
Statistics
View Statistics
School of History, Classics and Archaeology
Full Text Search:
Go
Advanced Search
Browse by
By Issue Date
Data Creators
Titles
Subjects
DataShare as institution's data repository
The School of History, Classics and Archaeology was created in 2002 from the former departments (now Subject Areas) of History, Economic and Social History, Scottish History, Classics and Archaeology (joined 2007) along with the Centre for the Study of the Two World Wars and the Scottish Centre for Diaspora Studies. The School has around 100 teaching and research staff, making it one of the largest units of its kind in the United Kingdom. It provides teaching for around 1,500 undergraduates and over 240 postgraduate students. All Subject Areas in History, Classics and Archaeology recruit strongly to their courses and offer students an enviable degree of choice. The School has a vibrant research culture. In History this reflects the approaches of both the Arts and the Social Sciences traditions, in Classics it covers the entire range of classical study, including classical archaeology, and in Archaeology it includes particular centres of strength in Mediterranean, Near Eastern, Prehistoric and Scottish Archaeology.
Collections in this community
Scottish History
Recent Submissions
Survey of Scottish Witchcraft, 1563 - 1736
Goodare, Julian; Yeoman, Louise; Martin, Lauren; Miller, Joyce (University of Edinburgh. School of History, Classics and Archaeology, 2010-08-18)
Database of dedications to Saints in Medieval Scotland
John Davies; Eila Williamson; Steve Boardman (University of Edinburgh, School of History, Classics and Archaeology, 2008-09-05)


OA-RJ project blog
The JISC-funded Open Access Repository Junction (OA-RJ) project is building a deposit broker tool to assist open access deposit into, and interoperability between, existing repository services
About
About: The Broker
About: The Junction
Contact
Junction API
It works....
Posted by: Ian | February 2, 2011
SEARCH BLOG POSTS
takes about four seconds for the Broker to transfer content via the following sequence of events:
1. broker creates .zip package;
2. broker transfers .zip via SWORD;
3. IR ingests package;
4. IR replies to broker;
5. broker writes summary page indicating results.
Theo Andrew & Ian Stuart (EDINA)
http://oarepojunction.wordpress.com/





OARJ Demonstrator with NPG

Est. Number Of Articles For Transfer During Six Month Period

Institutional partner	Articles in all NPG journals	Articles in participating NPG journals	Assuming author participation rate of 50%
Cambridge	1429	476	237
Oxford	962	321	160
Univ M**	499	166	83
Univ C**	275	92	46
Univ Y**	248	83	41
Edinburgh	194	65	32
Auckland	53	17	8
TOTAL	3660	1220	607

*based upon the number of papers published in journals in the NPG portfolio during Jan - June 2010, as recorded in PubMed Central and ISI Web of Knowledge. **Still to be confirmed as a participating institutions*

EDINA

http://www.rsp.ac.uk/start/software-survey/results-2010/

RSP (Nottingham) Repository Software Survey, Nov. 2010

Registered in OpenDOAR		OAI-PMH	**SWORD**	SWAP	RDF	OAI-ORE	RSS	Atom	Other
301	EPrints	Y	Y	Y	Y	Y	Y	Y	
690	DSpace	Y	Y	Y	Y?	O	Y	Y	REST, SRU/SRW, OpenSearch, WebDav
	Islandora Drupal/ Fedora	O	**3rd P**	O	Y	O	Y	Y	SRU/SRW
17	Fedora	O	Y	O	Y	O			
4	EQUELLA Repos	Y	Y	O	O	Y	Y	Y	
	Zentity	Y	Y		Y/RDFS		Y	Y	
16 Hosted	Open Repository	Y	Y	Y			Y		
	Intra Library	Y	Y				Y		SRU/SRW
15	DigiTool	Y	(Y)						Z39.50
81 Hosted	Digital Commons	Y					Y		
18	CONTENT dm	Y							Z39.50

Y = Yes; O = Optional; ($_{\omega}$) = future/pending; ? = I think it's there

25

Have JISC [programme managers] moved on?

"Dealing with institutional processes now, rather than repository technology. Depending on type of content, the projects would fit much more closely in:

- managing research data programme
- research information programme
- open educational resources programme

as they have much more in common with those projects than they do with each other."

"repositories have found their core business proposition via the REF and making sure Universities list research outputs to obtain research ratings

- have not succeeded in making the business case that IRs should be doing the job of archiving, a core library platform, or the job of an institutional demonstrator/poster space.

Repositories fit in the 'University Enterprise Stack' by virtue of being a system that delivers a business solution to a real financial problem."

EDINA

26

Re-stating our shared task, to (re-)include data:

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

27

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

28

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®

Linked Data … Is this shared understanding?

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'

EDINA®

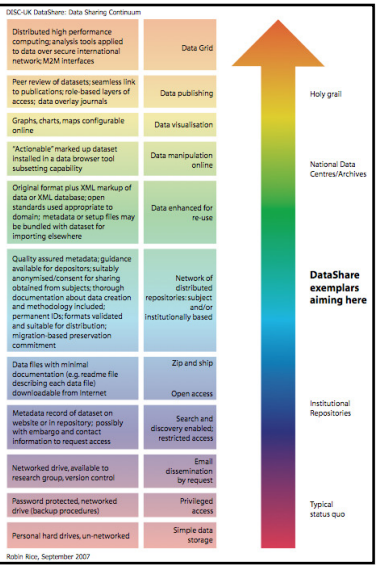
31

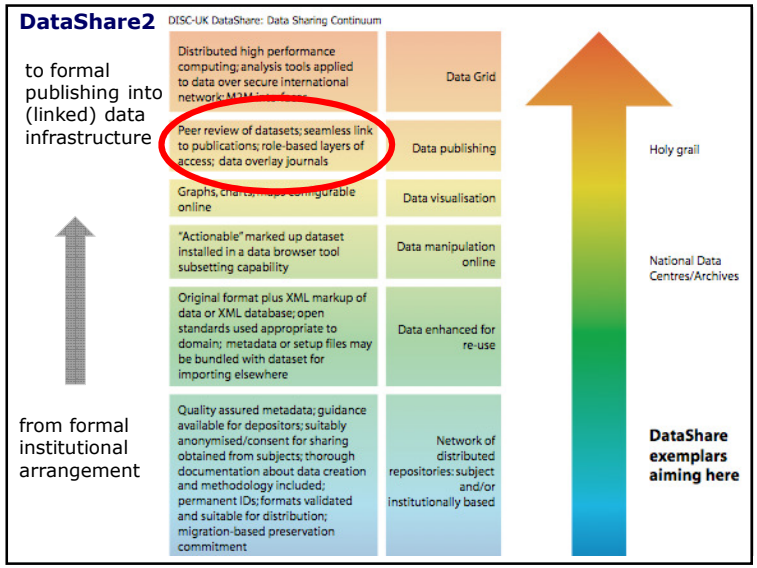
To formal publishing into data infrastructure

DISC-UK DataShare Project

(Edinburgh, LSE, Oxford, Southampton)

From informal storage and sharing





Time for me to stop

Hoping that I have left some space/place for questions

- Thank you

Acknowledgements

Theo Andrew & Ian Stuart, Pablo de Castro,
Gareth Waller & Robin Rice,
Dave Flanders & Andy McGregor

EDINA 34

Ease and Continuity of Access to Data in Difficult Times

End of an era? End of the R word?

1. Embedded in domain-specific processes, but with wider context
   a) Engage, connect and get leverage from Internet Engineering
      * W3C and the commercial/retail world
   b) Linkage and Identifiers really, really matter in m2m world
2. Moving from technology to policy & practice
   some domain-specific, some common to repositories
   a) Collection management: active curation & Linked relationships
      * versions, of data | articles | learning materials
      * Collections, 'see also' Curation as value-added linkage between items
   b) First point of public issue (availability); Take-down regimes
3. Institutional stewardship responsibility
   - for content that 'we need' for research and education
   - including data and other materials manufactured from within 'our world'
     * born-digital [and digitised] content
4. What of the (new) shared services imperative?
   - Who does what, at what level/scale?

EDINA 35

COAR: Confederation of Open Access Repositories

- 48 members drawn largely from Europe, but including both JISC & CNI, and also EDINA (University of Edinburgh)
- Work Plan for 2010/12, including
  1. Advocacy on behalf of OA ***and*** repositories (Rs) [both together?]
  2. **Populating (OA) Rs**
  3. Best practice documents
  4. **Facilitate and ensure data interoperability of (across?) Rs**
  5. **interoperability with other systems (such as CRIS systems)**
  6. Support national helpdesks
  7. Guidance on how Rs will form essential elements for global e-infrastructure
  8. Promote R manager profession
  9. Provide advice & guidance on suitable R infrastructure technologies
  10. Global (meta)data store
  11. Strategic partner other infrastructure-related initiatives worldwide

EDINA 36

Sound & Pictures: access to new data sources

- 20th Century is the first fully audio-visual century
  - With new forms of research material to use and to master
- EDINA as platform for downloadable film, video and audio
  - Licensed for use in learning, teaching and research
  - Wide range of subject coverage, including documentary film

1. Film & Sound Online
   - 600 hours of film, digitised for downloading
2. NewsFilm Online
   - 3000 hours of material from ITN & Reuters
   - Over 4TBs of clips to download

- Plus Education Image Gallery of still photography
- Visual and Sound Materials Portal
  - Discovering all sorts of audio-visual material

http://www.rsp.ac.uk/start/software-survey/results-2010/

Machine-to-Machine Interoperabilty

Feature	CONTENTdm	Digital Commons	DigiTool	DSpace	EPrints	EQUELLA Repository	Fedora	Islandora/ Fedora	IntraLibrary	Open Repository	Zentity
OAI-PMH:	current standard	current standard	current standard	current standard	current standard	current standard	current optional	current optional	current standard	current standard	current standard
OAI-ORE:	-	-	-	current optional	current standard	current standard	current optional	current optional	-	-	current standard
SWORD:	-	-	future standard	current standard	current standard	current standard	current standard	current 3rd party	current standard	current standard	current standard
SWAP:	-	-	-	current standard	current standard	current optional	current optional	current optional	-	current standard	-
RDF:	-	-	-	-	current standard	current optional	current standard	current standard	-	-	-
RoMEO Integration:	-	-	-	-	current 3rd party	current standard	-	current standard	-	future optional	-
OAI-PMH Harvesting:	-	-	current standard	current standard	-	current standard	current 3rd party	current 3rd party	-	-	-
Other:	Z39.50	-	z39.50	WebDAV, REST, SRU/SRW, OpenSearch	-	-	-	SRU/SRW	SRU/SRW	-	RDFS

Syndication

RSS:	-	RSS	-	RSS	RSS	RSS	-	RSS	RSS	RSS	RSS
Atom:	-	-	-	Atom	Atom	Atom	-	Atom	-	-	Atom

UK-CORR: UK Council of Research Repositories

individual rather than institutional, UKCORR-discussion@jiscmail.ac.uk

UK has 'rich heterogeneous repository landscape' (C.Awre); lurk

following comment from Dorothea Salo:

US mainly about OA full texts; UK mainly about … serving research assessment!

- Is there more to IRs than the REF: lots of bibliographic records & little full text?
- Should IRs only accept full text, not metadata only?

1. in absence of a CRIS, our IR had to do REF (Lancaster & Northampton)
2. was OA but then RAE2008, but should aim to include all (OU)
3. motive for IR was digital preservation, with different REF system; funder mandate compliance for OA; visibility via OA (Oxford/Bodleian)
4. RAE/REF is opportunity to engage institution-wide (Warwick)
5. Advent of CRIS (which don't manage outputs well) may be opportunity for IRs to have role, including use of 'metadata only' as lever to obtain full text (Hull)
6. REF & research management information allows IRs to be embedded as platform for OA (Southampton)
7. RAE/REF has different goals to OA and IRs with low % of full text may undermine OA movement (Nottingham)