Open Data:
projects, tools, initiatives

Stuart Macdonald
DISC-UK Datashare - http://www.disc-uk.org/datashare.html

EDINA National Data Centre & Edinburgh University Data Library

DataShare project

Seminario Sobre Datasets
Consorcio Madrono – 17
Nov. 2008

EDINA®

Overview

- EDINA National Data Centre
- Edinburgh University Data Library
- DISC-UK
- Data Deluge
- DISC-UK DataShare
- Data Audit Framework
- Harnessing Collective Intelligence
- Web 2.0 Data Visualisation Tools
- Citizens as Sensors

EDINA National Data Centre & Edinburgh University Data Library

- EDINA and University Data Library (EUDL) together are a division within Information Services of the University of Edinburgh.
- EDINA is a JISC-funded National Data Centre providing national online resources for education and research.
- The Data Library assists Edinburgh University users in the discovery, access, use and management of research datasets.

EDINA National Data Centre - *http://edina.ac.uk*

- Mission statement:

  "..to enhance the productivity of research, learning and teaching in UK higher and further education.."

- Networked access to a range of online resources for UK FE and HE

- Services free at the point of use for use by staff and students in learning, teaching and research through institutional subscription

- Focus is on service but also undertake R&D (projects ➢ services)

  - delivers about 20 online services
  - has about 10 major projects (including services in development)
  - employs about 75 staff, in two locations (Edinburgh & Merseyside)

Edinburgh University Data Library - first such service in the UK, started in 1983 – *http://datalib.ed.ac.uk*

Primarily within social sciences

Data Collection covering:

Large scale government surveys; Macro-economic time series; Financial time series; Population census data; Geospatial data

Specialised in data for Scotland and GIS

Teaching and (JISC-funded) projects

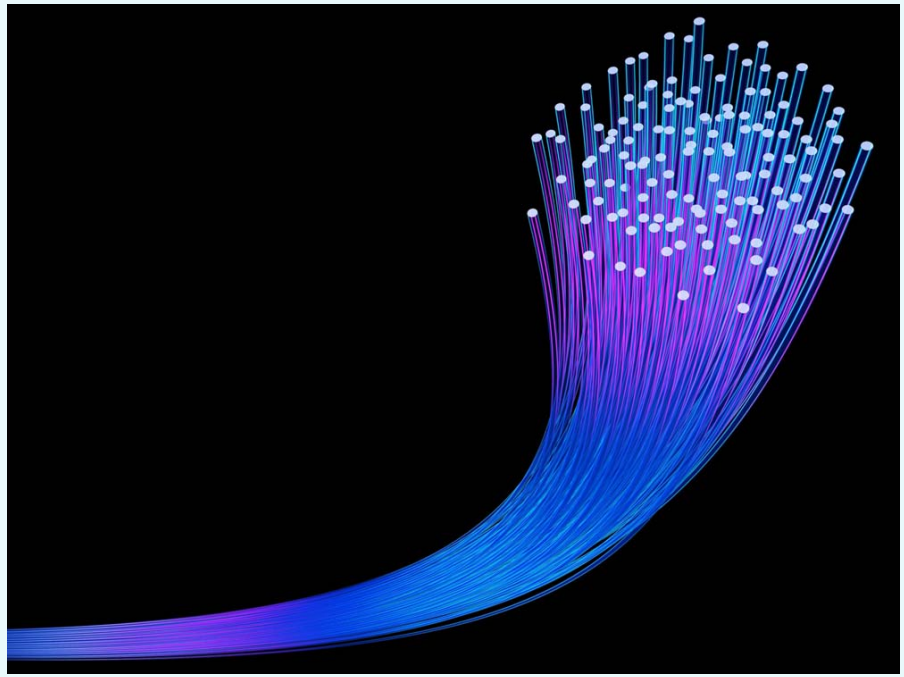
Institutional representative for national data services

Data Information Specialists Committee - United Kingdom (DISC-UK) – *http://www.disc-uk.org/*

- DISC-UK is a forum for data professionals working in UK Higher Education who specialise in supporting their institution's staff and students in the use of numeric and geo-spatial data.

- The aims of DISC-UK are as follows:-
  - Foster understanding between data users and providers
  - Raise awareness of the value of data support in Universities
  - Share information and resources among local data support staff

The Data Deluge



A recent IDC White Paper predicted that “between 2006 and 2010, the information added annually to the digital universe will increase more than sixfold—from 161 exabytes to 988 exabytes.”

*“The Expanding Digital Universe—A Forecast of Worldwide Information Growth through 2010"; www.emc.com/ahout/destination/digital_universe

“It is becoming increasingly clear that effective and efficient management and reuse of research data will be a key Component in the UK knowledge economy in years to come, essential for the efficient conduct of research ….”

*JISC (2008) “Identifying the benefits of curating and sharing research data” - http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/databenefits.aspx

DataShare project

EDINA®

DISC-UK DataShare

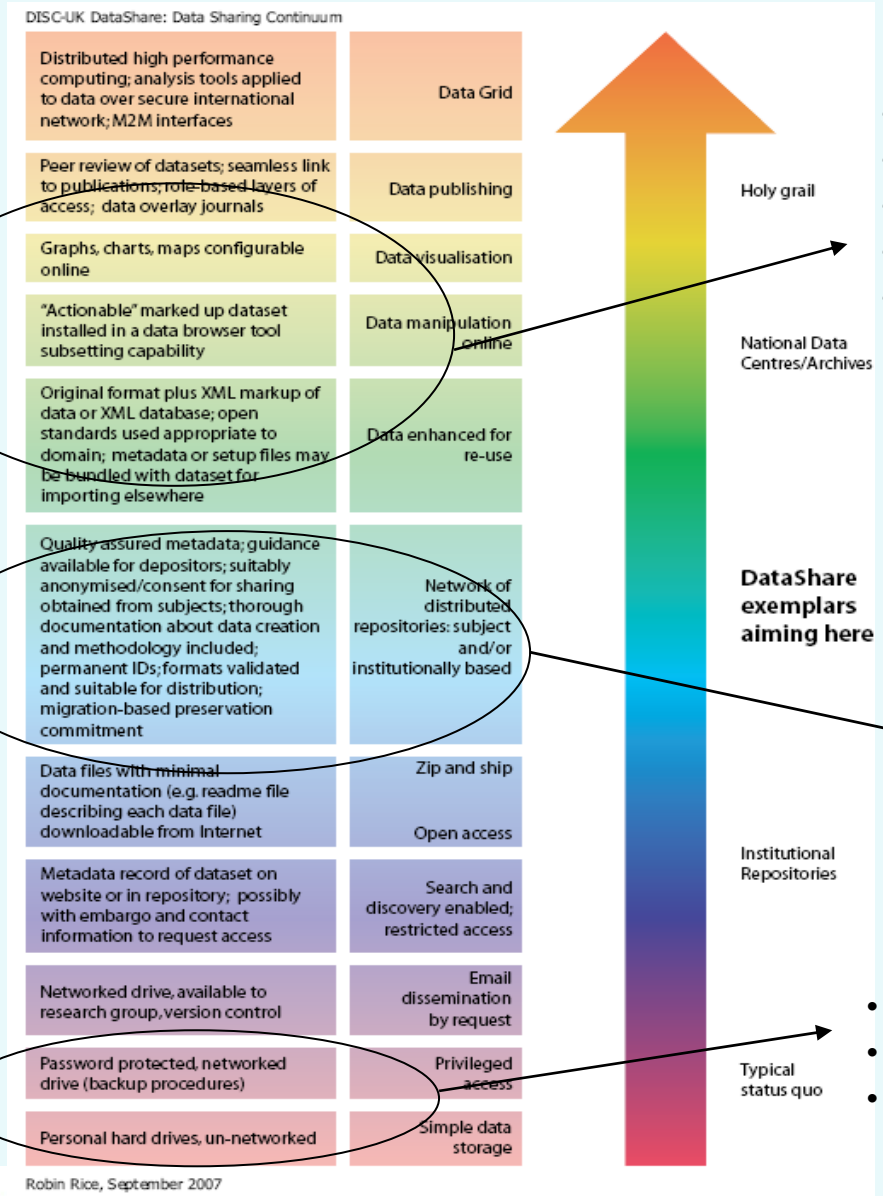
- Lyon* (2007) notes that, whilst many institutions have developed IRs over the last few years to store and disseminate their published research outputs, "…there is currently no equivalent drive to manage primary data in a co-ordinated manner."

- DISC-UK DataShare Project – funded by JISC (March 2007 – March 2009) - a collaborative project led by the University of Edinburgh, with the University of Oxford, the University of Southampton and the London School of Economics (Associate Partner).

- Investigate the legal, cultural and technical issues surrounding research data sharing within UK tertiary education community

- explore new pathways to assist academics wishing to share their data over the Internet via Institutional Repositories (IRs)

*LYON, L. (2007) Dealing with data: roles, responsibilities and relationships, Consultancy Report. Bath: UKOLN. - http://www.jisc.ac.uk/media/documents/programmes/digital_repositories/dealing_with_data_report-final.pdf



- Data visualisation and manipulation tools
- Learning & teaching materials
- Original format plus XML mark-up of data
- Open standards used relevant to domain
- Metadata & set-up files bundled with dataset

Data Sharing Continuum

- Quality assured metadata
- Guidance available for depositors
- Thorough documentation about data creation & methodology

- Data on flash drives, CD-ROMS, Hard drives
- Minimal metadata, local and trusted sharing
- No back-up plans

DataShare project

Seminario Sobre Datasets
Consorcio Madrono – 17
Nov. 2008

EDINA®

DISC-UK DataShare State-of-the-Art-Review

- Projects with related aims from which lessons may be drawn.
- Funders, publishers and institutions current requirements for data deposit and data sharing
- The current methods of research data storage/deposit/sharing.
- Intellectual Property Rights and issues arising from it.
- The benefits of and barriers to data sharing.

Barriers and Benefits to Data Deposit in IRs

- Barriers:
  - time
  - misuse
  - loss of ownership
  - IRs will cease to exist
  - unwillingness to change
  - IPR uncertainties
  - confidentiality

- Benefits:
  - reliable access to researchers own data
  - suitable environment to adhere to funders mandate
  - metadata increases exposure of individual's research within the community
  - preservation responsibility of institution rather than individual

Anticipated Outcomes and Dissemination

- Exemplars of the process, pitfalls and successful outcomes of setting up an institutional data repository service at each of the four institutions.
- Documentation and open source code for adapting DSpace, Fedora and EPrints repository software for handling datasets.
- Tools, presentations, briefing papers and other outputs to inform UKHE repository community about data management and research support.
- Web 2.0 dissemination of collected knowledge. (Social bookmarks, Tag Clouds, Blog & RSS)

Project Exit Strategy & Future Plans

- Expected to embed DSpace data repository in University's IR policy & procedures by end of project
- IS wanting to ramp up its research support
- Data-Audit Framework – JISC monies awarded to conduct an institutional data audit – 6 months project
- Data Library Service: traditionally for data users – try to do more for data creators - beyond soc sci?
- Link research data with associated ERA (Edinburgh Research Archive) citations
- Interface with ECDF / SAN - large scale compute and data facilities for the use of researchers within the University

Data Audit Framework - *http://www.data-audit.eu/*

- JISC funded a development project to create an audit framework and online tool and four implementation projects to test the framework and encourage uptake. (Apr – Oct 2008).

- DAF Development Project, led by Seamus Ross

  (Humanities Advanced Technology and Information Institute (HATII, Univ. of Glasgow)

- Four pilot implementation projects
  - King's College London,
  - University of Edinburgh
  - University College London
  - Imperial College London

DataShare project

EDINA®

Harnessing Collective Intelligence

Enabling cross-boundary & cross-disciplinary collaboration
Soft collision of ideas – create new knowledge to address global issues

Open Science and Open Research

- myExperiment – http://myexperiment.org/ – an open VRE enabling scientists to:
  - share digital items and workflows
  - build communities associated with their research
  - share expertise and avoid reinvention

- JoVE – Journal of Visualized Experiments - http://www.jove.com/ - publishing biological research in a visual format

- OpenWetWare - http://openwetware.org/ sharing experimental techniques and protocols, joining together labs and research groups in biology and biological engineering

- Open Science Project - http://www.openscience.org/ - mathematicians and engineers producing scientific software to encourage collaborative environments

Web 2.0 Numeric and Spatial Data Visualisation Utilities

- ‘Open’ ethos – anyone can upload/download or use data
- Commercial Services embracing Web 2.0 business models
- Collaborative
- Easy to use

But ...

- Ephemeral nature of web
- Not trusted repositories / archives
- if it’s computer generated and looks good it must be right!

- Palimpsest Project - http://research.google.com

Numeric Data Visualisation Tools:

- Data360 - (http://www.data360.org)
- Many Eyes - (http://services.alphaworks.ibm.com/manyeyes/home)
- Swivel - (http://www.swivel.com/)
- Gapminder – (http://www.gapminder.org/downloads/applications/)
- StatCrunch – (http://www.statcrunch.com/)
- Graphwise – (http://www.graphwise.com/)
- Numbrary – (http://numbrary.com/)
- Infochimps – (http://infochimps.org/home)
- Dabble – (http://dabbledb.com/)
- CKAN (incl. Open Economics) – (http://www.ckan.net/)

Spatial Data Visualisation - Tools

- Programmableweb - *http://www.programmableweb.com/tag/mapping*

 - map or spatial mash-up 'resource discovery tool'

 The following utilities empower the novice user by enabling the creation of maps and visualisations with a minimal knowledge of the underlying technologies

- GeoCommons – *http://geocommons.com/* - 'brings intelligence to the GeoWeb, unleashing tools and data' (downloadable as CSV, KML, Shape)

- OpenStreetmap – *http://www.openstreetmap.org/* - a free editable map of the world which allows users to view, edit and use geographical data in a collaborative way – user-generated content is free to all

- Mapufacture – *http://mapufacture.com/* - 'Helping build the geospatial web' – allows users to create a new map, add feeds and add data. Also hosts a large repository of spatial data including GeoRSS feeds, MKL, Shapefiles

- Platial – *http://www.platial.com/* - 'the people's atlas' – a combination of social networking and mapping APIs

- Others include: Mapmaker; mapbuilder.net; Wikimapia; Plazes; quikmaps; OnionMap

Spatial Visualisation and Research Organisations

Web 2.0 technologies and interactive earth viewers or geo-browsers have paved the way for research and governmental organisations to explore and expose their findings in new and innovative ways:

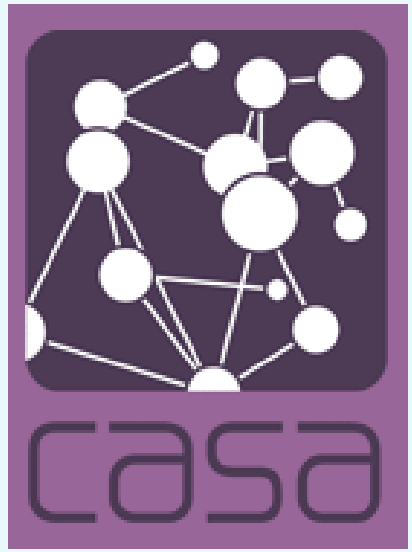
- Dutch Space Research Institute (SRON) & Royal Netherlands Meteorological Institute (KNMI)
- BODC
- U.S. National Snow and Ice Data Center
- USGS Earthquake Hazards Program
- Spanish General Directorate of Cadastre – use Google Earth to visualise spatial data it administers to government bodies, citizens and corporations
- UNHCR layer in Google Earth – mash together news, images, video, statistical databases relating to refugees across the globe via the “Google Earth Outreach programme” – *http://www.unhcr.org/events/47f48dc92.html*

Web 2.0 Mapping Utilities and Academic Research

Academic research groups and departments are utilising open spatial data, GIS and Web 2.0 technologies such as GeoRSS feeds, podcasts, wikis, tagging and commenting in innovative ways to enhance and report their findings.

- UCL Centre for Advanced Spatial Analysis (University College London)
  - *http://www.casa.ucl.ac.uk/*

    Aims to develop new technologies in several disciplines which deal with geography, space, location, and the built environment



DataShare project

EDINA®

GeoVue – one of the 7 nodes of the National Centre for e-Social Science set up to develop open and 'new kinds of virtual urban environments' (VUEs) through which users can participate in furthering their understanding of cities

Demonstrators developed at CASA include:

• MapTube *(http://www.maptube.org/) – a free resource for viewing sharing, mixing and mashing maps created with GmapCreator*

• London Profiler *(http://www.londonprofiler.org/) – enables users to search, visualise and create geogemographic KML profiles of Greater London from a variety of free statistical data resources*

Engage in other forms of collaboration e.g. Virtual London – OS MasterMap and Infoterra height data was piped into Second Life via ArcGIS to create a scrolling 3D map of London (3 million buildings) – removed on breach of copyright

- …and there's more
  - John Hopkins University's Interactive Map Tool
    - Supports digital field assignments allowing users to create custom mashups using a variety of digital media, text and data – http://www.cer.jhu/index.cfm?pageID=351
  - Minnesota Interactive Internet Mapping Project
    - A mapping application that provides maps and imagery similar to Google Maps – claims to be data rich, interactive, secure, easy to use, have analytical capabilities - http://maps.umn.edu/
  - Research at Pompeu Fabra University, Barcelona
    - Researchers mining spatial-temporal data provided by geotagged Flickr photos of urban locations – http://www.giradin.org/fabien/tracing/
  - Thematicmapping.org

    Thematic Mapping Engine (TME) enables you to visualise global statistics on Google Earth. The primary data source is UNData - http://blog.thematicmapping.org/

Citizen as Sensor

- Neo-geography – geographic techniques and tools used to create, assemble and share spatial information by a non-expert group

- Volunteered Geographic Information (VGI) – geographic data created by the public through readily available tools that use GPS (e.g. next generation mobile mapping, personal navigation devices, digital cameras) i.e. ‘citizen science’ or ‘citizen as sensor’

- Projects include:
  - geograph (*http://www.geograph.org.uk/*) – which aims to collect geographically representative photographic submissions for every 1km square of the UK
  - OpenAerialMap – (http://openaerialmap.org/) - ‘a non-profit, open access meeting place for theaerial photography community
  - eWorld project – (http://eworld.sourceforge.net/) – enables data imported from open mapping tools to be enhanced with environmental events such as weather, roadworks, traffic behaviour

Summary

- Tools, Plug-ins, APIs to enhance the data repository experience

- Collaboration and participation – democratize access to and use of data to address ‘grand challenge problems’

Muchas Gracias!!

Stuart.Macdonald@ed.ac.uk
DISC-UK DataShare

Seminario Sobre Datasets
Consorcio Madrono – 17
Nov. 2008