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
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
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**
2. research data & research publications


“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)
Preserve or Perish
International Polar Year

“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)
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)
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’
• **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
• **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](http://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
**DISC-UK DataShare Project** Edinburgh, LSE, Oxford, Southampton

<table>
<thead>
<tr>
<th>Informal Storage and Sharing</th>
<th>Formal Institutional Arrangement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data files with minimal documentation (e.g. readme file describing each data file) downloadable from Internet</td>
<td>Zip and ship</td>
</tr>
<tr>
<td>Metadata record of dataset on website or in repository; possibly with embargo and contact information to request access</td>
<td>Open access</td>
</tr>
<tr>
<td>Networked drive, available to research group, version control</td>
<td>Search and discovery enabled; restricted access</td>
</tr>
<tr>
<td>Password protected, networked drive (backup procedures)</td>
<td>Email dissemination by request</td>
</tr>
<tr>
<td>Personal hard drives, un-networked</td>
<td>Privileged access</td>
</tr>
<tr>
<td>Simple data storage</td>
<td></td>
</tr>
</tbody>
</table>

**Robin Rice, September 2007**

Robin Rice, Data Librarian, University of Edinburgh
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!
Unlock the potential in your data with our simple web services

places
Find the locations of place names

geocodes
Convert postcodes, grid references, etc to coordinates

extract
Extract placenames from text and find their locations

EDINA

banner logo and buttons - images © iStockphoto 2009
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)
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?)
There is no universal standard for citing data and computer files, but ...


  - ISO 690: 1987 Bibliographic references - Content, form and structure


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”

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](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!
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

- Identifiers for authors/creators
  - Open Researcher and Contributor ID (ORCID)
  - NAMES, EU Interparty, ISNI, VIAF
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

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)?

The stylistic diversity of the concert saxophone
by Rubino, Daniel J., 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 orchestra 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 genres. Specifically, the techniques of subtones, harmonics, and false fingerings are examined from both a performance and compositional perspective. Additionally, the instrument’s resemblance to 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 original 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 (6MB)
- Track_8.mp3 (3MB)

Indexing (document details)
School: York University (Canada)
School Location: Canada
Keyword(s): with Original composition
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 is not intended to 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 concert dance performances by 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 and 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 the same manner.

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 Movements of the 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 dealing with 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 encourage 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 Media Player or 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)
Advisor: Newhall, Mary Anne Santos
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
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), and URIs
- framework for representing information in Web; identifiers
  - http://www.w3.org/TR/rdf-concepts/
  - http://www.w3.org/TR/rdf-primer/

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
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’
Note that the man who gave you OpenURL ...


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

... is now into Linked Data


  http://arxiv.org/abs/0906.2135v1
A broker to discover nodes for deposit
• for long-term stewardship and added services
• for others to re-analyse for (secondary) research purposes
Research publications as research data

DataShare2

to formal publishing into (linked) data infrastructure

from formal institutional arrangement

<table>
<thead>
<tr>
<th>Peer review of datasets; seamless link to publications; role-based layers of access; data overlay journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributed high performance computing; analysis tools applied to data over secure international network; M2M interfaces</td>
</tr>
<tr>
<td>Graphs, charts, maps configurable online</td>
</tr>
<tr>
<td>“Actionable” marked up dataset installed in a data browser tool subsets capability</td>
</tr>
<tr>
<td>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</td>
</tr>
<tr>
<td>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</td>
</tr>
</tbody>
</table>

Data Grid

Data publishing

Data visualisation

Data manipulation online

Data enhanced for re-use

Network of distributed repositories: subject and/or institutionally based

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