



e-IRG Task Force on Integrated Data Management

Allan Foster, 11th April 2006

allan@allanfoster.co.uk



Members of the Task Force

- **Malcolm Read** (JISC – UK, Chair)
- **Mark Thorley** (NERC – UK)
- **Wojtek Sylwestrzak** (ICM – Poland)
- **Angelos Bilas/Manolis Marazakis** (ICS-FORTH, Greece)
- **Robert McGreevy** (ISIS Rutherford Appleton Labs – UK)
- **Dianne Rusch-Feja** (Knowledge Exchange)
- **Jan Windmüller** (MSTI – Denmark)
- **Donatella Castelli** (CNR – Italy)
- **Peter Rice** (EBI – UK)
- **Allan Foster** (JISC – UK)
- **Nike Holmes** (JISC – UK)



Terms of reference

Context: Data layer is third element of e-infrastructure alongside existing data network and Grid-middleware layers

- Identify key drivers and barriers to integrated data management across Europe
- Identify opportunities to promote and improve interoperability between digital data sources, across national, international and subject boundaries
- Identify where co-ordinated work across Europe can add value over next five years



The vehicle

- Chapter in the White Paper addressing the strategic issues in data management

(to be completed by May 23rd 2006 for submission to Editors)



Some key issues

- How can the results of publicly funded scientific research be optimally made accessible, internationally?
- How can we provide suitable integration and linkage between scientific publication and underlying data sets, incl. observational and experimental data?
- How can we encourage the proper curation of important data to ensure long term accessibility?
- How can we best work with partners in this endeavour, including funders, universities, researchers, users and publishers?



Some key issues (cont.)

- Humanities and social scientific data as well as STM
- Legal and IPR issues to be faced in these tasks?
- Very mixed picture across European states on advice/good practice on ownership of data, restrictions on use...
- Natural interest by funders to move on to new areas without sufficient attention to the established research record
- Mainly cultural, sociological and financial factors rather than technical
- Highlight data management as a profession and career

Ground rules in data management

- Ownership and IPR of the dataset must be clearly established, enabling its exploitation and access by 3rd parties
- Dataset must be catalogued to a defined level of detail and meet agreed standards
- Formal responsibility for custody must be agreed
- Data must be fully 'worked up' (calibrated, quality controlled etc) with sufficient documentation to be usable by 3rd party



Ground rules (cont.)

- Technical details of how data are to be stored, managed and accessed must be agreed and documented
- Technological implications must be established
- Resources needed to deliver these intentions over planned life of data – staff and IT infrastructure – must be estimated and sourced
- Review mechanisms must exist to reconsider periodically costs benefits of maintaining data
- Stakeholders should be informed in advance if data are to be destroyed



Data curation

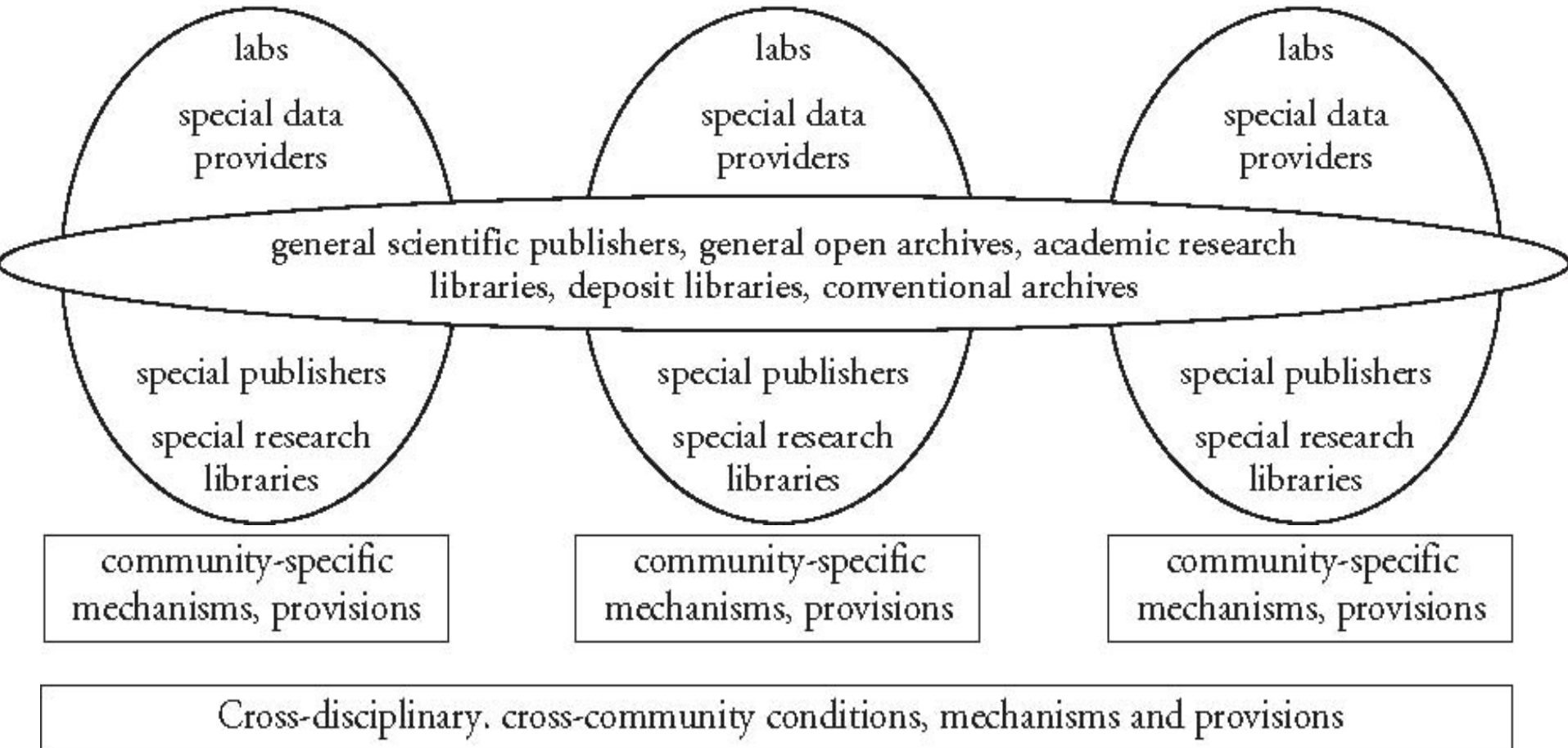
- Key to sustainability, 'reproduce-ability' and re-use of reliable and trusted digital resources
- Incentivising good practice: research funders requirements & career recognition of data management outputs
- Reproduce-ability sometimes requires a curation of a 'fixed' rather than dynamic dataset



Data curation (cont.)

- Software and versioning issues
- Need for new accounting model recognising long term costs of data curation and preservation
- European and international ongoing work such as by UK's Digital Curation Centre (DCC) to undertake research and disseminate good practice

Model of virtual infrastructure



Source: Task Force Permanent Access. Permanent access to the record of science: strategic action programme. ETFPA, 2005.

Disciplinary and sectoral differences



- Little understanding of heterogeneity of data management practices across disciplines
- Differences in
 - deposit arrangements
 - preservation arrangements
 - use of data and information
 - observational v experimental data
 - organisational and institutional factors
 - research process and methods
 - levels of workforce skills
- Need for cultural change in attitude towards data sharing



Metadata

- Descriptive, structural & administrative
- Includes preservation metadata – allows re-creation and interpretation of structure and content of digital data over time
- Needs to support various functions:
 - Discovery
 - Technical rendering of objects
 - Recording of contexts and provenance
 - Documentation of repository actions and policies
 - Rights management



Metadata (cont.)

- Metadata standards such as OAIS need to evolve over time
- Scientific datasets often require different set of solutions as they are often:
 - Stored in many different file systems and databases distributed around research organisations with no common way of accessing or searching them
 - Need to open files to understand what they contain
 - Little consistency between what is recorded; sometimes not online but in researchers' lab logbooks
- New scientific metadata models (eg CCLRC) should be mapped to Dublin Core to facilitate interoperability with digital libraries



Repositories

- Research publications, data and learning/teaching resources
- Institutional *and* disciplinary
- Inter-relationships between institutional repositories and specialist data centres
- “Scientific data is best handled by scientists in data centres” (NERC - UK)
- Open access embraces self-archiving in repositories *and* new business models for publishers



Repositories (cont.)

- Currently, how interested are researchers and institutions in repositories???
- Should deposit be a requirement by funders?
- Once again, cultural change is essential
- **D**igital **R**epository **I**nfrastructure **V**ision for **E**uropean **R**esearch (DRIVER)
 - Initially a common network of 51 existing repositories in 5 countries
 - Collective and enabling service layers
 - Using standards such as OAI-PMH, persistent identifiers and some technology standards (SOA, web services)



Any ideas, case studies and examples of
good practice on data management to ...

Allan Foster

for e-IRG Task Force on Integrated
Data Management

allan@allanfoster.co.uk