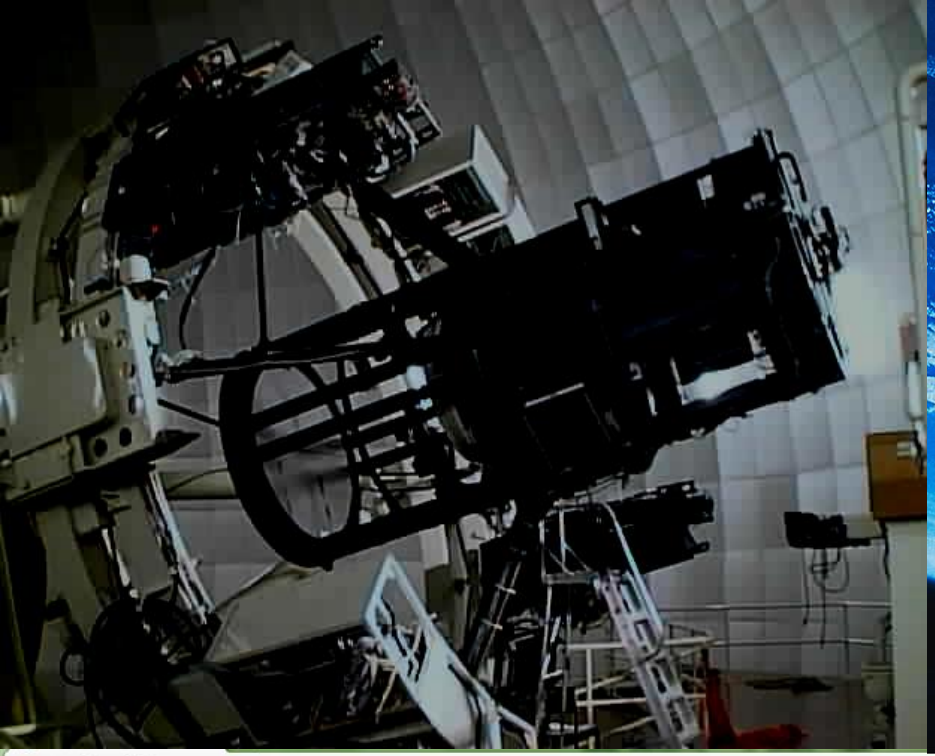
PREPARDE Cross Linking Workshop: Background, Introduction and Aims

Dr Jonathan Tedds
Senior Research Fellow
D2K Data to Knowledge
PI: PREPARDE project - Jisc Managing Research Data Programme
@jtedds

jat26@le.ac.uk

University of Leicester



onlinelibrary.wiley.com/doi/10.1002/gdj3.2/full

BADC – Trac | METAFOR | Home | Google Mail | BBC NEWS | News Fr... | Sorcha ní gCeallagh... | Add to Wish List

WILEY ONLINE LIBRARY

PUBLICATIONS | BROWSE BY SUBJECT | RESOURCES | ABOUT US

The Chadwick & RAL Libraries

REMEMBER ME

Home > Earth Sciences > General & Introductory Earth Sciences > Geoscience Data Journal > Early View > Abstract

JOURNAL TOOLS
Get New Content Alerts
Get RSS feed
Save to My Profile
Recommend to Your Librarian

JOURNAL MENU
Journal Home

FIND ISSUES

FIND ARTICLES
Early View

FOR CONTRIBUTORS
Author Guidelines
Submit an Article

ABOUT THIS JOURNAL
Society Information
Overview
Editorial Board
Permissions

SPECIAL FEATURES
Data Center FAQs
L F Richardson Award Prize Winners

RMetS Royal Meteorological Society

Geoscience Data Journal

SEARCH
In this issue
Advanced >

Data Paper

The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK

S. A. Callaghan*, J. Waight, J. L. Agnew, C. J. Walden, C. L. Wrench, S. Ventouras

Article first published online: 17 MAR 2013

DOI: 10.1002/gdj3.2

Issue

Geoscience Data Journal

Early View (Online Version of Record published before inclusion in an issue)

ARTICLE TOOLS
Get PDF
Save to My
E-mail Link
Export Citat
Get Citation
Request Pe
Share

Additional Information (Show All)

How to Cite | Author Information | Publication History | Funding Information

The research presented in this paper was funded by the UK's Ofcom as part of the Spectrum Efficiency Scheme and the support of Ofcom in providing the funding for the GBS experiment is greatly appreciated.

File Edit Image Catalog Overlay View Tool Help

Location 14:02:35.01 +54:2

Pixel unknown full

Object name resolver Ctrl+R
Simbad automatic pointer
Tooltip on sources
Auto-scroll on mouse pan
Script Console... F5
Macro controller...
VO tools
Remote tools
Plugins

VO tool controller...
VisIVO 3D visualisation tool [INAF/CINECA].
VOPlot plotting tool [VO-India]
TOPcat tabular data viewer [Starlink/Astrogrid]
SPLAT spectral analysis tool [Starlink/PPARC/JAC]

NED
Simbad
DSS1.E.POSSI

Zoom 1x

14.09' x 14.18'

14.17' x 14.17'

Search

MAIN_ID	OTYPE	RA	DEC	COO ...	COO ...	C...	PMRA	PMDEC
[H69] NGC 1969 146	HII	14 02 36.7	+54 21 55	3000	3000	171		
HGGK 179	HII	14 02 36.3	+54 21 46	3000	3000	171		
HGGK 180	HII	14 02 36.5	+54 22 00	3000	3000	171		
NGC 5451	Part...	14 02 36.1	+54 21 51	3000	3000	171		
HGGK 178	HII	14 02 35.8	+54 21 54	3000	3000	171		

27 sel / 3636 src 8Mb

Research data example - level 1:

- A typical example from physical sciences (astronomy) distinguishes between broad categories within the research data spectrum:

- **raw/initially auto-processed data** produced at a research facility such as an observatory
  - typically made publically available in this format after an embargo period of e.g. 1 year
  - in some cases available immediately - e.g. Swift Gamma Ray Burst satellite

Research data example – level 2

•**"research ready" processed data** which has been fully calibrated, combined and cleaned/annotated

- often produced by individuals or collaborations
- rarely available to anyone outside the collaboration except upon request/collaboration
- but needed if you want to reuse for science unless you have detailed sub domain specific knowledge and detailed contextual information to reproduce from raw
- considered to enable a competitive advantage for the researchers involved
- may well generate future additional samples and papers for the owning collaboration on top of the original published result(s)
- in some cases may be produced by dedicated data scientists on behalf of the community for major survey/missions e.g. ESA XMM-Survey Science Centre (Leicester), NASA…

Research data example – level 3

• **output dataset** – following detailed analysis of research ready datasets

- **forms the *data under the graph* in a journal publication** following analysis of research ready datasets
- *rarely available* to anyone outside the collaboration except upon request/collaboration
- may well generate future additional samples and papers for the owning collaboration on top of the original
- other researchers may request the data for their own research but may not get it!

….and STOP!

- Next project
  - Proposal long since written
  - Probably already underway
- Feel free to email ME if you would like to work on an idea using this dataset or code
  - As long as I'm a co-author on the paper!
  - You have to go through me to find out what you really need to know to reuse the data/code

Research data example – level 4

- **published catalogue** type representation of published output dataset

  - *NOT a "data paper"….but could be*
  - optional in many cases, mandatory for most major surveys
  - usually made available via project specific online resource
  - may be provided as table of parameters based on research ready dataset, usually linked from and associated with a journal
  - specifically produced in order for the wider community to reuse (cite!) and repurpose if wanted
  - The well-known Sloan Digital Sky Survey is a classic example or more recently the 2XMMi X-ray catalogue I have a close involvement with (largest X-ray survey of the sky).

The *XMM-Newton* Wide Angle Survey (XWAS)

P. Esquej[1,2,3,4], M. Page[5], F. J. Carrera[3], S. Mateos[3], J. Tedds[2], M. G. Watson[2], A. Corral[6], J. Ebrero[7], M. Krumpe[8,9,10], S. R. Rosen[2], M. T. Ceballos[3], A. Schwope[10], C. G. Page[2], A. Alonso-Herrero[3⋆], A. Caccianiga[6], R. Della Ceca[6], O. González-Martín[11], G. Lamer[10], P. Severgnini[6]

[1] Centro de Astrobiología (INTA-CSIC), ESAC Campus, PO Box 78, 28691 Villanueva de la Cañada, Spain
[2] Dept. of Physics and Astronomy, University of Leicester, Leicester LE1 7RH, U.K.
[3] Instituto de Física de Cantabria (CSIC-UC), Avda. de los Castros, 39005 Santander, Spain
[4] Departamento de Física Moderna, Universidad de Cantabria, Avda. de Los Castros, 39005 Santander, Spain
[5] Mullard Space Science Laboratory, University College London, Holmbury St. Mary, Dorking, Surrey RH5 6NT, UK
[6] INAF – Osservatorio Astronomico di Brera, via Brera 28, 20121 Milan, Italy
[7] SRON - Netherlands Institute for Space Research, Sorbonnelaan 2, 3584 CA, Utrecht, The Netherlands
[8] European Southern Observatory, Karl-Schwarzschild-Straβe 2, 85748 Garching bei München, Germany
[9] University of California, San Diego, Center for Astrophysics & Space Sciences, 9500 Gilman Drive, CA 92093-0424, USA
[10] Leibniz-Institute for Astrophysics Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
[11] Instituto Astrofísico de Canarias, (IAC), C/Vía Láctea, s/n, E-38205, La Laguna, Tenerife, Spain

Received ??, 2012; accepted ??, 2013

ABSTRACT

Aims. This programme is aimed at obtaining one of the largest X-ray selected samples of identified active galactic nuclei to date in order to characterise such a population at intermediate fluxes, where most of the Universe's accretion power originates. We present the *XMM-Newton* Wide Angle Survey (XWAS), a new catalogue of almost a thousand X-ray sources spectroscopically identified through optical observations.

Methods. A sample of X-ray sources detected in 68 *XMM-Newton* pointed observations was selected for optical multi-fibre spectroscopy. Optical counterparts and corresponding photometry of the X-ray sources were obtained from the SuperCOSMOS Sky Survey. Candidates for spectroscopy were initially selected with magnitudes down to $R \sim 21$, with preference for X-ray sources having a flux $F_{0.5-4.5\,\mathrm{keV}} \geq 10^{-14}$ erg s^{-1} cm^{-2}. Optical spectroscopic observations were made using the Two Degree Field of the Anglo Australian Telescope, and the resulting spectra were classified based on optical emission lines.

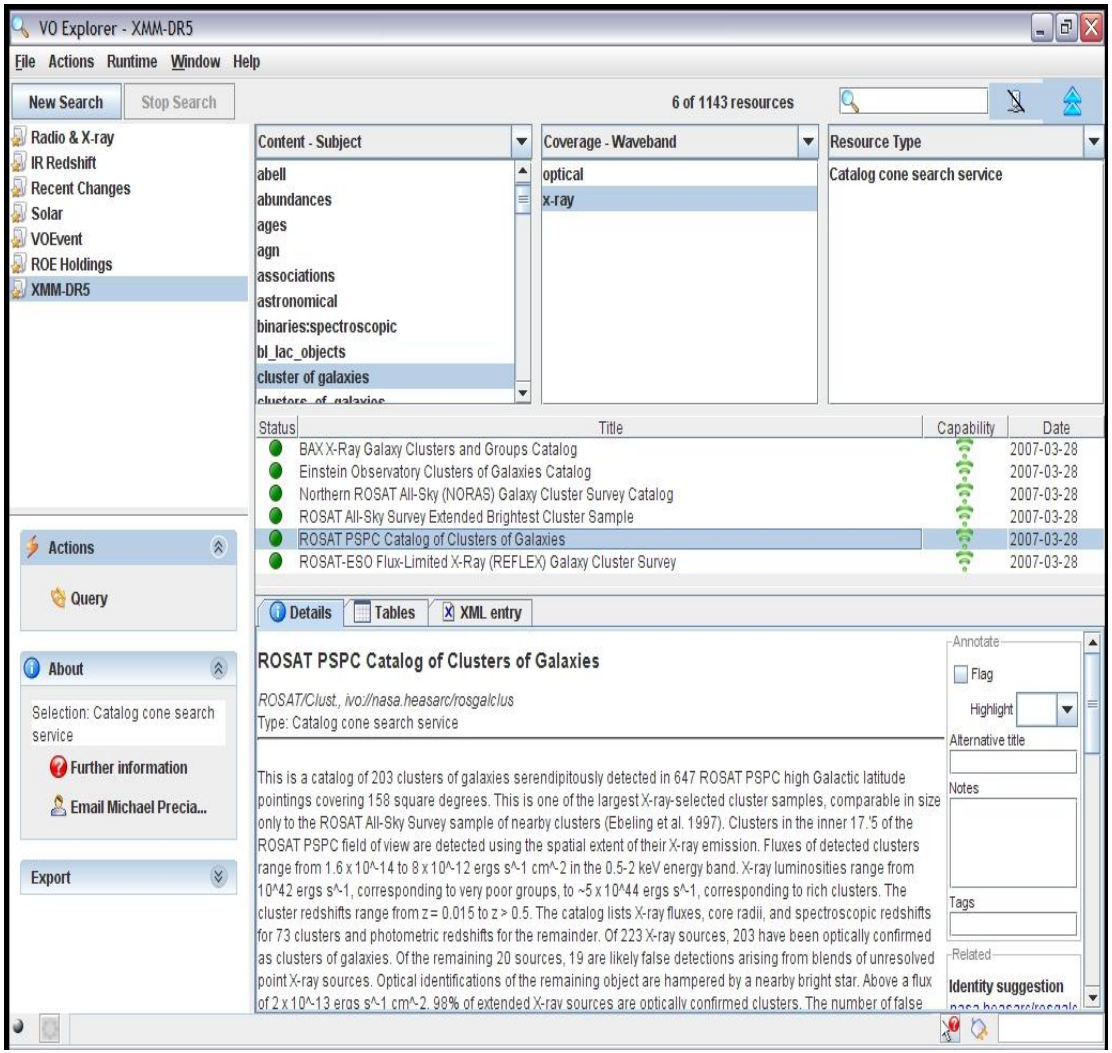
Results. We have identified through optical spectroscopy 940 X-ray sources over $\Omega \sim 11.8$ deg^2 of the sky. Source populations in our sample can be summarised as 65% broad line active galactic nuclei (BLAGN), 16% narrow emission line galaxies (NELGs), 6% absorption line galaxies (ALGs) and 13% stars. An active nucleus is also likely to be present in the large majority of the X-ray sources spectroscopically classified as NELGs or ALGs. Sources lie in high-galactic latitude ($|b| > 20$ deg) *XMM-Newton* fields mainly in the southern hemisphere. Owing to the large parameter space in redshift ($0 \leq z \leq 4.25$) and flux ($10^{-15} \leq F_{0.5-4.5\,\mathrm{keV}} \leq 10^{-12}$ erg s^{-1} cm^{-2}) covered by the XWAS this work provides an excellent resource for the further study of subsamples and particular cases. The overall properties of the extragalactic objects are presented in this paper. These include the redshift and luminosity distributions, optical and X-ray colours and X-ray-to-optical flux ratios.

Key words. X-ray: general – Surveys – X-rays: galaxies – Galaxies: active

http://adsabs.harvard.edu/abs/2013arXiv1302.5329E

http://www.astrogrid.org

- AstroGrid front end, part of VOdesktop
- *Resource-centric*
- Select a search-space
- Search for resources
- Filter these resources
- View selected resources
- Use the selection
  - Invoke it
  - Save/Bookmark/Tag it
  - Export it



Combining data from disparate sources

- '**New technologies for sharing data and for combining data from disparate sources** are particularly valuable in multidisciplinary fields such as earth science and nanoscience. ... **The challenge of federating, mining, analysing and interpreting these data will be a key focus in coming years.**'

http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/physical-sciences-case-studies-use-and-discovery-

Even the Chancellor says he gets it!

- "The next generation of scientific discovery will be data-driven discovery……"
- "We need to make sure we capture value from this mass of data – both for economic growth and for social advances, such as better health."
- "*This requires a transformation in data management*"

Speech by the Chancellor of the Exchequer, Rt Hon George Osborne MP, to the Royal Society – 9 Nov 2012

University of Leicester

RESEARCH COUNCILS UK

Excellence with Impact

Accessibility


Research and funding


Research Careers


Public Engagement with Research


Knowledge Exchange and Impact


International


Press and Media


Publications


About

- Home
- Research and Funding
  - Research Funding
  - Areas of Research
  - Cross-Council Research Themes
  - Research Infrastructure
  - Research Priorities
  - Peer review
  - Eligibility for Research Council funding
  - How to apply for research funding
  - Applications which may cross Research Council remits
  - Terms and Conditions of Research Council fEC Grants
  - Terms and Conditions of Research Council Training Grants
  - Open Access
  - **RCUK Common Principles on Data Policy**
  - Efficiency

Home > Research and Funding > RCUK Common Principles on Data Policy

RCUK Common Principles on Data Policy

Making research data available to users is a core part of the Research Councils' remit and is undertaken in a variety of ways. We are committed to transparency and to a coherent approach across the research base. These RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data policy.

Principles

- Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.
- Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.
- To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data.
- RCUK recognises that there are legal, ethical and commercial constraints on release of research data. To ensure that the research process is not damaged by inappropriate release of data, research organisation policies and practices should ensure that these are considered at all stages in the research process.
- To ensure that research teams get appropriate recognition for the effort involved in collecting and analysing data, those who undertake Research Council funded work may be entitled to a limited period of privileged use of the data they have collected to enable them to publish the results of their research. The length of this period varies by research discipline and, where appropriate, is discussed further in the published policies of individual Research Councils.
- In order to recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.
- It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds.

Search:

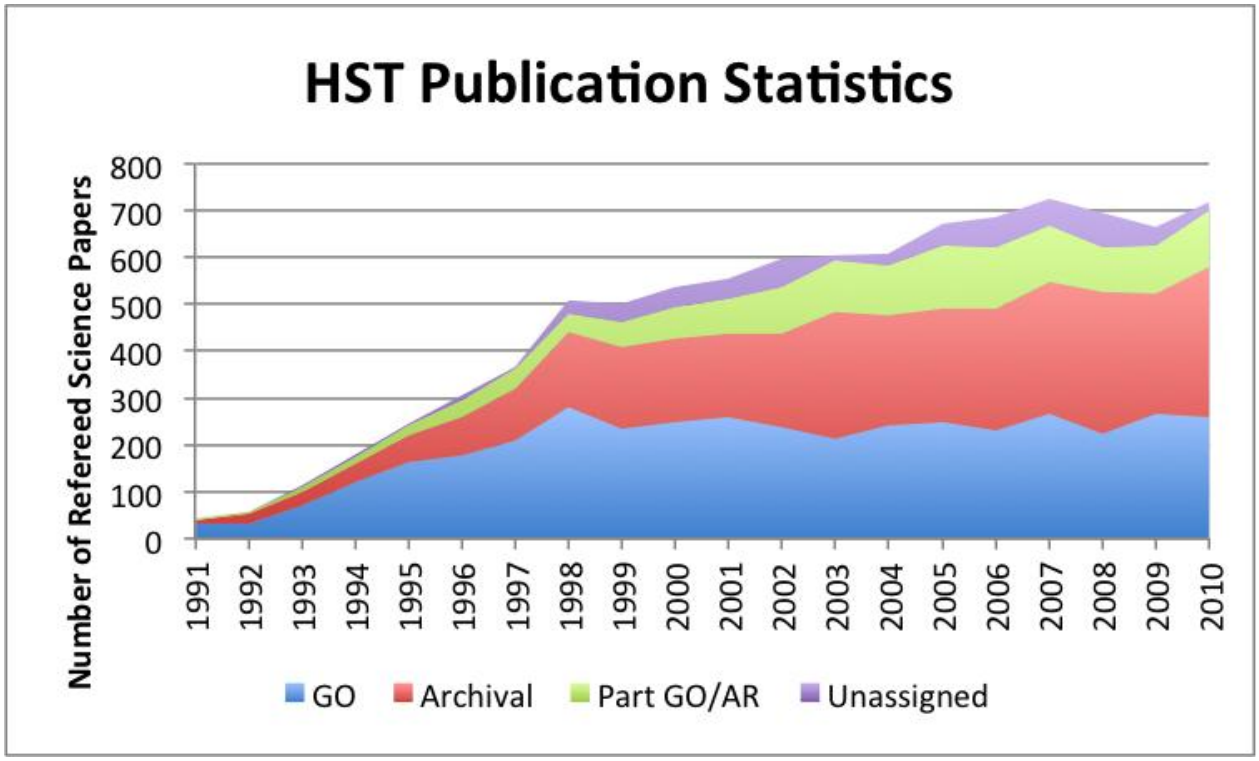
Go

This website

All Research Councils

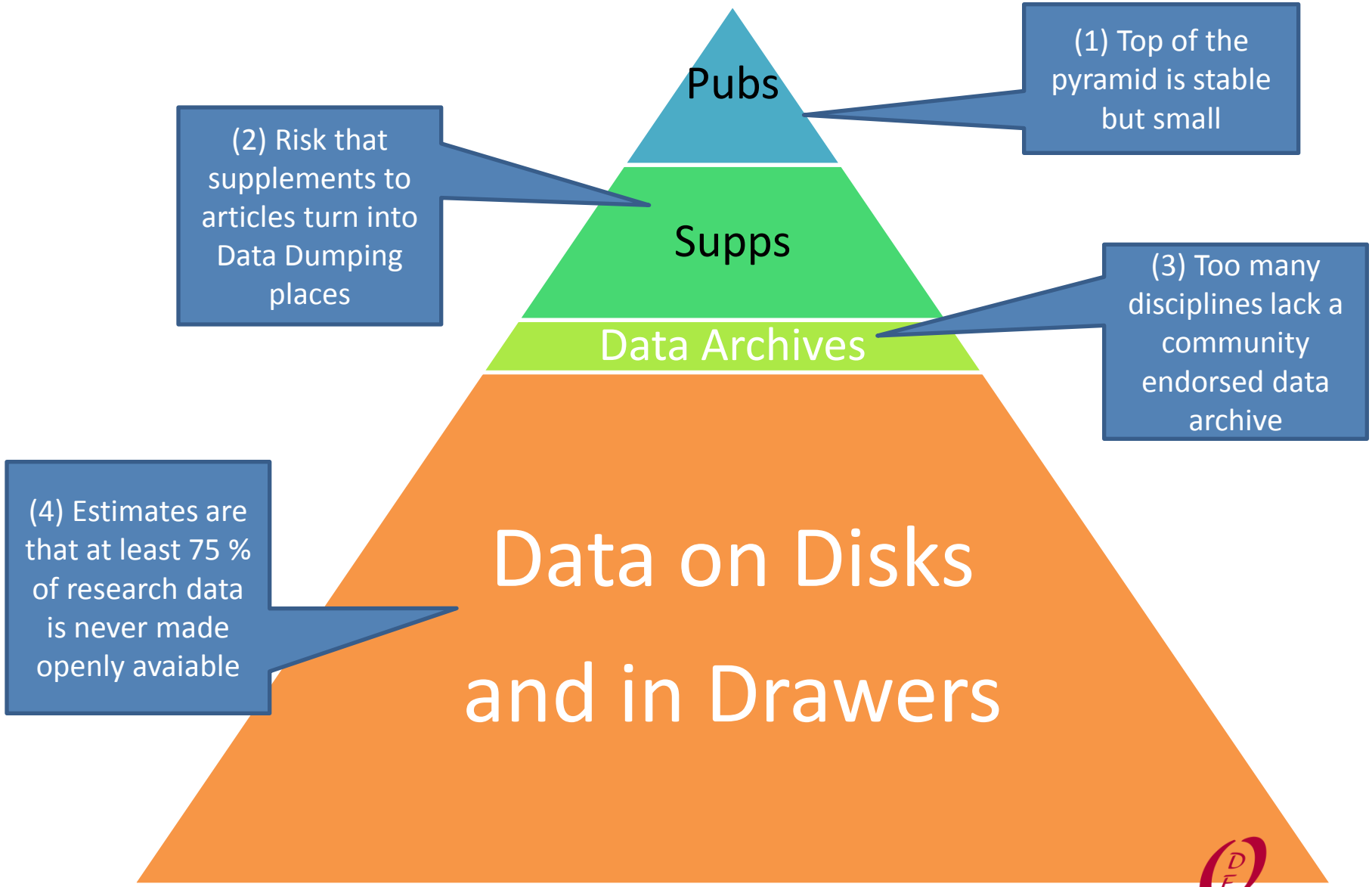
- Public good
- Preservation
- Discovery
- Confidentiality
- First use
- Recognition
- Public funding

Data Reuse: asking new questions



- Papers based upon reuse of archived observations now exceed those based on the use described in the original proposal.
  - http://archive.stsci.edu/hst/bibliography/pubstat.html

Data Publication Pyramid:

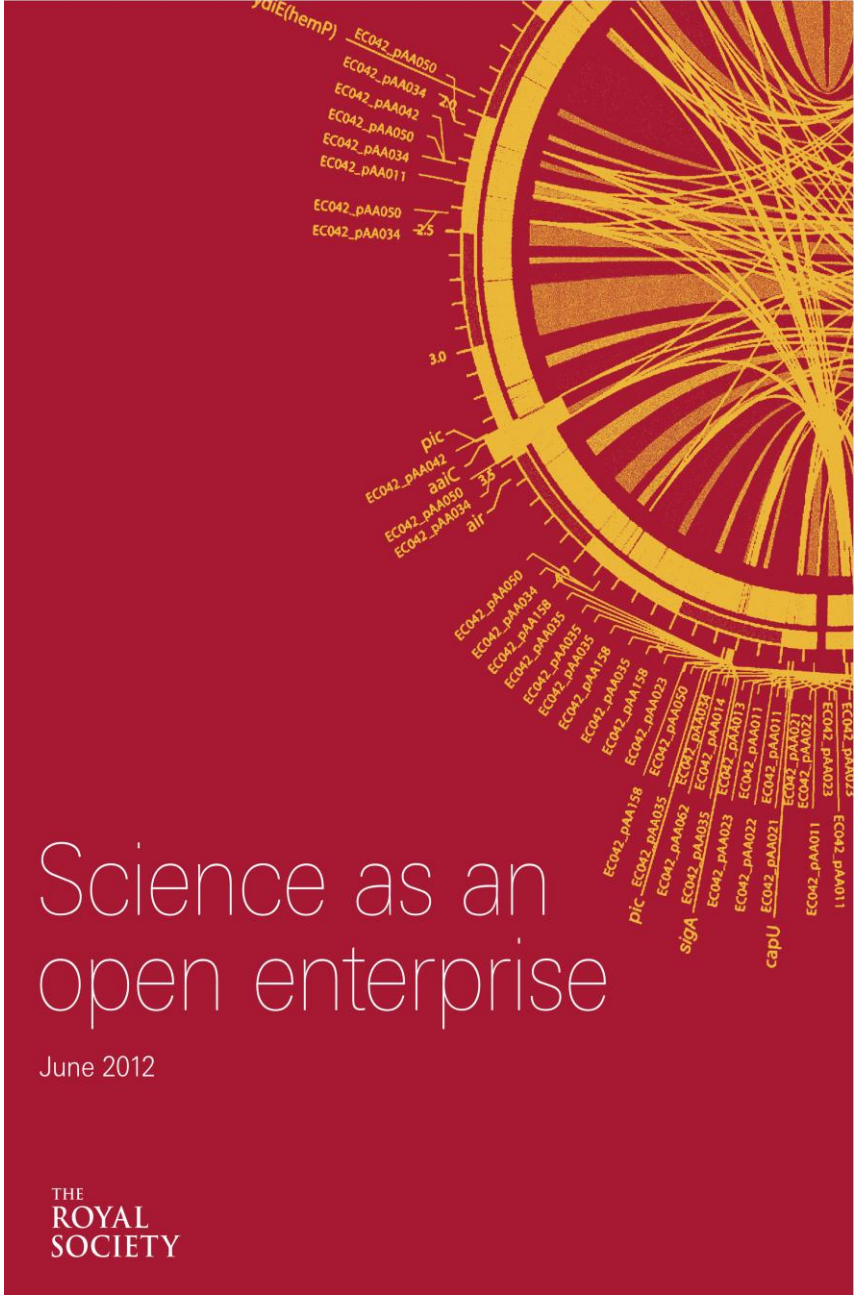


Opportunities for Data Exchange

Published research data – level 5?

- **published data paper** describing a research dataset
  - –Peer reviewed pre or post publication?

- **As a first step towards this intelligent openness, data that underpin a journal article should be made concurrently available in an accessible database.** We are now on the brink of an achievable aim: for all science literature to be online, for all of the data to be online and for the two to be interoperable. [p.7]
- Royal Society June 2012, *Science as an Open Enterprise*, http://royalsociety.org/policy/projects/science-public-enterprise/report/
- Issues linking data to the scientific record:
  - Data persistence
  - Data and metadata quality
  - **Attribution and credit for data producers**
  - … and many more



PREPARDE: Peer REview for Publication & Accreditation of Research Data in the Earth sciences http://www.le.ac.uk/projects/preparde

- **capture the processes and procedures required to publish a scientific dataset**
  - ingestion into a data repository
  - formal publication in a data journal
- address **key issues in data publication**
  - how to peer-review a dataset?
  - what criteria are needed for a repository to be considered objectively trustworthy?
  - how can datasets and journal publications be effectively cross-linked for the benefit of the wider research community?
- PREPARDE team includes **key expertise** in
  - Research
  - academic publishing
  - data management
- **Earth Sciences focus** but produce general guidelines applicable to a wide range of scientific disciplines and data publication types incl life sciences (F1000R)

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library

F1000 FACULTY of 1000 POST-PUBLICATION PEER REVIEW NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Repository accreditation

- Link between data paper and dataset is crucial!
  - How do data journal editors know a repository is trustworthy?
  - How can repositories prove they're trustworthy?

- What makes a repository trustworthy?
  - Many things: mission, processes, expertise, workflows, history, systems, documentation, …
  - Assessing trustworthiness requires assessing the entire repository workflow

- ***PREPARDE / IDCC13 Workshop – report in draft***
- Peer review of data is implicitly peer review of repository

And what does "trustworthy" mean, when you get right down to it?

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000 FACULTY of 1000 POST-PUBLICATION PEER REVIEW NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

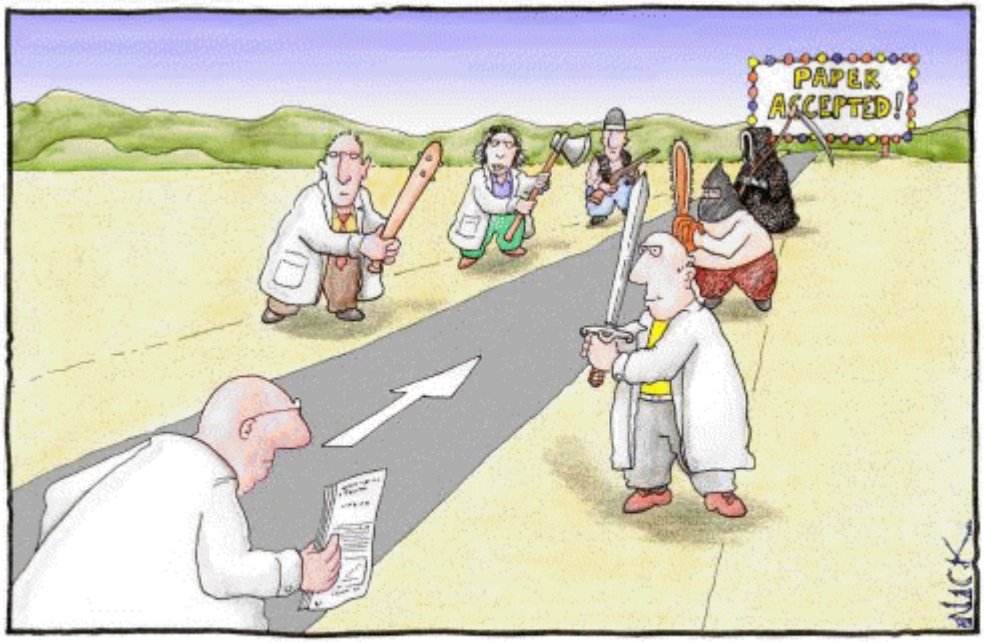
Peer-review of data

- Summary Recommendations from Workshop at the British Library, 11 March 2013
- Workshop attendees included funders, publishers, repository managers and other interested parties.
- Draft recommendations put up for discussion and feedback from audience captured.

Feedback from the community welcome!

Document at: http://bit.ly/DataPRforComment

Feedback to: https://www.jiscmail.ac.uk/DATA-PUBLICATION



Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'

http://libguides.luc.edu/content.php?pid=5464&sid=164619

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Proposal for RDA WG: recommendations on data peer review

Summary Recommendations from Workshop at the British Library, 11 March 2013

- **Connecting data review with data management planning**
- **Connecting scientific, technical review and curation**
- **Connecting data review with article review**
- 4-5 draft recommendations in each of above
- Assist Publishers, Journal Editors, Reviewers, Data Centres, Institutional Repositories, Researchers to map requirements for data peer review
- Matrix of stakeholders vs processes
  - Assist in assigning responsibilities for given context
  - New for most disciplines
  - Learn from disciplines where this already happens

Ok, let's talk about cross linking….

Critical issues the workshop aims to cover include:

- How can publishers and repositories collaborate to exploit and share metadata about related scientific outputs (i.e. datasets and articles)
- Where are the best points in the various journal and repository workflows to establish persistent links and share metadata for data publication?

Outcomes Participating in the workshop will help you to:

- Understand benefits and risks to stakeholders in scholarly publishing likely to arise from different data publishing models
- Shape PREPARDE project guidelines on:- a) workflows for data publication b) data and metadata standards for enabling cross-linking between data repositories and academic publishers

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library

F1000 FACULTY of 1000 POST-PUBLICATION PEER REVIEW NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Not shown

- Draft recommendations on peer review of research data in 3 categories on following slides

Connecting data review with data management planning

1. All research funders should at least require a “data sharing plan” as part of all funding proposals, and if a submitted data sharing plan is inadequate, appropriate amendments should be proposed.
2. Research organisations should manage research data according to recognised standards, providing relevant assurance to funders so that additional technical requirements do not need to be assessed as part of the funding application peer review. (Additional note: Research organisations need to provide adequate technical capacity to support the management of the data that the researchers generate.)
3. Research organisations and funders should ensure that adequate funding is available within an award to encourage good data management practice.
4. Data sharing plans should indicate how the data can and will be shared and publishers should refuse to publish papers which do not clearly indicate how underlying data can be accessed, where appropriate.

Connecting scientific, technical review and curation

1. Articles and their underlying data or metadata (by the same or other authors) should be multi-directionally linked, with appropriate management for data versioning.
2. Journal editors should check data repository ingest policies to avoid duplication of effort , but provide further technical review of important aspects of the data where needed. (Additional note: A map of ingest/curation policies of the different repositories should be generated.)
3. If there is a practical/technical issue with data access (e.g. files don't open or exist), then the journal should inform the repository of the issue. If there is a scientific issue with the data, then the journal should inform the author in the first instance; if the author does not respond adequately to serious issues, then the journal should inform the institution who should take the appropriate action. Repositories should have a clear policy in place to deal with any feedback.

Connecting data review with article review

1. For all articles where the underlying data is being submitted, authors need to provide adequate methods and software/infrastructure information as part of their article. Publishers of these articles should have a clear data peer review process for authors and referees.
2. Publishers should provide simple and, where appropriate, discipline-specific data review (technical and scientific) checklists as basic guidance for reviewers.
3. Authors should clearly state the location of the underlying data. Publishers should provide a list of known trusted repositories or, if necessary, provide advice to authors and reviewers of alternative suitable repositories for the storage of their data.
4. For data peer review, the authors (and journal) should ensure that the data underpinning the publication, and any tools required to view it, should be fully accessible to the referee. The referees and the journal need to then ensure appropriate access is in place following publication.
5. Repositories need to provide clear terms and conditions for access, and ensure that datasets have permanent and unique identifiers.