Data publishing, peer review and repository accreditation: everyone a winner?

Report from the PREPARDE Project IDCC Workshop

Annex to PREPARDE Deliverable D5.1 Report on requirements for data centre accreditation

Angus Whyte and Alex Ball, Digital Curation Centre, May 2013

Executive Summary

The report is based on presentations and discussion at the workshop of the same title as this report, held in Amsterdam on 17th January 2013 at the IDCC 2013 conference (see workshop page). The event drew 36 participants from 9 countries, with a spectrum of roles in data publication; data centres, publishers, research institutions, and libraries. The International Association of STM Publishers was represented, along with national organisations including the UK's Digital Curation Centre (DCC) and British Library, Netherlands' Data Archiving and Network Services (DANS), and South Africa's National Research Foundation.

The workshop was organised by DCC through the PREPARDE project (Peer REview for Publication & Accreditation of Research data in the Earth sciences). PREPARDE aims to establish policies and procedures for data publication in the *Geoscience Data Journal*, and to generalise those policies for application outside the Earth Sciences. The project addresses three issues: -

1. Workflows and cross-linking between a data journal and the repository source(s) of data published in an article.
2. Repository accreditation, from the perspective of journals recommending data repositories for authors to deposit data underlying their articles.
3. Scientific peer-review of data submitted to data journals, and its relationship to data management action by authors, repositories, research organisations and funders.

The workshop primarily addressed the second of these issues, i.e. repository accreditation. The presentations and discussion informed PREPARDE recommendations on repository certification [1] and peer review [2], which were made available for comment in March 2013.

The workshop began with context; first on the background to data journals (*Sarah Callaghan*), and then on recent developments in repository accreditation (*Peter Doorn*). *Michael Diepenbroek* described emerging infrastructure for data publication in the empirical sciences, drawing on the example of the ICSU World Data System. *Eefke Smit* then outlined lessons from integrating data and publications, from the perspective of the International Association of STM Publishers. A learned society view was given in *Richard Kidd*'s talk on the Royal Society of Chemistry's engagement in data publication. Data Centre and Institution perspectives followed from *Kerstin Lehnert* of the Integrated Earth Data

[1] PREPARDE Repository accreditation draft guidelines. Available at: http://bit.ly/ZhYHZl
[2] PREPARDE Peer-review draft guidelines. Available at: http://bit.ly/DataPRforComment

Applications (IEDA) centre, and from *Veerle Van den Eynden*, representing both the UK Data Archive (UKDA) and University of Essex 'Research Data @ Essex' project.

A full account of the discussion themes is included at the end of the report. Overall the workshop proceedings covered the following themes:

Data journals are changing the data publishing environment, but for how long?
Data papers describe the collection of data deposited in a repository, and expand on the metadata available there. In doing so they may give a bigger stake in addressing key issues to researchers and data managers who may otherwise be excluded from the author list of a journal article. Those key issues include quality assurance, persistence, reproducibility, discoverability, transparency and above all giving due credit. Alternative data publication models address some of these issues; notably data centres such as PANGAEA, IEDA and UKDA that perform extensive data review, and offer standard disciplinary metadata, and persistent, actionable links (e.g., DOIs, ARKs) to cite the data they hold.

Repository relationships with publishers also support a third model; data integration for 'enhanced publication' of journal articles. Variants of this model, exemplified by PANGAEA and Elsevier's collaboration, include the embedding of visualisation tools into articles, bidirectional citation between articles and data packages, and automatic highlighting of controlled vocabulary terms. An emerging infrastructure supports each of these models. It comprises Crossref, Datacite, ORCID and other third party 'linking services' which will support necessary exchanges between data repositories on the one hand and journals, catalogues and bibliometric services on the other. As this infrastructure becomes embedded in practice and data integration tools improve, some question whether the data journal will remain a useful concept in the long term. Evidence of community acceptance of data publication models is so far encouraging in the area of citation impact, but limited.

Repository accreditation offers assurances, but trust has many dimensions
The data journal model entails peer review of data, and any model involving a continuing relationship with a repository implicitly demands peer review of the 'trustworthiness' and reliability of that repository. Preservation is critical to the persistence of the research record. Trusted digital repository certification standards provide a self-assessment and audit-based approach that is relevant to this need. The European Framework for Audit and Certification of Digital Repositories is formalizing a three-level approach to suit different requirements:

- Data Seal of Approval, originating from DANS, is the basic 'bronze' level. It comprises 16 guidelines that may be self assessed or peer reviewed.
- DIN 31644, originating from NESTOR comprises 34 criteria, representing the 'silver' level when self assessed and externally reviewed.
- ISO 16363 originated from TRAC (OCLC, RLG), comprises over 100 metrics. The 'gold' standard is external audit performed using either ISO16363 or DIN 31644.

In earth science disciplines the ICSU World Data System (WDS) also offers certification procedures, informed by the Data Seal of Approval and ISO 16363 standard, and OAIS (Open Archival Information Systems) the reference model to which these relate. WDS membership criteria are intended to supply a transparent and objective base for periodic assessment of the WDS facilities and the overall performance of the system. The certification procedures should ensure the trustworthiness of WDS facilities.

Repository accreditation was seen to offer partial assurances for the peer review of data. It offers auditable checks that a repository performs technical reviews when data is deposited, that there are appropriate standards for checking metadata completeness, and the data authenticity and integrity are assessed. As well as these 'technical' aspects, peer review

entails academic assessment of data quality, and the repository's role here is contingent on the contribution to the assessment that researchers can make themselves.

The workshop noted the continuing debate on how far repository certification, which judges the quality of a repository, can provide a sound basis for trusting in the quality of particular datasets held in them. Some stakeholders involved in current debates e.g. in the Research Data Alliance, consider service level agreements to be a more relevant basis for trusted relationships. Even where there is certification of the repository itself, post-publication reviews of a repository's data holdings are also relevant in the peer review context. For example DANS has successfully piloted user reviews on data quality and five other criteria.

Workshop discussion also highlighted the essential need for repositories to demonstrate their sustainability and especially to provide a continuity plan in case they are unable to maintain access to the data in their archives. The repository should also provide clarity about the scope of the data it collects. Trust at least requires some evidence that a repository is actually used by the community it serves, e.g. information on its take-up and the usage of its holdings. Directories such as re3data.org and databib.org were looked to as a source of information to support decisions on trust.

Roles are evolving and collaboration is key to success

The critical role of data centres and repositories in supporting reuse and verification was evident throughout the workshop. The evolution of their role in response to community and policy needs was exemplified by IEDA, a community facility serving solid earth and polar sciences. IEDA has expanded its role in community outreach and governance, as well as enhancing data interoperability and utility for secondary research and learning. It supports funders to ensure compliance with data management policies, having recently developed a *Data Compliance Reporting Tool* offering information on datasets related to a NSF grant number.

The Royal Society of Chemistry illustrates the potential for learned and professional societies to support data publication, both directly as publishers and indirectly through support for standards. The society has promoted data sharing services including ChemSpider, and the UK National Chemical Database Service. RSC Journals invite data at peer review stage, but demand for verified data is much greater among 'downstream' research users than among peer reviewers. With the exception of crystallography, chemistry domains lack an established culture of data sharing, which is typically seen as too much effort and a hindrance to commercialisation. However RSC is advocating for greater reusability, and promoting electronic lab notebooks through e.g. the Dial-a-Molecule project.

Institutions also have a supporting role in changing the research culture around data sharing and publication. Support roles discussed in the workshop included advocacy, training and support for community standards in data management, e.g. for provenance metadata. Considering peer review of data, participants saw a need to separate institutional service roles to support 'technical' review, and any role of faculty members who have the expertise to judge a dataset's quality as evidence for academic claims. Researcher's informal sharing through their peer networks was also noted as a potential contributor to peer review.

Guidance on data appraisal/ review and selection is a key area for collaboration between data centres and institutions. This is exemplified by collaboration between the UK Data Archive and University of Essex on social science data collections held in their repositories. The trust implications of a tiered approach to curation need to be thought out; for example a publisher may need to check what tier the data is held in, rather than make trust judgements about the repository as a whole.

Workshop participants saw a need for further partnerships, as exemplified by the Memoranda of Understanding between Dryad and partner journals, and the Joint Data Archiving Policy. The Ubiquity Press agreements with repositories and DANS in particular were also cited. There is a need to promote such agreements across communities, perhaps with learned societies and professional associations recommending them as best practice.

Disclaimer
This report is based on the authors' notes from the workshop discussion. We have attempted to accurately convey points raised by the participants but these points do not necessarily represent the views of the authors, DCC or the workshop speakers named in the report. Their presentation slides are available from the workshop page at:

Presentations and Discussion

1. Workshop Aims and Introduction – Jonathan Tedds

Introduction. PREPARDE is looking at peer review in the Earth Sciences, but is also interested how it may apply in other disciplines. What are the implications for researchers, publishers, and repository management?

2. Data publication models: benefits, risks and peer review – Sarah Callaghan

Sarah traced a brief history of data in scientific publications. Journals were invented in the 17th Century: the first being the French journal *Journal des sçavans* (later renamed Journal des savants) shortly followed by *Philosophical Transactions of the Royal Society of London* in 1665. Data has traditionally been embedded in the research article, as illustrated by several examples of scientific data table and images published in mid-19th Century.

The key issues in data publication are about quality assurance, persistence, reproducibility, discoverability, transparency and above all giving due credit. The scientific community should be moving to change the 'data publication pyramid' (ref. ODE report)

1. Data in article (small, stable)
2. Supplementary data (considered 'a dumping ground' as publishers typically are not willing to support large volumes)
3. Data in repositories (not all disciplines well served)
4. Data held privately (75%)

Ideally, more data should be available from articles and repositories. Publishing a data paper is a relatively new option and more clearly a form of 'publication' than simply putting data on a website or 'in the cloud'. Submitting data as supplementary material to an article is an established option but one that journals may be increasingly reluctant to support. Publishing a data paper is an alternative that (like other models) brings disciplinary and institutional repositories into the relationship between author and publisher.

The 'data paper' model envisaged in PREPARDE is essentially quite simple; a data paper describes the data set, its collection, and structure, etc. rather than the data analysis. In this model an author submits a data paper to a journal, and the data underlying it to a repository. The peer review process includes checks on the data.

A list of data journals is available on the PREPARDE blog at: http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList. These cover a variety of disciplines although the list is still short. Each journal has its own policy on which repositories are recommended. Some e.g. the *Dataset Papers* from Hindawi Publishing Corporation [3] provide their own. Repositories may require particular licence/waiver conditions on the data.

A key risk that data publication needs to manage is the persistence of the links to data held in the repository. What if the link to a dataset breaks? It means loss of credibility, loss of key metadata, loss of reputation.

Different stakeholders in data publication have requirements to trust in each other and in the systems they use, but there are different definitions of "trustworthy". Some repository service options, such as dark archives, are problematic in terms of publication but may still be necessary for them to be 'trusted' e.g. with confidential data.

Journal editors need a quick and easy way to determine if a repository hosting a published dataset will meet their requirements. A minimum set of criteria was suggested; that repositories should:-

- Have long term data preservation plans in place for their archive.
- Actively manage and curate the data in their archive.
- Provide landing pages giving extra information about the dataset (metadata) and information on how to access the data.
- Use persistent, actionable links (e.g., DOIs, ARKs) to cite data held in their archive
- Resolve cited dataset links to landing pages

These requirements were offered for discussion. Does there need to be formal accreditation of repositories? Or open community review using a 'tripadvisor' approach?

Q & A

Provision of a review environment for unpublished datasets

- **Q:** There is sensitivity about when data is published. Example: one Australian group was scooped on publishing results from their data because they had published it early.
- **A:** As far as peer review is concerned, a possible solution might be for repositories to give publishers sets of credentials that reviewers can use to log in anonymously, in order to view otherwise restricted datasets.

Impact of overlay journals publication workflows on quality of data and metadats

- The list of data journals could include *ZooKeys* [4]. It has just published its first data paper generated from metadata (rather than being manually written).
- **Q:** You assume that one needs data journals to publish data – all you need is for the data to be reusable and comprehensible. Surely the ideal is to publish archived data directly?
- **A:** Not really assuming that, it's more that data journals provide an opportunity for an extra layer of review and a stamp of quality.
- But often researchers often only use subsets, having put the whole dataset into a QA'd database. The data paper model is not ideal for this case.

[3] Hindawi Dataset Papers. Available at: http://www.hindawi.com/dpis/
[4] Pensoft ZooKeys. Available at: http://www.pensoft.net/journals/zookeys/

- **Q:** Creating data paper from metadata: would this give the impression of low quality papers, and defeat the object of having a data article, as the metadata would already be in the repository?
- **A:** An automatically generated paper should only ever be the starting point for working up the paper towards publication.

Community take-up of data journals

- **Q:** Are data journals getting a lot of submissions?
- **A:** Most are only just getting going. *Earth Systems Science Data* [5]has been going 5 years; initially the submission rate was slow but it is believed to be growing.

Overlapping models

- **Q:** Couldn't data papers be published within traditional journal?
- **A:** Yes, why not?

3. Trusted Repository certification and its potential to improve data quality – Peter Doorn, DANS

Peter discussed the relevance of trusted data repositories for journals, giving background on trust and certification. He offered reflections on how to build trust and on the pros and cons of certification from his involvement in the developing European framework for certification and a Research Data Alliance Working Group on Certification.

He began with the new *Journal of Open Psychology Data*, which DANS is participating in as a recommended repository [6]. The Ubiquity Press workflow for this exemplifies the data journal approach. From the author's perspective the steps in this workflow are:

1. Register with journal.
2. Deposit data in repository.
3. Add data DOI (from repository) to paper.
4. Submit for paper for peer review.
5. Modify the paper in the light of reviewer comments.
6. Add paper DOI (from publisher) to data record in repository.

Trust is key for depositors, repositories, users and funders. Trust implies that in the face of uncertainty one can 'just take the first step without needing to see the whole staircase', to paraphrase Martin Luther King. Certification is a means of building trust.

Doorn questioned whether just saying "you can trust us" is enough for anyone to rely on. Things are not always what they say they are, and assurances given by an agent about itself might be false. So do we need certification?

The rationale for certification is that there is a need to guarantee trustworthy digital repositories, as an essential step for funders and depositors to rely on long-term archiving as

[5] Earth Systems Science Data. Available at: http://www.earth-system-science-data.net/
[6] Journal of Open Psychology Data. Available at: http://openpsychologydata.metajnl.com/about/editorialPolicies#custom-0

a basis for sharing and reusing data. Two main positions are currently being advocated around this:

- Trustworthiness of repositories is an illusion; on this view 'trust' is too complicated a concept to measure. Also the value of any certification wanes once given, and certification requires too much information, money and time. Objective and consistent auditing is also an illusion: " what happens to objectivity if auditing becomes a career?" is a view attributed to Helen Tibbo (see reports on panel session at iPRES 2012, Toronto), and it is impossible to guarantee global consistency in applying a standard.
- Trustworthiness is not illusory. There are three levels of certification available which are clear and balanced and suit different needs. Auditing is already a career in other areas, and some variation according to local needs is acceptable.

There is still need to raise awareness among funders, repositories, and communities about certification of trustworthiness. Data management/archiving is increasingly a condition of funding. There is recognition that RDM costs money, so it should be legitimate to add it to proposal budgets. Similarly financial support for certification should be provided.

Certification in Europe has developed around the emerging European framework[7], comprising three standards:

- Data Seal of Approval, which originated from DANS and is not managed by an international board. The approach is simple, lightweight and transparent: comprising 16 guidelines that may be self assessed and submitted for peer review, 8 'seals' have been awarded.
- DIN 31644, originated from NESTOR and comprises 34 criteria, with test audits carried out in 2013.
- ISO 16363 originated from Trusted Repository Audit and Certification (TRAC)[8] and the Open Archival Information System (OAIS). It consists of over 100 metrics, and is complemented by a full external auditing process. A self-audit on the criteria is also possible.

We can consider these to be a basis for bronze, silver, and gold levels of certification.

- Basic (Bronze) = DSA.
- Extended (Silver) = ISO/DIN done by self and externally reviewed.
- Formal (Gold) = ISO/DIN done entirely by an external auditor.

The European Framework for Audit and Certification of Digital Repositories is formalizing the three-level approach above. There is support for this framework from the European Commission, including in expected recommendations for the Horizon 2020 programme. Funders such as the Netherlands Organisation for Scientific Research also support it.

The EU APARSEN project has a Work Package on Trust. Some of the EU funded research infrastructures such as CESDA, CLARIN, DARIAH are all starting to implement trust requirements. A Research Data Alliance Working Group on Certification will report on the state of practice and write recommendations. The most pressing topics for the RDA Working Group are:

---

[7] European Framework for Audit and Certification of Digital Repositories. Available at: http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

[8] TRAC and TDR Checklists. Available at: http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0

- Should repositories develop Service Level Agreements instead of working towards certification?
- What is actually certified, and how can the certification address data quality?
- Who are the stakeholders? What is the role of the researchers?
- How to manage a worldwide certification effort?
- Which concrete actions can be taken? Can you go further than a report on the state of affairs and strategic recommendations? How to implement the recommendations?

Many issues have a bearing on these questions. For example few journals yet have policies on data availability. For those that do, practice varies across disciplines, and even within disciplines. In some cases journals require deposit to the journal, and in others to a repository? There are questions around who should be responsible for the policy: the editor, editorial board, or publisher? Also archiving only guarantees data can be checked, it does not guarantee its quality.

We can nevertheless be sure that a two stage approach is needed, according to Doorn;

1. When a paper is submitted for peer-reviewed publication, data access for reviewers is required
2. When the paper is published, readers should be able to access the data. In this is in a data archive it can implement a user review mechanism. For this open access and open data are preferred, but not required.

A user review approach has been successfully trialled by DANS, and Marjan Grootveld and Jeff van Egmond presented a report on this at the IDCC in 2011 (see [9]). The reviews were targeted at 6 criteria: -

1. Data quality
2. Quality of the documentation
3. Completeness of the data
4. Consistency of the dataset
5. Structure of the dataset
6. Usefulness of the file formats

Q & A topics

Journal data policy variations

- There were questions on the position of Elsevier journals. There is a variety of practice within Elsevier. It does permit data citations. But disciplinary differences are reflected in the variety of policies that exist across the different titles. Elsevier prefers to work with existing data repositories such as PANGAEA rather than archive everything itself.

Who can and should judge data quality?

- Monash is working with mass spectrometric data, released whether it is 'right' (error-free) or not; the quality of any given set has to be judged in comparison with existing data.

[9] Grootveld, M., & van Egmond, J. (2012). Peer-Reviewed Open Research Data: Results of a Pilot. International Journal of Digital Curation, 7(2), 81–91. doi:10.2218/ijdc.v7i2.231

- Participants from institutions experienced in working with researchers expressed concern about central institution roles judging data quality, seeing their role as to judge the completeness of provenance information.
- What is good and bad quality data can be very subjective and it's a 'slippery slope' for repositories to assume that role
- DANS can check that the file formats are right, that the metadata is in place: this is marginal checking. Data review is by peers. There will be a range of opinions, but it is most helpful to get comments by reusers. This is a good argument for user review approach.

4. Research data enters scholarly communication: towards an infrastructure for data publication in the empirical sciences – Michael Diepenbroek PANGAEA and ICSU World Data System

Michael reviewed the rationale for data publication and why it is a structural problem in empirical science. Societal benefits from data publication include support for government and commercial decision-making and other economic and commercial impacts. In some cases the data collection is irreproducible, so if data is lost, it is lost forever. Benefits to research include discoverability of results, their verification, and the fostering of large scale and complex science. And yet publications remain the currency of science.

The prerequisites for data publication can however be met without data journals. Firstly there need to be quality assurance procedures to support data review. There should also be standardisation of metadata and support for data interoperability standards, so these are preferably machine-readable. There are some advantages to putting data in RDBMS rather than a file repository. Licences and business models need to support open access.

There is certainly a need for more trusted and certified repositories. This was demonstrated by the ODE report[10] which showed that most researchers do not know whether there is a usable archive in their discipline or not, and of those who do only a small minority believe there is one. Also the Parse Insight report (3.4)[11] showed that a large majority of researchers want to see an international archive structure.

The principles to be followed have been set out in various forms recently; the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003)[12]. The Bermuda principles [13](1996) and the OECD principles and guidelines for access to research data (2007).

The building blocks for a broad international archive structure are being put in place. They include DataCite, its Metadata Registry for scientific data and collaborations with science publishers (Elsevier, Springer, Wiley, Thompson Reuters etc.) for linking and cross-referencing data & articles. Also important are the ORCID registry for researchers launched in 2012, and the Thomson Reuters Data Citation Index also launched last year.

[10] Opportinities for Data Exchange Final Report. Available at: http://www.dnb.de/EN/Wir/Projekte/Abgeschlossen/ode.html
[11] PARSE insight survey report. Available at: http://www.parse-insight.eu/publications.php
[12] Berlin Declaration. Available at: http://en.wikipedia.org/wiki/Berlin_Declaration_on_Open_Access_to_Knowledge_in_the_Sciences_and_Humanities
[13] Bermuda Principles. Available at: http://en.wikipedia.org/wiki/Bermuda_Principles

The ICSU World Data System (ICSU-WDS[14]) is an exemplar of what needs to be developed for long-term data stewardship and publication. ICSU offers a federated system whose core elements were reformulated in 2007 to include metadata and data services such as catalogues, certified repositories and data collection and processing facilities. These include QA/QC processes, support for developing data products, and data rescue services. The certification procedures are informed by the international standards in this area; OAIS, the Data Seal of Approval and the ISO 16363 standard.

WDS membership criteria are intended to supply a transparent and objective base for the evaluation and accreditation of WDS candidates as well as for periodic assessment of the WDS facilities and the overall performance of the system. The certification procedures should ensure the trustworthiness of WDS facilities in terms of authenticity, integrity, confidentiality and availability of data and services.

Currently there are links between PANGAEA[15] and catalogues, publishers, bibliometric services. There is a need to develop this infrastructure further, to build on the many one to one relationships currently formulated between these organisations. The community requires better linking infrastructure. Linking services could give more support to the necessary exchanges between data repositories on the one hand and journals, catalogues and bibliometric services on the other. ORCID, DataCite, CrossRef are the foundations of that, and will enable more dynamic cross-referencing between data repositories and publishers.

Q & A

Evidence of data publication impacting on article citation rates?

- Sharing data increases citation rates according to Piwowar et al.'s study (e.g.[16]) See also Jon Sears (AGU) study 1992-2011 published in a blog article.

5. Integration of data and publications - Eefke Smit, International Association of STM Publishers

STM is an association of over 125 publishers, both academic and professional, both big and small. Eefke drew attention to the changes that have occurred in how data and publication relates to each other.

A good example is the original paper on the structure of DNA: Crick, F., & Watson, J.,(1953). A structure for deoxyribose nucleic acid. *Nature, 171*, 737-738. This was 1.5 pages long with no data. The publication of the human genome in 2001 ran to 62 pages, including 27 tables and 49 figures. Ten years later (2010) an anniversary paper appeared on the iPad and other platforms: it relied on an immense amount of data provided through links to archives rather than reproduced in the paper.

---

[14] ICSU World Data System. Available at: http://www.icsu-wds.org/
[15] 'Elsevier and PANGAEA take next step'. Available at: http://www.reedelsevier.com/mediacentre/pressreleases/2010/Pages/elsevier-and-pangaea-take-next-step.aspx

[16] Piwowar, H and Vision, T.J Data reuse and the open data citation advantage. Available at: https://peerj.com/preprints/1/

There is now a variety of ways that data may be integrated into publications, including:

- Visualizing 'live' data from repositories within a paper; example of *Earth and Planetary Science Letters* paper showing PANGAEA data using Google Maps.
- Inverse citations from Dryad to *Nature*.
- *Gigascience* data repository and journal.
- *BioChemical Journal*, Portland Press, Semantic version[17]. Actionable PDF: users can jump to and use with data, get different representations.
- Elsevier data viewers: online article readers can interact with the data and change visualizations without leaving the page.[18]

In parallel, some Journals have begun to see supplementary data as a burden. *Neuroscience* was a high profile example of a journal deciding to stop accepting it because of the increasing volume of it. ES referred to the view of supplemental data as a 'dumping ground'; and cited *Cell* journal's move to introduce conditions on what would be accepted, such as a clear relationship to something in the article, restriction on numbers of additional figures. The key disadvantage of supplemental data is that it is not generally properly curated.

The visialisation of a 'data publication pyramid' developed in the ODE project has become a very well-used representation of how data is handled. It is likely that more data will become integrated into articles (the top of pyramid) as it becomes easier to do so. This is a good thing: 'the best metadata for data is the article itself.' Perhaps supplemental data and data journals are only temporary solutions. In due course data should be independently citable: the infrastructure will be in place to provide the 'guarantees' of quality and sufficient documentation; plus the integration of data into main articles will make data papers redundant.

APARSEN studied data peer review. Editors have a big fear about this: they think it is too much work, and would grind the process to a halt. They shy away from strict policies: they expect peer reviewers to at least check that the article's claims are supported by the underlying data, but not to perform full QA of the data. ES referred to Heather Piwowar's reported attempt to check ten PloS articles to see if she could find data. She could only do so for one, despite the publisher's mandate to make the data available.

Should there be special reviewers for data? Methodology, collection methods: needed but marginal. Real review, actually digging into the data, is almost impossible without watching the person do the research.

Certification of data repositories is a very important issue for publishers. It would mean they could plan how to integrate the data in the certified repositories. At the very least, it would mean they could ask for DOIs for datasets, and publish them with confidence, knowing that there will be associated registry records for them.

ODE offered the following recommendations to authors: -

[17] Biochemical Journal. Available at: ttp://www.biochemj.org/bj/semantic_faq.htm

[18] See e.g. Aalbersberg, I. (2011) Supporting Science through the Interoperability of Data and Articles. Available at: http://www.dlib.org/dlib/january11/aalbersberg/01aalbersberg.html

- Clearer editorial policies on the availability of underlying data
- Recommend reliable and trustworthy Data Archives to authors
- Enhance articles for better integration of underlying data
- Endorse guidelines for proper citation of data
- Launch and sponsor Data Journals
- Ensure persistent identifiers and bi-directional linking
- Partner with reliable Data Archives for further integration of data and publications, including interactivity for re-use.

Q & A

Persistent links to datasets

- **Q:** If you consider the article as being metadata for the dataset, if the link breaks between them, how do you get from one to the other? Shouldn't there be measures to preserve metadata alongside the dataset in the repository?
- **A:** Persistent IDs and registries are supposed to solve this problem.

Impacts of open access policy differences

- **Q:** There is also the issue that article are usually behind paywalls, whereas datasets usually aren't, so the link from the article will be less visible than the link to it.
- **A:** Publishers provide abstracts outside the paywall, and usually the link to the data is outside paywall as well (whether in the reference list or in/near the abstract). This is becoming common practice.

6. Learned Society perspectives on data review – Richard Kidd, Royal Society of Chemistry

[Via Skype]

Richard outlined his background in the RSC's publishing production. The Society has a role in its charter for defining standards in research practice for chemistry, and has extensive involvement in publishing.

RSC sees data mostly as supplementary material: usually in MS Word format or PDF, rarely raw data, usually figures, graphs and other derived data. Peer review of this? Supplementary Data (SD) is available to reviewers:

- they occasionally recommend swapping information between article and SD;
- they do find it interesting occasionally, but largely it is treated as second class and not reviewed as thoroughly.

Crystallographers are the exception. They have a well defined data format (CIF), extensive common toolset, domain repositories (e.g. Cambridge Crystallographic Data Centre, IUCr lists several more), and have an established culture of data sharing and integrating data with publishing. Crystallographers are even asking for more raw data to be published alongside articles.

Having the original data at the peer review stage? Theoretically this is great, and RSC journals do invite it but get very little. There is a concern that peer reviewers don't have time to review it, and the picture is complicated by varied formats and tools. Reviewers never ask for it (they don't think they need it for verification). It's the downstream reusers that clamour for it.

There have been attempts to address this. *ChemSpider* [19] supports community deposit of structures, spectra, reactions (with enough detail on how to reproduce/optimise). Depositors to this service should get credit from doing so.

Chemical science is a competitive domain, with no cultural norm of sharing. Lab group culture: stuff gets abandoned, left in cupboards. Few available standards. A sample of researcher attitudes towards data:

- Would bury it if wasn't too much effort.
- Technology transfer office only lets us release data if it has been proved worthless.

Funder actions: EPSRC ramped up requirements for data management. Projects: Dial-a-Molecule is looking at electronic lab notebooks for academics, so the information on failures – i.e. synthetic reactions that don't work – (in particular) is not lost.

UK National Chemical Database Service: EPSRC tender. Formerly operated by STFC's Scientific Computing Department, now RSC has taken it on. Developments planned: provide both core database services, plus a community domain repository for chemical science (acting as funder repository). It is starting with structured data. It is positioned as complementary to institutional repositories. RSC wants to use it to promote reuse (it can/will handle multiple levels of embargo – group, institution, the world – and data licences). See http://cds.rsc.org/; RSC will hold a meeting with academics imminently on the way forward for the service.

RSC wants data made available to feed initiatives like Dial-a-Molecule, to promote ELNs, to support model building and validation; repositories should provide APIs so data can be used in automated workflows.

RSC is thinking how to get the message out to researchers on why they should be doing Open Science (value that arises for the community).

Ultimately, the validation for services such as CDS comes from data being reused:

- Reuse of data (esp. in new contexts) shows it is of significant value, and indicates that the ways in which it was collected, coded and curated are acceptable.
- Reuse provides testing, validation, visibility, reward and recognition.

How do we validate a repository?

- Does it provide demonstrable value? This is necessary to build community acceptance.
- Is there community validation from funders and society?
- Is accepted good practice followed?

7. Data publication at IEDA – Kerstin Lehnert

[19] See 'Introducing ChemSpider'. Available at: http://my.rsc.org/chemspider

Kerstin provided background on IEDA (Integrated Earth Data Applications). This NSF-funded organisation is a partnership between the Marine Geoscience Data System and EarthChem, and is a member of the International Council for Science's *World Data System* (ICSU-WDS). IEDA is described as "… a community-based facility that serves to support, sustain, and advance the geosciences by providing a centralized location for the registry of and access to data essential for research in the solid-earth and polar sciences."

Objectives are focused on reuse and verification and include: -

- Increasing community productivity and capability.
- Providing public access to data.
- Enhancing data interoperability and utility for secondary research and learning
- Ensure compliance with data management policies.

Usage of IEDA is rising, as measured by citations and visitors. It maintains several data collections and systems (e.g. System for Earth Sample Registration (SESAR)). It performs data synthesis and offers products such as Global Multi-Resolution Topography (GMRT), PetDB, SedDB, and Geochron (allows subsetting, etc.). It also provides visualization tools (GeoMapApp, Virtual Ocean, EarthChem Portal) and portals to other data (ASP, EarthChem, USAP-DCC).

Kerstin advocated three main strategies for making data fit for reuse. These are identified in a recent NSF proposal (Zaslavsky et al.: "Managing community input to support cross--domain interoperability and fitness-for-use assessment", 2012), as follows: -

- There should be a record of the data's provenance.
- Data should comply with standards for collection/representation (formats, semantics).
- The precision of the data, any errors or missing data, and the workflows used for QA should all be documented.

The data in IEDA products have in the past been extracted from the published literature (primary, secondary, sometimes followed up with authors directly), and added manually to the respective databases (currently via an MS Excel proforma).

Data managers add value to data by pulling together metadata. This can range from extracting information from captions and article text to create metadata, developing metadata schemas and vocabularies that align with community standards, verifying data is readable, checking for metadata gaps in existing holdings.

There is a mismatch between the level of metadata that data creators, managers and consumers think is acceptable. For geospatial data, co-ordinates (of the coverage) are vital, but only 28% (2005) – 54% (2009) datasets referenced in *Journal of Petrology* had them.

Data standards have recently undergone some development, for example through community workshops for *EarthChem*. In 2008, an Editors Roundtable was convened over two conferences; American Geoscience Union and Goldschmidt, and produced a joint statement and recommendations, including: -

- Complete data should be accessible, not just analyses
- Essential metadata (e.g. sample location, lithology) should be provided
- Samples need unique IDs
- Data should be deposited in open access databases

Now IEDA is getting more data directly (still using MS Excel templates) and establishing better links with publications. This has been triggered by a co-operative agreement with NSF who wanted to point researchers to IEDA as a source of support for compliance with the NSF data policies on open access, moratorium periods and Data Management Plans.

In return for more stable NSF funding IEDA has introduced more formal community governance and guidance, including outreach activities and processes to demonstrate the centre's usage, utility and impact. There is an increasing focus on getting contributions from users. These changes have involved major improvements to infrastructure, management, and policies. The improvements include:

- Rigorous risk management
- Persistent identification of data & samples (DOI, IGSN)
- Long-term archiving via agreements with NGDC and Columbia University Libraries
- Cross-referencing with publishers, data citation index, etc.

The EarthChem Library data publication workflow offers an example of an entirely Web-based approached. Data managers perform QC, approve data then provide DOIs. However there are unresolved issues affecting IEDA's data publication, the four main ones being: -

- IEDA is getting requests to accept new data types where there are no community standards.
- How to deal with long narratives describing procedures; would these be better placed in a data paper?
- How to migrate data from the input (deposition) databases to synthesis databases?
- Recording feedback comments from data reusers.

IEDA is involved in improving the currently patchy monitoring of compliance with data sharing commitments made by researchers in their Data Management Plans. They have developed a *Data Compliance Reporting Tool* ([20]) This online service provides a list of datasets related to a NSF grant number. Each one retrieved identifies any citations to the data, alongside the data system, data types, datasets, instrument, and investigators.

Lehnert sees a slow change in the culture, driven by need for easy access to data, and supported by the evolving infrastructure (DataCite, etc.) In future, she envisages greater use of a digital lab book tool which is being developed. This will contribute metadata directly to IEDA collections, drawing these from lab profiles that researchers do not need to re-enter each time, and enabling more robust publication pipelines from these to Journals.

Q & A

Fragility of standards

- **Q:** Getting communities to agree on standards is fragile because of evolving needs. Like the editorial process.
- **A:** Yes. From 2009, IEDA has been careful to record provenance, track all changes made (cleaning, reformatting).

Level of data referenced from data publication – raw or derived?

- Synthesized data is not given an ID as it is a product, and may be different from submitted data.

[20] Data Compliance Reporting Tool. Available at: http://www.iedadata.org/compliance/report

8. Criteria and procedures for data review and quality control at the UK Data Archive – Veerle Van den Eynden

Veerle gave background on The UK Data Archive (UKDA) based at the University of Essex, which was until recently part of the *Economic and Social Data Service*. Its services now form part of the UK Data Service, which is largely funded by the Economic and Social Research Council (ESRC) with some additional funding from Jisc.

Her talk described the data review criteria and procedures that are being redesigned to fit the new organisation. These are also the basis for similar procedures being introduced for data offered for deposit in the institutional repository at the University of Essex, through the *Research Data @ Essex* project. This is one of 17 currently supported through the Jisc Managing Research Data programme.

The UKDA currently holds over 5000 datasets deposited mainly from ESRC funded research projects and government departments. Its purpose is to collect and make available data of use to UK researchers in social sciences. Currently around 230 datasets are accepted by the archive each year. These comprise government survey data, research data from individual grant-funded academic projects, public records and historical data. The archive's holdings include qualitative, quantitative and cross-disciplinary data types. It hosts international statistical time series data, and has links with other data archives worldwide. Popular datasets over the past 4 years have included long series, international macro data, crime survey, gender differences, and ancient parish boundaries.

The Research Data @ Essex project is piloting a research data management and sharing infrastructure for the University of Essex. Research groups involved in the project are from a much broader range of disciplines, including proteomics, bio-imaging, management, language acquisition, sociolinguistics and artificial intelligence. As well as establishing a data repository, the project activities include data policy and advocacy, and provision of support and training.

An Acquisitions Review Committee performed data review at the ESDS. The UKDS is splitting its functions into two independent groups; a Collections Development Committee and a Data Appraisal Group. The existing Acquisitions Review Committee has met every two weeks to evaluate data quality and potential re-use value.

Members decide a consensus rating using a scale (A*, A, B, C) depending on the condition of the data and documentation, and the anticipated level of re-use. The outcome of this is a decision to accept the data for long-term curation or not. However data that is not accepted for may still be self-archived in the ESRC Data Store (UKDA-store) for short-term management and access

UKDS will expand these options adding an additional two levels of curation:

- Long-term curation; this is expensive, so is reserved for collections with long-term secondary analysis potential. These collections will be processed, curated, preserved by UKDS.
- Short-term management; this means the data will backed up, accessible and discoverable but there will be no active preservation: researchers are expected to do that themselves.
- Delivery only; these will be hosted by third parties and accessed via APIs or services.

- Discovery only; these are held in other (institutional) repositories, harvest metadata

The (draft) appraisal criteria to be used by UKDS are similar to those mentioned in the DCC Guide *How to Appraise & Select Research Data for Curation* and those used by the Natural Environment Research Council (NERC) in their *Data Value Checklist*.

To be accepted for short-term management data must either be from an ESRC award, or covered by a contractual obligation between funder and researcher to archive with UKDS. Criteria that negatively influence acquisition are:

- Legal or ethical reasons for not sharing e.g. no consent gained, IPR problems, data protection requirements
- Restricted or limited sample size
- Lack of documentation and contextual information
- Format unsuitable for re-use

Any dataset affected by the above will also be less likely to be accepted for long-term curation. The criteria counting in favour of long-term curation are much more extensive. They comprise ‘essential’, ‘primary’ and ‘supporting’ criteria. These are as follows: -

Essential data criteria for long-term curation (relevance)

- Potential for secondary use, teaching and learning, or replication and validation of research use
- Needed for research and policy
- High quality, reliable and up-to-date
- Good temporal and spatial coverage
- Scientific value for social and economic sciences
- Historical value
- Data are well documented

Primary criteria

- New data source, e.g. transactions data (admin records, commercial records), tracking records
- Data with international value / for international research:
- Authoritative source of high quality statistics
- Reference databases used globally to support
- Decision-making and policy formation
- Harmonised and comparable between countries
- Longitudinal or consistent time series data
- Data needed for comparative research, e.g. across the UK and its devolved areas
- Disclosive microdata otherwise not available to the research community (e.g. business data, Secure Data Service)
- Contractual obligation for UKDS to acquire data
- Unique, unrepeatable data that are costly to reproduce
- Data directly requested by the dedicated user community

Supporting criteria

- Data in a suitable format for reuse (e.g. no specialised software needed or such software made available) or can be converted
- Data have sufficiently open access conditions
- Widely cited data

- Anonymised data

The discovery-only service will need institutional repositories and data producers to make the review decisions, with the UKDS providing training and online guidance. There will also be QA applying the same A*, A, B, C scale, depending on condition of data and documentation & anticipated level of re-use.

The primary criteria recommended for use by institutions and their researchers are:

- Data is cited / referenced in a publication
- Funder requirement to preserve/share data E.g. ESRC, BA, AHRC, MRC
- Institutional legal requirement to retain/share data E.g. Freedom of Information, Environmental Information Regulation
- Scientific value – potential for use in current or futureresearch
- Learning and teaching value
- No legal/ethical constraints prohibit preservation/ sharing E.g. consent, data protection, IPR

Supporting criteria:

- Good data documentation and metadata
- Suitable format for reuse (e.g. no specialised software needed or such software made available) or can be converted
- Openness of access conditions
- Easy to ingest without significant processing/ preparation
- Data in good condition (readable, undamaged,…)
- Cost of ingest

Q & A

Review costs and effort

- **Q:** How many sets do you evaluate? This seems a major task.
- **A:** It used to be the case that separate measures were in place for Government data: only the first (few) of a given type were appraised, then all subsequent data of that type was treated according to that decision. Data from all other sources were checked individually; in practice UKDA receives 20-25 data offers every fortnight. Maybe five of these will be selected for long-term curation.

Can usage metrics address difficulty judging long-term value

- **Q:** It is not always possible to tell what will be of long-term value; serendipity plays an important role. How does UKDA take account of this?
- **A:** Most ESRC-funded data will go into the Data Store, and be visible in the search interface. If something proves unexpectedly popular, it may move into long-term curation. Serendipity is another good reason to provide good discoverability metadata.

Discussion

Establishing trustworthiness

Roles of data repositories

What do data centres/repositories need to do to prove they are trustworthy, to users in general and to publishers in particular?

Repositories need continuity plan and evidence of sustainability

- Long term availability of data. Many projects are only funded for 2 years, and who maintains it after that? Repositories must demonstrate that efforts have been made to sustain the service (even if only through agreements with other repositories).
- Clear and reliable handover policy should repositories close.
- This need for a continuity plan is one of the WDS criteria for membership.

Trust requires information on repository take-up / usage

- Researchers have to say where they'll deposit their data as part of their data management plans. When looking them up, it's hard to tell whether some repositories are active (if, say, their Web pages have not been updated for 3 years). Even if a repository is not one of the main ones, researchers would still like some assurance of longevity. It's difficult to tell what criteria they might comply with – there's a need for some way of indicating where they fit.
- Need for clear documentation of how long the repository is funded for.

Depositors want opportunities for citation and recognition.

- What do users want? Long-term access, yes, but also opportunities for citation and recognition.

Roles of editors/ reviewers

What do editors/reviewers need to do to guide repositories that have no certification, and on what topics?

Recognise certification to an accepted standard if a repository has it

- Right direction is for repositories to get certification.
- Need to think about all the angles, and take into account perspectives from different disciplines and situations.
- We have some generally agreed standards, but they go beyond what is needed to make data available and protect it. There are different levels of standards.

Recognise clear collection policies and 3rd party commendations

- Certification implies everyone meets the same high standards, but needn't mandate it. More manageable approach would be for the repository to provide clarity over what it is providing, and to collect commendations from external bodies as and when they become ready to do so.
- Repositories should consider the different kinds of data, different ways of processing data they use e.g. the UKDS route of providing different levels (datastore versus higher layer of curation), should these be kept in the same repository or different ones.

Differentiate judgements about repositories from judgements about individual datasets

- There's a difference between certifying the management of a repository and certifying the usability of its data. It's more manageable if these are kept separate, so journals could apply criteria separately.
- The National Science Board report *Long-Lived Digital Data Collections* defined three levels of data collections: research, resource (or community), and reference. Different metadata would be required at each of these levels?
- It is a curatorial function to take data with highly specialised metadata and make it possible to use the metadata for general reference purposes.

Roles of institutions and faculty

What can institutions do to support data publication by identifying what data is worth keeping - at faculty and/or central repository level?

Recognise value of data publication in career progression

- Primary task for institutions is to acknowledge data publications when considering an individual's career progression.
- We should be cautious about institutions rather than academics having the power to make judgements about what data is worth keeping
- It's not that institutions should do the judging, but recognize where that judgement has been made, e.g. by the disciplinary community.

Recognise economies of scale in disciplinary specialisation

- In the long-term there might be repository for each disciplinary community, but not now. Sustainability of subject repositories is also doubtful, while institutions are more stable. So that means there is a role for institutions. Does the data substantiate the findings? If so, it should be somewhere and, where no subject repository exists, the institution is ideal place. But institutions are a long way from being able to do this and decide priorities.
- There's a role for research/finance offices to recognize that there need to be dedicated budgets for curation. Researchers need awareness of this too.
- Yes, finance is an important issue but disagree that institutions stabilize the system. There are 3 genome repositories worldwide. We'll see more of this (worldwide subject repositories) because of economies of scale. Small repositories are relatively inefficient.

Provide data management training for new researchers

- Work needs to be done to educate students and post docs: there need to be data management courses.

Embed archiving in scholarly communication workflows

- All digital content is in this boat. If we can reduce data archiving to a part of the established publication workflow (as we had with libraries for print copies), dealing with data is a matter of scaling up. Libraries provide LOCKSS physically. Applying LOCKSS to datasets would be robust against a few failures.

Roles of directories/ intermediaries

What can directories of repositories and data do as intermediaries e.g. in disseminating quality status?

Identify repository support for standards

- They could list Data Seal of Approval status next to the repository name...
- The Re3data directory intends to list whether a repository is certified or supports a standard
- Is this more about getting or depositing data? Probably the former is uppermost in most people's mind.

Provide information on usage or sustainability

- If academics can't find anything in a repository, they will be put off depositing their data there. Could you trust repositories to self-describe accurately? I can see how sustainability is the most sensitive issue.

Collaboration: benefits, risks and workflows?

How can repositories, publishers, institutions, researchers and societies work together effectively?

Form partnerships between repositories and publishers

- Partnerships are part of the landscape. Memoranda of Understanding such as Dryad has, and the Joint Data Archiving Policy that some journals have committed to support. The Ubiquity Press agreements with repositories and DANS in particular are an exemplar. There is a need to promote those agreements across communities, perhaps with societies saying these are best practice.

Develop cross-disciplinary consortia from disciplinary infrastructures

- (However) Not many repositories are known as good publishers of data; many are not aware of the new opportunities, and would have to evolve to take advantage of them. These agreements will lead to problems of maintenance. They are not balanced. So what we need is a global consortium comparable to STM, for data; the WDS would be well placed. It has as members not just large repositories but also large organizations (e.g. NASA): having agreements between members, having a WDS logo in the corner of a Website might become a status symbol.
- A global consortium could provide backup for repository losing its funding. Helps with sustainability. It would be about more than just certification.
- Earth science is spoiled with repository provision. We want to get things moving in other disciplines, not go in too low.
- PANGAEA is embedded into a larger research facility, so can do large investments, and provide guarantees for 3 decades. There's a scientific background for the repository: originally it only dealt with data generated in-house, then started taking data from other people. There are parallels here with how Oxford University Press developed as a publisher. The approach has been to build services based on what people wanted to do with stuff internally. Contrast this with Dryad, where researchers deposit and that's it. In a data centre like PANGEA the data gets compiled into databases and products with improved usability.

Cultivate depositors by providing evidence of community support

- If having data in a repository is seen as a status symbol, it solves the problem of giving researchers confidence to deposit there. Okay, not all disciplines will jump on that but it's a step along the way. Biology faculty have heard of Dryad now, and they can see which journals back it. The best infrastructure is invisible. All they want to know is, this is a good place to put data.
- Dryad Board saw Data Seal of Approval as a good thing but initially getting buy-in from the journals was more important. A journal editorial board recommending a repository is a stamp of approval.
- Once chemists know ChemSpider is backed by the Royal Society of Chemistry, they go ‘Oh, that's alright then’.
- Does it matter what the researcher thinks? Isn't it up to the journal? Researchers are only interested in accessibility and preservation.
- How do we factor in approval by the community (i.e. 'This service is valuable to me')? How do we turn this approval into a formal process concerned with security, preservation and other services?
- Usage stats would be useful for this.

Encourage institutional data repository standardisation

- Should institutions apply for the Data Seal of Approval?
- Institutions should think about what they're offering, and what demands will be put upon them. We wouldn't want to impose added conditions on them.
- Sustainable backup is what is looked for. Institutional repositories compete for which will hold data arising from cross-institutional collaborations.
- Even if an institution owns the research data, the culture among academics is for them to take it with them.
- Providing a tiered approach – curated data, managed data, discovery data – adds to complexity. A publisher would have to look at the data itself and not just the repository to know what was being 'guaranteed'
- (However) there is need to find the best strategy for the organization. We can't keep everything forever, so need some selection but want to avoid loss.
- When data is in the institutional repository, it is usually kept in a secure room, conforming to a better degree of security than the lab computers that it was on. So things are already better.

Recognise institutional roles in changing the research culture

- If archivists parachute into projects, researchers keep a handle on their data; it's a different experience to when they just dump it on strangers.
- We found, when working closely with scientists on a project-based approach, that we didn't want to be just an archive. We wanted a more dynamic role. Researchers don't know what their dataset is until they've used it. We often talk about local and remote repositories, without judgement on which is better, but the local ones are important because they're helping researchers prepare data for their own use. They often take too much data so they can pick from it. Now we're asking them to publish data for everyone else. It's a step harder.
- Training is very important for bringing about a cultural change. Sometimes we can persuade people, sometimes we need to force them. In UK, REF 2020 will require that the data underlying a paper be shared before the paper can be submitted in the return: this will drive a massive change in attitude. How will this data sharing be demonstrated? Don't know. Might be done on a sampling basis to save on effort.

Appreciate potential contribution of science studies/ history of science

- It seems there's a view not represented in the discussion: could we learn from a review of historical successes and failures from the history of science)? The digital divide issue, and research meeting the market: could we learn from examples from publishers in the past?
- The British Library commissioned a RAND report that took a broad perspective *Enabling long-term access to scientific, technical and medical data collections*, and it looked at requirements for long-term access.

What makes a repository recommendable to a journal?

If as a journal editor you were given lots of cash, how would you test if a repository was a good place to recommend to authors for data deposit?

- Point to subject repositories, because that's where people will look. People don't look for books by publisher.
- The risk is that the links will break between the paper and the data. Editors don't need to know the details, all they need to know is, does repository do X? If so, tick.

Recognition for social curation by the invisible college

- I'd want to know if the repository already holds quality data for my research community.
- How can you tell what your readership will think is quality?
- There's an underground culture of data sharing: researchers approve recipients before sending them a copy of the data.

Published criteria for collection and review

- By telling researchers to deposit their data, we're asking them to surrender this gatekeeper role to the repository. So they will be interested in their access criteria. For example, do they demand a statement of acknowledgement to the data creator, or that the data creator be listed as an author of any derived paper?
- I'd want to know if the repository actually rejects anything?
- UKDA makes its acceptance criteria very clear: that's great.
- Association with a publication 'ups the ante'. Data publication is interesting: it provides activation energy for researchers to think about the quality of their data. UKDS is changing its policy because very important data needs very thorough and expensive curation.
- *Journal of Open Archaeology Data* requires deposit in in the Archaeology Data Service. This is heavyweight curation; would an institutional data repository provide enough?
- Even an institutional data repository needs quality checks in place before accepting data. In UKDS, the data store data (short term management) is checked, even though it is self-service, and the archive reserves right not to publish the data if it is not up to standard. Do institutional repositories need to have someone doing similar QC checks, or should researchers self-regulate?
- Essex ensures there is no breach of copyright when papers are put in the institutional (eprint) repository – does the data repository need a parallel process? It's a matter of public reputation.
- Most researchers submit their paper and put a copy in the institutional repository. I don't see why they'd put inferior data out there as this would damage their reputation. We pick this up in talks with researchers: they are shy of sharing their data as they are embarrassed that it looks a mess.

- But if they don't see other shared data, they don't have an idea of what the standard should be.
- Things are changing so much, it always looks like we're in the early days. We have to strike a balance of moderation. The policy needs to be simple and accommodating, but with enough teeth to be meaningful. Stuff in Dryad is there because there's a paper involved, so that provides a surrogate quality statement.

Enabling reuse through contextual metadata

- What would researchers need to know in order to reuse data?
- See DIPIR project (http://dipir.org/) which is starting to answer this. Not enough is known about users of repositories: are they using data to check conclusions based on it, or acquiring the data for studies leading to new results?
- We need insights into what reusers are looking for; another case for collecting feedback.
- Metadata standards in this area are