

GRADE

Scoping a Geospatial Repository for Academic Deposit and Extraction

JISC DEVELOPMENT PROGRAMMES

Project

| | | | |
|--|--|-------------------|---------------|
| Project Acronym | GRADE | Project ID | |
| Project Title | Scoping a Geospatial Repository for Academic Deposit and Extraction | | |
| Start Date | 1 June 2005 | End Date | 30 April 2007 |
| Lead Institution | EDINA | | |
| Project Director | David Medyckyj-Scott, EDINA | | |
| Project Manager & contact details | Anne Robertson, EDINA University of Edinburgh Data Library, Main Library Building George Square Edinburgh EH8 9LJ Tel 0131 651 3874 email a.m.robertson@ed.ac.uk | | |
| Partner Institutions | University of Edinburgh, University of Southampton | | |
| Project Web URL | www.edina.ac.uk/projects/grade | | |
| Programme Name (and number) | <i>Digital Repositories</i> | | |
| Programme Manager | Neil Jacobs | | |

Document

| | | | |
|-------------------------------------|--|--|---|
| Document Title | <i>A baseline audit of geospatial asset management within Institutional Repositories</i> | | |
| Reporting Period | <i>Aug 2005 – Jul 2006</i> | | |
| Author(s) & project role | Pauline Simpson, National Oceanography Centre, University of Southampton Waterfront Campus, European Way, Southampton SO14 3ZH. Project Partner | | |
| Date | Jun 2006 | Filename | GRADE Survey Report DeliverableWP4.1 .doc |
| URL | www.edina.ac.uk/projects/grade | | |
| Access | <input checked="" type="checkbox"/> Project and JISC internal | <input type="checkbox"/> General dissemination | |

Document History

| Version | Date | Comments |
|---------|-------------|------------------------------------|
| 0.1 | 11 Jul 2006 | Part 1: James Reid |
| | 16 Jul 2006 | Comments incorporated |
| 1.0 | 30 Oct 2006 | SWOT Analysis added |
| 1.1 | 8 Jan 2007 | Reorder appendices: Anne Robertson |

Citation :

Simpson, P. 2006 GRADE Survey Report. Southampton, UK: National Oceanography Centre, 27pp. (GRADE Project Deliverable WP4.1)



Contents

Part 1: A baseline audit of geospatial asset management within Institutional Repositories

Part 2: Institutional Repositories vs Media Centric Repositories

Appendix 1: Questionnaire

Appendix 2: Questionnaire Responses

Part 1 :

A baseline audit of geospatial asset management within Institutional Repositories

Introduction

The [GRADE](#) Project is one of a cluster of projects in the Digital Repositories Programme funded by the [Joint Information Services Committee](#) of HEFCE investigating the interactions between data and institutional (publications) repositories, support for scientific lifecycle, storage and access requirements.

The JISC is bringing together a programme of work relating to digital repositories. Its aim is to bring together people and practices from across various domains (research, learning, information services, institutional policy, management and administration, records management, and so on) to ensure the maximum degree of coordination in the development of digital repositories, in terms of their technical and social (including business) aspects.

Within this context, GRADE is investigating the technical and cultural issues around the reuse of geospatial data within the JISC IE in the context of media-centric, informal and institutional repositories.

GRADE Work Package 4 has the following objectives:

- To determine to what extent Institutional repositories currently manage geospatial assets
- To determine to what extent Institutional repositories could or should manage geospatial assets
- To investigate the arguments for and against Institutional versus media-centric repositories for geospatial data assets

There is presently a large knowledge deficit in regard to how Institutional repositories currently manage geospatial content. It is patently preferable to understand how existing Institutional repositories approach geospatial asset management and to proceed from this baseline to determine best practice based on a firm understanding of the workflows and processes involved. To the author's knowledge, Institutional repositories do not routinely manage geospatial assets and it is unclear whether the software systems being used could handle the complexities afforded by geospatial datasets.

Generally, geospatial data are not embedded within institutional systems and processes as part of their information management strategy. There is also concern about costs and long term sustainability of repository infrastructure particularly for specialist media such as geospatial data. At the cultural level there is an underlying question as to whether users might identify more closely with an institutional repository or whether media-centric, subject-

centric or community-centric repositories will be a preferred source of information. In the research context, scholars are more likely to think along subject lines and to share data, and indeed be working with colleagues based at different institutions nationally and globally. However these views are based upon anecdotal information rather than any formal evaluation.

A baseline audit of Associate Partner's Institutional repositories and IRs in general has provided an evidence base for assertions on how Institutional Repositories deal with geospatial data in actuality and expose current practices, benefits and limitations. In Part 2 of this deliverable, we go on to weigh the merits of an Institutional vs. media-centric repository approach.

Methodology

To provide community feedback on the issues the GRADE Project organised a web based survey <http://edina.ac.uk/projects/grade/questionnaire.html> (see Appendix 1) initially to its Associate Partners and then wider to the Repository network. This was achieved by email canvassing to individual Repository Managers using the Registry of Open Access Repositories (ROAR), as well as making several calls on JISC-REPOSITORIES@JISCMAIL.AC.UK, and another to the SHERPA Project list.

Although aimed at UK institutional repositories, responses were received from Europe (2) and elsewhere (5) and from institutional repositories and data centres (2). However, the response of 35 replies (in full at Appendix 2) was disappointing even after several call exercises being made and a prize draw for 3 x £50 Amazon vouchers advertised. Responses to questions have been detailed in Appendix 2.

Results

The pre-conception within the GRADE Project was that few, if any, Institutional Repositories (IRs) were dealing with geospatial data, or data of any kind. This has been borne out by the survey responses, but also has something to do with the evolving repository landscape. Many embryo repositories are starting with a limited focus/horizon to ensure that the work involved in obtaining content is pointed at a realistic target. As can be seen from the replies, many repositories are dealing first with publication output since with support from the Open Access Movement they are likely to achieve some measure of success in obtaining content. It could also be resultant on the fact that the open source IR softwares do not have a ready made metadata schema to accommodate datasets. DSpace indicates that it is designed to describe datasets, but does not display the ability within its default metadata schema to accommodate them, If the IR software vendors developed a Dataset plug-in (as they have done for the Research Assessment Exercise) it is possible that Institutional Repositories would have already been challenged to manage them. Had open source softwares included functionality as demonstrated in the [GRADE Repository Demonstrator](#) or the Marine Environmental Data Inventory ([MEDI](#)) it is probable that a few datasets, at least, would have already appeared in Institutional Repositories. The question of whether IRs are the most suitable repository is debated in Part 2 of this report.

The survey responses indicate that IRs would be willing to consider managing this type of output, but have not yet been offered datasets as a deposit. From experience, repositories are certainly offered supporting data which is coupled with a published paper but this is often described very simplistically as a data appendix to the paper rather than as an object in its own right. Authors are very keen to use IRs to expose data particularly graphical data where, as is often the case, the publisher has simplified a diagram or graph in the published journal version. Making the original enhanced versions available through the IR is an opportunity seized upon by researchers wanting to make the detail of their observations globally available.

Where designated data centres exist it is unlikely that Institutional Repositories will be the archive of choice for datasets, but there are many disciplines where there is no formal data archive available, or the data centre has a strict scoping on size and subject of datasets they accept. In these cases rather than datasets existence being hidden on a personal pc, it is possible the IRs may have a role to play. (See Part 2)

PART 2 :

INSTITUTIONAL REPOSITORIES VS MEDIA CENTRIC REPOSITORIES

Subject-based repositories of e-Prints were pioneered in 1991 by Paul Ginsparg at the Los Alamos National Research Laboratory in New Mexico with a collection of preprints of articles in the subject area of high energy physics. This collection, known as arXiv, is now based at Cornell University (<http://arxiv.org/>) and has grown to include materials (but not datasets) in Atmospheric and Oceanic Physics, Mathematics, Computer Science and Quantitative Biology. However, despite the success of arXiv and others like RePEc for the Economics community (<http://repec.org>), there has been only varying success in other subject communities and some have even been terminated (e.g. Chemistry Preprints (<http://www.sciencedirect.com/preprintarchive>))

From 2000 onwards repositories have developed from being subject based to include the complementary institutional-based model and their growth has been fuelled by timely project funding from a variety of sources. Both the Registry of Open Access Repositories (ROAR, <http://archives.eprints.org/>) and the newly developed Directory of Open Access Repositories – OpenDOAR (<http://www.opendoar.org/>) now evidence the increasing number and diversity of repositories: subject, institutional, national, national/subject, international, regional, consortia, funding agency, project, conference, personal, media-centric, publisher and data archives. The dilemma for the researcher depositor is that the above are not mutually exclusive; there is a problem of repository choice.

In the UK the Joint Information System Committee (JISC) Focus on Access to Institutional Resources (FAIR) Programme. (http://www.jisc.ac.uk/index.cfm?name=programme_fair) project funded the implementation of a number of university institutional based repositories, but at the same time also supported the implementation of media centric repositories: e-theses; learning objects, images and museum artefacts. Projects to investigate the linking of text and data were funded in the JISC 2005 call (eg. CLADDIER <http://claddier.badc.ac.uk/trac> and StORe <http://jiscstore.jot.com/WikiHome>); in 2006/7 it is expected that JISC will expand the interest to fund a focus on data repositories.

One of the first policy decisions for Institutional Repository Managers is to scope the content of a repository. Within FAIR some opted to focus on research output only (Southampton) whilst others included any output from the organisation eg. Administrative documents. Another decision is whether to have one repository or maintain a number of repositories each one containing a distinct document; Glasgow or Nottingham are examples. Caltech's CODA Project has spawned some 17 repositories with more in development like Biological Imaging. It is often the case that as Repository

Managers become familiar with their initial content scoping, they find themselves canvassed to take more diverse material into the repository.

One of the initial constraints on scoping was how the repository software metadata schema could cope with specific document types. There are many IR softwares available now, but in the early days, EPrints identified only 'documents' metadata, and although EPrints can accommodate other object types it did not articulate this flexibility. It was not until DSpace in 2002 was distributed, with a great deal of PR, in which it overtly identified datasets as one of its object types that the debate really developed.

Institutional Repositories take responsibility for centralising a distributed activity and new Funder's mandates underpin their content acquisition. They provide the framework, infrastructure and permanence to sustain change. They have an acknowledged responsibility for stewardship, including preservation of their digital assets, and for providing a showcase for the research, teaching and scholarship of the institution. Whilst they maintain metadata quality they do not undertake any responsibility for the quality or correctness of the content. There are now projects investigating the migration and preservation of IR content eg PREServ (<http://preserv.eprints.org/>)

Data Centres, also enjoy mandate support from funders eg. NERC, MRC and international organisations like OECD, whereby datasets resulting from funded research must be deposited into designated data centres. NERC in particular funds environmental data centres <http://www.nerc.ac.uk/data/directory.shtml> (Polar Science; Atmospheric Science; Marine Science; Terrestrial & Freshwater Science; Earth Sciences; Hydrology, Earth Observation) and the UK Data Archive (<http://www.data-archive.ac.uk/>) focuses on digital data from the social sciences and humanities. They take responsibility for stewardship, including preservation and most process data through a rigorous quality assurance procedure. Until the last few years their metadata often followed internal requirements only and also up to fairly recently, their mission was not to showcase research, teaching and scholarship of the institution because they were mainly discipline based. Metadata standards have now risen high on their agenda for interoperability, discovery mechanisms and to raise their global visibility they are now looking at and using, OAI and other international standards.

The GRADE survey indicated that none of the respondents (except those that were data centres!) were dealing with datasets. However, no one ruled out accepting datasets into their repository and many were looking forward to the challenge of dealing with them.

Some repositories have implemented media specific metadata schemas eg. LOM (Learning Object Metadata), and draft standards are becoming available for digital images (NISO). There are also a number of geomatics standards: FDGC ; ISO 19115 for geospatial metadata (2003); UK eGMS etc

At present there is no dataset plugin for EPrint repositories, but there is no reason why there could not be. Clearly DSpace and EPrints need more than Dublin Core with better ways to describe bitstreams themselves. DSpace set

up a Special Interest Group listserv dspace-datasets in 2004 but closed it down in 2005 – was this lack of interest or just too early in the development of repositories? Many more IR softwares are being developed and it is only a matter of time before dataset metadata is an offered option.

Question 6 of the survey was designed to tease out attitudes on who should deal with datasets: IRs or Data Centres. In the main, responses felt specialist data centres were the most appropriate archive, but many felt that both could play a role. As mentioned in Part 1, there is data that is part of the published paper or thesis and most IRs would see that as a responsibility for the IR.

However, an opinion that comes through is that most IRs are general coverage and multidisciplinary, do not quality control content and at present do not have data management skills on their teams. For a detailed SWOT analysis of IRs dealing with geospatial data refer to figure 1. Data Centres are in the main, discipline specific and have the subject and data management skills and carry out QA procedures, but would they be prepared to manage datasets of any size? Large of course, but there are small datasets which at this moment in time may not be accepted by a data centre. Here an IR can play a part.

Another point was made that Repositories tend to be about bringing together institutional assets, whereas Data Centres tend to bring together subject specific datasets.

One response talks about “feature creep” with repositories, but this is true of data centres also. Whilst IRs are considering dataset inclusion, we find now that data centres are ingesting the publications derived from the datasets – the question could be asked, what skills do data centres have for dealing with document type objects?

Data Centres have preservation as a prime part of their mission, but now IRs are heavily involved in projects for preservation of their content. Do either of them have the definitive answer?.

A number of responses make the point that as long as the dataset is freely available does it matter where it is archived - the options are not mutually exclusive. Research funders *Position Statements on Open Access to Research* <http://www.sherpa.ac.uk/juliet/> require that publications resulting from public funded research should be deposited in an appropriate institutional repository. Supporting this, the JISC are setting up a repository that will be available for any researcher who does not have access to an institutional or thematic or funder repository. Is this a model that should be set up for datasets? The GRADE Repository Demonstrator offers one solution – multidiscipline but geospatially described data repository .

| STRENGTHS of an IR dealing with geospatial data | WEAKNESSES of an IR dealing with geospatial data |
|--|--|
| <ul style="list-style-type: none"> ■ One repository – less administrative and technical overhead ■ Linking text, datasets, images easier within one environment ■ Showcase for all institutional research ■ IR Software - Open Access – interoperability – visibility ■ Software based on International Standards ■ Metadata skills provided by Information community ■ Formal Dataset citation ■ Supports Citation analysis and metrics for research funding and personal promotion for data generators/managers | <ul style="list-style-type: none"> ■ Software not designed to cope with data ■ No IR metadata schema for datasets yet ■ IR staff without Data Processing skills ■ IRs do not quality control content ■ IRs not involved in production of information products ■ Storage – Preservation (all media types) ■ OA culture not yet extended to data altho OEDC, EU and some Research Councils etc. mandate deposit of data emanating from funding. |
| OPPORTUNITIES of an IR dealing with geospatial data | THREATS of an IR dealing with geospatial data |
| <ul style="list-style-type: none"> ■ Contribute to the design of an IR dataset metadata module ■ To offer a data archive (where non exists) ■ Treats ‘orphan’ datasets not accepted by DCs ■ Enhancement of IR staff skills ■ Showcase in one digital repository of all research output ■ Ready host when dataset deposit is mandated ■ Integration – joined up research ■ Additional Funding opportunities from e-Research projects ■ Input to the Data citation model ■ Data and Information communities working together ■ Collaboration between disciplines ■ Dataset harvesting from IRs to Data Centres | <ul style="list-style-type: none"> ■ Turf war between IRs and DCs ■ Will funding follow to IRs ■ Will funding stream for data management reduce? ■ Too large an undertaking for IRs ■ Data lost in publication ‘bucket’ ■ ‘Thematic’ datasets distributed ■ No migration/preservation policy ■ Datasets fall ‘between stools’ |

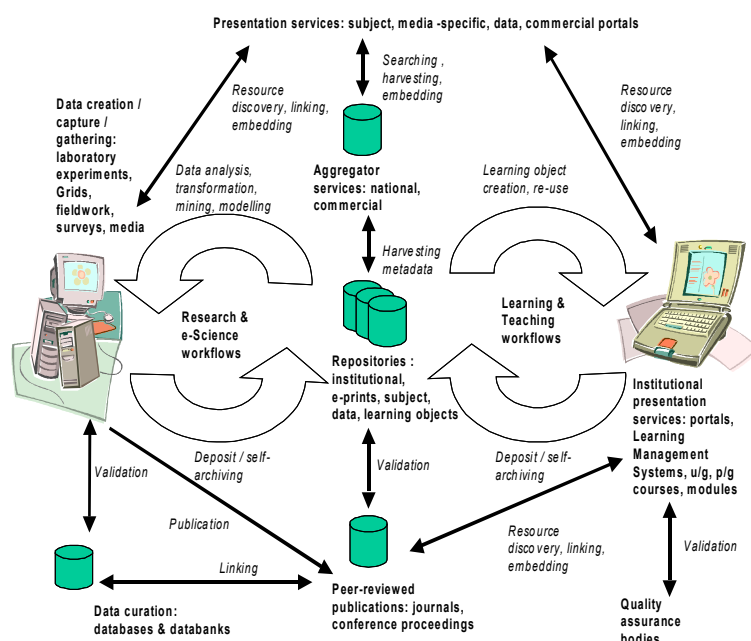
Figure 1: SWOT analysis of IRs dealing with geospatial data

In the world of e-Research, the ideal scenario is when data and documents and images etc will be linked and easily accessible.

The StORe Project conducted discipline-specific surveys, a comparison review is available at <http://jiscstore.jot.com/BusinessAnalysis>. It revealed that there is common ground in terms of a need for two way links between raw data repositories and academic publication repositories. Such links were considered useful by participants. Noticeable variations in the way that data are gathered, formatted, allocated metadata and subsequently shared (both between disciplines and within disciplines) were noted, and this needs to be taken into consideration. It is likely that the discipline-specific requirements will result in a need for customisation of any national geospatial data repository.

Open Archive protocols and metadata standards are a part of the Scholarly Knowledge Cycle (Fig 2), where repositories and data centres play symbiotic roles.

Fig. 2: The Scholarly Knowledge Cycle. Lyon et al 2004



The repository landscape is rapidly evolving. Perhaps a pragmatic approach is called for at present: If a researcher has access to an appropriate data centre to deposit the dataset then that should be the preferred route provided that the papers and publications resulting from the dataset are linked (The JISC CLADDIER Project is investigating linking IR and Data Centre content). However, if a researcher does not have access to an appropriate data centre, is it not better that the dataset is at least deposited in a trusted repository? Leaving the dataset on the researcher's pc, as is often the case now, will ensure it is 'lost' forever.

Acknowledgements

Thanks go to Anne Robertson for her comments on the initial draft survey and to EDINA services for hosting the survey and fielding the results.

References

GRADE Repository Demonstrator. Available:
<http://gradedemo.edina.ac.uk/dspace/index.jsp> [11 July 2006]

Lyon, Liz, Heery, Rachel, Duke, Monica, Coles, Simon J., Frey, Jeremy G., Hursthouse, Michael B., Carr, Leslie A. and Gutteridge, Christopher J. (2004) eBank UK: linking research data, scholarly communication and learning. In, *eScience All Hands Meeting*. Swindon, UK, Engineering and Physical Sciences Research Council.

<http://eprints.soton.ac.uk/8183/>

Reed, G. 2002 MEDI [Marine Environmental Data Inventory] : the IOC Metadata System. Paris: Intergovernmental Oceanographic Commission/International Oceanographic Data and Information Exchange
[PowerPoint](#)

Appendix 1 – Questionnaire

GRADE

Scoping a Geospatial Repository for Academic Deposit and Extraction

Grade Survey

GRADE will investigate and report on the technical and cultural issues around the reuse of geospatial data within the JISC IE in the context of media-centric, informal and institutional repositories.

A Work Package requirement is to carry out an audit of geospatial asset management within institutional repositories. The survey below only takes five minutes to complete; all completed responses will go into a prize draw for a £30 Amazon book voucher to be drawn during the week commencing 13th February 2006.

For the purposes of this survey geospatial data is defined as data explicitly containing coordinate geometry ie vector, raster, geo referenced images, text files, containing x,y coordinate values. (eg. Electronic maps, geo-referenced imagery, satellite data, data stored within a Geographic Information System (GIS))

Survey Questions

1. If you have a repository, what software do you use?

2. Is your repository publicly available, if not who are your depositors and users?

3. Do you accept the deposit of geospatial datasets into your repository? If so how much? If not do you plan to?

4. What special metadata fields do you offer in your repository to describe geospatial data - how can users search geographically for geospatial data?

5. If you have geospatial data, do you...

Receive supporting documentation from the depositor?

Have guidelines on file format required?

QA the data at all?

Require a declaration of ownership / copyright from depositors?

Confirm if derived from other datasets?

Have agreements to deal with issues of liability from use of the data?

6. Generally, do you think that archiving and providing access to research data is something institutions should do or specialist data centres?

7. What processes, if any, do you have in place to ensure that long term access to research data will continue e.g. migration of data so that it remains readable?

8. Please list any other Institutional Repositories you know that are managing geospatial datasets

COMMENTS :

Your Institutional Details

Organization:

Address:

Contact:

e-Mail ID:

Appendix 2 – Questionnaire Responses

1. If you have a repository, what software do you use

| | | | | | | |
|----------------|--|---------------|----------------|-------------|-----------------|-----------------|
| 1 | We do not. | | | | | |
| 2 | Eprints.org | | | | | |
| 3 | DSpace | | | | | |
| 4 | DSpace 1.2.2 | | | | | |
| 5 | GNU Eprints software | | | | | |
| 6 | ePrints.org | | | | | |
| 7 | eprints | | | | | |
| 8 | Home grown software | | | | | |
| 9 | N/A. Although there was the intention to use AcrSDE for departmental assets | | | | | |
| 10 | Eprints | | | | | |
| 11 | Digitool v3 from Ex Libris | | | | | |
| 12 | DSpace | | | | | |
| 13 | ePrints base install with considerable reconfiguration | | | | | |
| 14 | A Canadian implementation of D-Space | | | | | |
| 15 | DSpace | | | | | |
| 16 | DSpace | | | | | |
| 17 | Eprints | | | | | |
| 18 | I am building one to interoperate with interoperable repository services e.g.OAI with dSpace | | | | | |
| 19 | E-prints | | | | | |
| 20 | DSpace | | | | | |
| 21 | Fedora 2.1 - in development | | | | | |
| 22 | CDSware | | | | | |
| 23 | www.eprints.org | | | | | |
| 24 | E-prints | | | | | |
| 25 | DSpace (postgreSQL) | | | | | |
| 26 | eprints, probably moving to dspace soon | | | | | |
| 27 | D-Space | | | | | |
| 28 | DSpace | | | | | |
| 29 | Not yet established a repository is a lot of interest likely to happen in the next 12months | | | | | |
| 30 | ARNO | | | | | |
| 31 | Don't at the moment have recently acquired DSpace for other purposes, thinking of using | | | | | |
| 32 | Don't have one. | | | | | |
| 33 | Fedora | | | | | |
| 34 | Eprints | | | | | |
| 35 | Home grown. | | | | | |
| EPrints | DSpace | Fedora | CDSWare | ARNO | Digitool | In-house |
| 12 | 12 | 2 | 1 | 1 | 1 | 2 |

2. Is your repository publicly available, if not who are your depositors and users?

| | |
|----|--|
| 1 | N/A as we do not have a repository |
| 2 | All publicly available. |
| 3 | Publicly available to readers, deposit restricted to members of the University of Bristol |
| 4 | The repository is for the deposit of institutional material. All content is open access. |
| 5 | Yes it is publicly available |
| 6 | Yes |
| 7 | publicly available |
| 8 | Yes to reading. Deposit limited to members of CCLRC staff. |
| 9 | In principal this would have been for departmental staff and students |
| 10 | Depositors must be members of Royal Holloway. Access is open to all. |
| 11 | Not yet fully online. Material deposited will be digitised versions of the sound recordings, images, film and video held in our collections. Our users will be researchers and the communities from whom the material was originally collected |
| 12 | publicly available |
| 13 | Yes: ecrystals.chem.soton.ac.uk. Subject specific repository for Crystallography community. Also operate development repository for chemistry community to deposit analytical data under the JISC R4L project |
| 14 | Yes. |
| 15 | Publicly available |
| 16 | Content is almost all available to the world to see; depositors are affiliated with our institution. |
| 17 | Public use, private deposit |
| 18 | na |
| 19 | Public user access- Strathclyde university community can deposit. |
| 20 | Yes - OAI compliant |
| 21 | Not yet available. Eventually users will be University staff (academic, administrative, support) and students; some public access. |
| 22 | Depends how you define 'repository'. Assuming it is the whole database, then it is mostly available publicly. There exist some collections of items that are not made available off-site due to restrictions by the copyright owners (eg full-text of standards) and also some that are confidential items and retracted to selected groups of users (egg institutional financial documents). These collections are not preprints but other types of items. Deposit is limited to users on-site. |
| 23 | publicly available |
| 24 | Publicly available, our depositors are our academic and research staff. |
| 25 | Currently a pilot, but intended that depositors will, primarily be academics and postgraduate students. Repository will, generally, be open access. |
| 26 | Yes |

| | |
|---------------------------|---|
| 27 | Yes. |
| 28 | Not as yet, we are in the building stage but it will be available to the public and all areas of the University will have the ability to deposit material |
| 29 | most likely to be publicly available |
| 30 | YES |
| 31 | Don't have one as yet. |
| 32 | n/a |
| 33 | Yes, it will be when development is complete |
| 34 | Yes |
| 35 | Yes, but with access control dictated by individual datasets. |
| Publicly available | For defined community |
| 28 | 2 |

3. Do you accept the deposit of geospatial datasets into your repository? If so how much? If not do you plan to?

| | |
|----|---|
| 1 | N/A as we do not have a repository |
| 2 | Have not yet been asked to do so - but would be interested to try. |
| 3 | We accept any published or intended-for-publication academic research outputs, including journal articles, book chapters, presentations, conference papers and posters, etc. At the moment we have no geospatial outputs. |
| 4 | Not specifically, but if a researcher had nowhere else to archive it we would take it. |
| 5 | No. We don't plan to yet, but may review this in the future. |
| 6 | We haven't had any so far, but would accept them in principle if deposited. |
| 7 | No, there is no emphasis on or current plan to collect primary data from any subject discipline at the moment. Our repository targets journal papers, conference proceedings, book chapters and similar. We've had a job on our hands to get hold of this material. So there has been little discussion, to date, of holding datasets in general and no discussion of geospatial datasets in particular. So, no, it's not something we're planning at the moment. |
| 8 | No. Our IR is limited to documents and PowerPoint presentations. |
| 9 | In principal this would have been free. The benefits of re-use within the internal community (particularly the benefit to research) should outweigh the deposition costs |
| 10 | Not at present. We will consider it if any department has a need for it |
| 11 | We are thinking about a pilot project to examine this |
| 12 | We do, although none have been deposited as yet. |
| 13 | Do not accept geospatial data (or plan to) as our repositories are subject (chemistry) specific. |

| | |
|----|---|
| 14 | No. We do not accept any datasets in our repository. |
| 15 | Yes, but nothing has come in yet. |
| 16 | We would accept geospatial datasets; to date none have been offered. |
| 17 | No |
| 18 | could be but no requirements so far |
| 19 | not at this stage |
| 20 | At present, we are concentrating on published material. However, we are hoping to encourage other material in the future. |
| 21 | Not yet. Possibly in the future, relating to some image collections, geospatial metadata relating to the position a photo was taken. In the very long term, geographers and archaeologists (eg) may want to deposit datasets. My crystal ball doesn't see that far! |
| 22 | Not currently. Not aware of any plans. |
| 23 | yes, but none has done thus far. |
| 24 | No plans to at present as we are concentrating on bibliographic research output. |
| 25 | We do not at present accept such data (or at least we have no specific facilities to support their acceptance) and these are not a priority. We see it as a possibility in future. |
| 26 | We may accept it in the future |
| 27 | Not yet, but it is under discussion. Our first attempt will be with scanned historic maps. |
| 28 | We have had a request from the academic side for this and so yes it will be a service provided - but in Stage 2. Stage 1 is focusing on the traditional "paper" formats |
| 29 | This is not something we have yet considered, although there are researchers working with geospatial data in relation to health and so does require thought. We are about to undertake a digital asset scoping survey to try and establish what exists in the institution and hopefully this will encompass geospatial data |
| 30 | NO. WE PLAN TO MAKE REPOSITORY SUITABLE FOR DATASETS IN GENERAL |
| 31 | Have no repository as yet but we would consider it. |
| 32 | n/a |
| 33 | Yes, geospatial datasets are ingested as well as other kinds of data. |
| 34 | Not at the moment but we are currently considering whether our Institutional Repository could include datasets and if so in what form. |
| 35 | Yes, Most are geospatial |

| Yes | No | Interested |
|--------------------------------------|----|------------|
| 7 | 18 | 6 |
| Not easy to categorize these answers | | |

4. What special metadata fields do you offer in your repository to describe geospatial data - how can users search geographically for geospatial data?

| | |
|--|--|
| 1 | N/A as we do not have a repository |
| 2 | Not sure of the needs, so far. Doubtless we will get faced with this soon . . . |
| 3 | No special provision |
| 4 | None |
| 5 | - |
| 6 | This needs checking. We would work in collaboration with the data Librarian |
| 7 | N/A |
| 8 | Not applicable. |
| 9 | N/A not implemented |
| 10 | Our only metadata is the standard eprints set |
| 11 | undecided as yet |
| 12 | None |
| 13 | N/A |
| 14 | No. We do not accept any datasets in our repository. |
| 15 | To Be Determined. |
| 16 | We currently have no special metadata fields for geospatial data. |
| 17 | N/A |
| 18 | na |
| 19 | none |
| 20 | N/A |
| 21 | Not yet decided. Possibly extended Dublin Core - spatial? Image metadata is likely to use MIX but this doesn't seem to have a geospatial component. |
| 22 | n/a |
| 23 | n/a |
| 24 | n/a |
| 25 | None |
| 26 | Currently no specific metadata fields |
| 27 | We have not yet designed metadata for the maps. For GIS data, we will likely use a local variant of Dublin Core, with some elements mapped from FGDC. |
| 28 | Not there yet |
| 29 | N/A |
| 30 | NONE |
| 31 | Have no repository as yet. |
| 32 | n/a |
| 33 | When development is complete both DC and FGDC CSGDM will be used for discovery. |
| 34 | N/A |
| 35 | We use NetCDF with CF conventions and NASA-Ames file formats to force some metadata into the files. We only offer a very primitive geospatial search as yet. |
| Yes x 4 | |
| MIX; FGDC; FGDC CSGDM; NETCDF WITH CF CONVENTIONS; NASA-Ames | |

5. If you have geospatial data, do you ...

| | |
|--|---|
| Receive supporting documentation from the depositor? | 2 |
| Have guidelines on file format required? | 1 |
| QA the data at all? | 2 |
| Require a declaration of ownership / copyright from depositors? | 3 |
| Confirm if derived from other datasets? | 2 |
| Have agreements to deal with issues of liability from use of the data? | 2 |

6. Generally, do you think that archiving and providing access to research data is something institutions should do or specialist data centres?

| | |
|---|--|
| 1 | Both. The institution should be responsible for archiving their own data in a specialist data centre, which can provide access to the data to the appropriate parties based on restrictions defined by the producer/archiver |
| 2 | There is a role for both. I would see institutional repositories as serving the needs of academics in storing, exposing and allowing their data to be used. This might be best done at a local level, with ties into demonstrations of work being undertaken by a dept or research group, or learning systems, or supporting virtual research environments. With highly specialized data - or very very large volumes - a specialist data centre may be more appropriate. It also depends on the relationship between the data centre and the academics which it serves - are they responsive? Are they flexible? It may be that academics would like to keep thing local. There is a danger of "feature creep" with repositories, and I think that there should be clear plans when new data types or extensions to a repository's holdings are undertaken. |
| 3 | There is room for both. The important thing is to make research outputs freely available. The institutional structure sometimes helps to persuade researchers to comply. The institutional repository also has added value for the institution in providing a shop window for its research. |
| 4 | A specialist data centre would be preferred. I think that research data would possibly be better off being archived in specialist data centres rather than institutional repositories. Practically speaking as a repository manager, IRs already have enough overheads without worrying about collecting/preserving another type of content. I would rather |

| | |
|-----------|---|
| | focus and consolidate our original aims before diversifying. I don't think the institutional model for data sets is correct - each subject area has widely different needs that need to be catered for. Creating one repository to suit all would be problematic and may not address individual subject areas needs. Far better to create specialist centres that have the expertise to develop and maintain data-set repositories. |
| 5 | I do believe institutions should do this. Many institutions specialize in specific subject and regional areas and I think they are best placed to understand the finer points of their specialties. |
| 6 | Probably a combination. Specialist data centres should probably be the first port of call, but failing that, IRs can mop up anything not accepted by them. Also, institutional policy needs agreeing - if the IR is designed to preserve all institutional research output there is an argument for keeping the dataset (or a copy) in the archive. |
| 7 | I don't feel very confident answering this question. My initial inclination is to say that specialist data centres are likely to be better equipped to deal with research data. Institutions support such a diverse range of subject areas - each with its own data needs - that it may be difficult to cater effectively for multiple data types and the multiple metadata sets which may be required to handle these effectively. Therefore, data centres seem a more logical choice.. However, I can see a case for distributed data collection - along the lines of the institutional repository model - with an upper service layer collating, enhancing and manipulating the raw data as provided by individual institutions. On the whole, though, I'd say data centres. |
| 8 | I believe that data should be curated in specialist data centres and we should concentrate on providing links between documents (text) and data. I am part of the JISC funded CLADDIER project which is working on this type of approach. |
| 9 | Ideally this should be through specialist data centres so the costs can be minimized. However, only if the data can be accessed transparently at no/low cost. Furthermore, there may be copyright implications for a central repository that can be obviated by using an institutional repository |
| 10 | It is perfectly feasible for institutions to do it, but specialist centres may be better able to devote sufficient resources to preservation and accessibility. |
| 11 | Archiving is probably best done by specialist centres with expertise in handling such data, and the tolls and recurrent funding to sustain the data. In our case there is no alternative, as we are legally charged with preserving material deposited with us so we have to develop expertise in these areas |
| 12 | yes |
| 13 | It is imperative that research data underpinning a study is made publicly available for the purposes of thorough (peer) review, dissemination and availability for reuse. This should be facilitated and made available by institutions and/or subject specific data centres in a manner that is most appropriate to the discipline concerned. |
| 14 | Institutional repositories are great for things like those we store--reusable teaching material. While they are based on datasets, the datasets themselves are kept |

| | |
|-----------|---|
| | separately. The metadata are stored in DDI compliant form. IRs simply are not a feasible option if the data are to be preserved over the long term. In Canada, at least, there are no data specialists employed by IRs and the materials that are stored are for the most parts the results of research rather than primary research resources. |
| 15 | I don't see this as mutually exclusive. Both can take care of these datasets. |
| 16 | It really doesn't matter who does it as long as it is done somewhere, and the organization doing it is stable and has a chance of longevity. |
| 17 | Yes to both. Specialist centres can provide specialist services and can pioneer best practice. |
| 18 | institutions as this is their intellectual capital |
| 19 | both. Specialist centres should provide customized (GIS-based) services and interfaces to allow for complex data retrieval and other services. Institutional repositories should provide generic access to allow for university asset management and availability of raw data to support associated papers (even if the general institutional interface does not [and should not in my opinion] support retrieval based on specific GIS metadata fields). |
| 20 | I think it may be useful for institutions to work together with data centres. Institutions are in a good position to gather information and data centres are a good way to disseminate the data. |
| 21 | Institutions at the moment. |
| 22 | Both have their place. |
| 23 | Institutions should do this. |
| 24 | Probably specialist data centres would be better. |
| 25 | Probably both or either (!) depending on the nature of the data set. It seem sensible to store a data set upon which a thesis depends (and whose acquisition was part of the writer's work) should be stored alongside the thesis itself. Equally, specialized data might be better stored in a specialized repository. These approaches are, of course, not mutually exclusive. |
| 26 | Believe that there is a role for both institutions and specialist data centres to play, depending on whether the focus of the data gathering is institutional or subject/discipline based. |
| 27 | Institutions. |
| 28 | From a service manager's point of view I like the idea of "outsourcing" but academics are keen to retain strong branding and see the repository as an additional way to market the institution. I also don't actually have any additional money for outsourcing. |
| 29 | I think that there is a role for both, institutions need to collect, make available and preserve the data that is used by their researchers; data centres have a role in bringing together data in specific subject areas |
| 30 | INSTITUTIONS SHOULD INCLUDE DATASETS IN THEIR REPOSITORY |
| 31 | Specialist data centres would be best if coverage is good and perpetual open access guaranteed by contract. |

| | |
|-----------|--|
| 32 | Probably this is a role for specialist data centres not individual institutions. |
| 33 | Actually, we are a data center. |
| 34 | There are benefits to Institutions in terms of branding of research profile and linking to other outputs but specialist data centres may find it easier to cater for subject specific needs. There are advantages and disadvantages to both. |
| 35 | Specialist data centres - the tools and formats are too specific for institutions. |

7. What processes, if any, do you have in place to ensure that long term access to research data will continue e.g. migration of data so that it remains readable?

| | |
|-----------|--|
| 1 | none |
| 2 | Looking into preservation/access long term through SHERPA DP |
| 3 | We are committed to long-term preservation, but it is too early to be able to demonstrate processes or results. |
| 4 | We are involved in a project to implement the OAIS-reference model for all content in the repository. In the short term we follow normal DR procedures (back-ups etc). With a view to developing longer term practices we are participating in the SHERPA DP project which is currently investigating these issues |
| 5 | At the moment plans are in the discussion phase. Our repository server is shared with a number of other institutions and in the future we hope to build our own server. We have discussed issues about format. We currently accept PDF and ASCII. |
| 6 | We shall be looking into this over the coming year or so |
| 7 | N/A |
| 8 | Migration of data and associated metadata to ensure that not only is the file readable but there is enough information about the experiment to be able to analyze the results meaningfully. I think that there are more problems associated with the production & migration of meaningful metadata than migration of the bits and bytes. |
| 9 | Wherever practicable use of interoperable formats |
| 10 | None at present, but we are participating in the SHERPA DP project to study this. |
| 11 | Choosing well known file formats, well known physical data formats, and using publicly described and open formats where possible |
| 12 | We are currently developing a digital preservation program at our institution. |
| 13 | Working towards institutional support (akin to that provided by current |

| | |
|-----------|---|
| | ePrints repositories) to ensure long term archive and curation of data. |
| 14 | - |
| 15 | We have standard procedures for many types of files, which we will need to adapt to datasets. |
| 16 | oh dear. So far all we have is promises; we don't really know how we're going to do it. It still feels like we just started... |
| 17 | Putting it into a live repository is a first step - otherwise it would stay on degrading offline storage. |
| 18 | early days still in this respect |
| 19 | I'm not running a repository so I don't know. As far as I know Strathclyde is not currently accepting datasets. |
| 20 | N/A We have been limiting the formats we use within our repository to ease any future preservation problems. |
| 21 | Too early to say - however images (our earliest and so far only work area) have an archive datastream to hold the purest rendition we have (say, uncompressed TIFF) so that format migration could be achieved in the future. Other content models will incorporate a similar philosophy. |
| 22 | Data produced here is not currently made publicly available although there are plans to change this in future. It is available on special tapes and managed by the IT department, not by the library. Maybe there are plans in place but we are not part of them. |
| 23 | none except hoping that eprints will survive. |
| 24 | Nothing as yet |
| 25 | None, so far. |
| 26 | Regarding PDF as a primary long term format and are still evaluating processes to ensure long term access to other storage formats |
| 27 | Currently we promise to migrate PDF files if that format ever becomes obsolete. We also perform checksum verification to ensure data integrity. |
| 28 | Still early days, but we are working with the academics on a list of file types that we will migrate, geospatial may or may not be included again it is early days |
| 29 | This is rather patchy at the moment, although a JISC funded project on the development of a retention schedule for research data was carried out in 2003, which helped to raise awareness. It is an area that may be addressed during the digital asset scoping survey mentioned above. |
| 30 | NON AT THE MOMENT. FOR PUBLICATIONS LONG TERM PRESERVATION IS GUARANTEED BY THE DUTCH ROYAL LIBRARY. FOR DATASETS DISCUSSIONS ABOUT LONG TERM PRESERVATION ARE STILL GOING, POSSIBLE OPTION IS A ROLE FOR DUTCH NATIONAL SCIENTIFIC RESEARCH ORGANIZATION. |
| 31 | Have no repository as yet. |
| 32 | n/a |

| | |
|----|---|
| 33 | Multiple file formats are accepted and documented. Data conversion processes are being developed. We also have a long-term archive under development. |
| 34 | None currently but we would consider this as part of taking on any datasets |
| 35 | Media migration Dataset review procedures Format review Consumer and produces interaction |

8. Please list any other Institutional Repositories you know that are managing geospatial datasets

| | |
|----|---|
| 1 | My only knowledge of institutional repositories from Universities I have been at in NZ and the US are ad-hoc directories of who has what. The best or nearest to a "repository" that I have come across in a University was a collection of data on a server that was used for teaching and student projects. |
| 9 | Other institutional repositories with geospatial datasets = ADS; Maryland University; ESRI; Google; NASA; NERC |
| 13 | TIB World Data Centres (eCrystals is registering DOI's for datasets with TIB Hannover) |
| 19 | AHDS; GO-GEO; EDINA |
| 30 | AT THE MOMENT IN THE NETHERLANDS THERE ARE NO IR'S WITH GEOSPATIAL DATASETS. THE TECHNICAL UNIVERSITY OF DELFT RUNS A PROJECT CALLED DARELUX (http://www.library.tudelft.nl/darelux/index.htm) IN WHICH HYDROLOGICAL DATA ARE COLLECTED. |

COMMENTS :

| | |
|---|--|
| 6 | You may wish to contact our data librarian who is more knowledgeable about datasets. Contact details: Luis Martinez , l.martinez@lse.ac.uk |
| 7 | I can see the logic in asking researchers to deposit their data into their institutional repository as repositories are already asking for research outputs. It may be that IRs become a more attractive proposition for depositors if we are offering functionality which fits with what they would like to do i.e. not just disseminate their research outputs but capture, preserve and offer for re-use their primary data. I'm just wondering if the sheer variety of disciplines involved makes this an unlikely route. If GIS data is housed in this way, then why not a whole host of other data from other subject areas? Could the IR offer an appropriate range of specialist metadata fields to cater effectively for such a variety of requirements? There may be greater economies of scale to be had in specialist services. Leeds has been looking at support for e-research and exploring use of, for example, the |

| | |
|-----------|---|
| | White Rose grid, and this has generated some discussion of handling datasets - but discussion is about as far as this has got. Specific consideration of geospatial datasets may be taking place elsewhere in the University - but hasn't made it to the IR agenda yet. |
| | I'm not running or managing an IR (I just have a few opinions about metadata and repository services). |
| 30 | UNIVERSITIES AND INSTITUTES IN THE NETHERLANDS ARE WORKING TOGETHER IN THE SO-CALLED DARE (DUTCH ACADEMIC REPOSITORIES) PROGRAMME, COORDINATED BY SURF ('the Dutch JISC'). STORING DATASETS IN IR'S IS PART OF THIS PROGRAMME (DARELUX PROJECT). |
| 31 | While geospatial data presents technical problems, we might well be prepared to give it a home even if just in a raw format (where the responsibility would be on the user to solve the technical difficulties of usage). |