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 The University of Edinburgh Causewayside House 160 Causewayside Edinburgh EH9 1PR Anne.robertson@ed.ac.uk		
Partner Institutions	University of Edinburgh, Kingston Centre for GIS, University of Nottingham, University of Strathclyde Glasgow		
Project Web URL	http://gradedemo.edina.ac.uk/dspace/index.jsp		
Programme Name (and number)	*Digital Repositories*		
Programme Manager	Neil Jacobs		

Document

Document Title	*User based evidence for the requirements and functionality of a repository capable of managing licensed geospatial assets (public-version).*		
Reporting Period	*Jan-2006 – April-2007*		
Author(s) & project role	Rebecca Seymour GRADE Project Officer, EDINA Bex.seymour@ed.ac.uk		
Date	April-07	**Filename**	FormalGISRepositoryFeedbackFinal.pdf
URL	*www.edina.ac.uk/projects/grade*		
Access	☐ Project and JISC internal	√☐ General dissemination	

Document History

Version	Date	Comments

1	10/5/07	FormalGISRepository_FeedbackFinal.doc
2	10/5/07	FormalGISRepository_FeedbackFinal(public_version).doc

Contents

**Greyed out parts where data has been edited for confidentiality on public version of this report.*

Executive Summary

A geospatial repository set up for exploring the cultural and technical issues of sharing derived GIS research data in the UK HE-FE sector was created. A goal was to establish a framework guideline as to the components required for building a repository capable of absorbing complex GIS data (storage and retrieval and ingest), and based on direct-user feedback from the academic and GI-community.

The central geographical components highlighted were: to have a map-based search (with basic GIS functionality, zoom, pan, define a box, or click on boundary areas, and with detailed geographical elements to locate a position accurately). To search via one's academic Institution or within an established GIS/geographical peer group's (reflecting patterns of current informal data-exchange particular to the GI-community of researchers). To have explicit geographical collections grouping data via continental and country definition (the most important to UK researcher's being the UK, and EU), and to search via a gazetteer of place names. Finally, to search via research subjects (clear thought to geographical standards and ontology's used here was considered vital, particularly with the cross-over of disciplines using GIS/mapping data). In terms of discovery, basic Dublin Core metadata was considered sufficient, although there was desire for this to be extended and enriched with certain 'geographical' elements (such as, coordinate and projection datum, viewable extents, licensing/IPR rights, and data sources and lineage information). GI-community users were more 'visual' in there approach to the repository and the use of standard GIS icons for file types and thumbnail images was supported for this quick-discovery stage. In all cases downloading data straight to a desktop GIS was the favoured download method (online GIS environments and live streaming would only be used by advanced GIS users and only if they operated in a fast-network- this was a key). Downloading a GIS file in the same format as it was submitted was not considered a problem at this level (competent GIS users who are able to do file conversions at ease – this may not be the case with users from non-geography/GIS disciplines) At a download stage a full ISO19115 metadata to accompany all data downloads was considered superior. The issue of referencing a GIS dataset within a geospatial repository in a resultant publication is one that needs to be addressed.

In terms of ingest to a geospatial repository there were a number of key geographical elements that are essential. There is a need for a serviceable GIS file verification (GIS file types have complex arrays of component files and wide number of GIS file types and sizes, there needs to be a automatic control of this). Automatic ingest and metadata creation tools were considered vital for a durable long-term repository so that data input is not a burden, and to allow constant streaming of in situ live datasets (i.e. climate models for example – many GIS files are complex and multi-layered in nature). Further to this is a desire to have an ability to deal with upload of 'packaged' datasets, an ability to retain the internal integrity and links of multi-layered and inter-connecting data within a project (dividing this all up singularly can erode the purpose of the research). A geospatial repository will need to be able to deal with single data and retain these links in order to searchable at the bottom and still have hierarchy in the database. A repository would also have to be able to deal with GI-data-versioning. It was considered important that a repository was wider than just the UK (could allow non-UK data submit).

Further elements and issues that would be important for any derived **data** repository were also flagged. Issues of sustainability such as archival, longitivity of data within a repository (and of trust – that data will be looked-after, curated and maintained), of data harvesting from other sources, of machine to machine interoperability between repositories. And other more specific things, such as a need to search via date in numerous forms (data deposited, data created, date/duration/span-of-dates of research), searching via data-quality measures, in terms of submitting data to have set rules for title and date formats, more defined 'roles' of people connected to a data (depositor, institution, creator, source/contributors), offline functionalities.

Finally, more cultural issues of utilising a repository for GI-data sharing to promote research were considered. Issues connected to licensing and IPR rights of geospatial data was a fundamental issue. Willingness to share research data, and of quality pre- and post- publication and of usefulness of derived GIS-data (contribution to a GI-research community), were proven through an expanded user-base of the demonstrator.

Part 1: Introduction to the GRADE repository demonstrator, its purpose and usage statistics; and its 3 stages of technical development

Background: The ethos & purpose of the GRADE project and geospatial repository demonstrators.

The GRADE project (http://edina.ed.ac.uk/projects/grade) is one of a cluster of projects in the Digital Repositories Programme funded by the Joint Information Services Committee (www.jisc.ac.uk) 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 (Geospatial Repository for Academic Deposit & Extraction) is investigating the technical and cultural issues around the reuse of geospatial data within JISC IE in the context of media-centric, informal and institutional repositories.

The aim of GRADE work package 1 is to investigate the storage, management and access to geospatial data derived from licensed data within repositories. This was done using a 'learn-by-doing' philosophy to investigate issues, using a sandbox demonstrator media (i.e. subject specific) repository. This test bed environment, media centric repository demonstrator, will be used to evaluate issues relating to functionality and content packaging, digital rights and IPR, quality control processes, the types of support depositors and those extracting geospatial data will require, how repository activity and practices may be embedded within research and learning processes, and sustainability and business models required by the UK HE-FE GI-community.

The first activity in the work package was to evaluate existing repository software and applications, both commercial and open source (e.g. DSpace, Fedora, Intralex Intra-library) for suitability. It was recognised that software would need to support the following functions: upload – mechanisms and processes needed to get data into the system, search – data in the repository by subject, geographical extent, type and format and various other parameters, download – the ability to download geospatial data, metadata and support materials, workflow manager – including metadata and version management, and access control and DRM – users access to the data is managed/curated (specifically only registered users of Digimap will have access to the repository). DSpace was chosen as the demonstrator repository software[1].

This report will summarise the user-based evidence, outlining GI-user requirements for the sharing of geospatial data derived from licensed data, and document the issues raised

[1] See Report: "Grade Repository Demonstrator – strategic development plan", Edward Boyle, January 2006

by this pilot work, including: how the repository might be scaled to support the GI-user community within the UK HEFE sector, benefits, risks, and issues, and finally outline a business case synthesising the lessons learnt into best practice.

Aim of this Report & layout.

The aim of the body of this report is to summarise feedback on the GRADE geospatial repository demonstrator. And to describe in brief the working changes to the demonstrator in its 3 technical phases as a result of continuing feedback Jan06- April07.

Feedback has been divided into 5 mini-Reports that relate to different types of feedback and to the 3 stages of technical development. These are presented chronologically.

A quick summary of the reports and how we got feedback:

Stage 1:

1) An initial scoping of feedback from Telephone 'round-robin' and emails from the GRADE Partners in Feb06 after the first release of version 1 of the demonstrator.

Stage 2:

2) User Survey – located online at the repository webpage left menu – ran from Jan06-Jan07 (most entries came after the demonstrator cosmetic revamp to the interface in early Mar06).
3) Ad hoc comments from the public that came in from the free-text feedback form online, and emails to grade-support@ed.ac.uk, and verbal feedback from Go-Geo! Workshops (where a GRADE demo and exercise was given) Jun06-Jan07.
4) Results of the 4 Pilot Site Reports (Edinburgh, Nottingham, Kingston and Strathclyde University site representatives) Sept06-Dec06.

Stage 3:

5) A second User Survey to run with the release of our 2nd map-based repository demonstrator (modified in response to initial 1-4 feedback) conducted Feb07-April07. And emails, comments from the ICOSS experiment group, Sheffield 19th Feb (and the week of 12-16th Feb).

The third part of the reports aim is to summarise the lessons learned from the feedback, and making a business case for a nationalised geospatial repository that meets GI-community requirements.

The Technical 3 stages of the repository demonstrator & key characteristics:

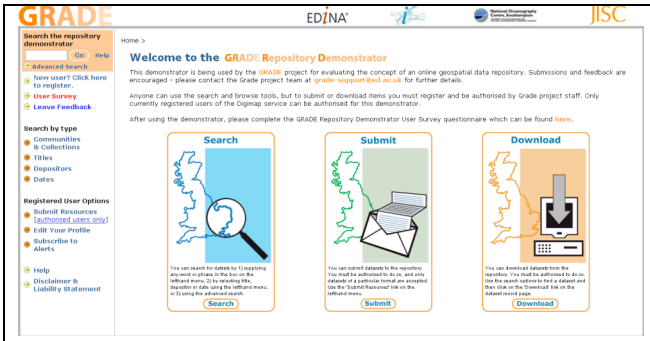
As mentioned above, the repository demonstrator had 3 technical stages of development through the lifetime of the project (Jan06 – April07). A quick summary of style and features of each is indicated below (details embodied later on in this report):

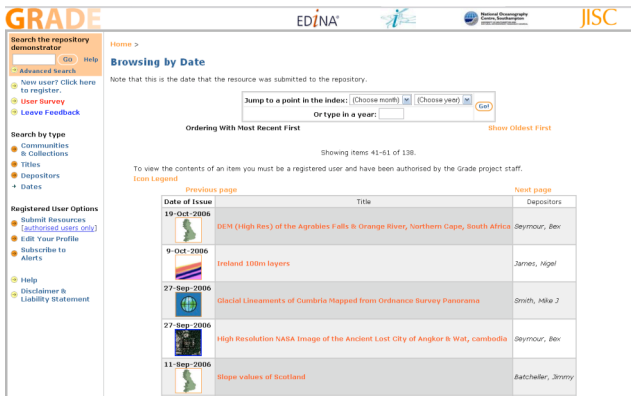
Stage 1: Original 'raw' DSpace repository solution

Original DSpace

- 'Out-of-the-box' DSpace basic function
- Athens Authentication/registration system added
- Text searching only (by Collection, Title, depositor, date)
- Basic GIS file types added into list of file types can submit.

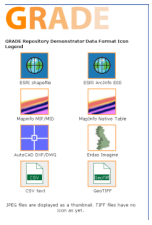
Stage 2: Cosmetic revamp, Geo-enabled repository



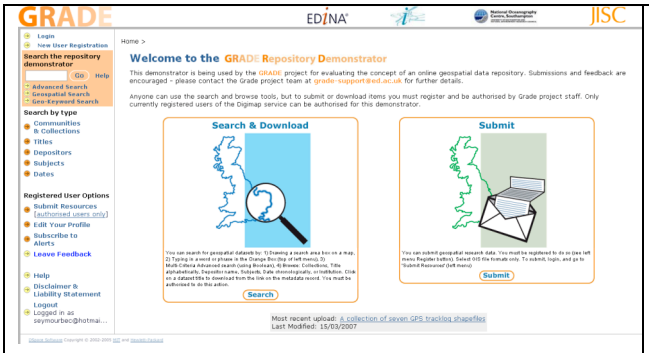


- Front page added
- Graphic redesign of logos, buttons, colours and layout (less text)
- Feedback form added

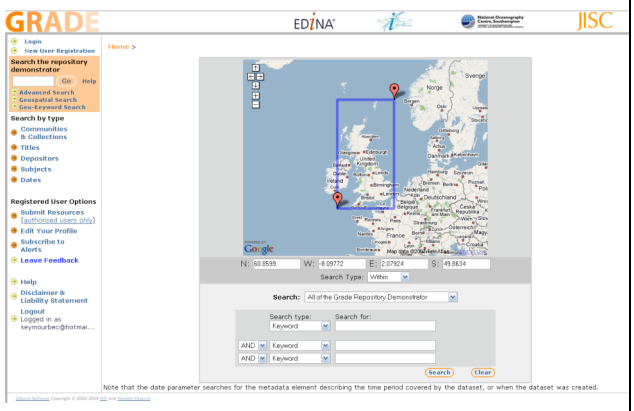
- GIS File type icons added
- Community collections divided up to add basic geography divisions (UK, Non-UK, Global datasets)



Stage 3: Final Map-based GIS validated repository, with added advanced search functionality



- Simpler front page (hyperlink all boxes)
- Login and new user registration buttons
- Re-ordering of the left menu (so that all search, and submit options are together)
- "Most recent upload" information box



- Map-based search (Google API used) "Geospatial search"
- Map & text search: "Geo-Keyword search"
- By Subject search (List of subject categories)
- Institutional search *(In development April-07)*

- Submit file validation stage (GIS file type, extents, and size)
- Automatically populates coordinate metadata fields, and GIS file-type fields
- Speedier submit process

<u>Performance</u>: The GRADE Geospatial Repository Demonstrator basic usage statistics (totals)

The formal GRADE geospatial repository became live at the start of February 2006. The URL to the site is: http://gradedemo.edina.ac.uk/dspace/index.jsp

There are currently 125 registered users of the GRADE geospatial repository, from a total of 38 HE-FE institutions in the UK[2] (and 6 non-UK Universities and other organisations[3] who are direct partners and special users).

[2] The Universities of: Edinburgh, Kingston, University College London, Cardiff, Strathclyde Glasgow, Southampton, Leeds, Nottingham, Oxford, Glamorgan, Liverpool John Mores, National University of Maynooth Ireland, Lancaster, University of Northumbria at Newcastle, Plymouth, University of Central Lancashire Preston, Leicester, Liverpool Hope University, Heriot Watt Edinburgh, Imperial College London, Sheffield, Glasgow, Manchester, Manchester Metropolitan, York, Portsmouth, Bournemouth, Scottish Agricultural College, Hull, Kings College London, City London, London School of Economics, Sussex, Liverpool, Cranfield University, Aberdeen, Glasgow Caledonian University, Bristol.

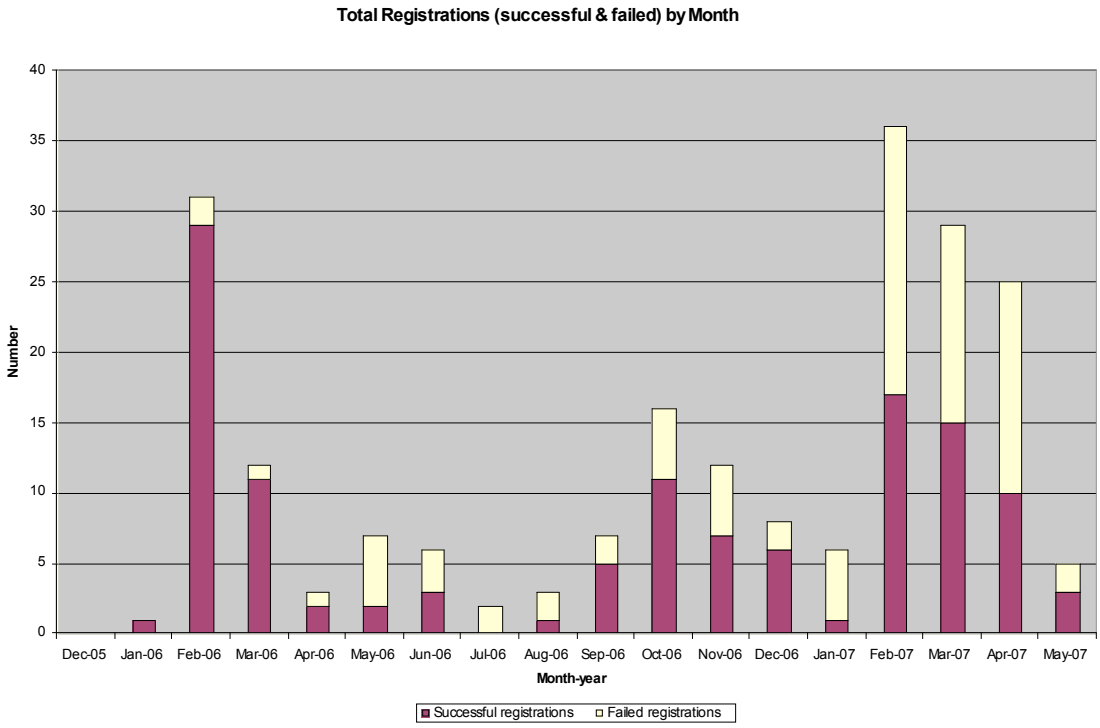
There have been 89 failed registrations. Many failed registrations were UK non-Digimap registered Institutions (13)[4] or from verifiable non-UK International HE-FE research Institutions (31)[5], &UK Government (3)[6] - it is worth mentioning these few as an illustration of the wider diversity and popularity of the service. Of the others 22 were unidentified and 20 Commercial organisations[7].

The number of registration requests during the term of the GRADE project has increased (see Graph 1) (now an average of 1 new registration per day). Although the pattern of registrations was artificial at the start (Feb-06 launch), & a forecast of lower registrations in exam-summer months, it has been since Sept-06 onwards (& pilots site launched) that registrations have steadily remained higher (minus a dip in Jan-07 presumably due to Christmas holidays).

Although the purpose of the GRADE geospatial repository demonstrator was never about achieving or targeting the number of users and GIS-datasets deposited, this remains an incredibly positive aspect (that the repository is being actively used & discovered despite its project-only status, i.e. it is not yet a JISC service/trial).

Graph 1

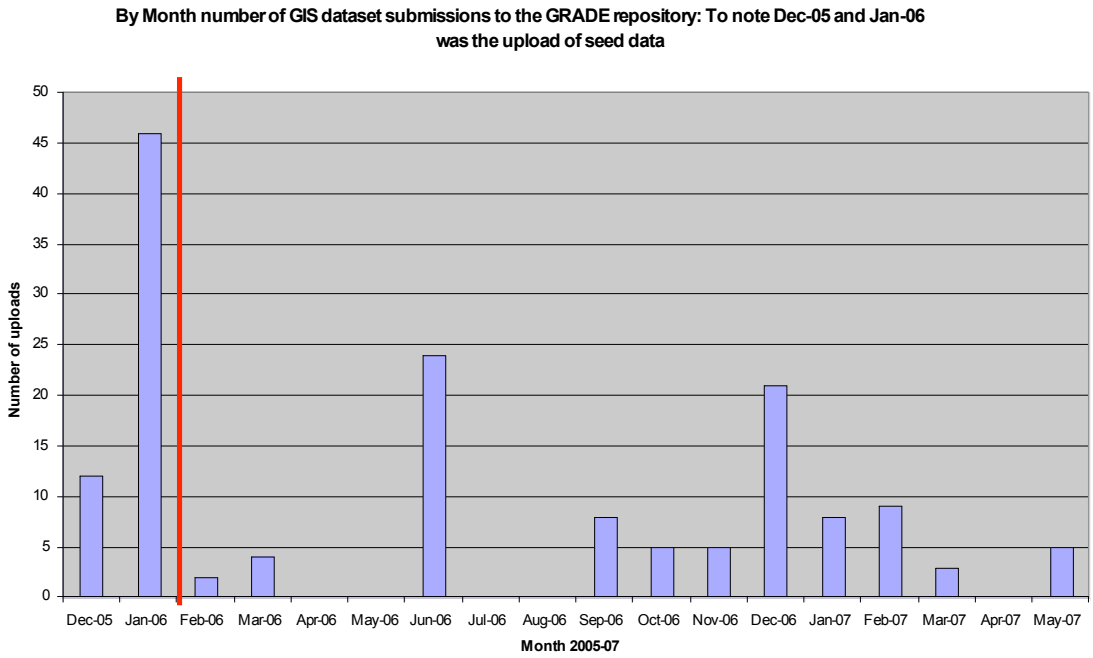
[3] Queensland University of Technology Australia, North Carolina State University USA, United Nations in Vienna, MIMAS Manchester, National Libraries of Scotland, Online-Archaeology UK.
[4] Users from *(in italics non-Digimap institutions)*: Kingston University, *Cornwall College*, Birmingham, Cambridge, The Robert Gordon University Aberdeen, *University of Ulster Ireland*, Durham, *City of Bristol College*, *Birbeck*, *University of Kent*, Napier University Edinburgh.
[5] Universities of: Buffalo-USA, Twente-Netherlands, Kyushu-Japan, Gotland-Sweden, Richard Stockton College New Jersey-USA, Sharif University of Technology-Iran, Tomar Institute-Portugal, National Council for Scientific Research-Lebanon, Institute AGranomique et Veterinaire Hassan II Rabat-Morocco, United Nations Vienna, The Wildlife Institute of India, Istituto Nazionale Geofisica e Vulcanologia-Italy, Diputación Provincial de A Coruña-Spain, US National Institutes of Health Fogarty International Centre, Minnesota-USA, National Institute of Rural Development-Hyderabad India, Michigan-USA, L'Universite Paris-Sud, University of Applied Sciences Bochum-Germany, Halle-Wittenberg-Germany, Mount Holyoke College –USA, Arizona State University –USA, McMaster University-Ontario Canada, Oregon State-USA, British Columbia Institute of Tachnology.
[6] Government-based: The Countryside Agency, Anglesey County Council, and The Crown Estate.
[7] Commercial(including for example): Geowise, SYPTE (South Yorkshire Passenger Transport Executive), Sempra Pipelines, Groundwork Caerphilly, POYRY (Global energy & Forest Industry Consultants), Delaggy Surveys, Hammac Noise Pollution mapping Glasgow.



There are a total of 151 GIS items in the repository to date. The repository hosts 93 UK geospatial datasets and 56 (2) in the Non-UK (& world collections)[8]. To date there have been 64 external[9] uploads (total of 30 depositors from 11[10] Institutions), and 87 seed geospatial datasets. See Graph 2 which shows submissions by month over the GRADE project duration. To note the zero submissions April06 – Aug06 (excluding Jul06) may be due to a couple of reasons: firstly, that the repository was only in its infancy at this time; it was compounded by the fact that this period covers academic exam and summer holidays (it would be expected in an annual cycle this time period would have low submissions). In the final leg of the project from Sept06 – April07 illustrates a general pattern of regular submissions to the repository, which is a really positive sign[11].

Graph 2

[8] To date at 16-Feb-2007 (International submissions were stopped with the release of the map-based repository demonstrator Feb-19-2007, a few ad hoc International entries via email only added).
[9] i.e. "external" to EDINA/MIMAS/JISC (although many of EDINA staff have uploaded their own academic research data completely independently this has been discounted)
[10] The 4 Pilot Sites (Edinburgh, Nottingham, Kingston, Strathclyde), plus: Oxford, Manchester, Sheffield, Leeds, UCL, York, & Kings College London.
[11] Zero uploads to the repository in April07 was a result registration technical issues (backlog of data uploaded in May07).



There have been **761** external downloads of GIS datasets in GRADE repository to date (Period 02/12/05 - 01/05/07)[12] – and this can be compared to the report done in Jan07 before the new map-based repository interface was introduced of 545 (Period 02/12/05 – 05/01/07)[13]. This is illustrating that in the most recent 3 months of the repository action downloads have been much higher (presumably reflecting both the greater diversity of content and number of registered users). These figures are again extremely rewarding[14].

For a summary of the main user 'activities' (Top geospatial data downloads, Top free-text search words, Top search methods, and Total user log-ins per month) on the repository site over the course of the last 9 months (period of time where the repository started to gain wider usage and pilot site started, sept-06 – April-07) see **Appendix 1.1.** Top free-text search words appear to be for place names: UK/England/Scotland (so very broad location search-terms), and for themes: Glacial/Geology/Soils/Terrain/Archaeology/Boundary/River/Landuse. Since its introduction Feb07 'Browse-By-Subject' has been the top method of searching, followed by 'Brose-By-Title' (although it should be noted that because the 'Geospatial' and 'Geo-keyword' search were added in functionality they are not picked up the DSpace Statistics page, and so are not represented – we believe that these have also been the most popular based on conversations with users). The average number of registered user logins per month (for the last 7 months) has been 15/month.

[12] http://nevis.ed.ac.uk:9200/cgi-bin/DSpace_stats2/stats1.pl
[13] http://nevis.ed.ac.uk:9200/cgi-bin/DSpace_stats/stats1.pl
[14] The Download statistics for the DSpace repository ***exclude*** all downloads by EDINA staff members from this tally, plus all downloads by the named depositor himself for his own dataset (the figures therefore represent external downloads by other repository users). This is an estimated figure however as this may also include instances of failed and attempted repeat downloads (however we believe this to be marginal).

Part 2: Presentation of feedback

Report 1: The first DSpace demonstrator, and initial 'First-thoughts' feedback from GRADE Partners only: The Starting point.

The first GRADE formal demonstrator became live on the 30th January 2006.

It represented a fairly 'raw' out-of-the-box DSpace with a few adjustments for geospatial data accommodation (see Stage 1 summary at start). It was seeded with a total of 57 GIS datasets. The sources were Plymouth University Historical GIS project ancient boundaries, Sample Environment Agency datasets, and two satellite images of Scotland[15].

The initial scoping exercise was to compile a list of initial design instinct/'gut-feelings' from GRADE's consortium of project partners[16] to try and establish what was the core functionality needed for GI-users that needed to be addressed at the start. The feedback comments were compiled following testing, gathered from a series of phone calls and emails in the month of February 2006 (this is summarised in full with quotes in **Appendix 1: 2**[17]).

A largely expected notion was the finding that GIS-map-users wished in a very first instance to be able to search to find a geospatial dataset on a map, or by querying a map. This was by far the strongest feedback in terms of search functionality:

> "I would like to see a map-based search, and not just keyword or Boolean (e.g. USGS ARcIMS, or Global Landcover Facility, or Journal of Maps). It would be better if a map was the first thing you see – would like to have a map grid and be able to click on a location (and be able to zoom & pan, and to see polygon extents of datasets"
>
> "As a geographer the easiest way I have been programmed to 'locate' is to look up on a map. Geospatial people understand maps. It also helps to double illustrate what the main tool of the repository is."
>
> "Map-based search is better, as location as a text string is more dependent on how this was entered (e.g. A UK wide dataset that includes Scotland may not be 'found' with a word-search 'Scotland' and yet would still be a valid dataset)"

One of the second strongest limitations that were directed at the repository was that there was no institutional search. It was certainly interesting that the consortium partners felt a strong sense to retain their institution affiliation, and a desire to set up exclusive working groups (perhaps mirroring the kinds of informal patterns already happening – as discovered in GRADE informal questionnaire[18]).

---

[15] This was a limited selection of datasets, to create an artificial basic testing environment only. The data the quality and coverages will always be an impacting factor on 'attractiveness' of a geospatial repository.
[16] Individual professors from Edinburgh GIS/geography department, Kingston Centre for GIS, Nottingham Geography department, Strathclyde Graduate School of Environmental Sciences, and HEAS Geography Earth and Environmental Science, and the Built Environment
[17] Extracts from a summary report (Mar-06) highlighting general issues.
[18] "Informal Geospatial Data Sharing - Current Practices amongst the UK Academic GIS Community", R. Seymour, Dec-2006.

> "An ability to search by institution would be useful– it is good to know what other parts of Edinburgh University are researching (presently unknown and a total failure of systems in place)"

> "The second most important search to me would probably be to search on 'depositor'. I think it needs to have 2 different fields to define the 'institution/company' and the 'individual'"

> "Would like to see a more 'federated' structure – where people 'own' different sections, i.e. an institution, or a collective of people, and can search within 'signed-up' sections"

> "To be able to search in the repository for work done in a certain field/study location or working group – e.g. Southern-Spain + desertification"

Other comments relating to 'search' functionality expressed a desire for a method of quality rating on a dataset, and to be ale to search by quality star system:

> "I think to have a quality self-rating system like has been done on Amazon is a really good idea"

To be able to search by a set of pre-defined subjects was also voiced (and, as noted, this in particular may become more profound as dataset volumes increase and data searching needs to be more advanced to deal with a complexity of subjects). What was interesting to note was that none mentioned a desire to search by file type or a wish to download files in a generic file type (example GML). Perhaps one can assume from this that GIS users are familiar with all GIS files types and formats and work with multiple file types and conversions.

In general, all the partners felt that the viewable metadata on a dataset (Dublin core, basic) was sufficient. Although a couple of members expressed a desire to see a thumbnail static image of a dataset.

In terms of submit/upload, and despite some initial confusion as to the instruction to do this in terms of recognising one must zip up files prior to submit (a learning curve occurred with new users), this was not later considered a problem:

> "I have no problem with the fact that you have to zip files before you upload, or unzip after download to use"

However there were many comments mentioning a future desire for a more complex solution to this – a wish to upload multiple project files bundled together, and not single files, perhaps even not just geo-files (i.e. also, photographs, license data). Content packaging (or wrappers)[19] standards such as METS[20] and IMS-CP[21] (or: MPEG 21

[19] http://wiki.osgeo.org/index.php/Geodata_Packaging_Working_Group

[20] From the digital library community: METS (Metadata Encoding Transmission Standard) is a framework of containers for packaging together descriptive, administrative and structural metadata about digital objects and the complex links between them. It does so using XML format encoding, and is made up of 4 sections (descriptive metadata, administrative metadata, file groups, behaviour, and a Structural map). METS is being developed as an initiative of the DLF (Digital Libraries Federation) and is being maintained in the Network Development and MARC Standards Office of the Library of Congress. (Extract from: UKOLN, eBank Project)

[21] From the learning technologies community: The IMS Content Packaging Information Model describes data structures that are used to provide interoperability of Internet based content with content creation tools, learning management systems (LMS), and run time environments. It has 2 elements: IMS Manifest file describing organisation of a package, and the physical files described in XML.

DIDL, XFDU)[22] are helpful here (how they can be complied with, and used for, GIS data needs to be considered[23]). This is also inter-linked to issues of persistent identifiers[24] for geospatial data deposited in a repository environment.

> "One of the main questions that I have (and that all partners did at the kick off meeting) was in terms of how data is loaded up, i.e. individually as singular datasets, or as a 'packaged' (multi-layered, or project, or so that you can retain linked datasets within one place, or just so you can have the raw dataset plus any other associated images/reports etc in a project). There is a question of maximum reuse. But also over ownership – most researchers do not want to break-up their research outputs (and may make it loose value – i.e. no longer make sense)"

One of the bigger issues in the upload was the post-submission editing, or inability to do this causing a big frustration:

> "An issue with the upload is that once you publish or finish your upload that you cannot go back and edit or change mistakes. Might need a 2-stage process to final publish – with an intermediary editing screen – perhaps for a week. And then after that, there is only ability to version."

> "I made numerous spelling mistakes – especially if you are uploading large volumes of similar datasets (like boundary data). It was extremely frustrating when I could not edit. I now have edit function – but this allows me to edit all of the datasets, not just my own. Ideal would be some limited edit functionality to change just your own uploads (or parts of)."

> "The full-waiver of my rights to my dataset upon depositing makes me uncomfortable. I would want to be able to remove my dataset at will."

But the general process of submit was seen as simple, step-by-step and easy.

> "Steps were easy to follow, and quite quick to fill in and do"

> "Would prefer some automation in terms of filling in metadata fields, i.e. name, institution, the process is a little laborious. Too many stages makes it slower".

> "Would like to see the upload process available as an offline series of pages – so can sort this out all in your own time"

In terms of entering geographic metadata extents/coordinate information directly by a user, certain learnt lessons can be taken from the experiences of the Journal of Maps[25]:

> "I do not think that the metadata needs to contain written extents or XY coordinates etc. I think this is moderately meaningless as words/ numbers to 'readers'. Would like to be able to see a geographical extent on a map (if done automatically when uploaded) – but the manual entering (as is my experience from JournalofMaps.com – which does require you to fill this in as metadata on upload) is that over 50% is entered incorrectly. Often depositors get even latitude and longitude mixed up! It depends how big an error check process is set up to validate entries."

One of the over-riding issues in the completion of metadata appeared to be in the defining 'roles' of people that contribute to the creation of a GIS dataset (and that this is generally complex). In the main, the query was about "who is 'depositor'?"

---

[22] OGC Strategic Plan, Archives & Geospatial Issues (Steve Morris, 2006).
[23] The international geographic community have not been thinking about content packaging strategically until recent months: the OGC Preservation working-group has been recently established to look at this, but no current recommendations have been made to date.
[24] A DOI (Digital Object Identifier), a fixed pointer or way of permanently attaching a unique code to a GIS data location/store (stays constant if a URL breaks for example).
[25] Journal of Maps website: http://www.journalofmaps.com/

> "This is confusing. I did not know if it was me, or who originally owned the data. I ended up deciding that this should be me. Others did not."

Interestingly only one person commented that they had been unaware of what to do to fill in the metadata licensing/rights box (and requested there be some extra guidance or examples for this). But in fact, a need for this might be assumed or further backed up as a clear issue, since all of the submitted datasets at this point left this box entirely blank.

There were no reported issues with the download.

There were numerous comments in relation to the user-interface and the design of the repository[26]. Initially much about the front page, and its wording:

> "Front page is unexciting. Text black on grey makes reading difficult. It is not clear what the demonstrator is. Would like to see a map/picture, and bigger fonts – perhaps some simple icons to show instruction."

> "It is not clear from the demonstrator home page what GRADE is or does"

> "I would like to see a map as the first page that you come to (with all the possible search methods clear)"

> "Looks a bit boring. It is not sexy. Bland colours. Not enough contrast (too much white space)."

> "The website looks good – very pleasing to the eye"

> "It was straight forward. I liked this. I would prefer a plain and straight forward interface even if duller – so this was fine – it met this criterion well."

> "The main problem is that I cannot see all of the left menu buttons on my screen"

> "I think that the advanced search box should be closer to the 'Browse' options – so it is clear that they are connected. And that the 'HOME' button should be at the top of the left menu."

> "The layout is not always logical for me. I would like to see all of the different search methods in the main window pain – at a start up main screen – not just on the left menu. The way it is organised – I also did sort of miss the 'advanced search' – I do not know why as it is at the top. It looks like it is a different thing to 'browse'."

> "Want a FAQ. Help is limited"

Overall, many of the comments were very 'visual' (cosmetic, about the website look and interface, and less about its functionality). This is perhaps a reflection of both the non-intuitive interface at this stage, but that also a factor that GIS-map-geography users may respond easier to icon/symbology, more visual rather than textural instruction.

In the main, initial comments were positive:

"Overall it was simple fast and effective, it worked well, and I liked it".

Following actions:

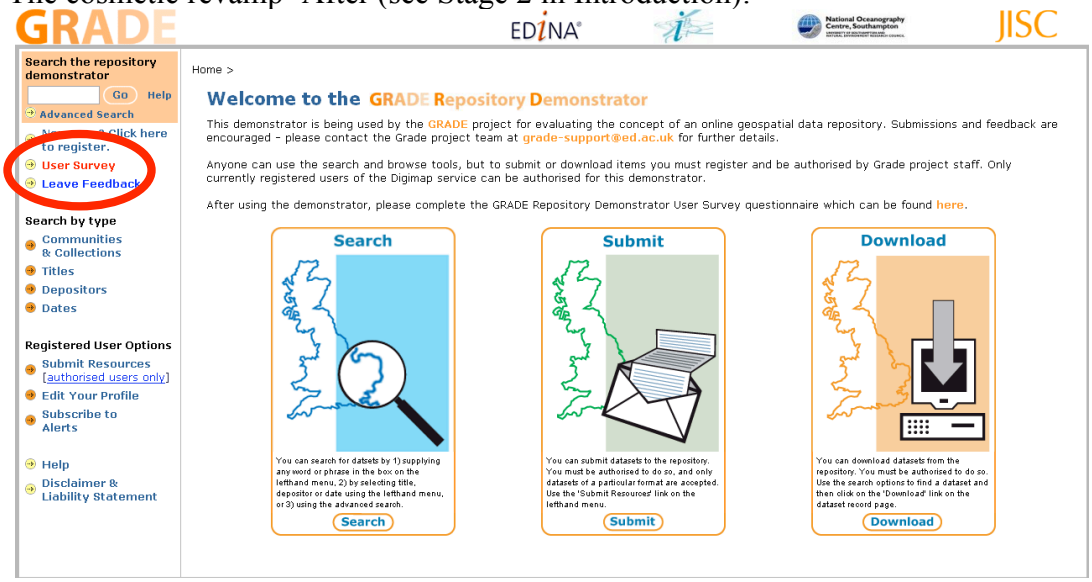
[26] This is somewhat in response to the raw presented interface at this stage.

The main reasons for conducting feedback at every stage of the geospatial repository technical development was to inform and direct and action about what steps need to be taken to fulfil GI-user requirements. A critical issues table was created resulting from the above discussions, and this timetabled in, targeted, and prioritised technical development work (some were rejected on a basis of complexity or difficulty in doing in the time frame of the project).

This worktable is in **Appendix 1.3** (Blue, left column, issues next to be addressed). This is a good summary table of all of the issues flagged by this initial scoping exercise.

<u>Report 2</u>: The User Survey Results: - Following the Cosmetic Revamp Interface of the Demonstrator, and an aim to increase diversity of users and gain more structured feedback.

The cosmetic revamp- After (see Stage 2 in Introduction):



The User Survey questionnaire (for a full copy see **Appendix 2.1**) ran from Feb06-Dec06. It was linked to the main GRADE geospatial repository demonstrator website left menu, as in the diagram above (circled red) (it is assumed that in the majority it is registered users that have been encouraged to complete the survey, but it was very much left to site visitors to fill in at their own wish. It was not extensively advertised).

A very respectable number of responses were had in - a total of 52[27], from 15 Institutions[28]. In the main respondents/repository-users were from GIS-Geography-Geosciences (with 37 of the total), but respondents covered a variety of other subject disciplines[29]. (A few key subject disciplines that are likely major GIS users and notably missing include: Archaeology, and Sociology/Demography – so preferences from these fields may be not represented it should be noted).

The aim of the questionnaire was to discover people's expectations about how a geospatial repository would function, and to critique the DSpace demonstrator version to test in the GRADE project. The main findings are summarised below (for a full numerical summary per question see **Appendix 1.4,** & full survey results **Appendix 3.1**).

In Question 1 a user was asked to nominate how they would most like to search to find geospatial data. Respondents rated a series of options (ranging from 1(essential) to 4(probably would not ever use)). The item(s) both 'scoring' highest & selected most frequently as 'essential' & 'would be good' (rated highly) was jointly "By drawing a box

[27] In fact a total of 55 replies were received in (the last 3 too late to be part of this analysis)
[28] No of replies per Institution: 16 Edinburgh, 14 Kingston, 5 Strathclyde, 5 Nottingham, 3 Manchester & Metropolitan, 2 Maynooth, 2 North Carolina State University, University College London, Keele, Glamorgan, Liverpool John Mores, Oxford, Brighton, UK Data Archive
[29] Breakdown by Subject/discipline of respondees: 37 Geography/GIS, 4 Environmental Science, 4 GIS Librarian/archivist, 2 Geology, 2 Computer Sciences, Civil Engineering, Physics, Biology

on a map" and "Entering a free text place name". This was closely followed "By clicking an area on a map, i.e. county boundary", and forth was "By entering a postcode" (which received fewer 'essential' ratings). It is significant to note that geographic methods of searching rated above other more traditional searching (keyword, Author – scoring more 3's and 4's- 'would not use'), illustrating a different approach is needed for GI-data. To note, this exact same pattern emerged in Question 1b (where respondents were asked to put their top-3 search methods only, the only difference being that 'keyword' rated higher in this list). The full scoring list in order, see below (**Fig. A**):

I would like to search for geospatial data by. (in order of preference)	**Total "score" Lower= higher rated**	**Number of times rated:**			
		1 - essential	**2**	**3**	**4 – would not use**
'By drawing a box on a map'	**87**	22	26	4	0
'entering a free text place name'	**87**	25	20	6	1
By clicking on an area on a map, i.e. county boundary'	**88**	20	28	4	0
By entering a postcode	92	24	18	8	2
By keyword (from controlled list)	97	20	21	9	2
By selecting a place name from a pre-defined list	101	16	27	5	4
By date	113	14	20	13	5
By entering min/max coordinates	117	12	22	11	7
By Author	136	6	17	20	9
Other:	By Institution/company, By Quality Rating, Type i.e. raster or vector (x2), By Subject, By Data Creator(not depositor), By a list of known geographies e.g. council boundaries/green field sites etc, Parson's Codes, Keyword list of geospatial topics (x3), By time period of data/experiment duration/collection, By a map (x2), By themes or data collections i.e. landuse/census data.				

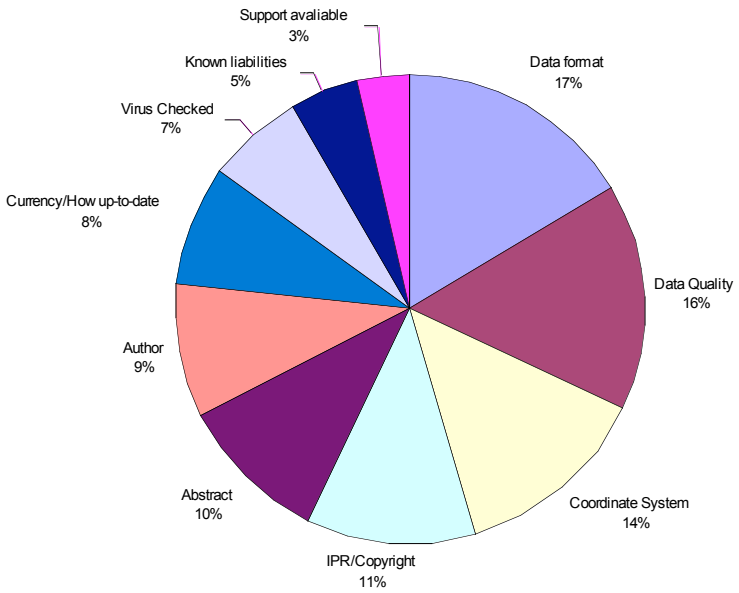
Fig. A: Score sheet totals, Question 1: User Survey questionnaire

Reflecting, at least partially, the initial scoping exercise by the partners, in 'Other' there were some individual interesting replies – for example search by data-creator (similar to via Institution/ publisher), and by subject collections and data 'themes' (i.e. Parsons), and by quality rating.

Question 2 was to discover what information is needed to assess if a dataset meets a users desired needs. Respondents could tick as many as wished from a list and the result is ordered as in popularity. The most needed information by nearly all of the respondents (47 of 53 required this) was to know the "data format". This was followed closely by "data quality /provenance" and "coordinate system". In forth and fifth place respectively was "IPR/Copyright" and an "Abstract" (just over half of the respondents were concerned with knowing these). See Pie Chart, **Graph 3**.

When asked about the importance of data quality and accurate metadata all most respondents said that this was 'Very important' to a repository (and all saying that this was very or fairly important).

Graph 3: Pie chart to show by % the most wanted information to assess a GIS-dataset need for purpose.



In order to confirm that a dataset meets a search criteria, the top voted essential was to "view the extent of the data on a contextual map", secondly, to "View a static image of the data", and thirdly to "interact with the dataset in a simple map viewer accompanied with contextual background information". Within an option of "other" a variety of alternate answers for things users would like to have at a 'data find' stage were mentioned and worthy of note, these include: would want all of the above but ONLY if this equalled a fast/speedy data retrieval and if it did not would rather none – speed is the priority; would want a file size and download estimation time (x2); would wish to see the attribute table (the data fields) prior to download; more accessible metadata viewing platform; and finally would like to see the mapping aspect (i.e. coverage, geospatial features, extents).

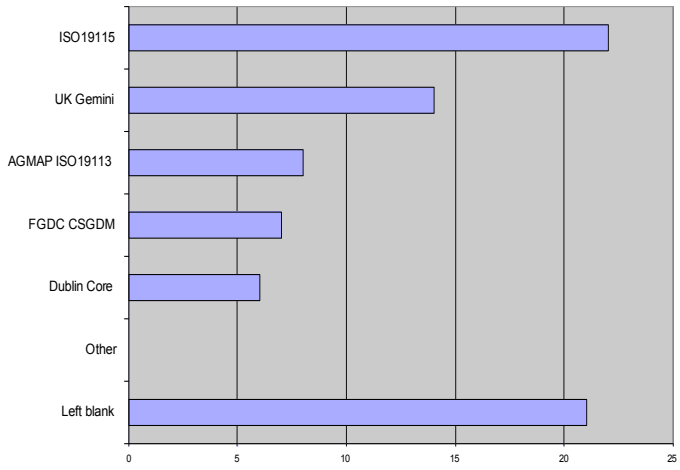
Question 5 asked respondents what they would most expect to be able to do when retrieving a GIS dataset from a geospatial repository. The by-far most popular answer was to be able to "Download the data directly for use within a desktop GIS".
One individual qualified their answer based on the estimated size and time of retrieval (favoured option as choice preference is whichever takes minimum of time). See **Fig. B**.

When retrieving a data set from a geospatial repository I would expect to. (in order of preference)	**Total "score" Lower= higher rated**	**Number of times rated:**			
		1 - essential	**2**	**3**	**4 – would not use**
Download the data directly for use within your desktop GIS	**64**	38	14	0	0

Have the data streamed direct to your desktop GIS or other internet-enabled application able to consume web services e.g. an OGC WMS or WFS	116	9	28	9	6
Interact with the data in an online environment with the aim of making a simple map for printing	118	12	21	8	11
Interact with the data in an online environment with the aim of carrying out simple spatial analysis e.g. on screen distance measure	128	11	17	13	11
Other	As online tends to be slow/limited functionality I'd prefer just a simple straight download, Simple visualisation with aim to print out map, Sub-setting/clipping services would be useful, A choice of download formats is my ideal answer- web services are good but only for advanced users.				

Fig. B: Score sheet totals, Question 5: User Survey questionnaire

Of the 52 User Survey replies, when asked about associated metadata accompanying a GIS-data-record download, the most wanted metadata form was to firstly be human readable, secondly machine readable, and thirdly compliant to a recognised geospatial standard. 31% (16 of 52) answers had all 3 as equal importance. Of the replies placing significance on having a metadata standard, they were asked to rate which standard they would prefer to work with (and select as many as appropriate). ISO19115 came top. This result can be quickly summarised by the bar chart, **Graph 4** below:



Graph 4

What types of information would you be willing to provide as discovery metadata if depositing a GIS-dataset in a repository. (in order of popularity)	**Total "score" Lower= more likely to be happy to supply this**	**Number of times rated:**			
		1 - Always	**2**	**3**	**4 – Never**
Dataset title	**57**	45	6	0	0
Geospatial format	**72**	33	14	4	0

Keywords	**73**	35	13	3	1
Coordinate system	**74**	34	11	5	1
Abstract	78	30	16	6	0
Contact details for dataset creator	87	23	20	6	2
IPR/Copyright Statement	98	22	9	17	2
Lineage or details about how dataset was created	111	15	14	21	2
Statement on positional or instrument accuracy	112	11	21	16	3
Other comments	▪ Positional & attribute accuracy should be compulsory even if this is unknown, ▪ work/datasets/journals connected to this dataset or list of related links, ▪ From point of view of receiver all of above are important – but lineage is complex and involves significant work and is probably unrealistic, ▪ I would only be willing to supply contact details if it did not result in a barrage of requests, ▪ Although a statement of IPR/copyright is desirable the process of creation of this needs to be simplified such as a drop down list of key IPR terms that apply.				

Fig. C: Score sheet totals, Question 7: User Survey questionnaire

In Question 7 (See score sheet above, **Fig. C**) asked about what types of information with a submission one would be willing to provide for portal-metadata discovery (on a scale: 1=Always 2=Most of the time 3=Possibly but would require extra work 4=Hardly ever). What was clear initially was that a high number (total of 8) individuals rated all (or nearly all, i.e. all but 1) of the choices as "Always willing to provide this". This was really encouraging a result as poor metadata creation has been noted at many UK HE-FE Institutions[30]. The top pieces of information are rather self-explanatory, basic key statements: Title, File format, Keywords, Coordinate system, Abstract. Less people were happy to provide information on IPR/copyright (presumably due to some confusion with this), Linage (complex in GIS data often), and statement of positional accuracy (this criterion might lead to some assumption of data-quality and self-criticism, presumably a reason why this might be less popular). However, what can also be said of the results is, that none stated that they would "Never" or refuse to provide *any* of the information stated in the question (in all, they would be willing to provide some discovery metadata information (only in 11 cases was it stated they would "hardly ever" wish to provide an option – so this is low, and very positive).

Furthermore, part B to this question in addition asked about willingness to provide extra metadata compliant to a metadata standard and also which metadata standard they were most comfortable with. The top answer was "ISO19115" and the pattern metadata standard of choice was exactly the same order as in question 6 (what standard metadata would like to receive in a download). The only difference was one added 'Other' which was 'AGI metagenie' (but this person also qualified that they would be unenthusiastic because it would mean extra work).

[30] Go-Geo! - Geo Data Portal: Pilot study & Business Model Report (Tony Mathys, 12-Dec-2006) (2.1 Spatial Data Audit)

In the final question users were asked about geospatial data longitivity in a repository: - Once deposited in a repository should geospatial data persist indefinitely? Just tipping the barrel came "Yes" as winner, but it was extremely close, hinting that this is indeed a very hotly debated topic for repositories.

YES – 28 Votes	NO – 24 Votes

Of those who voted "No" they were asked what the lifespan should depend upon, more said "usage" and as secondary was "perceived usage to others". A few ticked other, quoting:

> "Data should be updated, so latest version only"
>
> "Even if data is not used often it may still be very useful. Maybe two separate lists are needed: the most "popular" and the least "popular". People could ask to list the most popular first and only by request could they look deeper into the depository for other data"
>
> "Level of Usage; but also, lifespan agreed by depositor or of control of depositor (it may get superseded by other work- or be disproved research)"

Additional Comments:

"Very useful online tools. Especially for up coming projects in same field"

"When making data usable for others, it is important to include metadata and the reason for the data use so others can use it appropriately and reliably"

"The GRADE repository has a good amount of support material around the core of the site to assist first time users. I would suggest that for the page that provides a list of Depositors that as well as there name being listed the organisation the depositor is from is also included. This might be of use if there was a particular question needing to be answered about the dataset"

"I found the search facility difficult to use. it would be more helpful to be able to browse by subject area and date of the dataset, rather then date the dataset was obtained. Also, I found the depositor field confusing, you need to be able to search by data creator."

"It would be useful for depositors to be able to access an expandable list of keywords from which to choose or to which to add, to help in searching"

"While a laudable initiative I see various problems with the suggestion that data is deposited. (1) May would argue that it is the research paper which is the final product not the data, and there are ethical / competitive reasons why data should not be deposited. Take an archaeologist; they won't deposit data relating to a site because it may be used as the basis of a sequence of papers over a number of years and a team doesn't want to give competitive advantage to others by making the family silver available. (2) It can be a lot of work packaging up datasets; a complex model in a relational database could take weeks to document and package the data up in a suitable form - there may be dozens of tables and equally it make take those who wish to use the data weeks to reassemble the data in a usable manner; the whole issue of effective (and reusable) data modelling needs to be considered and there are precious few standards here (3) much will be derived from other data like OS data, which is commercial / copyright and simply can't be deposited - and indeed would you want it to be deposited if it is little changed from the original (4) data volumes could become enormous, esp. if student projects are expected to deposit data - some of our students use Gbs of data - yes too much, but that's what they use. What do you do with intermediary datasets, or is it just final datasets? FYI I purposefully stopped doing ESRC grants because the bureaucracy in terms of lodging papers and datasets, and the consequent documentation, was a tedious diversion from the research itself and something that often fell on the shoulders of the PIs and the RAs had by that stage left for another job/contract."

"Nice simple and easy to use service - quite a pleasant surprise to find Irish historical datasets here. More importantly the site seems to b easy navigate and seems to be aimed at a sort of WikiMap approach in a better protected environment which can't be a bad thing"

Conclusion.

Geospatial search came out as a top priority for searching for GIS datasets. Equally so was for entering a free place name, and having a gazetteer. For quick assess judging fit-for-purpose of a GI-dataset, users would wish to know data format, quality, and coordinate information. Having accurate and standard compliant metadata was deemed 'very important' (most said they would prefer to work with ISO19115 and UK Gemini standards). To confirm a dataset meets 'a search criterion' it would be most important to be able to view extents on a map. Downloading direct into a desktop GIS was the most popular form of download. It appeared most people were willing to provide all basic metadata/information about the data they are submitting. More people (although this was much divided) thought that if data is deposited in a repository is should exist indefinitely – of those who did not then lifetime was considered to be a factor of 'usage'.

Report 3: Further comments (ad hoc) via email, the Feedback Form online, from the GRADE/Go-Geo! UK-Wide workshops held April06-Dec06, and Blog entries.

As per the ad hoc nature of comments (emails/blogs) and time span (the lifetime of the project) this report is harder to cohesion. For a chronological summary of all emails see **Appendix 1.5.** To tie and display the main points see **Table 1** listed summary of the common feedback themes and issues via email, feedback form[31] and blog[32]. It is worth noting that authors are not *all* registered users of the GRADE repository (although would request to be) – and so this may provide a different flavour to comments in this section[33]. A key contribution was provided in a detailed feedback from a led debate at NCSU[34].

Table 1: Summary of main email/feedback form/blog comment themes.

Theme	Comments (either commonly mentioned/ particularly noteworthy)
Search Issues	Orange box not as useful Wish to search by Institution/Organisation Wish to search by date of coverage (really important issue for finding temporal GI-data) as well as date deposited. Thought needed to GIS-ontology/terminology used Geospatial search good Like Google maps API
Discovery	Abstract is most important section (important this is completed well) Description/Abstract really important (must get in full texts here) Would like a measure of quality of the data. Would wish to see more metadata
The GIS data	Environment Agency exemplar dataset really useful (like to see more like this) Pleasantly pleased to see Irish data in the repository GIS crime data and sensitivities (controls and specified access restriction ability needed) Interested to use GRADE to host planetary mapping data
Download Issues	Would only wish download to desktop GIS, no more. Zip file-size is displayed rather than the raw file-size (which would be more useful) GML download not there
Upload Issues	Zip upload not compatible with non-windows, restriction/limit of access to GRADE for heavier-duty GIS people who work on UNIX platforms. ISO19115 should be an option i.e. drop down list of extra metadata at each section should you wish to add more (example, could have a fast process and a full/detailed submission process) – ability to choose is key. Long process. Interface at the beginning about the rules could be made simpler Need a 'best practice' hint for title naming Need a forced standard date format (consistency) Licensing box – needs guidance, a check box of what to fill in would be

[31] The "Feedback form" was a digitised form on the GRADE geospatial repository website submitted to grade-support@ed.ac.uk

[32] http://www.journalofmaps.com/cgi-bin/blosxom.cgi?_start=61 (Mike Smith) and http://www.geog.leeds.ac.uk/people/a.turner/personal/blog/archive/2007/02/ (Andy Turner)

[33] Submit & download functions are **not** available to non-registered users, however the content of the geospatial repository and metadata is publicly viewable on the web, as is all search functionality. Feedback from non-registered users was therefore considered valuable and was therefore not omitted in this Report 3.

[34] North Carolina State University, USA.

	ideal. Need automatic file validation to filter non-working files Needs a metadata update procedure (an editing suite for errors/spelling mistakes & version updates to accompany data updates) File size limit is limiting Add GML as a file type submersible (veto IDRISI) Bulk injest for large/many file uploads from Institutions (more automated – or links to external servers where global datasets updated regularly).
Using the repository in real life/sustainability	If using GIS data from GRADE repository how would one reference this dataset in a published article (given that there is no source information). Would only commit an upload if GRADE had longitivity (would not want to spend time doing metadata for it to be deleted at some future point – would need to know my dataset was safe) – sustainability for archival.
General	GRADE would be useful for sharing Landmaps (MIMAS) GRADE would be a sensible repository for recycling derived OS data released via Digimap. Links to other GIS portals would be good. GRADE could act like a GI-data repository for academia, government, and public – different filters (but all in one place) Would not wish to share unpublished data/materials as this would represent piratable research (this would also assure quality worthy of archival) Would wish data in GRADE to be peer reviewed. The focus on metadata creation is of huge importance
Overall	Great Idea This project is a good thing. Like it Valuable resource Suitable GI-data sharing facility Simple straightforward interface I am really interested in the idea This is a good thing, I hope that it takes off The GRADE catalogue: most impressive A potentially fantastic GIS repository Holds some pretty useful datasets not accessible elsewhere.

Conclusion

Again, key searching would be by Institution, Date and Map-based search. In terms of discovery metadata an important emphasis was put on having full-Abstracts. Quite a few issues connected to the submit process were flagged – namely, limit of windows application for UNIX GIS users, and need for a bulk ingest for high-volumes of datasets, that ISO19115 is not supported should a user wish to provide more metadata fields, need for naming standards and greater guidelines. Sustainability issues were highlighted, and the question of how you reference a GIS-datasets held within GRADE repository in a print-publication.

Report 4: Comments and feedback from the Pilot Site Reports (Edinburgh, Nottingham, Strathclyde and Kingston)

The Pilot Site Experiment work was conducted Sept06 – Dec06. The four sites were chosen for their excellence in the subject field and diversity, and geographical spread across the country. They are the Universities of Edinburgh, Nottingham, Kingston Centre for GIS, and Strathclyde at Glasgow. The aim was to conduct an intense series of detailed testing of the repository. Site representatives conducted interviews with Geography/GIS-based staff and students, and encouraged the use and deposition of geo-data and cross institution sharing. A set of criteria were laid out by the GRADE team in directing the form of feedback on the repository. Each Institution submitted a Site Report (see **Appendix 2.2 template**), which also included detail about current sharing practices and failings (covered in a separate report). The aim of this section of the paper (Report 4) is to summarise the results of findings and formal repository feedback from the Pilot Sites.

The version of the demonstrator being tested during the Pilot Site phase was: Stage 2 repository (See Introduction).

What we asked the Pilot Site to do was a little different. They were asked to directly interact with the repository with 'download' and 'upload' tasks and provide feedback, asked to envision what an 'ideal' geospatial repository might look like and what their particular research 'needs' were versus what was currently in existence at their own institution.

The contributors are as follows:

	Edinburgh	Kingston	Nottingham	Strathclyde
Participants	8	17	5	5

For a full summary of all recorded activities see **footnote reference[35].**

The major part of the investigation here was the Interviews and Feedback as a result of 'upload & download' activities. For this in full see **Appendix 1.6.** Below, (**Table 2**) is a Quick summary of the main feedback, grouped by Institution (& discussed below).

Geospatial data mining 'Activities'

Sites were also asked to document what datasets they chose to download (reflects in some ways patterns of interest, and discovery, and suggests what might make a repository more appealing to a user, or encourage dataset deposit). This is summarised in a Table in **Appendix 1.7**. Boundary data[36] and Syria datasets were the most commonly downloaded.

In combination with the above, Sites were asked additionally to make a 'wish-list' of 'desired datasets' (a top 10) that would love to find in a geospatial repository like GRADE. This is summarised in the Table in **Appendix 1.8**.

[35] 'Pilot Site Reports: Testing the GRADE geospatial repository' (Dec 2006), J. Batchelor, P.Adderley, V. Kasovera, and M. Smith.
[36] GB Historical GIS Project Ancient boundaries and Environment Agency boundaries