Discipline Variations

A comparison review of the discipline-specific surveys 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 in the surveys and interviews across the disciplines and potential obstacles to sharing of data in such a way were also generally consistent. 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 when establishing a Source to Output repository interface. It is likely that the discipline-specific requirements will result in a need for customisation of a generic Source to Output model. The disciplines investigated were Archaeology, Astronomy, Biochemistry, Biosciences, Chemistry, Physics and the Social Sciences. *The draft version of this section of the business analysis does not include biochemistry specific reference as the individual report is not yet complete.*

Identities:

Astronomy	**Archaeology**	**Biosciences**
64 Astronomers responded to the questionnaire, following which five interviews were conducted at the University of Edinburgh and a workshop was held at Johns Hopkins University.	65 responses to 721 questionnaires (9%) of whom just over half were University academic staff.	70% of respons there was a str and Bioinforma amongst those r of cross disc replies 12 interv
Chemistry	**Physics**	**Social Sciences**
Higher response from postgraduate research students than from academic staff. 47% post grad, 39.5% academic, remainder from postdoctoral, research assistants and contracted researchers. 65% of survey respondents claimed not to have used a repository before and were not familiar with open access repositories in general. However of those interviewed, once terminology was explained, most indicated they had used such repositories with particularly emphasis to the Cambridge Structural Database (note: where researchers claim not to have used repositories, across the disciplines this has turned out to be unfamiliarity with the terminology rather then them indeed not having deposited or accessed deposited data. 38 responses 17 interviews.	63 Physics researchers responded to questionnaire and 13 agreed to participate in an interview.	61 questionnai grad students, 1

Project Aims:

The development of a pilot demonstrator is the key deliverable from the StORe Project, it will consist of a set of middleware designed to demonstrate the function of bi-directional links between source and output repositories.

Astronomy	**Archaeology**	**Biosciences**

Astronomers thought that agreeing to a set of standards and web services for accessing, organizing and disseminating data within their discipline would be an essential component. They were generally supportive of the projects aims, there was a minority of respondents who were opposed to the aims, and one going as far as saying that linking would be a dangerous development with reference made to protecting ones data from premature release.	60.0% selected 'significant advantage to my work' with reference to source to output linkage, and 64.6% output to source. Respondents from archaeology seem far more enthusiastic about the issue of source to output repository linkage (in both directions) than for other disciplines. Archaeologists are looking at the potential of improving speed within the research process. Linking repositories would enable more efficient scrutiny of methodology and research process. Possibility of enhanced research profile was also a reason for enthusiasm.	More then 80% the StORe proje between source repositories wo prove extrem Improvements search function researchers.
Chemistry	**Physics**	**Social Sciences**
Academic staff were interested in linking the primary research data to the published outcome, PhD and Postdoctoral researchers were more interested in navigating from the published outcome to the primary data sets. 67% of academic staff indicated that they would find such linkage from primary data to published work useful but not a major significance to their work. 73% found the reverse to be of use. Chemists are concerned with increased functionality, searching, and quality assurance of data, sustainability, and a service that could compete or complement commercially available data sources.	The principal aim for project StORe was well received. 60% thought that source to output linkage and 67% thought output to source linkage would be either a significant advantage to or useful but not of major significance to their work.	Social Scientist generally favor between source advantage to the a useful but viewed from th 60 found sign useful but not generally high among the soci of the StORe Pr

Source Data:

Astronomy	Archaeology	Biosciences
Astronomy data is unconstrained, in the sense that it doesn't contain private, legal and commercial parameters that affect the other disciplines. Astronomers are happy for their source data to be used as long as it is credited. In instances where research is publicly funded, there is an obligation after a propriety period to share data. Source repositories monitor how much they are used, especially if usage figures are likely to be useful in garnering additional funding or support.	Archeologists tend to produce highly complex data sets, and these are often but not always linked into GIS (Geographical Information Systems) which forms part of the way that the information is stored and presented. 74.4% of overall respondents that use GIS are archaeologists and archaeologists produce more maps, plans, plots and images then other disciplines surveyed.	Wide range of videos, images gene/protein se array image dat jpg, tiff, bmp, formats such a protocol) used. combinations favored portabl rather then accessed data access other res
Chemistry	**Physics**	**Social Sciences**

Many variations in data produced, and its recording and storage. Spectra Data, represented by drawings, spreadsheets and image files. Spreadsheets, Word Processed files and image files are the most utilized document formats used, thou discipline specific software's such as .CIF(crystallographic data), binary data files, cdx, xwin nmr, chemdraw, Chemdraw Word, Chemical Markup Language, spectrometer specific code and Fourier induction decay files generated from Bruker and Varian NMR instruments are used.	Many felt that the source data that may be most useful to link to is the final Physics results produced towards the end of a particular analysis and that in most cases linking to 'raw' or 'unprocessed' data would be of little use to others. Physics researchers produce a wide variety of electronic source data and hold this in a variety of formats. Known formats are used, but physicists also write their own analysis software, particularly in the case of high energy physics. Data can range from kilobyte file size up to petabytes (10^15 bytes!). Many researchers do not access other researcher's data.	Extensive use c software, willin qualitative or q statistical data instances) or C instances) and instances). 85.7 came from t Qualitative qu Quantitative qu included HTMI saved MSN con

Source Repositories:

Astronomy	Archaeology	Biosciences
Strong culture of citing sources (thou should be the case across disciplines). Facilities to link source to output repositories are in operation but these are not yet comprehensive. *"If a standard feature of such repositories was the ability to identify and link to the publications that had been developed from these data, how advantageous would you find it?"* *Significant advantage to my work 45%* *Useful but not of major significance 34%* *Interesting but not particularly useful 13%* *Of no interest to me 2%* *Not sure at this point 3% Other 3%*	64.9% of Archaeologists had already either deposited with the Archaeological Data Service or were 'intending to do so soon'. Only 54% of those who had submitted data to an online depository had done so with ADS, indicating use of other depositories.	50% of the res source reposito GenBank (25% GenBank (or Pl submission of a
Chemistry	**Physics**	**Social Sciences**
More then half replying to the questionnaire (65%) claimed not to have used a repository before, but as outlined above once terminology was explained at the interview stage most had been long term and consistent users of such repositories such as the Cambridge Structural database. Quality control of the data in such repositories, comprehensiveness and maintenance were considered to be of primary importance.	Many researchers do not use source repositories: the notable exception being High Energy Physics, where their use is the norm, thou access is often restricted. CERN was the most popular.	Relatively low repositories b questionnaire r never deposite those that had, (8). Individual Global Entrepr reference to the were made. Of with the UKD repository but agenda and 8 unaware of its e

Metadata:

Astronomy	Archaeology	Biosciences

Astronomers should define standard methods to refer to the same objects; there is currently a degree of disparity when objects are viewed through different spectra. This will be of particular importance when data is to be deposited into output repositories. Additional Metadata gathering through automated functionality (automated weather information linked to telescope data etc) would be useful.	High level of metadata awareness. Many expressed frustration at the difficulties of searching accurately and reliably for resources, mainly down to differences in keyword usage or inadequate information on the datasets for the discipline. High degree of enthusiasm for a standardized word list and thesauri. Of those that had deposited data sets, 66.2% had decided on and assigned metadata themselves. Of those that hadn't deposited, awareness of metadata was often vague. Main concerns were that the process of data depositing, especially the assignment of metadata was perceived as a time consuming and complex process and had deterred them from doing so.	Some research metadata. Main data and projec on links and d metadata. An e research was do inconsistent sta research data is Lack of familiar
Chemistry	**Physics**	**Social Sciences**
Author/Creator was considered the most important metadata element for 89% of the chemists. Other important considerations were Project Description (68%) project title (68%), and subject keyword assignment (58%). The least important metadata was considered the funding source (13%). More than one third of the respondents (37%) indicated that metadata is assigned to resources during file saving which indicates the involvement of software for automatic assignment. 53% noted they themselves decided on the terms to use and the assignment of metadata; however 29% did not know who assigned the metadata to their resources.	Metadata most commonly assigned during file saving as part of the indexing process of source files. Most commonly defined and assigned by the researchers themselves or is done automatically by the software. Researchers believed the most important data to assign consist of generic keywords and a number of terms specific to the physics field of interest; the type of metadata assigned also varies according to the stage of analysis.	Generally soci have assigned third claiming used project titl creator names project descri period, and num said they didn admitted they t often given to instance or how access it. Inco storage media w

Data Access and Sharing:

Astronomy	**Archaeology**	**Biosciences**
Due to the unconstrained nature of the data astronomers and librarians can build systems in an open manner and generally ensure that data is widely available. There are no controls on the information due to confidentiality, ethical constraints, concerns over premature broadcast or lost commercialization opportunities etc.	Most respondents are happy to share their data widely (64.9% had or intending to deposit with ADS and 13.8% had or intended to deposit with another source depository. There were still significant levels of concern regarding public data access. There was concern over the illegal looting of archaeological sites if such data offered up a geographical location. Others felt that collaborative projects, especially those working in conjunction with overseas teams, result in shared data ownership and such data couldn't be disseminated without others approval. There was limited understanding of access control methods.	High level of t Most data is sh publication or th stated they hav research data av visibility would were concerned commercializati constraints did willingness to s expressed hosti of information other researche request, not p 37% provide su any formal restr
Chemistry	**Physics**	**Social Sciences**

There was a spread of responses and no single key factor that appears of significant import that would encourage the respondents to share access to their data. Those that were broached included, potential benefits to the research community and demonstratable benefit to research profile. The threat of loss of ownership and premature broadcast were considered hurdles to sharing data. Academic staff and postgraduate research students did not apply any formal restrictions to their data but judged each request on its merits as opposed to proactively publishing data. Academics preferred an 'ownership retained - request acknowledgement on reuse' control. Contracted researchers tended to secure data on pass worded systems or standalone terminals.	Over a third of respondents said that they take no measures to make their research data available. Many would be encouraged to share data if it was for collaborative research purposes or would benefit the research community and raise their own research profile. They were deterred by premature broadcast of results and a thread of loss of ownership. Time spend facilitating the data sharing was also of concern. Many were not against the idea in principle but considered there to be practical obstacles in doing so.	Social Scientist principle, althou could be more maximum use making it ava currently large within closed r requests from c individual perc social science make their data attempt to activ 19 claimed to u portable media.

Output repositories:

Astronomy	Archaeology	Biosciences
"How advantageous to you would it be if it were possible to go directly from within an online publication (electronic journal article or other text) to the primary source data from which that publication was developed?" *Significant advantage to my work 36%* *Useful but not of major significance 55%* *Interesting but not particularly useful 6%* *Of no interest to me 0%* *Not sure at this point 0%* *Other 2%*	Archaeologists do not make as much use of Output Repositories as researchers in other disciplines. 2.1% claimed not to use them to gain access to published papers, compared to just 8.1% claiming this overall. Over 41.5% do not deposit, compared to 20.4% overall. That said, the results are misleading as interviews suggest that researchers initially misunderstood the definitions used for output repositories by the StORe Project. All those researched has used them.	Generally brow as a general inf were used more of all researche an advanced sea logic. None use
Chemistry	**Physics**	**Social Sciences**
Commercial sector output repositories managed by journal publishers were those most commonly accessed. Academic staff used institutional, discipline, publisher and 'other' repositories. Prefer simple search terms thou a wide use of search methods is utilized. Subject specific thesauri and Boolean logic are only mentioned in searched institutional and discipline repositories.	The vast majority of Physicists make use of output repositories for their research. All three types of repository: publisher, discipline and institutional were cited as being used. Publisher repositories were the most commonly used. Most were supportive of the idea of an open source repositories but had concerns about appropriate peer reviewing occurring before depositing.	Out of 61 Sc questionnaire, output reposito claimed not to cited 'other' typ level understan types of repos indication that journals depo automatically a none meant the information the the disciplines claim not to hav

Support:

Astronomy	Archaeology	Biosciences

Astronomers are more likely to seek assistance with Metadata and Preservation related to datasets then they are to seek help from librarians or informational professionals with regards to navigating the systems.	Most were not aware of the support available to them, and relatively few make much use of online help. Despite initial hesitance to ask for help, those that had done, found that they had benefited a great deal from doing so and that they could carry that enhanced awareness into future repository use.	Personal suppo seems importan claiming to use that they receiv utilized was training/docume knowledge ma output reposito
Chemistry	**Physics**	**Social Sciences**
It was felt that the availability of a prototype that would illustrate what the StORe project proposes would have made it easier to understand and comment upon the advantage and barriers to use. Academic staff were familiar with existing level of support mechanisms available at repositories they use, thou this was not the case of postdoctoral and research assistants.	Mainly self sufficient, 1/3rd having used no support, of those that do use support, the repository enabled support is the most popular. Where assistance is provided by librarians or other knowledge management support, the provision of documentation along with online or telephone held are popular services. There exists a clear lack of awareness of what assistance is available from such staff by a significant proportion of physicists.	23 of 61 So questionnaire support or guida 23 had receive either in respect provided by a support or 'othe know what supp

Cross Disciplinary:

It was discovered that within the context of the Biosciences discipline, there were substantive cross disciplinary access to information. Researchers working within the field of biosciences regularly accessed and referenced data from other disciplines, such as chemistry and mathematics and any portal for linking different data depositories would need to enable this. There was substantively more cross disciplinary access to information that apparent in the other disciplines.

Hilary Beedham, UKDA, November 2006