Exploring Research Data Hosting at the HKUST Institutional Repository

Abstract
The profusion of data created by modern research has brought about increasing discussion on the practice of data management and the development of such infrastructure. Many universities and libraries are exploring how to address this issue using institutional repositories, the HKUST Library being one of them. As part of the exploration, the author conducted a case study of datasets at major institutional repositories.

This paper visits the issues relating to data archiving, reports common practices of data treatment at institutional repositories, and presents a list of recommendations for experimenting data archiving at the HKUST Institutional Repository.

E-Research and Data
Technology has been forming and transforming the nature and scale of e-research.

The advance of Web technology opens up new ways for researchers to communicate. The capability to collaborate online encourages cross-discipline research. Researchers from different parts of the world, carrying different expertise, readily get in touch with each other and engage in online discussion and exchange. They share files in a wide variety of format, containing both text-based and non-text-based research findings. Such dynamic interaction creates synergies that were not possible before the Web 2.0 era.

On the other hand, new computing tools are reforming operations and procedures in laboratories and research facilities. They allow easy generation and ready capturing of large amount of data via simulation, modeling, networks of sensors and instruments. While file size may have become less and less an issue when storage space and computing speed are improving, the volume of data that scientists are creating is also growing at a daunting speed that poses real challenge to data management. The data flood is particularly prominent in large scale projects such as those in genomics, climate research, high energy physics and astronomy. Digitally-born data from such “Big Science” is usually highly organized on the front end to make it easier to handle and to archive; while data from “Small Science”, relatively smaller projects performed by researchers and research teams in universities, is much more heterogeneous. Although much discussion focuses on research data in science and technology, from a broader perspective, data management is an issue concerning research in all disciplines, from physical, biotechnological, environmental, to social sciences and humanities.

Well-managed data in digital form has great potential to be searched, accessed, mined and reused. Data may be examined to validate research results; it could be consulted by researchers of related interest and save time and resources in data re-collecting; data may
even be re-purposed to answer questions unrelated to the context in which it was first generated or gathered. The value of data grows significantly as it forms collections and become more accessible.

The awareness of the data deluge phenomenon, the potential impact of data reuse, and the desire to maximize the return of investment of research funding, all led to increasing amount of discussion on research data management. Some research funding bodies, information agencies and institutions have established policies on data management, or set up working groups to investigate the issues. The National Science Foundation (NSF) of the United States was moving toward requiring that grant proposals include data management plans for data generated by sponsored projects; the Australian government formed the Australian National Data Service (ANDS) as a cooperative centre with expertise in research data management; in the United Kingdom, the Joint Information Systems Committee (JISC) funded the Digital Curation Centre (DCC) to facilitate digital curation best practice within the rapidly changing environment of e-Research. A number of reports and guidelines facilitating data sharing have been released. For instance, the National Science Board (NSB) published the report “Long-lived Data Collection” in 2005; the UKOLN consultancy report funded by JISC explored the roles, rights, responsibilities and relationships of data stakeholders; a 2008 Research Information Network (RIN) report surveyed researchers’ attitudes and their practices in relation to the data they created and gathered in the course of their research.

Studies like these and other discussion of data archiving issues seem to point to one general direction: effective data management is a business with high complexity; it involves technological issues, intellectual property concerns, and entangles interests of various stakeholders of scholarly communication. A good amount of innovation, negotiation and discussion is needed before widely acceptable and practicable frameworks come to being.

**Data Archiving**

**Objective**

In a 2008 document that provides guidance for building policy framework, RIN states the objective of data management as ensuring that “ideas and knowledge derived from publicly-funded research should be made available and accessible for public use, interrogation, and scrutiny, as widely, rapidly and effectively as practicable.” The document sets out five principles as a starting point for drawing up policy framework, with responsibilities shared among all stakeholders: universities, research institutions, libraries and data services, publishers, research funders and researchers.

**Open Data**

To maximize the potential value of data, it should be openly accessible for ready access by human users as well as data mining by computers. While open access has been discussed widely, and implemented by repositories in different parts of the world, researchers’ response to archiving their publications for open access has not been as warm as hoped. The situation for open data is even more challenging. Intimately linked to the altruism of the
open data movement are technological, cultural, and legal issues that need to “be addressed in order for what is now a global research community to reap the full benefits.” The RIN objective quoted above alludes to open access, but in reality, time delay and other access restrictions would be legitimate and necessary in order to balance conflicting interests and rights.

Brody points out two hurdles for open data as “that not all researchers want to make their data OA and that the online infrastructure for [data]-OA still needs additional functionality.” The RIN report identifies a number of reasons for researchers’ reluctance to making data public:

- lack of career rewards is a major disincentive
- wish to retain exclusive use of data until all the publication value is extracted
- lack of time, resources and expertise to handle the data management
- legal and ethical constraints, such as data ownership issue and confidentiality issue when personal data is involved
- lack of an appropriate archive service
- fear of exploitation or inappropriate use of the data

Henty succinctly remarks that “[a]t present there are not the policies, attitudes, understanding, commitment or mechanisms in place to allow data deposit to occur as a matter of course and it is here that institutional policies and advocacy have a major role to play.”

**Institutional Repositories for Research Data**

The present picture of research data management can be described as patchy. Although some research communities have established norms and standards for depositing data, with well-structured data repositories in place, the reality remains that the majority of data in most disciplines is left unattended after the publication value is extracted, regardless of whether the data is in digital form or not.

Data from research projects in universities, being an essential component of research output, naturally falls into the purview of institutional repositories. Many academic libraries are already using their repositories for purposes beyond text-based publications. With the experience of storing images, audio files, software codes and others, it appears that hosting datasets is a simple step further. Technically speaking, most repository software accommodates datasets; for instance, DSpace accepts all manner of digital formats. However, considering the goal to make the data discoverable, accessible, and usable, libraries cannot simply treat datasets in the same manner as published papers in repositories. Data differs fundamentally from text-based research output in a number of aspects, including the nature of the information content, the role it plays in the scholarly communication cycle, metadata required for effective discovery, its potential use, intellectual property right ownership, and last but not least, contributors’ perception to archiving.
Despite these constraints and uncertainties, institutional repositories do offer themselves as attractive venues for experimenting data management. Repositories have been established for a number of years in universities worldwide, with good amount of experience and expertise in research output management accumulated, they are well-positioned to take on the challenge to explore data management. Data archived in a repository immediately enjoys a number of benefits: value-adding metadata that is accessible via standard searching and harvesting protocols, including the Open Archives Initiative (OAI-PMH); unique persistent identifiers via the Handle System for easy referencing and citing; and aggregation with related research output deposited in the repository. More importantly, institutional repositories become the valuable option for researchers whose discipline does not have any well-established data dissemination system.

**Experiences and efforts at universities**

Many universities are implementing data archiving at their institutional repositories with different approaches and levels of ambition.

Purdue University adopted a “distributed approach”. *E-Data*, the designated repository for research datasets built with Fedora, is under development by the Distributed Data Curation Center (D2C2). It serves as a platform for experimentation in data curation.14 Disciplinary librarians took up the responsibility to solicit research datasets; a list of standard interview questions was produced to assist them to talk with their faculty. In addition, the D2C2 also works with the university IT unit to explore a data archiving service. Purdue’s approach of partnership with researchers to best understand and meet their data needs has laid the groundwork for future, higher-level work to formalize data curation services for the institution.

At Johns Hopkins University, data archiving is bringing up new ways of looking at the model of “institutional repositories”. As data has become a new form of publication, the Sheridan Libraries explore repositories as a mechanism for housing data as part of a compound object publication,15 which is being modeled using the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) protocol.16

Monash University, Australia, realizes that different kinds of objects cannot easily be accommodated from a common repository infrastructure.17 It adopts a multi-dimensional view of information – data curation continuum. Two repositories were created to address different ends of the continuum: the collaboration and publication repository, and the preservation repository. The former is characterized by having less metadata, more items, larger objects that are often continually updated, research management, less preservation, mostly closed access and less exposure. The latter is for finished projects in the sense that some results are available for public viewing. This is characterized by having more metadata, fewer items, smaller objects that are usually static or derived snapshots, organizational management, more preservation, mostly open access and more exposure.
The HKUST Institutional Repository

The HKUST Institutional Repository (IR) was launched in February 2003. Powered by DSpace, the IR was built with the aim to create a permanent record of the scholarly output of HKUST in digital format, and to make the Repository globally and openly accessible. In other words, the defining characteristics of the HKUST IR are:

- being fully open access;
- focusing on scholarly output. Other community output such as students work and teaching objects are not included.

In March 2009, the IR hosts over 3,000 items, most of which are text-based research papers.

Similar to the challenges faced by other universities managing and promoting institutional repositories, the support from faculty as well as university administration at HKUST to such an open access repository has not been as strong as desired. To augment the IR service to cover research data, the Library would have to adopt a bottom-up approach. The Library administration initiated an exercise to study the technical issues of data archiving, and to find out how to facilitate the discoverability, accessibility and usability of data hosted at the IR. The author, acting as the IR Coordinator of the Library, was charged with the exploration by reviewing data issues discussed in the literature, and performing a case study to survey how datasets are presently hosted in institutional repositories.

The Case Study Sample

The author conducted a case study to survey the common treatment of research data at institutional repositories. The study aims to find out:

- What kinds of data are currently being archived at institutional repositories
- How data is accommodated in different repository software, especially in a DSpace environment
- What measures are effective in making the data discoverable, accessible and usable

The author used OpenDOAR, a major directory of open access repository, as a starting point. In February 2009, a search in the directory specifying “Institutional” as the repository type, in English language, and “Datasets” in the content type, retrieved 43 records. During the course of the investigation, the author came across another group of 10 repositories that also host data but were not listed in the OpenDOAR search result. Using this set of 53 repositories, steps were taken to locate data items:

1. If there was a browsing structure, looked for any collections that might contain research data;
2. Looked for relevant search options such as “document type” or “media type”, and then searched for “data” or “dataset”;
3. If none of the above options was viable, searched the repository with the keyword “data” and “dataset”.

Step 1 yielded a variety of data types including images and audio streams. On the other hand, what step 2 could retrieve depended a lot on how the repository defined “data”, and the result leaned towards numerical data. At some repositories listed on the OpenDOAR search results, no research data could be found after all three steps were taken.

From the items discovered, the author focused on the selection of research data; teaching objects were excluded. Some repository items selected did not appear to represent research output generated directly from the corresponding institutions; it might be, for instance, a dataset digitized from a special collection item. Since the study focused on the technical treatment of research data at repositories, where the data was generated was not a concern.

Consequently, the author selected 55 repository items for further study. They come from 32 repositories of institutions in 6 countries (Appendix). 46 items are hosted in DSpace repositories. They can be categorized into 5 groups in terms of data genres (Table 1).

The author would like to emphasize that the sample by no means represents the data collections in the repositories visited. This investigation did not intend to evaluate the policy and practice of those repositories either.

<table>
<thead>
<tr>
<th>Data Genres</th>
<th>Number of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>2</td>
</tr>
<tr>
<td>Dataset</td>
<td>37</td>
</tr>
<tr>
<td>Database</td>
<td>7</td>
</tr>
<tr>
<td>Image</td>
<td>6</td>
</tr>
<tr>
<td>Program code</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
</tr>
</tbody>
</table>

Discussion

Repository terms
For effective discussion, the author uses the terminology of DSpace when referring to repository structure and components:

- Organization structure – a repository is divided into communities; each community contains collections, which are groupings of related content.
- Items – the basic unit of a repository. An item consists of a set of metadata and bundles of bitstreams.
- Bitstreams – computer files deposited in a repository. Each bitstream is associated with one Bitstream Format for interpretation.
Data types

To set up sensible policy for data management in repositories, a repository manager must understand the different types of data and their characteristics. This distinction between types is crucial to choices made for archiving and preservation.

Research activities use, gather and process data in different ways and different forms. Data may be numerical, descriptive, audio or visual. For instance, for an oral historian, her primary raw data is likely to be audio recording; for a computer scientist, her research data may easily be executable programming codes or computer programs. Data may be born-digital, or digitized from analogue or print format. Digitized data may come in different file structure and file format. Although the term “data” and “datasets” are sometimes used interchangeably, “datasets” usually represent files that contain data in tabular or delimited format. The data in a dataset is often alphanumerical entities, other forms such as audio or video streams are also possible.

The NSB report describes three kinds of data according to the ways data is gathered, and how the nature of the gathering process affects the preservation policies:\(^{20}\)

- **Observational data**, such as direct observations of ocean temperature on a specific date, the attitude of voters before an election, or photographs of a supernova are historical records that cannot be recollected. Thus, these observational data are usually archived indefinitely.

- **Experimental data** such as measurements of patterns of gene expression, chemical reaction rates, or engine performance present a more complex picture. In principle, data from experiments that can be accurately reproduced need not be stored indefinitely. In practice, however, it may not be possible to reproduce precisely all of the experimental conditions, particularly where some conditions and experimental variables may not be known and when the costs of reproducing the experiment are prohibitive. In these instances, long-term preservation of the data is warranted. Thus, considerations of cost and reproducibility are key in considering policies for preservation of experimental data.

- **Computational data**, such as the results from executing a computer model or simulation. If comprehensive information about the model (including a full description of the hardware, software, and input data) is available, preservation in a long-term repository may not be necessary because the data can be reproduced. Thus, although the outputs of a model may not need to be preserved, archiving of the model itself and of a robust metadata set may be essential.

The RIN report adds these two categories on top of the three above:\(^{21}\)

- derived data, resulting from processing or combining raw data
- canonical or reference data relating, for example, to gene sequences, chemical structures, or literary texts

The report continues to give two reasons of data gathering, which have implications on the long term value of the data:
• for the benefit of those engaged in a specific project, where some or all of the data may or may not retain a value beyond the life of the project
• for the benefit of a wider group within a discipline, or across disciplines, to provide reference information, or a basis for evidence-based policy-making

In the case study sample, one can find observational, experimental as well as computational data. Regarding whether they are raw or derived types, except a few that were not determinable, most represent raw data, with a few cases of derived data. Majority of the cases are project-based data, while a handful is reference data.

At HKUST, the author anticipates that the most voluminous group would be project-based, raw data.

Apart from these data varieties, another characteristic unique to datasets that affects management decision is data updating. Unlike published papers that exist as a definitive, archival version, data may be a “living” entity. Some datasets are produced through longitudinal studies with new data being added from time to time. How an institutional repository handles growing data is a matter of project mission and policy. At the HKUST IR, it is likely that datasets will be archived as snapshots; new versions will be created as new items in the repository.

**Organization of items**

**Datasets as individual items vs. datasets as accompanying files**

For datasets that lead to paper and thesis, some repositories treat them as self-standing items, while some others put them as accompanying files to the text-based output. In the sample, two cases were found with datasets treated as accompanying files. The rest of them possess separate items and hence their own metadata. Without its own metadata, it was extremely difficult for such datasets to be discovered. That may be explaining why only two such cases were encountered.

**Browsing structure**

“Communities” and “collections” in DSpace, and similar organization structure in other repository platform, facilitate browsing.

Among the sample items, some are organized by collections carrying project names. An example is the Kommos project at *T-Space*, University of Toronto. Under that collection, users can find unpublished data such as numerical tables and illustrations of the archaeological sites, alongside with the published papers.

Another frequent practice is to have a collection designated for “data”, one example is the collection “Datasets, School of Psychology” at the *Leicester Research Archive*.

DSpace@Cambridge allows browsing by “Document Type”. This feature is particular helpful to show the variety of document types and highlight the availability of datasets collection.
Descriptive information
Datasets are much more versatile than text-based research output. Adequate description is crucial for data management to facilitate data discovery, access and use. Productive descriptive information for data should cover:

- the origin – the context in which the data was generated or gathered, and the methods
- the “table of content” – the list of bitstreams, and what they contain
- the use instruction – how to approach the datasets, what software is required
- use policy – any use or license agreement

The core of descriptive information appears as metadata of the repository items.

Metadata
Standard metadata schema significantly enhances the effectiveness and efficiency of searching. In some fields, professional databanks set up standardizations by imposing rules for depositors to comply with. The specific file formats and metadata contain all details that are relevant and integrable for other researchers in the same field.

On the contrary, an institutional repository needs to cater for heterogeneous data from different disciplines. DSpace uses Dublin Core as its metadata schema. How to tailor this generic schema for effective data description requires further deliberation.

The ANDS proposal lists three types of metadata relevant to the discussion of data archiving:

- descriptive (this describes and identifies information resources, and is often used for discovery);
- structural (this describes the relationships between resources, and is sometimes used to facilitate navigation and presentation of electronic resources); and
- administrative (this can include provenance, access control, and preservation information, and is used to facilitate both short-term and long-term management and processing of resources).

From the case study sample, all three types of metadata are presented at different levels of details and comprehensiveness. An outstanding example is Database of dedications to Saints in Medieval Scotland at DataShare of the University of Edinburgh; the metadata provides details about the origin of the data at the “abstract” field (descriptive), relationships between files in the “TOC” field (structural) and provenance (administrative). A data archive from Discovery 247, a process study of the Faroe Bank Channel overflow at WHOAS does not have much descriptive metadata, but clearly stated steps for opening the data files (administrative) considerably enhance the data usability.

Generally speaking, the descriptive information in the sample is rather inadequate for users to effectively approach the datasets. The observation indicates that metadata specification in current practice requires further strengthening to enhance discovery and usability by both computer and human users.
In DSpace environment, a number of Dublin Core qualifiers that are less commonly used for research papers may be important for data. Those demonstrated as functional in the sample items include:

- dc.coverage.spatial
- dc.coverage.temporal
- dc.description
- dc.description.abstract
- dc.description.provenance
- dc.description.tableofcontents
- dc.relation.hasversion
- dc.relation.ispartof
- dc.relation.ispartofseries
- dc.relation.isreferencedby
- dc.relation.isversionof
- dc.rights.uri
- dc.type

No repository visited indicates distinctively the data types (observational, experimental or computational).

**Other descriptive information**
Apart from item metadata, descriptive information can be presented at various locations in and around the items. The sample items demonstrate these variations:

- as supplementary files; such as the “readme” and “introduction” for the water quality data at eCommons@Cornell;\(^28\) the deposit agreement at University of Edinburgh DataShare,\(^29\) and the disclaimer in pdf at Leicester Research Archive.\(^30\)
- as description of the collection; such as the detail explanation for the 21 datasets of the Dobe !Kung project at T-Space.\(^31\)
- using bitstream description and bitstream format. The description is specific for the bitstream, not the repository item. For instance, for a preprint at KU ScholarWorks,\(^32\) the metadata at the item level describes the preprint, the dataset attached does not possess its own Dublin Core metadata, but relies on the bitstream description. Many repositories leave the bitstream description field blank. Bitstream format is usually populated; some repositories use plain language description such as “Microsoft Excel”, some keep the MIME type such as “application/vnd.ms-excel.”\(^33\) Plain language is obviously easier to understand for human users.

The author finds it important to have adequate but precise description at the item metadata for users to decide whether he or she should download a bitstream. More detail instruction can be attached as supplementary bitstreams.
Usability issues
Even when a dataset is found and accessed, it has to be readable and interpretable in order to be usable. A few issues need to be considered for satisfactory data usability.

File formats
Research data comes in many file formats, which may limit the usability of the data files. Pdf format, although carries a lot of merits for text-based output, imposes restrictions on further manipulation of the data embedded. Some file types require specific software to read. How common that particular software is, and whether it is free for download, become an issue for reuse. Even when a dataset can be read by basic software, it may not reveal its meaning until the data is visualized by additional software application, which may be proprietary or developed specifically by the data creator.

For data that requires software for viewing, manipulation and analysis, the issue of long-term availability of software application comes to fore. Software may have new versions or may become obsolete in “close-enough” future. This poses another layer of complication for repository management.

The 55 repository items are presented in a good variety of file types:

- Microsoft Excel files (xls)
- Adobe PDF (pdf)
- Microsoft Access databases (mdb)
- Audio files (mp3)
- Graphic files (jpg)
- SPSS files (sav)
- Comma Separated Values (csv)
- HTML pages (html, htm)
- Compressed files (zip)
- Microsoft Word files (doc)
- Executable files (exe)
- An external website
- Others, including software program code, Lotus 1-2-3, plain text, Filemaker Pro, and xml

Most of these file types are readily readable by commonly used software, except SPSS, Filemaker, and perhaps Microsoft Access. Zip files usually contain packages of files that may include specific software for further manipulation.

Repository managers need to consider the support level of various file types in the long term. After all, an archive full of unusable datasets does not bring any value to the community.

Citation standards
Researchers need to cite resources that they use. Data is no exception. Similar to publication citation, data citation is essential for these purposes:
• other researchers can locate the data for validation and reuse
• data creator can receive the recognition and acknowledgement of her contribution
• data creator and her mother institution may use citation tracking to gauge the level of use

While citation standards for traditional research publications are well-established, a robust data citation standard is not yet in place. The issue has drawn attention and being discussed in a number of initiatives. The major social research data repository ICPSR provides clear instructions for users to make citation. Altman and King have proposed a standard for citation of quantitative data; while NISO called a Research Data Thought Leader Meeting in October 2008, with the goal to “incubate new standards initiatives by discussion issues and areas where standards can help address pain points, push forward reuse of data, or drive application of systems in research and information exchange.”

Deriving a data citation standard is more complicated than publication citation. It needs to address citation cases that are specific to datasets, such as citing a portion of a large dataset, and citing a particular version of a growing dataset. Repository managers may help to promote usage and deposit rate by providing guidelines for data users on proper data citing.

**Connection with related items**

Datasets seldom exist as standalone research output. To facilitate proper use, a dataset should be clearly linked to its related files, such as derived data, research papers resulting from the data and previous or later version of the dataset.

In DSpace, related items can be presented in a number of ways. One or more description fields may be employed to store the information, either in plain text or hyperlinks to other items in and beyond a repository.

The observation from the sample indicates that not many data items in institutional repositories list out related research output clearly; and even fewer make direct hyperlinks to those related items.

These are a few cases in the sample that point to related items. They compose a demonstrative set of the kinds of related items for datasets:

• data and manuscript of the same project; listed in a description field, with handle numbers given but not hyperlinked.
• theses or research papers supported by the data; citation may be stated, hyperlink is usually absent.
• earlier version and later version.
• two file formats for the same set of data, hosted as separate items.
• the same dataset at a subject databank, with hyperlink provided.

The OAI-ORE protocol currently being developed defines standards between distributed resources such as articles and datasets. It has the potential to become a more appropriate solution than manually adding and maintaining hyperlinks in repository items.
Way forward for the HKUST Institutional Repository

With current institutional culture and human resources available to the Library, the pilot for hosting data at the HKUST IR would be most appropriate to adopt a bottom-up approach. Datasets would be ingested into the existing IR, with an additional set of policies and measures to address their special requirements.

Policy

The RIN report 42 is an excellent starting point for drawing up policy framework. Under each of the five principles listed, the report presents a series of key issues and questions to be addressed. A number of these issues are particularly relevant to institutional repositories; to name a few:

- Are procedures in place to determine ownership of data?
- Are metadata machine-understandable, and in compliance with recognized standards?
- How to ensure that the facilities, software and other tools necessary to access the data available to users in the long term?
- How to ensure that the data underpinning research findings reported in publications are made accessible to readers, and that references to data are included in relevant papers, along with appropriate links?
- Have clear protocols been established for the citation of datasets?
- Are there sustainable procedures to ensure that the data are protected through explicit security protocols from unauthorized modification, damage or destruction?

These questions would direct the Library to identify strengths and weaknesses of the university community regarding its awareness and current practice of research data management; they also help the Library to define its role and responsibility in the effort to preserve data. The answers to some of the questions may readily be translated into practical service guidelines. Naturally, not all issues can be, nor need to be, addressed imminently before the IR starts accepting data, but the awareness of them would help drive the development of the service and promotion effort to the university research community.

In particular, the policy for data archiving within this open access IR should cover:

- the role, responsibility and limitation of the Library in the process of data curation, selection, management and preservation
- the types of data to be accepted, or preferred
- treatment of different versions and related research output
- the level of support to different file formats and their long term usability
- the level of metadata required from the contributors
- what kind of deposit agreement would be desirable

Guidelines to contributors should be devised to bring about:

- the awareness about the potential usefulness of data, and how to perform the evaluation
• the awareness of intellectual property issues
• the awareness of any confidentiality concern in the data
• how file types affect usability
• how to provide descriptive information as metadata
• whether any use policy or restriction should be specified, and how
• requirement to provide codebooks as supplementary files
• suggestion to effectively organize multiple files into a single IR item
• how to organize research output into collection and delineate connection between items
• technical advice on file manipulation such as conversion, zipping and unzipping

Structural measures
To facilitate discoverability, accessibility and usability, the IR should consider specifying measures within the repository structure regarding these four areas.

• Organization of items – Considerations include the criteria for treating datasets as individual items, as groups or as supplementary to published papers. The key point is in what circumstance a dataset deserves its own set of metadata. Contributors may be encouraged to aggregate research output into project-based collections to enhance browsing, and thus leverage the visibility of the projects.
• Descriptive information - The “Document Type” field is crucial for datasets. Although it is common for repositories to use “Data” or “Datasets” to label the document type, the author recommends that the new type “Research Data” be added to the HKUST IR. Regarding the desirable level of metadata, one criterion is that there should be enough information for the user to decide if he/she should download the files without actually downloading them.
• Usability concerns – Contributors tend to have weak awareness of the usability issue. The repository may provide guidelines and assistance regarding the choice of file formats, citation style, use instruction and use policy.
• Connection to related items - Contributors would need assistance and encouragement to identify research output related to a particular dataset. The IR should make an effort to provide the linkage as much as possible.

Content recruitment
To start up the pilot of data archiving from bottom-up, the Library may encourage data contribution into the IR via these means:

• Dissertations - The HKUST IR also hosts dissertations of HKUST graduates. Students may be encouraged to submit their research data at the time of dissertation submission through the Library.
• Communication with data-intensive research centers - The Library may approach research centers and projects at the university that produce large amount of data, and explore how the IR may help them with their data management.
The Next Steps
Data management is not a library-only issue. Libraries may initiate projects as start-up attempts to address the issue, but the ultimate success depends on support from other stakeholders in the scholarly communication cycle. Efforts and projects that aim to elevate awareness of research data management among researchers, faculty, university administration and funding agencies would be conducive.

Mutual understanding is the core of successful collaboration. Although it would be instrumental for the HKUST Library to perform surveys to study data production and management practice at the HKUST, such effort would call for human resources that may be beyond what the Library can currently afford. Nevertheless, establishing communication with faculty on this regard is a very crucial start for the project.

Furthermore, data management requires a wide range of skills that is currently lacking among researchers and librarians. Relevant staff training would be one of the investments that the Library may consider while pilot testing data archiving at its IR.

Conclusion
Research data management has become a pressing issue. While the need for effective data management is rising everyday with the incessant production of data in all research fields, a myriad of issues are yet to be resolved before a satisfactory system comes into being. However, the hurdles have not deterred institutions and interested agencies from developing different systems and models that attempt to address the data issue. Universities that are operating institutional repositories are well-positioned to place their repositories on the testing ground. The HKUST Library is ready to take a start on this endeavour.

Appendix
1. ARROW Repository, Monash University, Australia
2. Demetrius, The Australian National University, Australia
3. Adelaide Research and Scholarship, The University of Adelaide, Australia
4. UQ eSpace, The University of Queensland, Australia
5. Sydney eScholarship Repository, The University of Sydney, Australia
6. Digital Repository, University of Melbourne, Australia
7. AUSpace, Athabasca University, Canada
8. Qspace, Queen's University, Canada
9. SFU Institutional Repository, Simon Fraser University, Canada
10. Dspace at University of Calgary, University of Calgary, Canada
11. T-Space, University of Toronto, Canada
12. Open Access LMU, Ludwig-Maximilians Universitat Munchen, Germany
13. CURATOR, Chiba University, Japan
14. Kernel, Kobe University, Japan
15. Kurenai, Kyoto University, Japan
16. Spiral - Imperial College Digital Repository, Imperial College London, UK
17. Loughborough University Institutional Repository, Loughborough University, UK
18. SAS-Space E-repository, School of Advanced Study, University of London, UK
19. ROSE, University of Bristol, UK
20. Dspace@Cambridge, University of Cambridge, UK
21. DataShare at the University of Edinburgh, University of Edinburgh, UK
22. Leicester Research Archive, University of Leicester, UK
23. SOAS Research Online, University of London - School of Oriental and African Studies, UK
24. eCommons@Cornell, Cornell University Library, US
26. Jscholarship, Johns Hopkins University, US
27. WHOAS, Marine Biological Lab and Woods Hole Oceanographic Institution, US
28. Dspace@MIT, Massachusetts Institute of Technology, US
29. KU ScholarWorks, The University of Kansas, US
30. University of Delaware Library Institutional Repository, University of Delaware, US
31. IDEALS@Illinois, University of Illinois, US
32. ResearchWorks, University of Washington, US

Notes

http://chronicle.com/free/v52/i42/42a03501.htm (access April 15, 2009)

2 “Cyberinfrastructure vision for 21st century discovery, 2007,”

3 “Towards the Australian Data Commons, October 2007,”

4 “About the DCC,” http://www.dcc.ac.uk/about/ (accessed April 15, 2009)

5 “Long-lived Data Collection: Enabling Research and Education in the 21st Century, October 2005,”


7 “To share or not to share: Publication and quality assurance of research data outputs, June 2008,”
http://www.rin.ac.uk/data-publication (accessed April 15, 2009)


11 “To share or not to share”, 26

12 Margaret Henty, “Dreaming of data: the library’s role in supporting e-research and data management,” (paper presented at the Australian Library and Information Association Biennial Conference, Alice Springs, Australia, September 2-5, 2008), http://hdl.handle.net/1885/47617 (accessed April 15, 2009)


20 “Long-lived Data Collection,” 19
21 “Stewardship of digital research data,” 6


23 “Figure 2: Plan of Minoan hilltop houses,” http://hdl.handle.net/1807/9604 (accessed April 15, 2009)

24 “Dilution as a model of long term forgetting,” http://hdl.handle.net/2381/3850 (accessed April 15, 2009)

25 “Towards the Australian Data Commons,” 41

26 “Database of dedications to Saints in Medieval Scotland,” http://hdl.handle.net/10283/16 (accessed April 15, 2009)

27 “A data archive from Discovery 247, a process study of the Faroe Bank Channel overflow,”
http://hdl.handle.net/1912/950 (accessed April 15, 2009)

28 “Manuscripts and Water Quality Data for Watersheds and Lakes in Central NY, 1972-2003,”
http://hdl.handle.net/1813/2547 (accessed April 15, 2009)

29 “Refractive indices (500-3500 cm^-1) and emissivity (600-3350 cm^-1) of pure water and seawater,”
http://hdl.handle.net/10283/17 (accessed April 15, 2009)

30 “Separate and dissenting opinions in the European Court of Human Rights: an Access database,”
http://hdl.handle.net/2381/1406 (accessed April 15, 2009)


32 “Is there reduction in disease and predispersal seed predation at the border of a host plant’s range?
- field and herbarium studies of Carex blanda,” http://hdl.handle.net/1808/1507 (accessed April 15, 2009)


34 “ICPSR: Citing Electronic Data Files,” http://www.icpsr.umich.edu/ICPSR/org/citation.html
(accessed March 18, 2009)

35 Micah Altman and Gary King “A Proposed Standard for the Scholarly Citation of Quantitative Data,”
(accessed April 15, 2009)
Maureen C. Kelly “NISO Thought Leader Meeting on Research Data draft report,”
http://www.niso.org/topics/tl/NISOTLDataReportDraft.pdf (access March 18, 2009)

“Water quality data for southern tributaries to Cayuga Lake (Tompkins County, NY): 1987-1989,”
http://hdl.handle.net/1813/9336 (accessed April 15, 2009)

“Results files for organic solid-state PM6 calculations from Nick Day's PhD thesis, “
http://www.dspace.cam.ac.uk/handle/1810/197582 (accessed April 15, 2009)

(accessed April 15, 2009)

“England Wills Project,” http://www.dspace.cam.ac.uk/handle/1810/131687 and
http://www.dspace.cam.ac.uk/handle/1810/131688 (accessed April 15, 2009)

“1SNG - Structure of a Thermophilic Serpin in the Native State,”

“Stewardship of digital research data”