Research Data Management Seminar

~ Data Visualization and Data Management ~
Prof. Huamin QU (Director, Sino Software Research Institute, Department of CSE, HKUST)

~ Big Data: Recent Development, Benefit and Usage ~
Prof. Raymond C W WONG (Director, Computer Engineering Program, Department of CSE, HKUST)

URL: http://lbcone.ust.hk/sc-sem
Agenda

- Research Data Management (RDM), why matters?
- Data visualization and data mining
  
  Prof. Huamin QU (Director, Sino Software Research Institute, Department of CSE, HKUST)
- Big data: recent development, benefit and usage
  
  Prof. Raymond C W WONG (Director, Computer Engineering Program, Department of CSE, HKUST)
- Introduction and resources on data management
  
  Data Management Plan (DMP) tools, data repositories, data publishing, data sharing, data citation, funder requirements, and best practices
- Demonstration of DataSpace@HKUST (data repository and workspace service for HKUST research community)
- Q&A
RDM, why matters?

- **Research Data Management (RDM)** is part of the research process, and aims to make the research process as efficient as possible, and meet expectations and requirements of the university, research funders, and legislation.

  What is Research Data Management — University of Leicester
  http://www2.le.ac.uk/services/research-data/rdm/what-is-rdm

- Good RDM includes the following aspects:
  - Data are stored and backed up
  - Data files are preserved and documented
  - Data are discoverable and citable
  - Date are in compliance with journal data policies
Sharing session

Data Visualization and Data Management

Prof. Huamin QU
Director, Sino Software Research Institute, Department of CSE, HKUST
 Sharing session

Big Data: Recent Development, Benefit and Usage

Prof. Raymond C W WONG
Director, Computer Engineering Program, Department of CSE, HKUST
Introduction and resources on data management

Jacky Leung
Reference Librarian, HKUST Library
Funder Requirements

Many funding agencies have guidelines and policies for research data management on proposal submission. Proposals submitted to NSF must include a supplementary document of no more than two pages labeled "Data Management Plan" (DMP). This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. Proposals that do not include a DMP will not be able to be submitted.

“One of these principles states that it is “appropriate to use public funds to support the management and sharing of publicly-funded research data”. - RCUK
Data Management Plan (DMP)

- DMP is a document that describes what data will be collected or generated and how they will be handled during or after the project.

- Some common components in a DMP include:
  - Types of data produced or collected
  - Access, sharing and re-use policies
  - Confidentiality and intellectual property

- Some software tools can help on compiling DMP
  - DMPTool (by University of California)
  - DMPonline (by UK Digital Curation Centre)
Many journal publishers now require article authors to provide well-documented data used in the research and make these data easily available to other researchers.

Examples of major publishers and their policies:

- Elsevier - Research Data
  - a platform to support researchers to store, share, discover and use data

- Nature - Availability of data and material and methods
  - ... authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications

- Taylor & Francis - Enhancing your article with supplemental material
  - a number of initiatives to ensure supplemental material is effectively included, or linked to, within the article abstract.
Data Repositories

**Specialist repositories**

GenBank, NeuroMorpho.org, HepData, EarthChem, PubChem, ICPSR (openICPSR), Harvard Dataverse, Protein Databank, etc.

**Generalist repositories**

Dryad, figshare

**Institutional (project-based) repositories**

DataSpace@HKUST

More examples: [http://www.nature.com/sdata/data-policies/repositories](http://www.nature.com/sdata/data-policies/repositories)
Data Citation

- Citing data brings the **same benefits** as citing articles and books because it
  - helps data creators and users **find, access, verify and reuse** data
  - allows **data impact** to be tracked
  - gives proper **recognition** to data producers

- **Examples:**

  Creator (Publication Year): Title. Publisher. Identifier

  Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo.  
  [http://dx.doi.org/10.1594/PANGAEA.726855](http://dx.doi.org/10.1594/PANGAEA.726855)

  Creator, Publication Year, "Title", Identifier, Publisher, Version, [Universal Numerical Fingerprint]

  Campbell, Cameron Dougall; Lee, James; Dong, Hao, 2015, "Longitudinal Links to Construct the Korean Multi-Generational Panel Dataset – Tansung (KMGPD-TS) from the Tansung Household Registers", [http://dx.doi.org/10.14711/dataset/IVIDZV](http://dx.doi.org/10.14711/dataset/IVIDZV), DataSpace@HKUST, V1 [UNF:6:dSzr9Zp14Qhmtl7RtXBRg==]
Data Discovery and sharing

- Adding DOIs to finalized data sets
  - Makes data publications more acceptable for CVs
  - Creates more recognition for data sharing
  - Provides a means of tracking publications

- Creating your own ORCID iD for publishing
  - ORCID iD is a unique researcher identifier for distinguishing your data/article publications from those of others with similar names
  - You can establish clear ownership of your scholarly output
  - 492 HKUST staff and 660 students have already registered their ORCID iDs
  - You can use your HKUST login to easily create yours in a few clicks via http://repository.ust.hk/orcid/
Information on data management

- DMP
- Data repositories
- Data Citation
- Funder requirement & publisher policies

http://library.ust.hk/sc/
Research Data Management (RDM) Service Kit

- Data publishing
- DOI persistent link
- ORCID ID

- Data citation
- Metadata support
- Data repository

- Research coaching w/ subject librarians
- Create DMP

- Secure Workspace
- Easy manage / share Data

Source: https://www.jisc.ac.uk/guides/research-data-management
License: http://creativecommons.org/licenses/by-nc-nd/3.0/
Contact our Subject Librarians at

http://library.ust.hk/about-us/contact-us/subject-librarians/
Demonstration of DataSpace@HKUST

Jacky Leung
Reference Librarian, HKUST Library
Library Research Support Faculty Survey 2015

- Responses on data preservation and management
  - 65% of the respondents build up collections of research data in the course of their research.
  - 56% of the respondents build up collections of image or media
  - Most important data sources are those collected by the researchers themselves and those that are freely available online
  - Most important data management features are
    - Ability to track academic research citing my data
    - Ability to store multiple versions of my data in the same place
    - Ability to store my data with my corresponding published outputs
  - Most of the researchers manage their data on their own computers
  - More than one half of researchers manage their data on the cloud storage service (such as Google Drive, Dropbox, Flickr, etc.)
  - Majority perceived data support provided by
    - Freely available software
    - University Library
  - Majority preserve data by themselves, using commercially or freely available software or service

http://library.ust.hk/about-us/user-engagement/ithaka/
DataSpace – Research data portal

Dual Functions

Repository of published datasets

Workspace for data management
Street Level Air Quality Data Collected by the Mobile Real-time Air Monitoring Platform (MAP)

Fang, Ming; Yao, Xiaohong; Chan, Chak Keung; Lau, Ngai Ting; Lau, Arthur P. S., 2015, "Street Level Air Quality Data Collected by the Mobile Real-time Air Monitoring Platform (MAP)."
http://dx.doi.org/10.14711/dataset/I5VZXZ

Repository - for you to publish datasets
- Service for all HKUST researchers
- Openly accessible perpetually
- DOI assigned – persistent identifier facilitating discovery
- Standardized citation and metadata – created by the Library
- Trusted repository - with HKUST branding

http://dx.doi.org/10.14711/dataset/I5VZXZ
<table>
<thead>
<tr>
<th>Dataset Persistent ID</th>
<th>doi:10.14711/dataset/f5VZXZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication Date</td>
<td>2015-12-07</td>
</tr>
<tr>
<td>Title</td>
<td>Street Level Air Quality Data Collected by the Mobile Real-time Air Monitoring Platform (MAP)</td>
</tr>
<tr>
<td>Author</td>
<td>Fang, Ming (HKUST)</td>
</tr>
<tr>
<td></td>
<td>Yao, Xiaohong (HKUST)</td>
</tr>
<tr>
<td></td>
<td>Chan, Chak Keung (HKUST) - ORCID: 0000-0001-9687-8771</td>
</tr>
<tr>
<td></td>
<td>Lau, Ngai Ting (HKUST)</td>
</tr>
<tr>
<td></td>
<td>Lau, Arthur P. S. (HKUST)</td>
</tr>
<tr>
<td>Contact</td>
<td>Use email button above to contact.</td>
</tr>
<tr>
<td></td>
<td>Chan, Chak-Keung (HKUST)</td>
</tr>
<tr>
<td>Description</td>
<td>Data collected by the Mobile Real-time Air Monitoring platform (MAP) from 2003-2006. This project was done by the research team of the Institute for Environment and Sustainable Development and the Department of Chemical Engineering of the Hong Kong University of Science and Technology. See the readme file for information about the data. The data can be visualized by the GIS software developed by the data contributors. Their contacts are available in the readme file.</td>
</tr>
<tr>
<td>Subject</td>
<td>Earth and Environmental Sciences; Medicine, Health and Life Sciences</td>
</tr>
<tr>
<td>Keyword</td>
<td>Air quality</td>
</tr>
<tr>
<td></td>
<td>Hong Kong</td>
</tr>
<tr>
<td></td>
<td>Street level air quality</td>
</tr>
<tr>
<td></td>
<td>Pollution</td>
</tr>
<tr>
<td></td>
<td>Mobile Real-time Air Monitoring Platform</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Depositor</td>
<td>DataSpace Team</td>
</tr>
<tr>
<td>Deposit Date</td>
<td>2009-05-06</td>
</tr>
<tr>
<td>Kind of Data</td>
<td>Geographic information data; Meteorological data; Air quality data</td>
</tr>
</tbody>
</table>
**Workspace for data management**

- For faculty members and their research project groups (including their research postgraduate students)
- Store, manage and share data within research group in a **restricted** environment

1. **Create workspace**
2. **Add dataset**
3. **Upload files**
4. **Share internally**
5. **Publish dataset**

- Manage and share your data internally
  - **Data storage** throughout the research life cycle (50GB per dataset)
  - Share findings with collaborators
  - Student assignments
  - **Fulfill data management plan** specified in grant proposals
  - Replication of data readily available to go with the **manuscript submission**
Creating a dataverse for your research project data

Provide information about your research project, and assign roles to your team members: Contributors can add, update or delete datasets and change dataset access permission; Managers have the authority of contributors, plus the permission to edit your dataverse.

- **Project Information**
  - **Project Title**: China economic history research group
  - **Alias**: econ-han
    - The Alias will be part of the URL of the dataverse. Use all lower case with no space; e.g. alias “project5” will produce the URL "https://dataspaces.ust.hk/dataverse/project5"
  - **Project Description**: Datasets collected and generated by the CEHR Group

- **Project Team**
  - **Manager(s)**
    - **ITSC Account**
      - Lam
      - Email: lbkt@ust.hk
    - **Non-HKUST**
      - Department: LIB
      - Institution: HKUST
      - Delete/Clear

- **Contributor(s)**
  - **ITSC Account**
    - Cheung
    - Email: frankycheung@ust.hk
  - **Non-HKUST**
    - Department: LIB
    - Institution: HKUST
    - Delete/Clear

- **Get Started by creating your workspace (dataverse)**

- Specify the project and team members
- Managers – manage the workspace
- Contributors
  - Can upload and view datasets
  - Allow your non-HKUST collaborators to join
Managing dataset in workspace

1. Tabs - Files, Metadata, Terms of use, Versions
2. Manage permission – i.e. determine who can access this dataset
3. Citation – readily available [Note: no DOI, as it has not yet been published]
4. Upload, update, delete and set restrictions to files
5. Download files – restricted to members you specified in 2.
6. Request Library to publish the dataset
Data services provided by the **DataSpace Team**

- Help HKUST researchers to deposit (publish) datasets
  - Create standard metadata
  - Assign DOI
  - Format / normalize data files
  - Upload big data files
  - Make datasets accessible via ** SPD (Scholarly Publications Database)**
- Have questions or request a demonstration, contact the Team at **lbds@ust.hk**
Library Services Supporting Research Output

Electronic Theses
http://lbezone.ust.hk/rse/electronic-theses

DataSpace
http://dataspace.ust.hk

Scholarly Publications Database
http://repository.ust.hk/sp

Institutional Repository
http://repository.ust.hk/ir
Q & A