Data Discovery & Citation Workshop

April 9, 2019  10:30 -11:30am
April 10, 2019  12 noon – 1pm

Victoria Caplan  lbcaplan@ust.hk

https://libguides.ust.hk/data-curation/store-share-outside-hkust
Goals

• Will be able to search for & find data sets relevant to your research interests in order to be able to:
  • **Explore** and **Use** those data in your research and study

• Use the data ethically and properly by understanding and following “terms of use”

• Cite the data properly
Outline

1. Tips & methods for Data Discovery in your research area.

2. Tips & Methods for finding & understanding "terms of use" for data sets

3. Basics of citing data.
What do I mean by “Data”?

**data** noun singular or plural  
\ˈdā-tə, ˈdat-əl\ also ˈdāt-\-

1: factual information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation

2: information in numerical form for use especially in a computer
Exercise 1. Where does Data come from? (5 min)

- Who creates & who collects data & why?

• 1 minute think & write

• 1 minute share with partner

• 3 min class discussion
Lots of data collected across disciplines & across the world

• Some **scholars** will make data available to others
  • Love to share & want to further human knowledge
  • Funding Requirement

• Some **Government & Non-Organizations** will make data available
  • Love to share & want to further human knowledge
  • Legally required

• Some **Companies** compile & sell others’ data
  • It’s a profitable business
  • Examples: CDMNext, Euromonitor (HKUST pays for HKUST students & staff to use)
Private-Public Data Sets
Example: Twitter

Twitter = Private Social Media Company

There are 4 primary ways of acquiring Twitter data (and I'm not including “cutting and pasting” from the Twitter website):

- Retrieve from the Twitter public API.
- Find an existing Twitter dataset.
- Purchase from Twitter.
- Access or purchase from a Twitter service provider.

https://gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data

https://www.docnow.io/catalog/
Exercise 2 - How do I usually look for data?

2 min: by self
• I usually look for data by....

• My biggest success in searching for data has been...

• My biggest difficulty in searching for data is...

3 min class share:
Data Repositories & Discovery Tools

https://libguides.ust.hk/data-curation/store-share-outside-hkust

Data Repositories Outside HKUST

These repositories are places where you can put your own data, or search for and use other's data.

**Data Directories**

- **re3data** - A global registry of research data repositories in many disciplines.
- Directory of Open Access Repositories (OpenDOAR)
- Harvard Dataverse
- Stanford University Libraries > Domain-specific Data Repositories
- University of Minnesota > Discipline-Based Data Archives

**General Sciences**

- Dryad
- figshare - Accept a wide variety of file types, such as figures, datasets, images, and videos.

**Life Science**

- Genbank - The NIH genetic sequence database
- Protein Databank - A worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids

**Social Science**

- ICPSR (Inter-university Consortium for Political and Social Research) - Maintains and provides access to a vast research archive of more than 500,000 data files in social science

**Earth Science**

- DataONE | ONEMercury - A network of data centers, etc., focusing on Earth observational data.
- IRI/LDEO Climate Data Library - Climate-related datasets from different countries and regions.
Tip 1: Search Data Repositories and Archives

- Data repositories or archives offer a secure long-term home for research data
- Re3Data.org offers a searchable catalogue of repositories

http://service.re3data.org/
Exercise 4—Search R3DatadataOrg (a)

http://www.re3data.org/
Exercise 4—Search R3DataDataOrg (b)

Tip 2 – look for “open access”
Exercise 4—Search R3DatadataOrg (c)

Example: Nucleic Acid Database

<table>
<thead>
<tr>
<th>Nucleic Acid Database</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subject(s)</strong></td>
</tr>
<tr>
<td>Basic Biological and Medical Research</td>
</tr>
<tr>
<td>Physical Chemistry of Molecules, Interfaces and Liquids - Spectroscopy, Kinetics</td>
</tr>
<tr>
<td>Microbiology, Virology and Immunology</td>
</tr>
<tr>
<td>Biological Chemistry and Food Chemistry</td>
</tr>
<tr>
<td>Biology</td>
</tr>
<tr>
<td>Natural Sciences</td>
</tr>
<tr>
<td><strong>Content type(s)</strong></td>
</tr>
<tr>
<td>Standard office documents</td>
</tr>
<tr>
<td>Raw data</td>
</tr>
<tr>
<td><strong>Country</strong></td>
</tr>
<tr>
<td>United States</td>
</tr>
</tbody>
</table>

The NDB is a resource for nucleic acid research and education. The NDB assembles and distributes information about the three-dimensional structures of nucleic acids through a variety of resources, including a searchable database, Atlas, and software.
Exercise 4– Search R3DataDataOrg (d)

Tip 3: Look over record before visiting site

1. Mouse over the summary of use
2. Look at subject tags
   Can help you search pertinent data
3. Visit site
Exercise 4– Search R3DatadataOrg (e)

Tip 4 – Look at “about” & explore site (tabs or sitemap)

http://ndbserver.rutgers.edu/ndbmodule/about_ndb/about.html
Exercise 4—Search R3DatadataOrg (f)

Exploring can help – this site also has software & tools
Exercise 3: Search R3Data

Tip 5 – Look for words like “Download”

Tip 6 - Pay attention to format! No sense in downloading if you can’t use it.
Exercise  Search R3DatadataOrg (h)

Tip 7 – From beginning, think of how to cite it

DataCite’s Recommended Format:
Creator (PublicationYear): Title. Publisher. Identifier

Example of Victoria’s try:
(2019?). NA-NMR: Copies of the Coordinate Files Released by the PDB for Nucleic Acid-containing Structures Determined by NMR. The Nucleic Acid Database Project, Rutgers, The State University of New Jersey.
http://ndbserver.rutgers.edu/files/ftp/NDB/coordinates/na-nmr/

Not sure?  Ask!   Look for email address in “about” or elsewhere
ndbadmin@ndbserver.rutgers.edu
Tip 7 – From beginning, think of how to cite it

DataCite’s Recommended Format:
Creator (PublicationYear): Title. Publisher. Identifier

Example of Victoria’s try:
(2019?). NA-NMR: Copies of the Coordinate Files Released by the PDB for Nucleic Acid-containing Structures Determined by NMR. The Nucleic Acid Database Project, Rutgers, The State University of New Jersey.
http://ndbserver.rutgers.edu/files/ftp/NDB/coordinates/na-nmr/

Not sure? Ask! Look for email address in “about” or elsewhere
ndbadmin@ndbserver.rutgers.edu
Citation = an **ART**, not a science

Purposes:
- Another person can find what you used (use it themselves)
- Giving credit (manners, politeness)
- Using authority
- Often required in terms of use

But – each data set is different and can be tricky

Do Your BEST, give it a good try
Exercise 4.  **Search R3DatadataOrg (i)**

- Search R3.data.org yourself. Take notes.

(3 min)
Tip 8 – Discover Data from Good lists (a)

University of Minnesota: https://www.lib.umn.edu/datamanagement/datacenters
Tip 8 – Discover Data from Good lists (b)

Tip 9  Smart Googling

About 191,000 results (0.44 seconds)

Major River Basins of the World – GIS Layer - Datasets - European ... https://www.europeandataportal.eu/data/.../fb19590d-a17e-49f3-8e29-d5d1eb43b7 • Jul 31, 2018 - The "Major River Basins of the World" is an ongoing GIS project of the Global Runoff Data Centre (GRDC) that aims at the provision of a set of ...

OSNI Open Data: River Basin LIDAR 2007 - DTMs and DSMs - data ... https://data.gov.uk/dataset/.../osni-open-data-river-basin-lidar-2007-dtms-and-dsms • Apr 14, 2016 - Terrain (DTM) & Surface (DSM) elevation models of river basins derived from airborne LIDAR survey systems. A Digital Terrain Model (DTM) is ...


About 201,000 results (0.50 seconds)

Data repositories - Open Access Directory oad.sims mons.edu/oadwiki/Data_repositories • Aug 29, 2018 - TROLLing is designed as an archive of linguistic data and statistical ... provides access to data sets contributed by investigators funded by the ... Biology. Environmental sciences. Geosciences and ... Linguistics

Phonological Data and Analysis | Linguistic Society of America https://www.linguisticsociety.org/content/phonological-data-and-analysis • As an online open access journal published by the LSA, Phonological Data and ... that use rich data sets to document sound patterns in individual languages; ...

Anthropology; Open Access Repositories for Data & Research https://www.lib.umn.edu/libdate/page_print.plhtml?page_id=3955 • Directory of Open Access Journals covers free, full text, quality controlled ... Users can search tDAR for digital documents, data sets, images, GIS files, .... psychologists,
Tip 9  Smart Googling  (data producers or collectors)
Exercise 4 - Own search  (5 min)
Data Repositories & Discovery Tools & Smart Google

1. Search on your own for a set that is open or terms of use that let you use it
2. Find a data repository you’re interested in
3. Note down:
   - Creator (person, lab, research group?)
   - Date
   - Title of data or data set
   - Publisher (where it comes from?)
   - Identifier or URL or FTP or other access info

https://libguides.ust.hk/data-curation/store-share-outside-hkust
# Data Citation - Basics

[https://libguides.ust.hk/data-curation/cite](https://libguides.ust.hk/data-curation/cite)

## How to Cite Data

The basic format of data citation should include:

**Creator (Publication Year). Title. Version. Publisher.**  
**Resource Type (as appropriate). Identifier**

- Version & Resource Type are optional  
- Identifier could be a DOI or a persistent URL

### Examples (via DataCite)


## Cite Data from DataSpace @ HKUST

### Citing DataSpace@HKUST

For datasets published in DataSpace@HKUST, the system provides citation in this format:

**Creator(s), Publication Year, "Dataset Title", Identifier, Data Repository, Version, [Universal Numerical Fingerprint]**

**Example:**

Hong Kong University of Science and Technology Library, 2015, "Wi-Fi Connection in HKUST Library", [https://dx.doi.org/10.14711/dataset/RQXRQ8, DataSpace@HKUST, V1 [UNF:6:vtJGn8nA+sEACfFisPHbA==]]

Users can also download citation to EndNote XML or RIS Format for ease of management.

## Guides to Data Citation

- Dataverse Project – Data Citation  
- Digital Curation Centre – How to Cite Datasets and Link to Publications  
- Australian National Data Service – Data Citation Awareness  
- DataONE – Tutorial on Data Citation  
- MSU Libraries – How to Cite Data Style Manual Examples
Data Citation:
Some Data Repositories will tell you (a)

Data Citation: Some Data Repositories will tell you (b)

[Image of ICPSR database]

Citation

The established bibliographic reference for the data collection. The citation is dynamically generated from other metadata fields and follows this format: Principal Investigator(s), Title, Place-of-Distribution and Distributor, Date-of-Distribution, DOI.


Export Citation:
- RIS (generic format for RefWorks, EndNote, etc.)
- EndNote

https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/36755
Data Citation
Referencing Software – Use “Generic” entry

- RefWorks & Mendeley – no specific format for data
  - Creator (named person or lab, if possible)
  - Date published (publication year)

Data Cite’s Recommended Format:

Creator (Publication Year): Title. Publisher. Identifier
Exercise 6 - Make a citation for the data you found
(3 min)

Make your data Citation using this format:

Creator (Publication Year): Title. Publisher. Identifier