Disciplinary data issues
Nanyang Technological University

Jonathan Rans
Digital Curation Centre

This work is licensed under the Creative Commons Attribution 2.5 UK: Scotland License.
What can be provided centrally?
What needs to be provided at a faculty level?
What disciplinary infrastructure already exists?
What are the key preoccupations of researchers in these areas?
Issues with hard science data

- Data volumes
- Non-standard
  - Descriptions
  - Formats
- Software preservation
Reliance on specialist research software

Slide from Neil Chue-Hong, Software Sustainability Institute

Do you use research software?

- Yes: 92%
- No: 8%

What would happen to your research without software?

- Would be impossible: 68%
- Possible, but difficult: 21%
- No effect: 11%
- Develop their own software: 56%
- Have no formal software training: 71%

Survey of researchers from 15 UK Russell Group universities conducted by SSI between August - October 2014. DOI: 10.5281/zenodo.14809
CERN data

- Non-standardised, ad-hoc formats
- Highly complex models
- Particle collision data often unique
- High volume
  » 30PB/year

http://www.stfc.ac.uk/stfc/cache/file/1C507A82-3D6B-4C28-95C66FC261E6F936.jpg
Large survey data at CASU

Prof. Janet Drew, Director of the Centre for Astrophysics, U. of Hertfordshire

- Large image files
- Catalogues of measurements using FITS standard
- Large files transferred using DVD
- Small images shared via European Southern Observatory archive
- Photometric catalogues publicly available

www.herts.ac.uk/rdm/training/case-studies/rdm-in-astrophysics
Issues in Life Sciences

- Increasing data volumes
- Increasing number of data producers
- Secure access and governance of human data
  - Data safe havens
  - Governance support – UCL IGT support
- Interoperability
  - Standards
  - Contextual data – Electronic lab notebooks
Creating a taxonomy of interactions

- Text from Encyclopedia of Life (EOL) entries accessed using APIs
- Mined for names using GNRD tool
- Names resolved to EOL
- Interaction taxonomy constructed

Thessen, A. and Parr, C.
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089550

Formatting + punctuation were responsible for most errors
Social sciences and humanities data

http://archaeologydataservice.ac.uk/archives/view/southamp2-SOU1615

http://firstfolio.bodleian.ox.ac.uk/download/xml/F-tgv.xml
Disciplinary issues

- Consent and ethics
- Controlling access to data
- Rights issues when reusing third party data
- Production of online resources
- Notion of data as a life’s work in humanities
- Science-focused RDM language
Degrees of access

Open data
- Data under an ‘open licence’ which are not personal and have relatively few restrictions to use. Neither login nor registration are required for these data collections.

Safeguarded data
- Data where the owner considers there to be a risk of disclosure resulting from linkage to other data. These are provided under the UK Data Service’s End User Licence. The safeguards include knowing who is using the data and for what purpose. Conditions may be attached e.g. limiting to non-commercial or academic users.

Controlled data
- Data which may be identifiable and thus potentially disclosive. These data are only available to users who have been accredited and their data usage has been approved by the relevant Data Access Committee. Access will be through a physical or virtual secure environment.

Three tiers of access as implemented by the UK Data Service
Social science repositories

- Long history of data archiving in the Social Sciences

- CESSDA is the consortium of European Social Science Data Archives and points researchers to national data services.
A longitudinal public health study

- Twenty-07 followed 4510 people from West of Scotland over 20 years to investigate the reasons for differences in health.
- Undertook interviews, questionnaires, physical measurements, blood samples etc so wide range of data types.
- Strict access controls and guidelines for data collection.
- Data managed within the MRC Social and Public Health Sciences Unit and accessible under a data sharing agreement - http://2007study.sphsu.mrc.ac.uk/information-on-data-sharing.html.
Engineering and computing data

http://schools-training.com/articles/neutral-formats-computer-assisted-design.html

http://anabim.com

http://www.myexperiment.org/workflows/192.html

http://schools-training.com/articles/neutral-formats-computer-assisted-design.html
Engineering issues

- Challenges preserving CAD files
- Intersections with other communities e.g. bioinformatics, computational physics, data science
- Less information sharing in engineering
  - Commercial sensitivities preclude sharing
  - Consultancy-style research can lead to internal-only results
  - Data generated from private systems, so less applicable to others
- Lack of specific repositories means a reliance on general or institutional offerings
Engineering data sharing – Sphere project

- University of Bristol EPSRC project
- Monitoring food preparation actions via:
  - Headcam footage
  - Environmental sensors
  - Camera
- Ontology developed for actions
- Model developed and tested against simulated data

https://data.blogs.ilrt.org/sharingdata/
Not everything is available – partial publication

Data underlying the paper
   » Annotated head camera footage
   » Action ontologies
   » Simulated data

Data access statement and DOI included in paper

Participant consent is for research purposes only
   » 3rd party users must be affiliated to an HE institution
   » Must sign a data access agreement
# SPHERE Unscripted kitchen activities

<table>
<thead>
<tr>
<th><strong>Creator(s)</strong></th>
<th>Majid Mirmehdi, Thomas Kirste, Samuel Whitehouse, Adeline Paiement, Kristina Yordanova</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Publication date</strong></td>
<td>22 Sep 2016</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>eng</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>University of Bristol</td>
</tr>
<tr>
<td><strong>Licence</strong></td>
<td>Non-Commercial Government License for public sector information</td>
</tr>
<tr>
<td><strong>DOI</strong></td>
<td>10.5523/bris.raqa2qzai45z15b4n0za94toi</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Majid Mirmehdi, Thomas Kirste, Samuel Whitehouse, Adeline Paiement, Kristina Yordanova (2016): SPHERE Unscripted kitchen activities. <a href="http://dx.doi.org/10.5523/bris.raqa2qzai45z15b4n0za94toi">http://dx.doi.org/10.5523/bris.raqa2qzai45z15b4n0za94toi</a></td>
</tr>
<tr>
<td><strong>Total size</strong></td>
<td>7.8 GB</td>
</tr>
</tbody>
</table>

Environmental sensor data of several people performing unscripted cooking activities in the Sphere kitchen.

Access to this dataset is restricted and is available to bona fide researchers only. Please email data-bris@bristol.ac.uk to apply for access.

**Access to this dataset is not open**

In order to access this data please contact the University of Bristol Research Data Service: contact page.

[https://data.bris.ac.uk/data/dataset/raqa2qzai45z15b4n0za94toi](https://data.bris.ac.uk/data/dataset/raqa2qzai45z15b4n0za94toi)
What purposes do different types of metadata serve?

» Citation

» Discovery

» Understanding

What can/should be mandated?
Any questions?

Jonathan Rans
J.Rans@ed.ac.uk
@JNRans

Image Credits
Wetlab: http://www.sbes.vt.edu/behkam/img/LabPan.png
Research Dictionary: http://www.saquiresearch.com
Himalayas: www.cntraveller.com