Why Data Management?

Wolfgang Müller
(with plenty of Slides by Carole Goble and others from the FAIRDOM project)
You know that one?

Measurement3Algorithm4
D20_final.csv
You know that one?

3µC4
Yes, it’s true:

You will remember the meaning in the next coupla months
Big question:

Will you remember
5 years from now?
Purpose of project data management

- Organisation
- Communication
- Dissemination

of data associated with a project

- Helps you find your way
- Reuse later
- Enable team to reuse
- Reuse with new partners
- Helps others find out
- Tell more and take credit

Organisation
Drive for open, reproducible, long-term maintained "FAIR" data
H2020 Programme
Guidelines on FAIR Data Management in Horizon 2020

Mandatory Data Management Plans

Realising the European Open Science Cloud
First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud

Data stewardship resourced. Data stewardship capacity and capability
Findable (Citable)  
Accessible (Trackable)  
Interoperable (…for machines)  
Reusable (Reproducible)  

Record All  
Automate All  
Contain All  
Access All
• Findable
• Accessible
• Interoperable
• Reusable

• Data
• Operations
• Models
Catalogue of distributed resources

- Standards
- Personal Data
- Local Stores
- Models
- SOPs
- Articles
- External Public data
- External Patient-related data + samples
Investigation
Study
Analysis
Data
Model
SOP
(Assay)

Biosamples system, very flexible, link samples to data

https://fairdomhub.org/investigations/56
Determination of (cytosolic) enzyme activities

Preparation of cell free extracts for enzyme determinations

Freeze buffer (1 liter)
- Make 1 liter of 10 mM K$_2$HPO$_4$
- Make 500 ml of 10 mM KH$_2$PO$_4$
- Take about 800 ml of 10 mM K$_2$PO$_4$
- Adjust pH to 7.5 by adding 10 mM KH$_2$PO$_4$
- Add 0.75 g EDTA to 1 liter
Community Standards

http://biosharing.org/

Source: Susanna-Assunta Sansone (University of Oxford, UK)
Linking content to catalogue using the tools of the researcher...

Excel spreadsheets enriched with ontology annotations

Embed guidelines and ontologies into Excel templates

Upload, extract metadata and catalogue

http://www.rightfield.org.uk
JERM templates and domain ontologies

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asset Title</td>
<td>The name of the data file</td>
</tr>
<tr>
<td>Uploader</td>
<td>The person submitting the asset to SEEK</td>
</tr>
<tr>
<td>Project SEEK ID</td>
<td>If you add your own SEEK ID, this will help us link this asset with your profile</td>
</tr>
<tr>
<td>Project</td>
<td>The project that the asset belongs to</td>
</tr>
<tr>
<td>ASSAY</td>
<td>If referring to an existing Assay, you can link to it via the Assay SEEK ID.</td>
</tr>
<tr>
<td>Assay Title</td>
<td>The title of the Assay</td>
</tr>
<tr>
<td>Assay SEEK ID</td>
<td>The SEEK ID of the Assay</td>
</tr>
<tr>
<td>Assay Type</td>
<td>The assay type, e.g., proteomics</td>
</tr>
<tr>
<td>Technology_type</td>
<td>The technology type, e.g., mass_Spectrometry</td>
</tr>
<tr>
<td>Description</td>
<td>The description of the experiment</td>
</tr>
<tr>
<td>Experimentalist</td>
<td>The name of the people who carried out the experiments. Those can either be SEEK numbers or external scientists</td>
</tr>
<tr>
<td>Date</td>
<td>The date of the experiment</td>
</tr>
<tr>
<td>SOP (protocol)</td>
<td>The steps of the experiment</td>
</tr>
<tr>
<td>SOP Type</td>
<td>The type of SOP used</td>
</tr>
</tbody>
</table>

**Human Disease Ontology**

- Summary
- Classes
- Properties
- Notes
- Mappings
- Widgets

- breast cancer
- breast carcinoma
- breast granular cell tumor
- breast large cell neuroendocrine carcinoma
- breast lymphoma
- breast malignant eccrine spiradenoma
- breast malignant phyllodes tumor
- breast myoepithelial neoplasm
- breast sarcoma
- estrogen-receptor negative breast cancer
- estrogen-receptor positive breast cancer
- female breast cancer
- Her2-receptor negative breast cancer
## Excel Gymnastics

### Table 1: Experiment Details

<table>
<thead>
<tr>
<th>Experiment No</th>
<th>Cells</th>
<th>Starvation</th>
<th>Stimulus</th>
<th>Treatment (nM)</th>
<th>Time Points (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>Hepatic WT</td>
<td>4h</td>
<td>TGFβ</td>
<td>1</td>
<td>0, 60, 120, 240, 600</td>
</tr>
<tr>
<td></td>
<td>Hepatic S7-Flag</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Protein Detection

<table>
<thead>
<tr>
<th>Protein Detection</th>
<th>pSma3</th>
<th>pSma2</th>
<th>Snor</th>
<th>Sma4</th>
<th>totSma2</th>
<th>totSma3</th>
<th>Smad7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Sample Loading Stage

<table>
<thead>
<tr>
<th>Lane</th>
<th>Cells (time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hepatic (6h)</td>
</tr>
<tr>
<td>2</td>
<td>Hepatic (1h)</td>
</tr>
<tr>
<td>3</td>
<td>Hepatic (2h)</td>
</tr>
<tr>
<td>4</td>
<td>Hepatic (4h)</td>
</tr>
<tr>
<td>5</td>
<td>Hepatic (6h)</td>
</tr>
<tr>
<td>6</td>
<td>Hepatic (1h)</td>
</tr>
<tr>
<td>7</td>
<td>Hepatic (2h)</td>
</tr>
<tr>
<td>8</td>
<td>Hepatic (4h)</td>
</tr>
</tbody>
</table>

### Quantification Stage

<table>
<thead>
<tr>
<th>Lane</th>
<th>Gene Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gene1</td>
</tr>
<tr>
<td>2</td>
<td>Gene2</td>
</tr>
<tr>
<td>3</td>
<td>Gene3</td>
</tr>
<tr>
<td>4</td>
<td>Gene4</td>
</tr>
<tr>
<td>5</td>
<td>Gene5</td>
</tr>
</tbody>
</table>

### Analysis Stage
Data integrating Jupyter in SEEK

- Use pre-stored notebook templates for processing data in SEEK
- Download notebook with data
- Extend locally...
Clinical Data Problem

- Shared
- Confidential
- Confidential
- Full data
Clinical Data Problem

Shared

Confidential

Confidential
Organisation – Collaboration - Dessimination

Self-managed spaces

Project area

Project partner area

Project partner area

Personal area

Personal area

umcg

LiSyM

dkfz.

GERMAN NETWORK FOR BIOMEDICAL INFRASTRUCTURE
We help with

Standards compliance
- Systems and Synthetic Biology standards support
- Support in finding standards & project wide standardisation

Consistent reporting
- Structured using ISA
- Our specially made Just Enough Results Model

Metadata curation
- Spreadsheet templates for omics data and samples
- Data and model annotation tools

Integrate with existing systems
- Integrated tools for modelling, parts, ELNs, LIMS
- Custom plugins for your tools
- REST API to plugin to your systems

Export
- Package and export into other repositories
- Export into other FAIRDOM installations
- COMBINE Archive export
Data management planning

What data?

How much data?

Shared with whom?

Shared when?