How can we combat heterogeneous, unFAIR and disparate data in digital chemistry?

ChemSpider Webinar 3: Challenges & Opportunities
7th December 2023
Dr Samantha Pearman-Kanza
University of Southampton
About Me & PSDI

- Senior Enterprise Fellow at University of Southampton
- Pathfinder Lead & Researcher for PSDI Project: Process Recording
- Research Interests: Semantic Web Technologies, IoT, Research Data Management, Digitisation, Lab of the Future, Paperless Labs, Re-use of Technology
- @SamiKanza

Physical Sciences Data Infrastructure

An Integrated Data Infrastructure for the Physical Sciences

PSDI aims to accelerate research in the physical sciences by providing a data infrastructure that brings together and builds upon the various data systems researchers currently use.
How can we combat heterogeneous, unFAIR and disparate data in Chemistry?

- Understand the environment and the challenges
  - Barriers & Challenges to Digitisation
- Process Recording
  - Digitisation Requirements
  - Choosing your tools for process recording
- Producing FAIR Data AND Research AND Code
  - Considering all aspects of FAIR and going beyond the guidelines
  - Establish common vocabularies and practices (data and metadata)
Barriers & Challenges to Digital Research

- Logistical Barriers
  - Cost
  - Time

- People Barriers
  - Attitude & Adoption Factors
  - Training

- Data Barriers
  - Un-FAIR Data
  - Metadata/Provenance
  - Size of data

- Standards Barriers
  - Too Many Standards
  - Proprietary formats

- Software Barriers
  - Oversaturated Market for ELNs, Notebooks & Domain Based Software
  - Software Integration/Compatibility
  - Trust in Software

- Hardware Barriers
  - Data Storage
  - Legacy Equipment
What do Users want from ELNs?

<table>
<thead>
<tr>
<th>Notebooking Features</th>
<th>Domain Specific Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Alternative input methods (voice/handwriting/text recognition)</td>
<td>• Integration with Chemical Equipment</td>
</tr>
<tr>
<td>• Searching/Tagging/Indexing</td>
<td>• Integration with Chemical Data</td>
</tr>
<tr>
<td>• Colour Coding/ Personalisation</td>
<td>• Attach and view characterization data in ELN directly</td>
</tr>
<tr>
<td>• Links with reference management software</td>
<td>• Setup for multiple domains</td>
</tr>
<tr>
<td>• Collaboration features</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Features</th>
<th>Technical/Logistical Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data Management features</td>
<td>• Integration with Hybrid Devices</td>
</tr>
<tr>
<td>• Version Control</td>
<td>• API Access</td>
</tr>
<tr>
<td>• Linking between records</td>
<td>• More Storage</td>
</tr>
<tr>
<td>• Archiving old data</td>
<td>• Open Source / Development Capabilities</td>
</tr>
<tr>
<td>• Store structured data</td>
<td>• Cost</td>
</tr>
<tr>
<td>• Flexible data export/data portability</td>
<td></td>
</tr>
<tr>
<td>Notebooking Features</td>
<td>Domain Specific Features</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>• Alternative input methods (voice/handwriting/text recognition)</td>
<td>• Interface with Chemical Structure Editor/have features inbuilt</td>
</tr>
<tr>
<td>• Create/Use Templates</td>
<td>• Pasting Chemdraw Structures</td>
</tr>
<tr>
<td>• Add schemas/diagrams/images</td>
<td>• Integrate with ELN</td>
</tr>
<tr>
<td>• Searching/Tagging/Indexing</td>
<td></td>
</tr>
<tr>
<td>• Collaboration features</td>
<td></td>
</tr>
<tr>
<td>• “Be just like paper”</td>
<td></td>
</tr>
<tr>
<td>• Integrate with Project Management Software (ToDo lists/Gantt Charts)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Features</th>
<th>Technical/Logistical Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Linking between records</td>
<td>• Mobile Support</td>
</tr>
<tr>
<td>• Flexible data export/data portability</td>
<td>• Interoperability between devices</td>
</tr>
<tr>
<td>• Excel features to work with data/plot graphs</td>
<td>• Speed</td>
</tr>
<tr>
<td>• Link to external data sources</td>
<td>• Cost</td>
</tr>
</tbody>
</table>
## What do Users Want from a Digital Research Environment?

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td><strong>API Access</strong>, Automation, GUI, Localisation, Remote Access, Synchronisation</td>
</tr>
<tr>
<td>Notebooking</td>
<td><strong>Content Support</strong>, Interaction/Access, File Links, Organisation/Reconfiguration, <strong>Paper Integration</strong>, <strong>Referencing/Literature</strong>, Word Processing</td>
</tr>
<tr>
<td>Data</td>
<td>Access, Conversion, Exchange, <strong>Integration</strong>, <strong>Management</strong>, Quality, Retention, Security, <strong>Standards</strong>, Support, FAIR, Identifiers, Provenance</td>
</tr>
<tr>
<td>Publishing &amp; Sharing</td>
<td>Documentation &amp; Instructions, DOIs, Export, Licensing, Open Access, Publishing, Sharing, Social Media, Researcher Attribution, Repositories</td>
</tr>
<tr>
<td>Collaboration &amp; Management</td>
<td><strong>Auditing</strong>, <strong>Comments</strong>, <strong>Notifications</strong>, <strong>Subscribe</strong>, <strong>Team Management</strong></td>
</tr>
<tr>
<td>Domain Based Features</td>
<td><strong>Chemical/Molecules</strong>, Default Lists, <strong>Equipment Interface</strong>, Experiment Planning/Recording, Health &amp; Safety, <strong>LIMS/ELN</strong>, Link to Domain based databases &amp; software</td>
</tr>
<tr>
<td>Coding Support</td>
<td>Coding, <strong>Versioning</strong></td>
</tr>
<tr>
<td>Metadata, Semantics &amp; AI</td>
<td>AI Tools/Integration, Metadata, Semantics</td>
</tr>
<tr>
<td>Searching</td>
<td><strong>Search By</strong>: Domain, Characteristics Search, Keyword/Concept via Content Types, Literature &amp; Notebook, Indexing</td>
</tr>
<tr>
<td>Customisation &amp; Extension</td>
<td>Personalisable, <strong>Templates</strong></td>
</tr>
<tr>
<td>Training &amp; User Support</td>
<td>Training, User Documentation</td>
</tr>
</tbody>
</table>

---

Choosing tools and methods for Process Recording?

- What data are you recording?
- How are you recording it?
- Where are you recording it?
- What data is not being recorded?
- What are the pain points?
- What is the actual problem you are trying to fix?
The ELN Finder helps you to search and select a suitable Electronic Lab Notebook (ELN) for your purposes.

- More than 40 filter criteria available.
- Filter criteria clearly divided into categories.
- Result list of the identified ELN tools displayed in an overview.
- Brief descriptions of the individual tools included.

- Detailed hierarchical criteria catalogue created, defines and describes the metadata structure for the ELNs (Excel):
  - > 40 criteria and associated values, attributes (e.g. name/URL).
  - Summary of criteria in categories
  - Fully functional first version developed on the basis of the open source software DSpace 7:
  - External ELN information collection created for individual ELNs
  - Entering data from the information collection
  - 35 ELNs entered

- APIs
- Automation
- Collaboration
- Compliance
- Controlled vocabulary
- Customizable user interface
- Data access
- Data export
- Data import (formats)
- Data import (method)
- Data input
- Data storage location
- Device connection
- Laboratory management functions
- Languages Support

- License
- Location of provider
- Offline functionalities
- Operating system
- Plug-Ins
- Preservation of evidence
- Pricing
- Project management tools
- Search functions
- Standard interfaces
- Subject
- Templates
- Usage option
- Usage statistics
- Versions
- Workflows
Let's talk about FAIR

From ‘The FAIR Guiding Principles for scientific data management and stewardship’¹

- **F** – Findable
- **A** – Accessible
- **I** – Interoperable
- **R** – Reusable

F is for Findable

To be Findable:
- It needs to exist
- But existing ≠ findable
- Provide your users with pointers!

Are all your code/data/lab book/notes actually there?
A is for Accessible

- What should and shouldn’t be accessible?
- What is the use case?
- If access is restricted or complex, have you provided relevant information?

Technically accessible != Easily accessible
I is for Interoperable

- Consider your data standards
- Use Common and Shared Vocabularies
  - For Data and Metadata
- Use Ontologies/Knowledge Graphs to the best of their potential

Even standards need standards

https://www.pinterest.co.uk/jaci_mize/metadata/
This isn’t JUST about the data
You need to consider:
  - Data, Tools, Code, Methods, Context
  - How could/would your work be re-used, replicated, reproduced or repurposed
    - Re-use – re-use the data (or run the software) in the same manner
    - Replicate – repeat entire research from scratch including data collection and analysis
    - Reproduce – reanalyse the existing data in the same manner
    - Repurpose – use existing data or software for a new purpose

https://www.cartoonstock.com/directory/s/scientific_method.asp

This is only the tip of the “R” Iceberg
FAIR Details

Data
- Do your data file names make sense?
- Do your data headings make sense?
- Are your files understandable?

Code
- Do your code files make sense?
- Is your code all there?
- Is it commented?

Lab Books
- Does your lab book fully detail your reagents, samples, experiment parameters?

CC BY-ND 4.0 Errant Science - https://errantscience.com/
FAIR Pre-requisites

- Performing any of our 'R' operations on data of software is complex
- Data
  - Is this stored on outdated media?
  - What tools/software/dependencies do we need to use the data
- Databases:
  - How do we use these? Are there database dumps? Schemas? Instructions?
- Software:
  - What coding libraries are required?
  - Are there dependencies?
  - What installations and drivers are required?
  - Is all the underlying data included and accessible
- Lab Books
  - What were the experimental conditions?
  - What was the experimental setup?
  - What context exists for the experiment that you haven’t recorded
Be clear
- Do not assume prior knowledge
- Include all steps from start to finish (which means documenting as you go along)
- How was the data collected?
- What scripts/parameters were used?
- How did you get your database to interface with your code?
- How do you access the data?
- How do you run the software locally?
- If someone had your lab book and all your data could they re-run your experiment?
- Could someone else really re-use, reproduce, replicate or repurpose this?
Conclusions

- There are still many barriers to overcome
- But the community is working towards solutions
- We need to remember the following:
  - Ask the right questions, about your data, your tools, your situation
  - FAIR is a FOUR letter word, but it has many many nuances
  - Collaboration is key - This is as much a human endeavor as a software/data one
  - We must all strive to be better

To the well organised FAIR dataset, re-use, replication, reproduction and repurpose are but the next great adventure
Relevant Talks


Kanza, S., 2018. What influence would a cloud based semantic laboratory notebook have on the digitisation and management of scientific research? (Doctoral dissertation, University of Southampton). https://eprints.soton.ac.uk/421045/


Acknowledgements

PSDI Team: Simon Coles, Jeremy Frey, Nicola Knight, Cerys Willoughby, Colin Bird, Ray Whorley, Mark Anderson, Stephen Gow, Samuel Munday, Thomas Allam, Hannah Gittins (University of Southampton), Juan Bicarregui, Barbara Montanari, Brian Matthews, Vasily Bunakov (Science & Technologies Facilities Council)
PSDI & Personal Details - Questions

www.psdii.ac.uk

@PSDI_UK

@PSDI_UK

linkedin.com/company/psdiuk

Mailing List: https://www.jiscmail.ac.uk/PSDI