Bridging the Gap

Molecular Biology DMP catalog creation

Yasmin Demerdash & Sarah Wettermann
Who are we?

- Data Manager since March 2023
- PhD in Molecular Biosciences

- Data Manager since June 2023
- 3.5 years research experience in Polymer Physics
- MSc in Computer Science
Effective Data Management: Essential for Successful Research

- My data is disorganized and hard to interpret
- Lost crucial research due to lack of backup
- Funding was denied due to unclear data management

My data is disorganized and hard to interpret
The Need for Better Guidance

Data management as a fundamental role

- Equal importance to data generation and analysis.
- Embrace data management as a researcher.
- Acknowledge that data management is a fundamental part of your role.

Lack of support and training

- There is a prevailing assumption that researchers should navigate these areas independently.
- Researchers often lack adequate support and training in project and data management.
- Need for detailed and relevant guidance
Promoting RDM Practices and Collaboration in SFBs and IMB

Increase awareness about RDM in SFBs and IMB.

Make RDM practices a regular aspect of scientists' work.

Provide all scientists with comprehensive RDM workshops.

Address data-sharing challenges by implementing comprehensive strategies.

Carry out streamlined data management using DMPs.

Train scientists to create effective DMPs using RDMO tool.
Navigating Challenges and Enhancing DMP Support Through Collaboration

**C**arry out streamlined data management using DMPs.

**T**rain scientists to create effective DMPs using RDMO tool.

How?

- Part of the institutional and SFB policy to submit DMP
RDM Supports Researchers Throughout the Data Lifecycle

DMPs serve as the first step in the RDM life cycle.

Reuse
Plan
Collect
Share
Preserve
Process
Analyze
DMP Realities: Scientists' Expectations vs. Practical Insights

Expectations

• Data handling solutions in research projects
• Aiding researchers' daily tasks
DMP Realities: Scientists' Expectations vs. Practical Insights

**Expectations**
- Data handling solutions in research projects
- Aiding researchers' daily tasks

**Reality**
- Generic and lacks specific guidance
- Tedious and bureaucratic
- Requirement for grant approval

I don't know how to fill it out

GOOD THING I HAD A PLAN!
DMP Realities: Scientists' Expectations vs. Practical Insights

Expectations

• Data handling solutions in research projects
• Aiding researchers' daily tasks

Reality

• Generic and lacks specific guidance
• Tedious and bureaucratic
• Requirement for grant approval

Aim

• Discipline-appropriate guidance and improved answering options
• Assists researchers in their day-day work
• Greater satisfaction with DMPs

GOOD THING
I HAD A PLAN!

I don't know how to fill it out!

Great! Now I have helpful guidance!
RDM Working Groups with Tailored DMPs

Adapted DMPs of some existing initiatives in various disciplines, for example in Chemistry, Biodiversity and Engineering using RDMO.
DMP Tools: An Overview

Table 1: Evaluated DMP tools. Discipline-specific tools are marked in light green.

<table>
<thead>
<tr>
<th>DMP tool</th>
<th>Discipline</th>
<th>Hosting/Developers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARGOS</td>
<td>Interdisciplinary</td>
<td>OpenAIRE AMKE, EUDAT CDI, Europe</td>
</tr>
<tr>
<td>ARJADNE</td>
<td>Archeology</td>
<td>Vast-Lab, Italy</td>
</tr>
<tr>
<td>Clarin-d DMP</td>
<td>Humanities/social sciences</td>
<td>Eberhard Karls Universität Tübingen, Germany</td>
</tr>
<tr>
<td>Data Stewardship Wizard</td>
<td>Interdisciplinary</td>
<td>Czech Technical University, Dutch Techcentre for Life Sciences, Czech Republic, Netherlands</td>
</tr>
<tr>
<td>DataWiz</td>
<td>Psychology</td>
<td>Leibniz Institute for Psychology Information, Germany</td>
</tr>
<tr>
<td>DMP Canvas Generator</td>
<td>Life sciences</td>
<td>Swiss Institute of Bioinformatics, Switzerland</td>
</tr>
<tr>
<td>DMPlanline</td>
<td>Interdisciplinary</td>
<td>Digital Curation Centre, University of Edinburgh, United Kingdom</td>
</tr>
<tr>
<td>DMPTool</td>
<td>Interdisciplinary</td>
<td>California Digital Library, University of California, USA</td>
</tr>
<tr>
<td>easyDMP</td>
<td>Interdisciplinary</td>
<td>EUDAT, Finland, Norway</td>
</tr>
<tr>
<td>eoDMP</td>
<td>Interdisciplinary</td>
<td>Columbia University, Rutgers University, University of Illinois, USA</td>
</tr>
<tr>
<td>GFBio</td>
<td>Biodiversity</td>
<td>GFBio, Germany</td>
</tr>
<tr>
<td>NFID4Plants DataPLAN</td>
<td>Plant science</td>
<td>Eberhard Karls Universität Tübingen, Germany</td>
</tr>
<tr>
<td>QUT</td>
<td>Interdisciplinary</td>
<td>Queensland University of Technology, Australia</td>
</tr>
<tr>
<td>RDMO</td>
<td>Interdisciplinary</td>
<td>Leibniz Institute for Astrophysics Potsdam, University of Applied Sciences Potsdam, Germany</td>
</tr>
<tr>
<td>RDMD NFID4ng</td>
<td>Engineering</td>
<td>University and State Library Darmstadt, Germany</td>
</tr>
<tr>
<td>TUDD DMP</td>
<td>Interdisciplinary</td>
<td>TU Dresden, Germany</td>
</tr>
<tr>
<td>TUM Workbench</td>
<td>Interdisciplinary</td>
<td>TU München, Germany</td>
</tr>
<tr>
<td>UWA-DMP</td>
<td>Interdisciplinary</td>
<td>University of Western Australia, Australia</td>
</tr>
</tbody>
</table>
Challenges with Current Data Catalogs in Molecular Biology

Lacks informative guidance and overly generalized

- The content is too generic and lacks specificity
- Absence of helpful explanations and guidance
- Specific examples are needed for better understanding

Discipline vs. data type classification

- Current classification is based on research disciplines
- Our approach aims to classify based on data types
- Molecular biology involves numerous subdisciplines; projects often span multiple areas
Interdisciplinary research projects encompass various sub-disciplines with unique research focuses.
Different sub-disciplines have distinct data requirements

**Engineering Sciences**
- **NFDI4DataScience**: NFDI for Data Science and Artificial Intelligence
- **NFDI4Energy**: National Research Data Infrastructure for Interdisciplinary Energy System Research
- **NFDI4ing**: NFDI for Engineering Sciences
- **NFDI-MatWerk**: National Research Data Infrastructure for Materials Science and Materials Engineering
- **NFDIxCS**: National Research Data Infrastructure for and with Computer Science

**Life Sciences**
- **DataPLANT**: Plant research data
- **FAIRagro**: FAIR Data Infrastructure for Agrosystems
- **NFDI4immuno**: National Research Data Infrastructure for Immunology
- **GHGA**: National Research Data Infrastructure for Immunology
- **NFDI4Biodiversity**: Biodiversity, Ecology and Environmental Data
- **NFDI4BIOIMAGE**: National research data infrastructure for microscopy and bioimage analysis
- **NFDI4Health**: NFDI personal health data
- **NFDI4Microbiota**: NFDI for Microbiota Research

**Natural Sciences**
- **DAPHNE4NFDI**: Data from PHoton and Neutron Experiments for NFDI
- **FAIRmat**: FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids
- **NFDI4Cat**: NFDI for sciences related to catalysis
- **MaRDI**: Mathematical Research Data Initiative
- **NFDI4Chem**: Chemistry consortium for the NFDI
- **NFDI4Earth**: NFDI Consortium Earth System Sciences
- **PUNCH4NFDI**: Particles, Universe, NuClei and Hadrons for the NFDI
Involves a broad spectrum of data types and methodologies

- The specificity of sub-disciplines within molecular biology results in even greater diversity of data types and methodologies.

- Diverse data types
  - Quantitative data: laboratory measurements
  - Qualitative data: textual information and images
  - Physical data: biological specimens, genetic sequences, and experimental records
  - Statistical data: analysis results and statistical models
Diverse Data Landscape in Molecular Biology

High-Throughput Experiments:
1. Genomic data (e.g., DNA sequences, gene expression profiles, ChIP-seq data)
2. Transcriptomic data (e.g., RNA-seq data, microarray data)
3. Proteomic data (e.g., mass spectrometry data, protein-protein interaction data)
4. Metabolomic data (e.g., metabolite profiles, metabolic flux data)
5. Epigenomic data (e.g., DNA methylation data, histone modification data)
6. High-resolution microscopy images (e.g., confocal microscopy, super-resolution microscopy)
7. Flow cytometry data (e.g., single-cell analysis, cell sorting data)
8. Next-generation sequencing data (e.g., whole-genome sequencing, targeted sequencing)
9. Single-cell sequencing data (e.g., single-cell RNA-seq, single-cell ATAC-seq)
10. CRISPR/Cas9 screening data (e.g., knockout screens, genetic interaction screens)
11. Protein structure data (e.g., X-ray crystallography, NMR spectroscopy)
12. Time-lapse imaging data (e.g., cell migration, cell division)
13. Comparative genomics data (e.g., phylogenetic analysis, gene family evolution)
14. Protein-DNA interaction data (e.g., DNA-binding assays, chromatin immunoprecipitation)
15. Network analysis data (e.g., protein-protein interaction networks, gene regulatory networks)
16. DNA methylation data (e.g., methylation profiles, Cpg island analysis)
17. Single-nucleotide polymorphism (SNP) data (e.g., SNP genotyping, association studies)
18. Metagenomic data (e.g., microbial community composition, functional profiling)
19. Pathway analysis data (e.g., enrichment analysis, pathway mapping)
20. Computational modeling and simulation data (e.g., molecular dynamics simulations, protein folding predictions)
21. Bioinformatics analysis outputs (e.g., gene annotations, sequence alignments)
22. Protein expression localization data (e.g., subcellular localization, organelle-specific markers)
23. DNA-protein crosslinking data (e.g., chromatin conformation capture, Hi-C data)
24. RNA secondary structure analysis data (e.g., RNA folding predictions, RNA-RNA interactions)
25. Single-molecule imaging data (e.g., single-molecule fluorescence, single-particle tracking)

Low-Throughput Classical Approaches:
1. Western blot data (e.g., protein expression levels, post-translational modifications)
2. PCR and RT-PCR data (e.g., gene amplification, gene expression analysis)
3. Gel electrophoresis data (e.g., DNA, RNA, or protein separation)
4. Immunohistochemistry and immunofluorescence data (e.g., tissue staining patterns, cellular localization)
5. Histological data (e.g., tissue sections, staining intensity)
6. Enzyme activity assays (e.g., enzyme kinetics, substrate specificity)
7. Cell viability and proliferation data (e.g., cell counting, MITT assays)
8. Reporter gene assays (e.g., luciferase assays, beta-galactosidase assays)
9. Protein purification data (e.g., protein yield, purity, activity)
10. Microbial growth data (e.g., growth curves, colony-forming units)
11. Tissue culture data (e.g., cell passage number, cell morphology)
12. Cell migration and invasion assays (e.g., scratch assays, transwell assays)
13. DNA footprinting data (e.g., protein-DNA interaction analysis)
14. Mutagenesis data (e.g., site-directed mutagenesis, functional characterization)
15. Cell cycle analysis data (e.g., DNA content analysis, cell cycle phase determination)
16. Knockdown or knockout data (e.g., siRNA experiments, CRISPR/Cas9-mediated gene knockout)
17. Enzyme kinetics data (e.g., Michaelis-Menten analysis, Lineweaver-Burk plots)
18. Ligand-receptor binding data (e.g., binding affinity, dissociation kinetics)
19. DNA/RNA hybridization data (e.g., in situ hybridization, northern blot)
20. Cell signaling pathway analysis data (e.g., phosphorylation cascades, signal transduction)
21. Metabolic flux analysis data (e.g., stable isotope labeling, metabolic network modeling)
22. Protein-protein interaction data (e.g., yeast two-hybrid, co-immunoprecipitation)
23. Cell adhesion and migration data (e.g., scratch wound healing, transmigration assays)
24. Ion channel electrophysiology data (e.g., voltage-clamp recordings, current-voltage relationships)
25. Hormone or ligand response data (e.g., dose-response curves, signal transduction pathways)
Data handling at different levels. The lack of existing guidelines creates standardization issues for scientists.

High throughput Core facilities at IMB
Addressing Challenges in Molecular Biology Catalog Creation

Interdisciplinary research projects encompass various sub-disciplines with unique research focuses.

Different sub-disciplines have distinct data requirements.

Involves a broad spectrum of data types and methodologies.

Data handling at different levels. The lack of existing guidelines creates standardization issues for scientists.
First Steps to Molecular Biology and Institutional-Specific DMP Templates

- Work with team of PIs at the institute to discuss current DMP and propose modifications
- Follow the same steps as the discipline specific WG guidelines
Offer Guidance by Providing Suitable Answer Options

Datasets

Current question

The following questions collect information on the data that is produced or used in the project. They also help to estimate the value of the data in terms of potential re-use and long-term preservation. Before data is newly created, it is advisable to check if there is existing data that could be re-used. This way, redundant collection or creation of research data is prevented. This saves efforts and costs.

Please fill in the form for each dataset. The different datasets will be referred to in following questions. You can add a new dataset using the green button. Once created, you can edit or delete datasets using the buttons in the top right corner.

What kind of dataset is it?

Please briefly describe the data type and/or the method used to create or collect the data, for example: * genomic sequencing data (e.g., whole genome, exome) * 3D model/digital reconstruction of an ancient settlement * software developed within the project

- Genomic sequencing data (e.g., whole genome, exome)
- Transcriptomic data (e.g., RNA-Seq, microarrays)
- Proteomic data
- Metabolomic data
- Structural biology data (e.g., X-ray crystallography, NMR)
- Functional assays (e.g., reporter assays, CRISPR screens)
- Imaging data (e.g., microscopy, flow cytometry)
- Biophysical data (e.g., isothermal titration calorimetry, surface plasmon resonance)
- Computational models and simulations
- Pathway and network data
- Others (please specify):
Outlook

Modify first draft of DMP -> aim for the next 6 months
Distribute DMP to researchers
Adjust DMP based on feedback
Share the template publicly

Great! Now I have helpful guidance
Enhancing the Role of Data Management Plans in Molecular Biology

**Current perception**
Considered a recent requirement in molecular biology.

DMPs are viewed as paperwork for grant application/approval.

**Existing challenges**
Current DMPs lack practical usefulness for scientists.

Seen as tedious and bureaucratic tasks.

**Proposed solutions**
Develop a new DMP model tailored to molecular biology.

Provide practical, project-specific guidance and support.

Enable scientists to effectively manage their research activities.

**Expected Benefits:**
Empower scientists to navigate their projects more efficiently.

Foster a positive perception of DMPs as valuable tools.

Improve the overall research experience in molecular biology.
Thank you for listening
References

- [https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-dmp.html](https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-dmp.html) - adjusted the original image
- [https://gerbi-gmb.de/2022/06/13/nfdi4bioimage_community_survey/](https://gerbi-gmb.de/2022/06/13/nfdi4bioimage_community_survey/) , [https://twitter.com/DrHenningFalk](https://twitter.com/DrHenningFalk) - adjusted the original image
- [https://rdmkit.elixir-europe.org/images/data_life_cycle.png](https://rdmkit.elixir-europe.org/images/data_life_cycle.png)
- [https://www.mls-phd.uzh.ch/en/Programme-Overview.html](https://www.mls-phd.uzh.ch/en/Programme-Overview.html)
- [https://www.rd-alliance.org/](https://www.rd-alliance.org/)
- [https://www.nfdi.de/](https://www.nfdi.de/)
- [https://nfdi4ing.de/](https://nfdi4ing.de/)
- [https://www.nfdi4chem.de/](https://www.nfdi4chem.de/)
- [https://www.mardi4nfdi.de/about/mission](https://www.mardi4nfdi.de/about/mission)
- [https://fodako.nrw/](https://fodako.nrw/)
- [https://www.gfbio.org/](https://www.gfbio.org/)
- [https://fair-wizard.com/](https://fair-wizard.com/)
- MS pictograms
- [https://www.uni-mainz.de](https://www.uni-mainz.de)
- [https://sfb1551.com](https://sfb1551.com)
- [https://www.sfb1361.de](https://www.sfb1361.de)
- [https://www.imb.de](https://www.imb.de)