A discipline-specific catalog for molecular biology

RDMO Community-Treffen

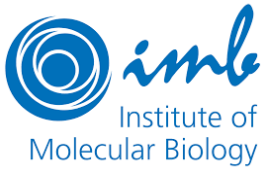
8.10.2024 - virtual

Yasmin Demerdash, PhD

Institute of Molecular Biology, Mainz

Who am I ?

- **Data Manager** since March 2023
  - Employed by Institute of Molecular biology (IMB), Mainz.
  - I support Research Data Management at the SFB1361.



SFB 1361: REGULATION OF DNA REPAIR & GENOME STABILITY

Data management policy

- **Data Management Plans (DMPs) are now compulsory** for all PhD students and all SFB projects.

DMP Realities: Scientists' Expectations vs. Practical Insights

Expectations

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

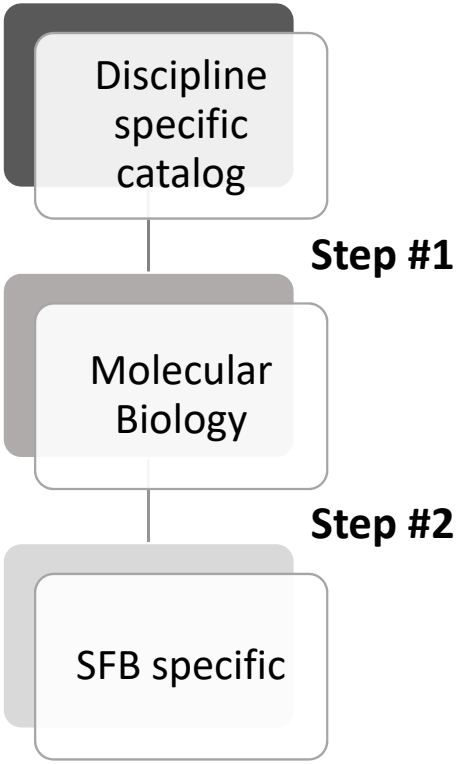
Reality

- Generic and lacks specific guidance
- Tedious and bureaucratic

Shortcoming of current catalogs

- **Novelty of Data Management**
  - Data management is a new field for many researchers
  - DMPs often introduced through funding requirements
- **Perception Challenge**
  - Many don't recognize DMPs as useful project tools
  - Need to demonstrate practical value beyond compliance
- **Irrelevance and Lack of Guidance**
  - Many sections were irrelevant to specific disciplines
  - Abundance of free-answer options without proper guidance
    - Students unsure how to respond effectively
- **Standardization Issues**
  - Difficult for data managers to review DMPs
  - Wide variety of answers made standardization necessary

First Steps to Molecular Biology and SFB-Specific DMP Templates

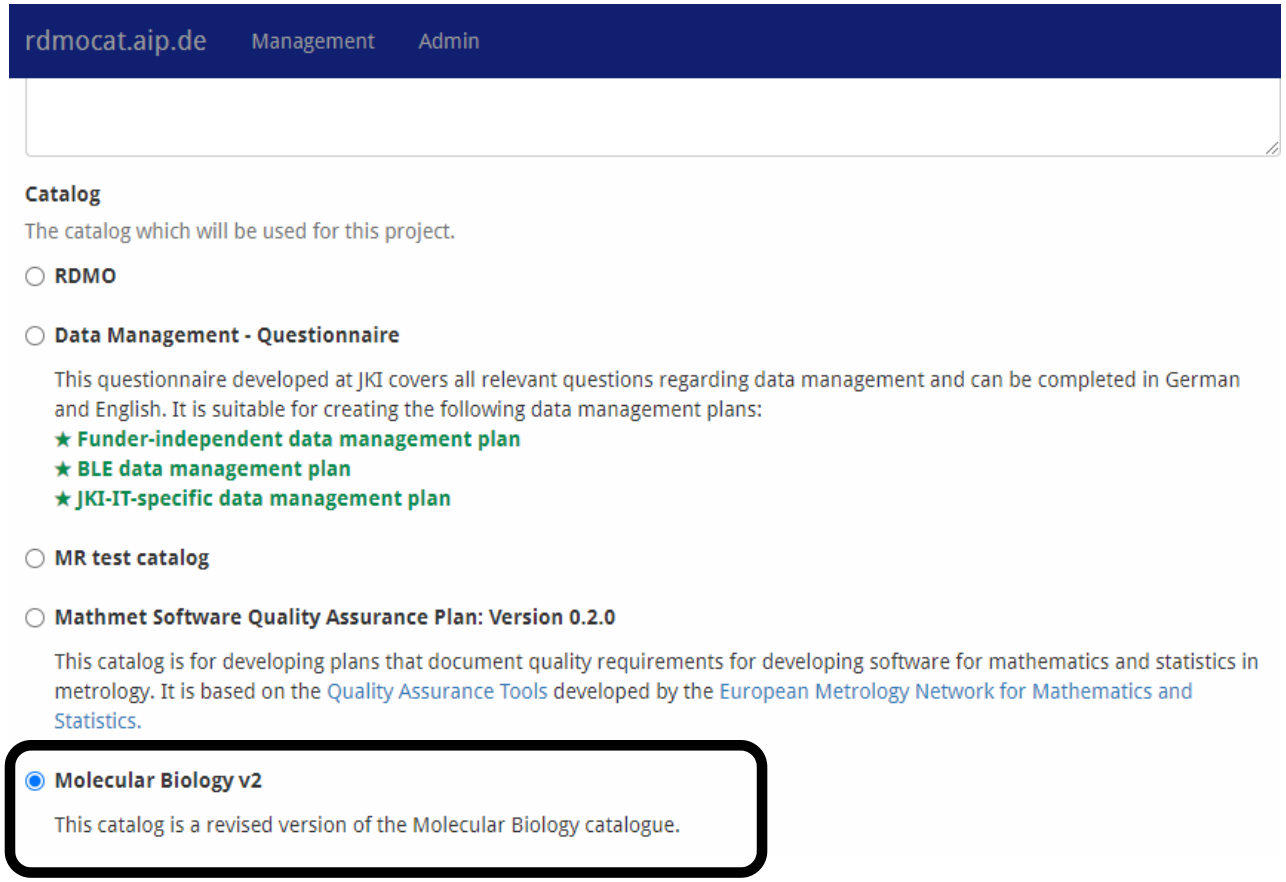


- Worked with team of group leaders - Collaboration project with SFB1551 and a team from IMB (Sarah Wetterman, Björn Kruspig, Bastian Hülsmann, Katja Luck, Anton Khmelinski, Frank Rühle, Fridolin Kielisch)
- Meetings over 3 months

- The template was designed based on the RDMO and the DFG-V5

- Distributed DMP to researchers in a training workshop

Molecular Biology Catalog is available on RDMOcat



- Removed all questions not relevant to the field
- Covered all core data management components
- Made guidance short and focused
- Improved guidance to be discipline-specific
- Reduced free text and increased answer options

Researchers feedback

Were the sections of the catalog appropriately structured to guide you through the process of data management planning?

More Details

Yes, very well-structured	2
Yes, adequately structured	12
No, somewhat disorganized	3
No, completely disorganized	0

Researchers feedback

Did the catalog adequately cover the key aspects of data management relevant to molecular biology research?

More Details

Yes, completely	6
Yes, to some extent	9
No, it missed important aspects	2
No, it didn't cover anything rele...	0
I'm not working in this field	0

Researchers feedback

Overall, how would you rate the usefulness of this discipline-specific catalog for managing research data in your field?

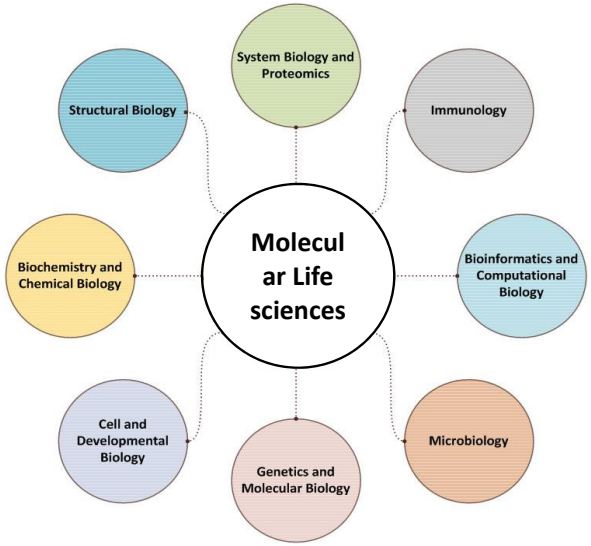
More Details Insights

Highly useful	6
Moderately useful	8
Not useful	3

Challenges in Molecular Biology Catalog Creation

1- **Interdisciplinary research projects encompass various sub-disciplines with unique research focuses**

- Molecular biology is a branch of biology that focuses on understanding life at the molecular level
- Overlaps with genetics, biochemistry, and other biological sciences
- Combines techniques from various fields to understand cellular processes



Challenges in Molecular Biology Catalog Creation

2- 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

nfdi Nationale Forschungsdaten Infrastruktur

Life Sciences

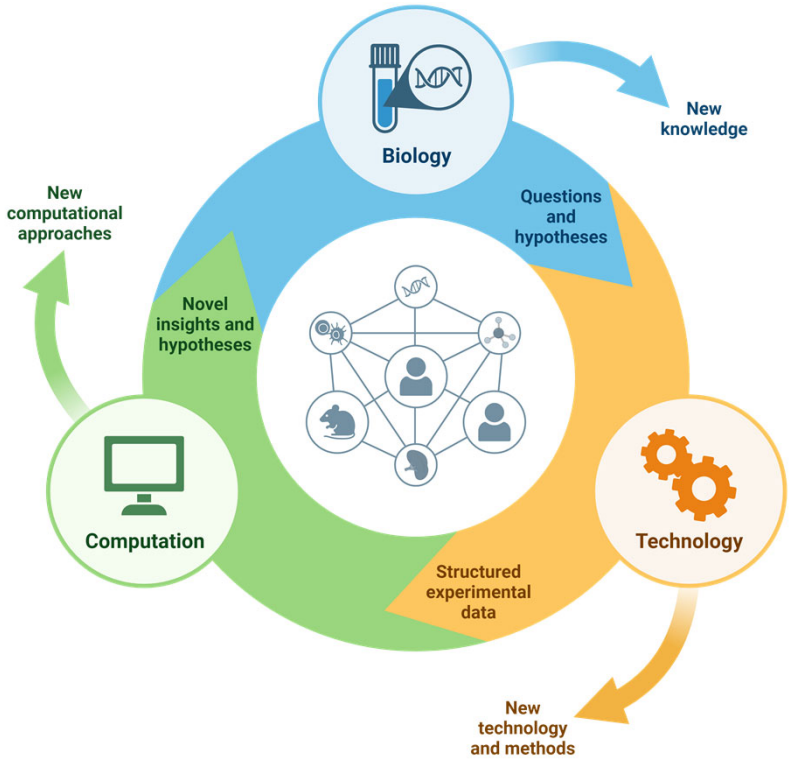
- DataPLANT: Plant research data
- FAIRagro: FAIR Data Infrastructure for Agrosystems
- NFDI4Immuno – National Research Data Infrastructure for Immunology
- GHGA: National Research Data Infrastructure for Immunologyv
- NFDI4Biodiversity: Biodiversity, Ecology and Environmental Data
- NFDI4BIOIMAGE: National 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

Challenges in Molecular Biology Catalog Creation

3- Involves a broad spectrum of data types and methodologies



It is very rare that a DMP would have one dataset

- 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, MTT 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)

Challenges in Molecular Biology Catalog Creation

4- **Data handling at different levels.**
The lack of existing guidelines creates standardization issues for scientists.

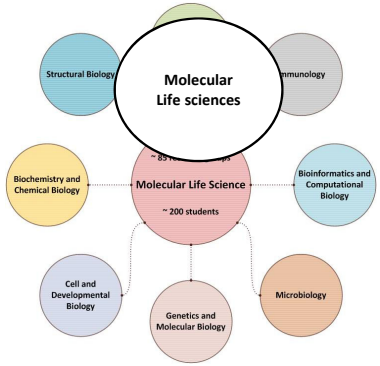
High throughput Core facilities at IMB

DIRECTOR OF CORE FACILITIES	BIOINFORMATICS	FLOW CYTOMETRY
GENOMICS	HISTOLOGY	MICROSCOPY
PROTEOMICS	PROTEIN PRODUCTION	MEDIA LAB
E-LAB	CORE FACILITIES LECTURES	CORE EQUIPMENT

Challenges in Molecular Biology Catalog Creation

1

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



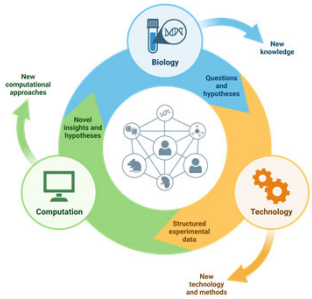
2

Different sub-disciplines have distinct data requirements.

nfdi Nationale Forschungsdaten Infrastruktur

3

Involves a broad spectrum of data types and methodologies.



4

Data handling at different levels. The lack of existing guidelines creates standardization issues for scientists.

Outlook

- Evaluate it if it is following the quality criteria published by the infra-DMP
- Adapt according to researchers feedback
- Have a github page
- Create a comprehensive guidance document
- Publish a research paper
- Adapt to include SFB guidance.

Call for a New Working Group: "Discipline-Specific Catalogs"

- Focus on molecular biology as our first project
- Anyone interested in joining? Y.Demerdash@imb-mainz.de