SDC BioDATEN

Final workshop

12.06.2023

11.00-15.30

Alte Aula, Tübingen

Agenda

10.00 - 11.00 Arrival and reception with coffee at the Alte Aula

11.00 - 12.45 Start of the official part

- Welcome and greeting by Thomas Walter and Renke Siems
- General overview by Holger Gauza
- Developed metadata schema by Claus Zinn

12.45 – 13.45 Lunch break

13.45 – 15.00 Presentations

- Metadata harvesting and annotation by Jan Kaltenbach
- Research data publication using InvenioRDM by Jonathan Bauer
- Demonstration of the publication workflow by Maximilian Müller

15.00 Open discussion and closing remarks

15.30 End

General overview

- The meta perspective
  - Timeline
  - Some numbers
  - Output, outcome & impact
- Sustainability (in a very broad sense)
- Highlights

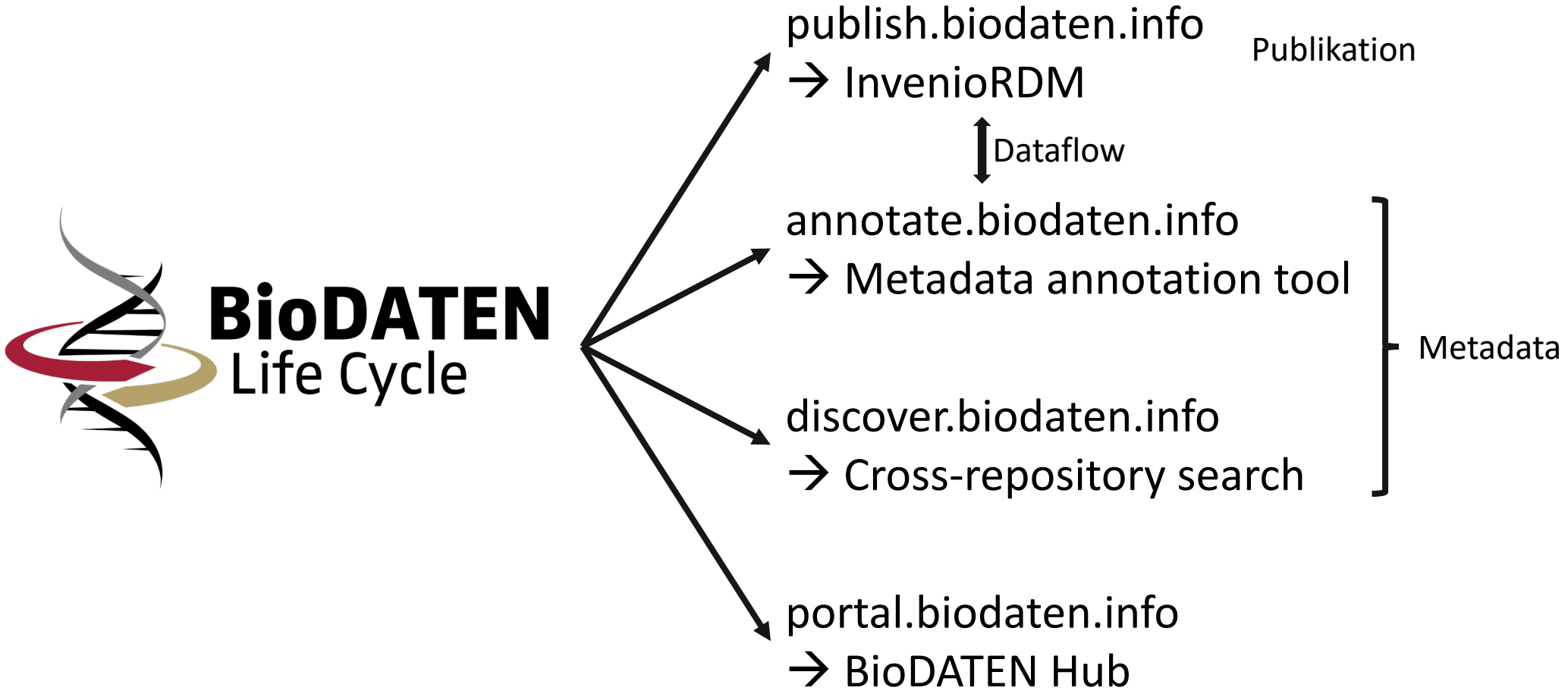
Timeline

- June 2018: Call for Proposals (SDCs to prepare the NFDI launch)
- July 2019: Start of BioDATEN
- **June 2020: DataPLANT & GHGA funding granted**
- February 2021: BioDATEN very positive midterm evaluation
- January 2023: Start of TRR356 PlantMicrobe
- June 2023: End of BioDATEN (no-cost extension granted)

Some numbers

- 2.5 m € funding and in-kind contribution over four years
- 14 partners
- 130+ regular weekly meetings focusing on metadata
- 60+ meetings with other SDCs & working groups
- 40+ regular coordination calls every three weeks
- 16+ publications (O-Bib, Bausteine, Zenodo, various proceedings)
- 28+ trainings (infrastructure, RDM …)
- 12+ workshop / conference contributions

Some numbers



Some numbers

Michael Haferkamp (https://commons.wikimedia.org/wiki/File:Eisberg-diskobucht.jpg), „Eisberg-diskobucht“, https://creativecommons.org/licenses/by-sa/3.0/de/legalcode

Visible: New projects kickstarted, talks, infrastructure, services online, publications …

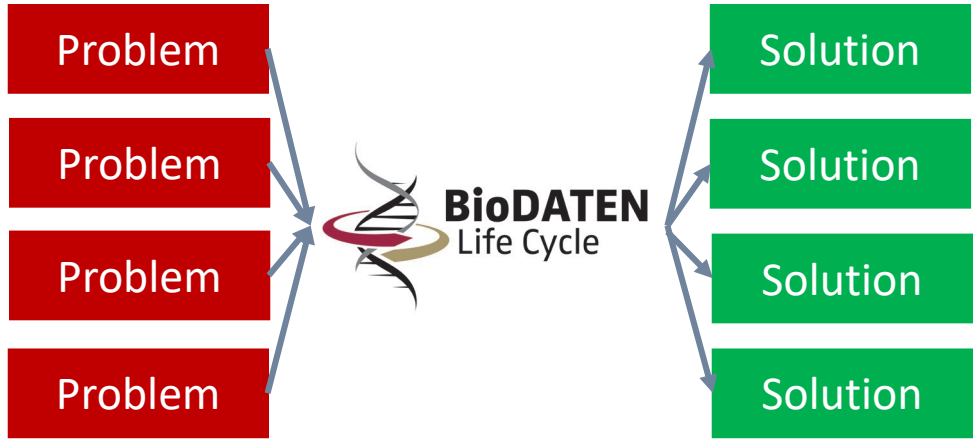
Invisible: Discussions, hard work, trying things, starting over, getting feedback, knowledge gained, people pushed in the right direction, counselling people, networking …

Output, Outcome & Impact

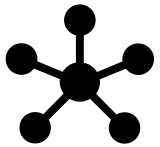
„Classic“ project

1 project tackles 1 problem and provides 1 solution

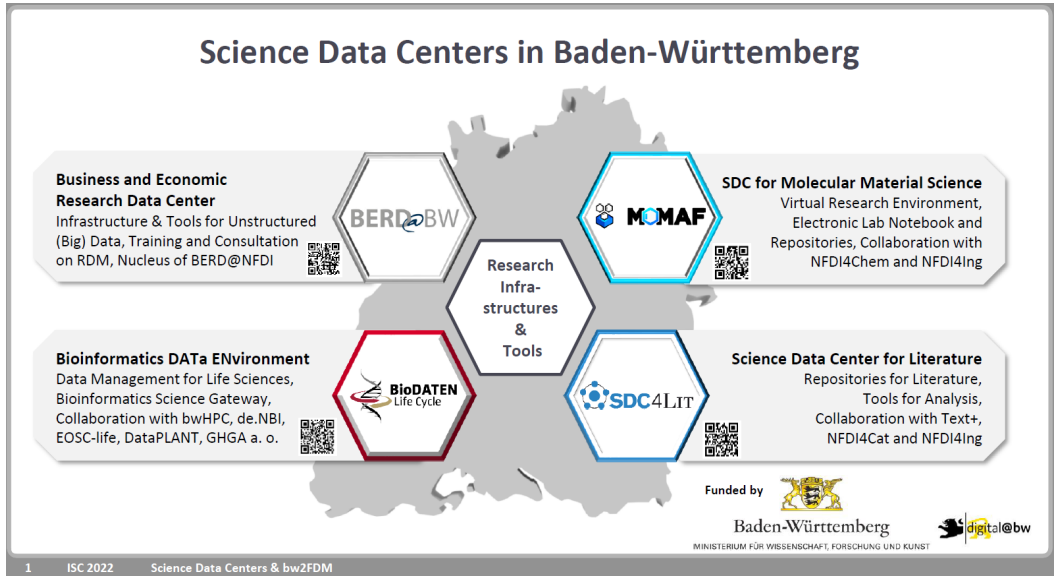
SDC project



BioDATEN tackles a lot of problems, creates a lot of solutions and paves the way for subsequent projects



Output, Outcome & Impact



12.06.2023 BioDATEN Final workshop 9

Output, Outcome & Impact

- Output:
  - Publications & Talks
  - Tools & infrastructure
- Outcome:
  - Pushing RDM on institutional levels
  - Building know-how
  - Fostering RDM within and outside the BioDATEN community
  - Kickstart of several other projects
- Impact:
  - Strengthening Baden-Württemberg as research site (NFDI, TRR, IRTG other projects)

Output, Outcome & Impact

- Collaboration across projects and institutions
- People active in more than one project e.g. DataPLANT & TRR356, BioDATEN & DataPLANT, de.NBI & BioDATEN
- Community feedback
- Collaboration with bw2FDM and forschungsdaten.info

Sustainability (in a very broad sense)

- Use of existing infrastructure
  - de.NBI Cloud → ISO-certified environment for services
  - bwSFS → storage backend for published data
- Re-use of existing components
  - CMDI Framework → from language to life science
  - RDMO → Re-use / re-mix and external hosting
  - RDM FAQ → Links to fd.info, re-use of material
  - VuFind → Search engine, UB Tübingen active part of community

Sustainability (in a very broad sense)

- Work with large institutions
  - InvenioRDM → Developed at CERN, technology scouting (e.g. S3 capability & integration into bwSFS), contribution to development process. Also used within DataPLANT
  - Integration in institutional repositories → organizational aspects such as DOI registration, CTS-certification
- Adaptability by design for developed components
  - bwHPC → workflow integration and genericity
  - Ease of operation → change and adapt for other communities, good documentation, exchange and integration of schemata, integration of external vocabularies via API, barrier-free, responsive design

Sustainability (in a very broad sense)

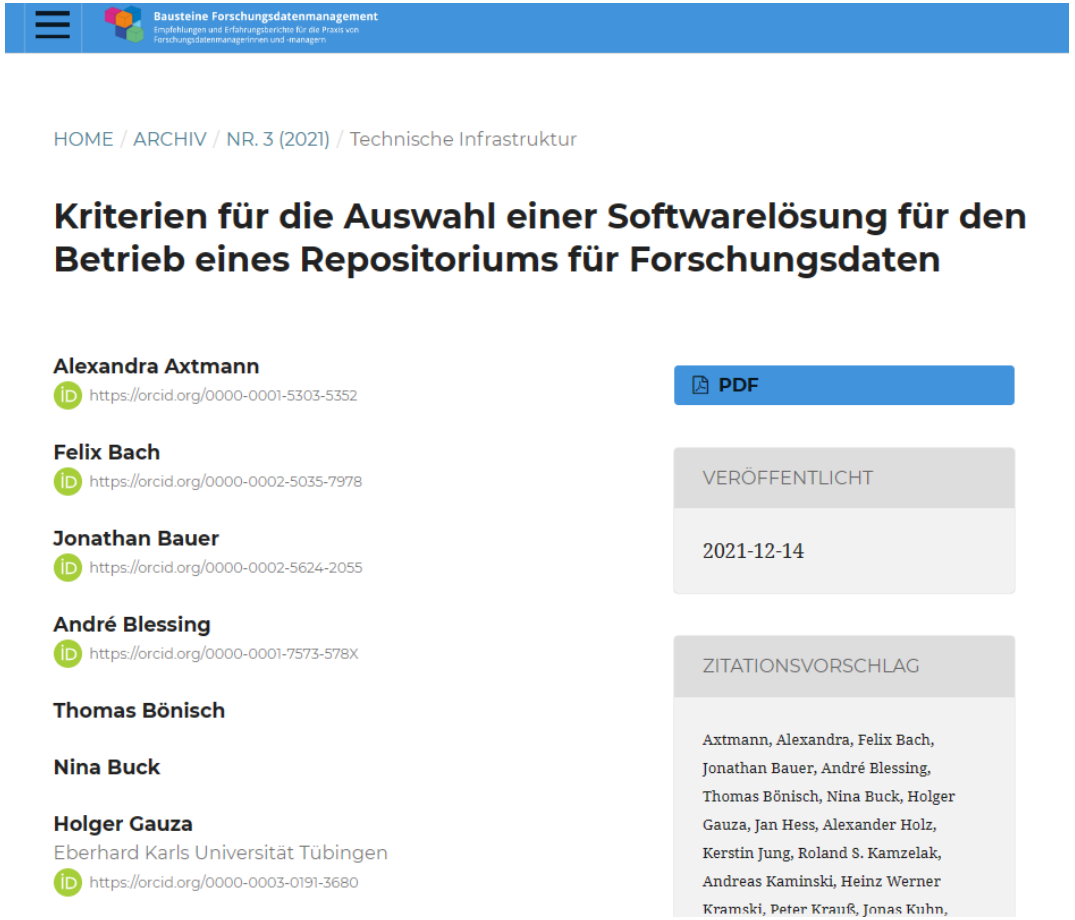
- Knowledge transfer
  - DataPLANT
  - GHGA
  - FAIRagro, NFDI4Bioimage, and beyond

- Open-source policy
  - Use of open-source components → XSLT processor, CMDI framework, InvenioRDM, VuFind etc.
  - Public source code on GitHub → AGPL 3.0 for the metadata schema and annotation tool

Sustainability (in a very broad sense)

Highlights

- Data publication and infrastructure as cross-cutting topic of all four SDCs
- Joint publication of all SDCs and supplemental material
- 27 authors
- Initiated by the SDC infrastructure working group
- Coordinated by BioDATEN in collaboration with bw2FDM



Highlights

- Talk on the CARE-Principles
- Collaboration with forschungsdaten.info live
- Presented within the SDC community
- Not directly related with BioDATEN but broadens the RDM spectrum
- Adding an ethic perspective on RDM

Highlights

- Contribution to the Tübinger RDM Days (2021-2023)
- 2021: RDM in life science
- 2022: RDM in general
  - NFDI4Culture (Torsten Schrade)
  - Text+ (Andreas Witt)
  - SDC4Lit (Jan Hess, Roland S. Kamzelak)
- 2023: RDM in general
- Work on “FDM Zertifikatskurs”

Highlights

- Poster at the E-Science-Tage 2023 at Heidelberg (also 2021 and other contributions)
- Example of excellent teamwork
- Networking opportunity
  - CMDI framework
  - Exchange with ZB MED / FAIRagro → helped to improve our service

Ein Werkzeug zur XSD basierten Metadatenannotation

BioDATEN Life Cycle

Olaf Brandt, Holger Gauza, Jan Kaltenbach, Maximilian Müller, Gabriel Schneider, Claus Zinn

Annotationstool

Die Annotation der Metadaten findet über ein ausfüllbares **HTML-Formular** statt.

Das Annotationstool verwendet einen **XSLT-Prozessor**, um aus den Schemadateien (XSD-Format) das Formular zu generieren.

Schema als XSD-Datei — XSLT - Prozessor — </>

Hierfür können **beliebig viele Schemadateien** hinterlegt werden. Für jedes dieser Schemata wird ein Reiter im Annotationstool-Frontend erstellt, in dem die einzelnen Felder des Schemas als Eingabefeld angezeigt werden.

Anpassungen im Schema

Um das Formular möglichst **übersichtlich** und **minimal** zu gestalten, besteht die Möglichkeit, über spezielle **Parameter innerhalb der Schemadateien** Abhängigkeiten zwischen mehreren Feldern zu definieren. Dies führt dazu, dass einzelne Eingabefelder nur dann angezeigt werden, wenn weitere Voraussetzungen, wie z.B. eine spezifische Auswahl in einem Dropdown-Feld, erfüllt sind.

Beispiel: Die Auswahl für die Art der analytischen Methode wird nur dargestellt, wenn zuvor als Methodentyp auch analytische Methode ausgewählt wurde.

Eigene Datensätze

Eine **Übersicht aller Datensätze** und deren **aktueller Status** ist im Benutzermenü einsehbar. Nach erfolgreicher Anmeldung an der Anwendung erscheint hier eine tabellarische Übersicht, in der alle Datensätze dargestellt werden. Die Anmeldung ist über ELIXIR AAI / ORCID möglich.

Ontologien

Um die bestmögliche **Qualität der Metadaten** sicherzustellen, gibt es innerhalb des Annotationstools die Möglichkeit, für einzelne **Eingabefelder Vokabulare zu hinterlegen**. Diese Vokabulare werden aus ausgewählten Ontologien exportiert und innerhalb der Administrationsoberfläche mit dem entsprechenden Eingabefeld verknüpft.

Das **Eingabefeld** wird anschließend um eine **Autocomplete-Funktion** erweitert.

Sind innerhalb der Ontologie **Beschreibungstexte** zu den Begriffen enthalten, können diese innerhalb des Autocomplete-Fensters über einen **Info-Button** angezeigt werden.

Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

digital@bw

Highlights

- Participation at IWSG 2022 in Trento
- Colleagues from de.NBI, GHGA, DataPLANT, HPCC
- Two publications
  - Distributed but Integrated
  - Developing the German Human Genome-Phenome Archive

Highlights

- Workshop on legal aspects in RDM
- Guests: researchers, administrators, RDM experts, legal expert
- Highlighting importance of legal aspects in RDM and data ownership
- Licences and copyright

Rouge PhD student
A PhD was employed by the University. The relevant research data were stored under the PhD student's identity on the BinAC (HPC resource with storage and compute). The BinAC is administered by employees at the ZDV who also have access to the files stored on it. After graduating and a fight between the PhD student and its supervisor, the latter fears that the PhD student deletes the data. The supervisor asks the BinAC administrators to take actions to prevent deletion and data loss. What can the supervisor do in such a scenario? What may the administrators legitimately do?

Usage monitoring
To use the BinAC resources, it is required to apply for a "Rechenvorhaben (RV)". The PI of an approved RV is responsible for the work done by members of the RV. But, from a technical perspective, it is not possible for the PI (nor intended) to monitor, leading to a sort of dilemma. What kind of organisational measures are legitimate to tackle this issue?

Questions

Who owns the data produced by employed PhD students? As stated in the introduction, data production in bioinformatics is a rather technical process.

How do data ownership and data copyright / "Urheberrecht" interact?

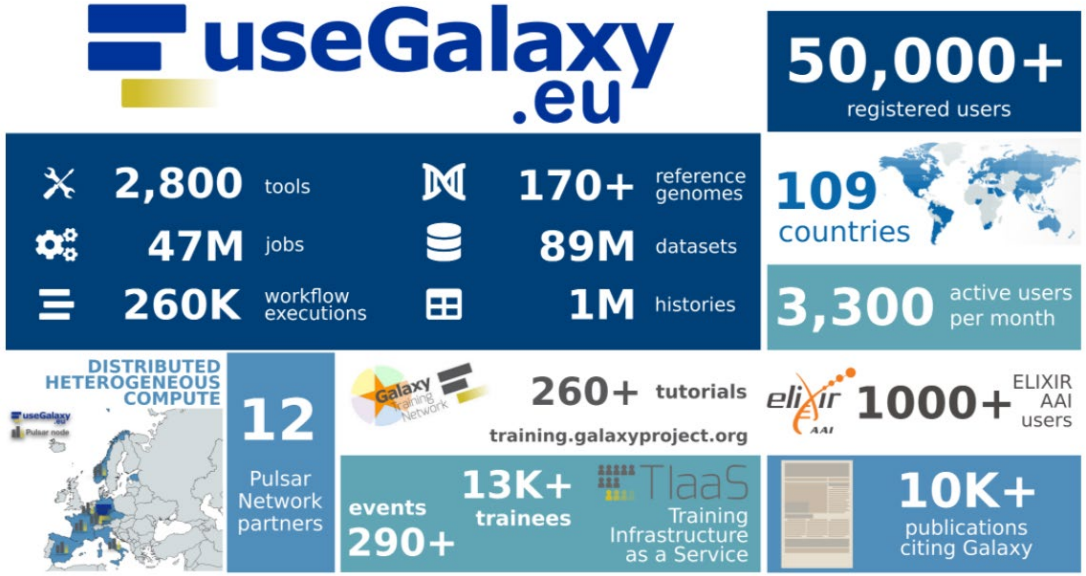
Current approaches seem to recommend CC licenses for research data. Are there any alternatives in case researchers don't know / follow the recommendation?

Quobyte uses usernames and no UID Numbers internally. Even after people leave and user account are removed, files who were owned by them could provide identifiable metadata on them if the username encodes their actual name. Is this critical / should one delete / reown / use different usernames?

Could data in different locations be treated differently: e.g. project folders more owned by PI and $HOME folders more private to the user?

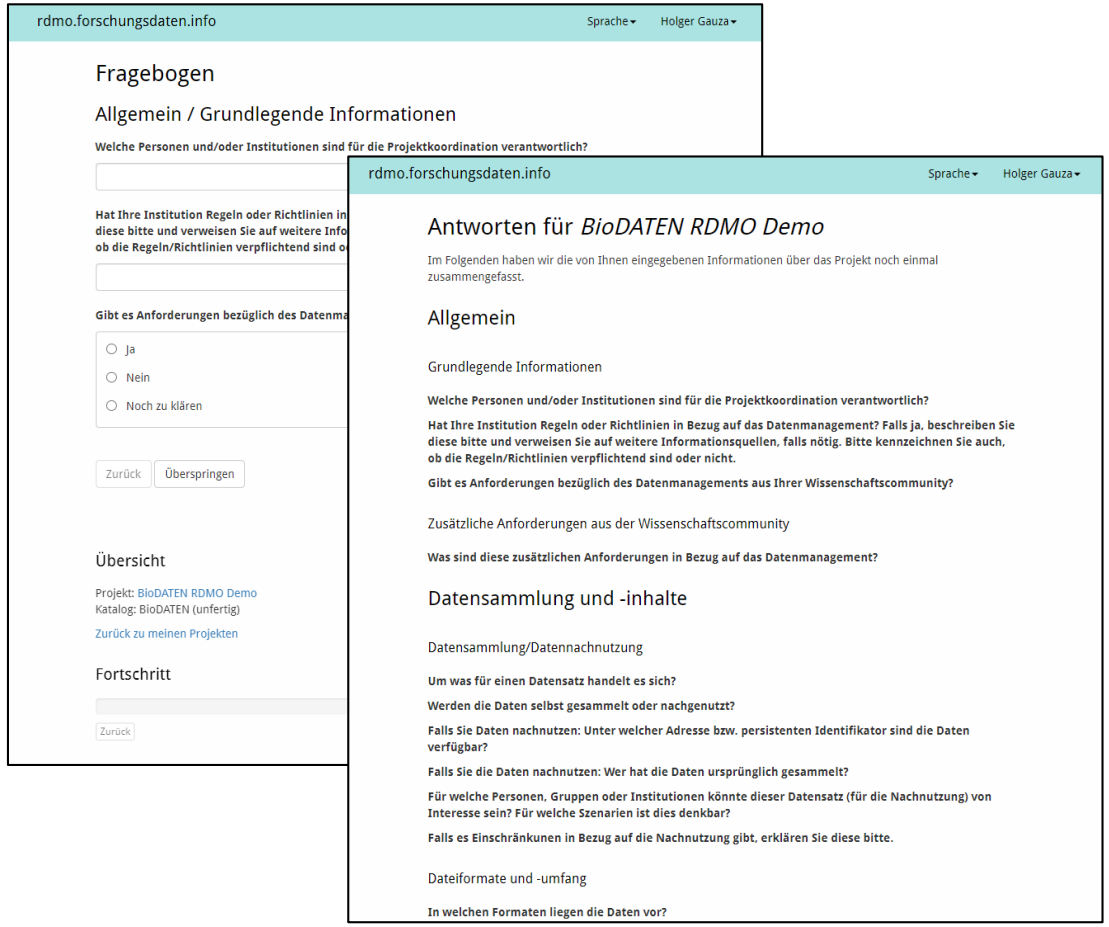
Highlights

- 50.000 registered users
- Constant delivery of new tools
- Constant creation of learning material and training infrastructure
- Workflows for DataPLANT



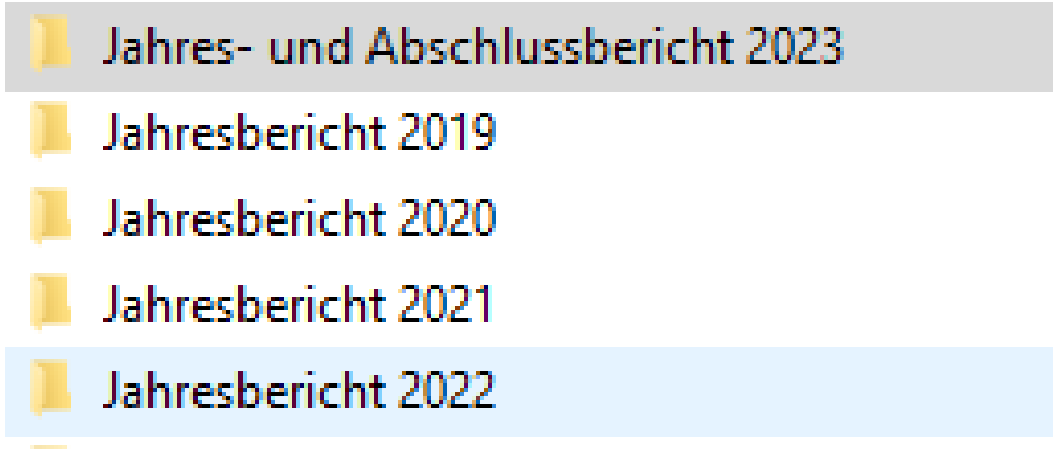
Highlights

- Data management plan implemented in RDMO
- Accessible via rdmo.forschungsdaten.info (Currently offline due to migration)
- Katalog | BioDATEN
- Presented to DataPLANT and the SDCs



Not so highlighty

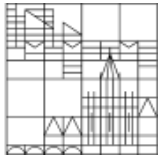
- Not everyone is part of the no-cost extension period
- Let’s start with the final report after the workshop and complement updates as addendum after the no-cost extension



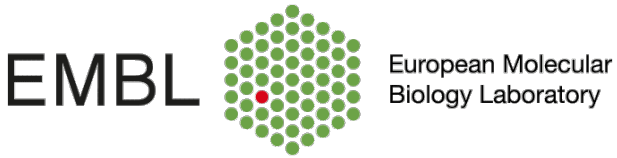
Thank you for your attention!

UNIVERSITÄT HOHENHEIM

Universität Konstanz



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



BioDATEN is sponsored by:

Baden-Württemberg

MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST