Plan Overview

A Data Management Plan created using DMPonline

Title: NWO Open Science 203.001.121: BridgeDb

Creator: Egon Willighagen

Affiliation: Other

Funder: Netherlands Organisation for Scientific Research (NWO)

Template: Data Management Plan NWO (September 2020)

ORCID iD: 0000-0001-7542-0286

ID: 94084

Last modified: 16-02-2022

NWO Open Science 203.001.121: BridgeDb

General Information

Name applicant and project number

- Title: BridgeDb and Wikidata: a powerful combination generating interoperable open research
- Dossiernummer: 203.001.121
- Main Applicant: Egon Willighagen

Name of data management support staff consulted during the preparation of this plan and date of consultation.

No support staff was consulted. Our group develops FAIR solutions and provides DMP solutions to EC-funded projects.

1. What data will be collected or produced, and what existing data will be re-used?

1.1 Will you re-use existing data for this research?

If yes: explain which existing data you will re-use and under which terms of use.

- Yes

Data output from this project is based on data that comes from other sources. Some of this has an open license, some of it we have a personal license to reshare. For gene/protein ID mappings data comes from the Ensembl project, a generally open European database. For metabolites we use three sources: HMDB (personal permission), ChEBI (a CC license), and Wikidata (CCZero waiver/license). We except additional mapping files to be created, particularly from Wikidata (CCZero), but ideally also other openly-licensed sources.
In all cases, whether enforced or not, attribution will be given, a normal scholarly expectation.

1.2 If new data will be produced: describe the data you expect your research will generate and the format and volumes to be collected or produced.

All data will be made available for reuse, and our data derivations will include some level of new contribution (e.g. interpretation). Where possible, we echo the license of the source we used. Therefore, mapping files from data in Wikidata will be released as CCZero too.
While not immediately planned, we do not rule out creating mappings ourselves. These will also be available under Open Science principles: release soon, release often (and Open, obviously).

1.3. How much data storage will your project require in total?

- >1000 GB

Current mapping files continue the grow in size and are typically in the size of 200MB to 3GB right now. We hope that changing technical solutions in BridgeDb we can reduce the file size. The number of files is currently in the order of 20 files. We also plan to release mapping files two or more times, multiplying the total size of one release. Likely, we will release more than 1000 GB of data over the course of the project.

2. What metadata and documentation will accompany the data?

2.1 Indicate what documentation will accompany the data.

There is documentation about how to use the mapping files in various place, mostly with the tools using them. There is an overview at https://bridgedb.github.io/pages/docs.html and we have, as an ELIXIR Recommended Interoperability Resource, material indexed in the ELIXIR TeSS, findable via https://tess.elixir-europe.org/search?q=bridgedb.

2.2 Indicate which metadata will be provided to help others identify and discover the data.

There is minimal metadata is provided in the BridgeDb ID mapping files (which is also available via the BridgeDb Webservice), along with the data releases (and Zenodo and Figshare happily share this metadata, making it visible in DataCite), and futhermore on the BridgeDb mapping file download page at https://bridgedb.github.io/data/gene_database/. This download page has embedded BioSchemas JSON-LD, which allows search engines to find the data easily and Google exposes this in Google's Dataset Search. However, a VoID header file with more and machine-readable information can already be embedded and the plan is to upgrade this to the W3C HCLS Community Profile for Dataset descriptions.
Metadata includes

3. How will data and metadata be stored and backed up during the research?

3.1 Describe where the data and metadata will be stored and backed up during the project.

- Other (please specify)

Data will be release via Zenodo and Figshare.

3.2 How will data security and protection of sensitive data be taken care of during the research?

- Not applicable (no sensitive data)

4. How will you handle issues regarding the processing of personal information and intellectual property rights and ownership?

4.1 Will you process and/or store personal data during your project?

If yes, how will compliance with legislation and (institutional) regulation on personal data be ensured?

- No

4.2 How will ownership of the data and intellectual property rights to the data be managed?

Openly-licensed data also has a copyright owner and IP applies here as well; it just happens that people get a license to reuse, modify, and redistribute it. This information will be shared as part of the metadata, where applicable.

5. How and when will data be shared and preserved for the long term?

5.1 How will data be selected for long-term preservation?

- Other (please specify)

Data will be release via Zenodo and Figshare, which both guarantee availability for 20 years (last time I asked).

5.2 Are there any (legal, IP, privacy related, security related) reasons to restrict access to the data once made publicly available, to limit which data will be made publicly available, or to not make part of the data publicly available?

If yes, please explain.

- No

5.3 What data will be made available for re-use?

- All data resulting from the project will be made available

5.4 When will the data be available for re-use, and for how long will the data be available?

- Data available as soon as article is published

Actually, it will be made available when ready, before the article is published, and during the project.

5.5 In which repository will the data be archived and made available for re-use, and under which license?

Figshare and Zenodo. License of the data will be open and follow that of the upstream ontology or arrangement.
For gene/protein ID mappings data comes from the Ensembl project, a generally open European database. For metabolites we use three sources: HMDB (personal permission; HDMB is in a collaboration project with our group), ChEBI (a CC license), and Wikidata (CCZero waiver/license). We except additional mapping files to be created, particularly from Wikidata (CCZero), but ideally also other openly-licensed sources.

5.6 Describe your strategy for publishing the analysis software that will be generated in this project.

The software around the data, both the BridgeDb Java library and the tools to create BridgeDb ID mapping files will be shared during the project (and already is for multiple mapping files) via GitHub. The Java library is archived using the GitHub/Zenodo integration. We will do the same for the code during this project and all code will be archived at Zenodo at least once (at the end) during the project.
Current code repositories (with software license):

- BridgeDb Java Library, Apache License 2.0, github.com/bridgedb/BridgeDb
- Metabolite ID (HMDB, ChEBI, Wikidata), Simplified BSD License, github.com/bridgedb/create-bridgedb
- Interaction ID (Rhea), Simplified BSD License, github.com/DeniseSl22/create-mapping-database-bridgedb-interactions
- Disease ID mapping (Wikidata), Simplified BSD License, github.com/DeniseSl22/create-bridgedb-diseases
- Gene/Protein ID (Ensembl), custom (open) license, github.com/bridgedb/create-bridgedb-genedb
- Protein complexes, virus proteins, journal articles (Wikidata), Apache License 2.0 github.com/bridgedb/Wikidata2Bridgedb

6. Data management costs

6.1 What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

The DMP is already part of the development models and no additional resources are needed.