BridgeDb and Wikidata: a powerful combination generating interoperable open research (BridgeDb)

Egon L. Willighagen[‡], Martina Kutmon[‡,§], Marvin Martens[‡], Denise Slenter[‡]

‡ Dept of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, Netherlands
§ Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, Netherlands

Corresponding author: Egon L. Willighagen (egon.willighagen@maastrichtuniversity.nl)

Abstract

Like humans have a unique social security number and different phone numbers from various providers, so do proteins and metabolites have a unique structure but different identifiers from various databases. BridgeDb is an interoperability platform that allows combining these databases, by matching database-specific identifiers. These matches are called identifier mappings, and they are indispensable when combining experimental (omics) data with knowledge in reference databases. BridgeDb takes care of this interoperability between gene, protein, metabolite, and other databases, thus enabling seamless integration of many knowledge bases and wet-lab results. Since databases get updated continuously, so should the Open Science BridgeDb project.

Keywords

BridgeDb, Wikidata, open science, identifie

Dutch public summary

Net zoals mensen een uniek Burgerservicenummer (BSN) hebben en verschillende telefoonnummers van diverse telecomaanbieders, zo hebben eiwitten en metabolieten een unieke structuur maar andere identificatiecodes in verschillende databases. BridgeDb is een interoperabiliteitsplatform die het combineren van databases mogelijk maakt op basis van gelijkwaardige identificatiecodes. In het Engels heten deze *identifier mappings* en ze zijn essentieel in analyse van biologische data. BridgeDb zorgt ervoor dat experimentele data over genen, eiwitten, en metabolieten eenvoudig gekoppeld kan worden aan kennis over biologische processen opgeslagen in andere digitale bronnen. Omdat deze databases regelmatig veranderen, zal het Open Science project BridgeDb dat ook doen.

Project proposal

The vision for your project

Linking any two or more databases always requires linking identical entities described in those databases. Unfortunately, the identifier used for the same entity in one database is often different from the identifiers for the same entity in the other database. BridgeDb was created to make the bridge between databases by providing uniform access to mappings between different database identifiers for the same entities. This is why BridgeDb is a *Recommended Interoperability Resource* (RIR) of ELIXIR, a collaboration of leading life science organisations, and has been supporting projects like the ELIXIR-NL WikiPathways resource (Slenter et al. 2017).

The vision of this project is to improve the foundation of BridgeDb, to allow us to widen the scope in the future and enhance the support of currently unsupported, but important data sources. This will open up the road to wide adoption in the European Open Science Cloud (EOSC). To reach this vision, we aim to

1. modernize the project by updating the library and accompanying build system,
2. extending the functionality of the webservice to deploy identifier (ID) mapping databases effortlessly, by extending the support of creating ID mappings databases from Wikidata (Waagmeester et al. 2021, Waagmeester et al. 2020), and
3. by updating the tools to create ID mapping databases, along with new archived and citable releases for the genes, proteins, protein complexes, metabolites, nanomaterials, adverse outcome pathways, and journal articles ID mapping databases.

The first output of this project is an improved BridgeDb Java library (Batchelor et al. 2014, van Iersel et al. 2010), using the stable build system Apache Maven and following its practices, higher test coverage, including automated testing of the MySQL backend, and higher coverage of JavaDoc (see **WP1** below). Second, the project will produce a new version of the live BridgeDb Webservice (webservice.bridgedb.org), with support for modern FAIR data standards, like Compact Identifiers (Wimalaratne et al. 2018), DataCite (Anonymous 2021), and W3C's HCLS Community Profile for Dataset descriptions (HCLS Community 2015) (see **WP2**). Third, this project will output various ID mapping databases and streamline the tools to create them (see **WP3**). All ID mapping databases in this project will be released under a CCZero waiver whenever possible. All source code will be released under an Apache License 2.0 or a more liberal open license.

Currently, BridgeDb has been an important project to link multiple life science databases, e.g. genes, proteins, metabolites. With clear open licenses, FAIR approaches (Jacobsen et al. 2020), and collaboration with open projects (WikiPathways, Open PHACTS (Williams et al. 2012, Batchelor et al. 2014), EpiLipidNET, FNS-cloud, COVID-19 Disease Maps (Ostaszewski et al. 2020)), we have demonstrated how key infrastructure can be free and open by design. BridgeDb is essential for omics data analysis and links corresponding entries between databases, whether these databases are open or closed. All output will be made available as Docker Images, allowing repurposing for any other ID mapping need.

Project plan

The project plan is organized in three work packages (WP1, WP2, WP3), following the three output themes. Work package 1 (**WP1**) intends to upgrade the BridgeDb Java library. Currently, the main Java library is already built with Apache Maven, however, the build system should also be applied to related tools, and we will extensively use GitHub Actions for automation. Second, only a subset of library modules is currently available as OSGi bundles, which is essential for reuse in various third-party tools, like PathVisio (Kutmon et al. 2015) and Cytoscape (Kutmon et al. 2013, Shannon et al. 2003). Therefore all modules will be extended to support OSGi bundles, something that is already done for five core BridgeDb modules. Furthermore, to improve maintainability, WP1 will continue extending the unit tests and integration tests. Particularly, the testing of the database backends that hold the ID mapping data (Apache Derby and MySQL) needs to become more comprehensive.

Work package 2 (**WP2**) focuses on the BridgeDb Webservice. This continuously running service is an ELIXIR *RIR* and daily supports projects like WikiPathways and Cytoscape to assist data analysis of omics datasets (transcriptomics, proteomics, metabolomics, etc.). The Webservice will be extended to support Compact Identifiers (Wimalaratne et al. 2018) as a new input and output format, in order to support persistent, machine-resolvable citation of research data in written material. Furthermore, we will introduce support for JavaScript Object Notation (JSON) as a serialization format for multiple application programming interface (API) calls. The OpenAPI (Swagger) interactive documentation will be updated accordingly. Furthermore, the Webservice itself will become even more FAIR, by adopting the DataCite standard, and providing provenance in the HCLS Community Profile for Dataset descriptions.

The last work package (**WP3**) translates the new functionalities to practical use cases. In this WP, existing ID mapping databases will be updated, using the new releases of BridgeDb Java library and tested in applications using the new BridgeDb version. We intend to widen the scope of ELIXIR resources supported in the ID mapping databases, to make more resources interoperable (and therefore more FAIR). Here, we will increasingly use Wikidata and its international scientific collaborations (Waagmeester et al. 2021, Waagmeester et al. 2020). These mapping databases will continue to be released via public archives (e.g. Figshare, Zenodo) under open licenses, and indexed on the BridgeDb website annotation at bridgedb.github.io/data/gene_database/. To do so, WP3 will develop a tool that takes DOIs of the mapping databases as input to extract metadata from the respective repositories and generate this indexing website. WP3 will test the resulting mapping databases with downstream tools (PathVisio, WikiPathways, Cytoscape, etc.). Docker Images of the various tools will be developed to simplify dissemination and reuse. Practically, this work will involve two hackathons involving the senior scientific employees

(Slenter, Kutmon, Martens) and the full-time non-scientific personnel (see the Section *Team members* and Table 1).

Team members

The funding will be used to employ a **scientific programmer**. Additionally, from the Dept of Bioinformatics (BiGCaT), the following people will be involved for WP3 for testing the upgraded BridgeDb library to create updated ID mapping databases. **Denise Slenter** (orcid:0000-0001-8449-1318) will work on the metabolite, disease and interaction ID mapping databases, **Dr Martina Kutmon** (orcid:0000-0002-7699-8191; assistant professor) on the gene and protein ID mapping database (with Ensembl as source), and **Marvin Martens** (orcid:0000-0003-2230-0840) will work on a gene and protein mapping databases for *Daphnia magna* and *Daphnia pulex* (relevant model species for toxicology, but currently not in Ensembl). Slenter, Kutmon, and Martens have all been previously involved in the BridgeDb projects in their research projects (e.g. created the Docker Image for BridgeDb and using Wikidata as a source of ID mappings), and are experts in the fields relevant for these mapping databases: chemistry and metabolism (Slenter); systems biology and data analysis (Kutmon); toxicology and Adverse Outcome Pathways (Martens).

Open Science track record of the main applicant

Dr Egon Willighagen has been active in Open Science for over 20 years, for example, contributing to projects like JChemPaint (since 1998; doi:10.3390/50100093), WikiPathways (since 2011; doi:10.1093/NAR/GKV1024), and (temporarily) leading projects like Jmol and coordinating the science in the EU FP7 project eNanoMapper (doi:10.3762/BJNANO.6.165), and co-founded the Chemistry Development Kit (in 2000; doi:10.1021/ci025584y). He is recognized for his work with the international Blue Obelisk Award (2007) and a national runner-up Open Initiative Trophy (2021). From 2016 to 2021 he has been one of two Editor-in-Chief of the fully CC-BY, highly ranked *Journal of Cheminformatics* (issn:1758-2946), which promotes Open Science in chemistry. At various National Plan Open Science events and meetings, Willighagen has provided input from a researcher's perspective and is co-founder of the *Open Science Community Maastricht*. A more complete list of his Open Science work can be found in his publication list: orcid.org/0000-0001-7542-0286.

Data management

Will this project involve re-using existing research data?

Yes. Where existing data is reused, these will have an open license or a public domain waiver (like the American *public domain* or the international CCZero waiver). Any license, including open licenses, constrain the reuse. License information will be clearly provided, following the FAIR principles.

Will data be collected or generated that are suitable for reuse?

Yes, reuse is the aim of the BridgeDb project, where downstream users are, for example, WikiPathways, PathVisio, and Cytoscape.

After the project has been completed, how will the data be stored for the long-term and made available for the use by third parties? Are there possible restrictions to data sharing or embargo reasons?

Data will be archived during the project in public repositories, like Figshare and Zenodo, which have committed themselves to availability of 20 years or more. The open licenses allow other repositories to archive a copy of the data.

No restrictions (other than the open license terms) and no embargoes are anticipated.

Will any costs (financial and time) related to data management and sharing/preservation be incurred?

No: All the necessary resources (financial and time) to store and prepare data for sharing/ preservation are or will be available at no extra cost.

Software sustainability

Will software be generated during the project?

Yes.

How will the software be licensed and be made available for re-use?

All BridgeDb software is available under an OSI-approved license on GitHub. This includes the Apache License 2.0-licensed BridgeDb library as well as the existing source code to generate ID mapping databases, available under other open licenses (see Table 2).

What measures are needed to make the software appropriate for long-term (re-)use by third parties?

WP1 will improve the maintainability and portability of the software. The main BridgeDb Java library is developed on GitHub and disseminated via Zenodo (using the GitHub-Zenodo integration) and via Maven Central (search.maven.org/search?q=g:org.bridgedb).

How large do you expect the community that will potentially use the software to be, and do you expect outside contributors to the software?

The size of communities is hard to accurately estimate, but with the highly cited WikiPathways (monthly 15,000 unique website users) and Cytoscape projects as daily users and being an ELIXIR *Recommended Interoperability Resource*, we estimate a few thousand daily users. The gene/protein ID mapping database is downloaded more than 14 thousand times for local use, and the Bioconductor R package for BridgeDb (doi:10.18129/B9.bioc.BridgeDbR) is downloaded 50-100 times each month (rank 774 out of 1974).

BridgeDb has been used in EU projects like OpenPHACTS, OpenRiskNet, and NanoSolveIT. A full list of past contributors can be found on GitHub for each of the subprojects, e.g. at github.com/bridgedb/BridgeDb/graphs/contributors.

What expertise do you expect to be needed to make the software appropriate for long-term re-use by third parties? Is this expertise available?

The main applicant has more than 20 years of experience in the development of open data, open-source, and open standards projects, and the BridgeDb project already exists for over 10 years. As Editor-in-Chief of a journal that has reuse and Open Science as strong editorial standards, the required expertise is available.

Other grant applications with overlapping content

No overlapping grant applications.

Acknowledgements

This project is funded by NWO grant 203.001.121.

Funding program

Open Science (OS) Fund 2020/2021

Grant title

BridgeDb and Wikidata: a powerful combination generating interoperable open research (BridgeDb)

Hosting institution

Maastricht University

Conflicts of interest

References

- Anonymous (2021) DataCite Metadata Scheme. https://schema.datacite.org/
- Batchelor C, Brenninkmeijer CA, Chichester C, Davies M, Digles D, Dunlop I, Evelo C, Gaulton A, Goble C, Gray AG, Groth P, Harland L, Karapetyan K, Loizou A, Overington J, Pettifer S, Steele J, Stevens R, Tkachenko V, Waagmeester A, Williams A, Willighagen E (2014) Scientific Lenses to Support Multiple Views over Linked Chemistry Data. The Semantic Web – ISWC 201498-113. https://doi.org/10.1007/978-3-319-11964-9_7
- HCLS Community (2015) Dataset Descriptions: HCLS Community Profile. W3C. URL: http://www.w3.org/TR/2015/NOTE-hcls-dataset-20150514/
- Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, Courtot M, Crosas M, Dumontier M, Evelo C, Goble C, Guizzardi G, Hansen KK, Hasnain A, Hettne K, Heringa J, Hooft RW, Imming M, Jeffery K, Kaliyaperumal R, Kersloot M, Kirkpatrick C, Kuhn T, Labastida I, Magagna B, McQuilton P, Meyers N, Montesanti A, van Reisen M, Rocca-Serra P, Pergl R, Sansone S, da Silva Santos LOB, Schneider J, Strawn G, Thompson M, Waagmeester A, Weigel T, Wilkinson M, Willighagen E, Wittenburg P, Roos M, Mons B, Schultes E (2020) FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2: 10-29. https://doi.org/10.1162/dint_r_00024
- Kutmon M, Kelder T, Mandaviya P, Evelo CA, Coort S (2013) CyTargetLinker: A Cytoscape App to Integrate Regulatory Interactions in Network Analysis. PLoS ONE 8 (12). https://doi.org/10.1371/journal.pone.0082160
- Kutmon M, van Iersel M, Bohler A, Kelder T, Nunes N, Pico A, Evelo C (2015) PathVisio 3: An Extendable Pathway Analysis Toolbox. PLOS Computational Biology 11 (2). https://doi.org/10.1371/journal.pcbi.1004085
- Ostaszewski M, Mazein A, Gillespie M, Kuperstein I, Niarakis A, Hermjakob H, Pico A, Willighagen E, Evelo C, Hasenauer J, Schreiber F, Dräger A, Demir E, Wolkenhauer O, Furlong L, Barillot E, Dopazo J, Orta-Resendiz A, Messina F, Valencia A, Funahashi A, Kitano H, Auffray C, Balling R, Schneider R (2020) COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. Scientific Data 7 (1). https://doi.org/10.1038/s41597-020-0477-8
- Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13 (11): 2498-2504. https://doi.org/10.1101/gr.1239303

- Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL (2017) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research 46 https://doi.org/10.1093/nar/gkx1064
- van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, Conklin BR, Evelo CT (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics 11 (1). https://doi.org/10.1186/1471-2105-11-5
- Waagmeester A, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, Hanspers K, Hermjakob H, Hudson TS, Hybiske K, Keating SM, Manske M, Mayers M, Mietchen D, Mitraka E, Pico AR, Putman T, Riutta A, Queralt-Rosinach N, Schriml LM, Shafee T, Slenter D, Stephan R, Thornton K, Tsueng G, Tu R, Ul-Hasan S, Willighagen E, Wu C, Su AI (2020) Wikidata as a knowledge graph for the life sciences. eLife 9 https://doi.org/10.7554/eLife.52614
- Waagmeester A, Willighagen E, Su A, Kutmon M, Gayo JEL, Fernández-Álvarez D, Groom Q, Schaap P, Verhagen L, Koehorst J (2021) A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses. BMC Biology 19 (1). https://doi.org/10.1186/s12915-020-00940-y
- Williams A, Harland L, Groth P, Pettifer S, Chichester C, Willighagen E, Evelo C, Blomberg N, Ecker G, Goble C, Mons B (2012) Open PHACTS: semantic interoperability for drug discovery. Drug Discovery Today 17: 1188-1198. https://doi.org/10.1016/j.drudis.2012.05.016
- Wimalaratne S, Juty N, Kunze J, Janée G, McMurry J, Beard N, Jimenez R, Grethe J, Hermjakob H, Martone M, Clark T (2018) Uniform resolution of compact identifiers for biomedical data. Scientific Data 5 (1). https://doi.org/10.1038/sdata.2018.29

Table 1.

Gantt diagram of project work timeline. In the months M5 and M10, two two-day hackathons (H) will be organized.

	M1	M3	M5	M8	M10	M12
WP1	🟩	🟩	🟩	🟩		
WP2			🟩	🟩	🟩	🟩
WP3			H	🟩	H	🟩

Table 2.

The BridgeDb project comprises of multiple independent code bases, of which a few are listed here.

Name	**Source of mappings (where applicable)**	**Source Code License**	**Code repository**
BridgeDb Java Library		Apache License 2.0	github.com/bridgedb/BridgeDb
Metabolite ID mapping database	HMDB, ChEBI, Wikidata	Simplified BSD License	github.com/bridgedb/create-bridgedb-metabolites
Interaction ID mapping database	Rhea	Simplified BSD License	https://github.com/DeniseSl22/create-bridgedb-interactions
Disease ID mapping database	Wikidata	Simplified BSD License	https://github.com/DeniseSl22/create-bridgedb-diseases
Gene/Protein ID mapping database	Ensembl	see this issue report	github.com/bridgedb/create-bridgedb-genedb
Protein complexes, virus proteins, journal articles	Wikidata	Apache License 2.0	github.com/bridgedb/Wikidata2Bridgedb