Integrating Data Treasures – The First Knowledge Graphs of the DSMZ Digital Diversity Databases

Julia Koblitz[1,†], Lorenz C. Reimer[2,†]

[1]a Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Inhoffenstr. 7B, Braunschweig, Germany

Abstract

The DSMZ (German Collection of Microorganisms and Cell Cultures) hosts a wealth of biological data, spanning microbial taxonomy, enzymes, rRNA genes, cell lines, cultivation media, and more. To make these diverse datasets accessible and interoperable, the DSMZ Digital Diversity initiative provides a central hub for integrated data and establishes a framework for linking and accessing these resources (https://hub.dsmz.de). At its core is the DSMZ Digital Diversity Ontology (D3O), which standardizes and connects data from all databases to enable seamless integration and advanced exploration. In this work, we present the first knowledge graphs of two major databases: BacDive, which provides detailed microbial strain information, and MediaDive, which focuses on data on microbial cultivation. Both knowledge graphs are accessible via SPARQL endpoints at https://sparql.dsmz.de, allowing researchers to query and analyze the data in a standardized way. These initial steps lay the groundwork for integrating additional databases, such as BRENDA, SILVA, LPSN, and StrainInfo, into a unified, queryable knowledge graph. Our goal is to connect this vast diversity of datasets and foster collaboration toward a more open and connected future for biological databases. By sharing our approach and results, we aim to inspire others to explore the potential of linked data in the life sciences.

Keywords

Data Integration, Life Science Databases, Semantification

1. Introduction

The German Collection of Microorganisms and Cell Cultures (DSMZ) is home to a diverse portfolio of biological databases, covering a wide range of domains such as microbial strains (BacDive) [1], cultivation media (MediaDive) [2], microbial taxonomy (LPSN) [3], enzymology (BRENDA) [4], bacteriophages (PhageDive) [5], cell line data (CellDive) [6], rRNA data and taxonomy (SILVA) [7], and StrainInfo, a service that integrates strain identity information from multiple sources. Together, they represent a treasure trove of biological knowledge, serving as essential resources for the life sciences, which has also been acknowledged by the Global Biodata Coalition and ELIXIR by selecting four of the resources as Global Core Biodata Resources and three of them as ELIXIR Core Data Resources, respectively.

Despite their significance, accessing and integrating these datasets across platforms remains a challenge due to the diverse structures and formats. The DSMZ Digital Diversity initiative (https://hub.dsmz.de) addresses this challenge by providing a central hub for integrated data and establishing a framework for linking and accessing diverse datasets. At its core, this effort leverages the DSMZ Digital Diversity Ontology (D3O) to standardize and interconnect data, laying the foundation for a comprehensive knowledge graph that enables seamless interoperability and advanced exploration across multiple domains. The D3O aims to connect a vast diversity of datasets (Figure **??**) spanning microbial taxonomy, enzymology, microbial strains, cultivation media, ribosomal RNA sequences, and cell line data. By linking these domains, the ontology provides a unified framework for integrating information from all resources of DSMZ Digital Diversity. By adopting Semantic Web technologies, we aim to create a unified, queryable knowledge graph that fosters collaboration and innovation in biological research.

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
julia.koblitz@dsmz.de (J. Koblitz); lorenz.reimer@dsmz.de (L. C. Reimer)
0000-0002-7260-2129 (J. Koblitz); 0000-0002-7805-0660 (L. C. Reimer)

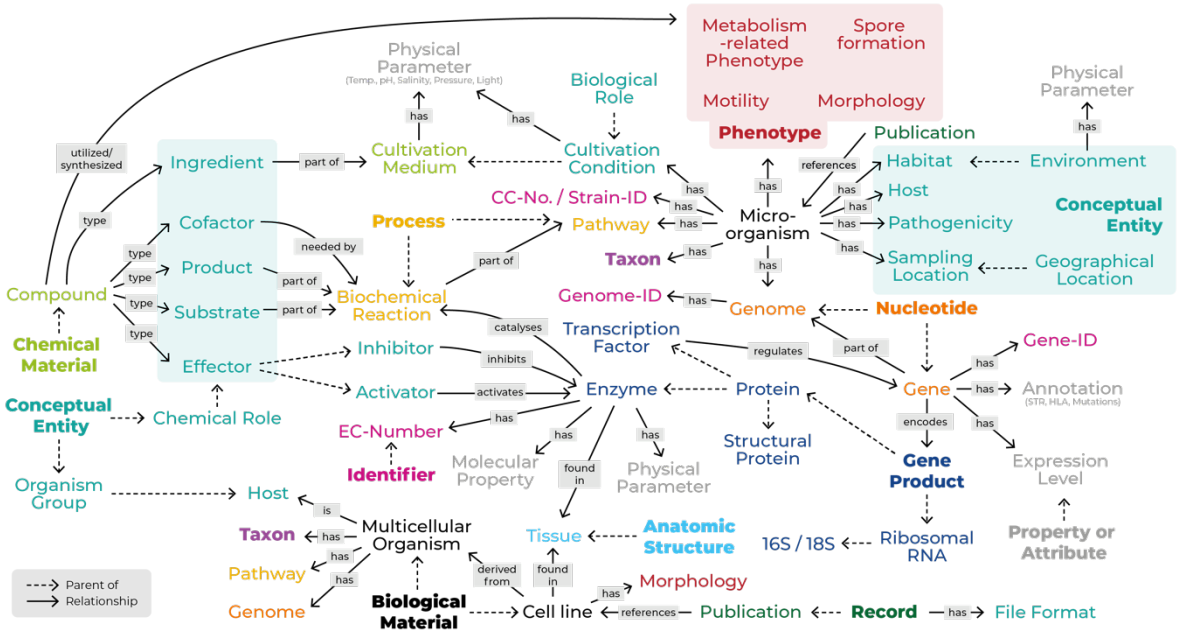

Figure 1: Overview of the DSMZ Digital Diversity Ontology (D3O). This diagram illustrates the interconnected entities within the databases of DSMZ, showcasing the relationships between chemical, biological, and conceptual entities. The highest-level entities are shown in bold and entities belonging to the same class are of the same color.

2. The first knowledge graphs of DSMZ Digital Diversity

The DSMZ Digital Diversity initiative is in its early stages, with the ontology and knowledge graphs being developed in parallel. This iterative approach allows us to refine both as new requirements and use cases arise. As a first result, we present the BacDive knowledge graph, available at https://sparql.dsmz.de/bacdive, which provides standardized access to high-quality, curated information about prokaryotic strains.

The BacDive knowledge graph currently contains over 16.5 million triples, representing 26 key entities such as Strain, CultureMedium, IsolationSource, and GenomeSequence. By leveraging the DSMZ Digital Diversity Ontology (D3O), these entities are semantically structured, facilitating interoperability with other data sources. Researchers can query BacDive directly using a SPARQL endpoint powered by the QLever query engine, enabling tailored searches and advanced data analysis.

In addition, the MediaDive knowledge graph, also accessible via the same interface, provides a complementary resource focusing on cultivation media. Together, these knowledge graphs represent the first steps toward a comprehensive RDF-based infrastructure that connects diverse datasets within the DSMZ Digital Diversity framework.

Future efforts will focus on extending the ontology, adding more datasets (e.g., BRENDA, LPSN, StrainInfo), and exploring opportunities for federated queries that integrate external data sources such as UniProt or PubChem. These developments aim to build a foundation for a more open and connected future of biological data.

Acknowledgments

We gratefully acknowledge the broader DSMZ Digital Diversity Team for their continuous support, valuable discussions, and contributions to the development of the data resources, infrastructure, and collaborative environment that made this work possible.

We appreciate the opportunity provided by BioHackathon Japan 2024 to advance our work on the

BacDive Knowledge Graph and would like to thank Jerven Bolleman and Daniel Fernández-Álvarez for their guidance on RDF best practices. We are grateful for funding through the following projects: Federal Ministry of Education and Research (BMBF) [de.NBI 021A539C]; Leibniz Association [SAW project DiAS-Pora, Funding No. K280/2019]; German Centre for Infection Research (DZIF) [8005512901, 8005512001]; Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [NFDI4Biodiversity; project number 442032008; NFDI 5/1, NFDI4Microbiota; project number 460129525; NFDI 28/1].

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] I. Schober, J. Koblitz, J. Sardà Carbasse, C. Ebeling, M. L. Schmidt, A. Podstawka, R. Gupta, V. Ilangovan, J. Chamanara, J. Overmann, L. C. Reimer, Bacdive in 2025: the core database for prokaryotic strain data, Nucleic Acids Research 53 (2024) D748–D756. doi:10.1093/nar/gkae959.

[2] J. Koblitz, P. Halama, S. Spring, V. Thiel, C. Baschien, R. Hahnke, M. Pester, J. Overmann, L. Reimer, Mediadive: the expert-curated cultivation media database, Nucleic Acids Research 51 (2022) D1531–D1538. doi:10.1093/nar/gkac803.

[3] J. P. Meier-Kolthoff, J. S. Carbasse, R. L. Peinado-Olarte, M. Göker, Tygs and lpsn: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes, Nucleic Acids Research 50 (2021) D801–D807. doi:10.1093/nar/gkab902.

[4] A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal, D. Jahn, D. Schomburg, Brenda, the elixir core data resource in 2021: new developments and updates, Nucleic Acids Research 49 (2020) D498–D508. doi:10.1093/nar/gkaa1025.

[5] C. Rolland, J. Wittmann, L. C. Reimer, J. Sardà Carbasse, I. Schober, C.-A. Dudek, C. Ebeling, J. Koblitz, B. Bunk, J. Overmann, Phagedive: the comprehensive strain database of prokaryotic viral diversity, Nucleic Acids Research 53 (2024) D819–D825. doi:10.1093/nar/gkae878.

[6] J. Koblitz, W. Dirks, S. Eberth, S. Nagel, L. Steenpass, C. Pommerenke, Dsmzcelldive: Diving into high-throughput cell line data, F1000Research 11 (2022). doi:10.12688/f1000research.111175.2.

[7] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, F. O. Glöckner, The silva ribosomal rna gene database project: improved data processing and web-based tools, Nucleic Acids Research 41 (2012) D590–D596. doi:10.1093/nar/gks1219.