BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37412
doi: 10.3897/biss.3.37412

OPEN ACCESS

Conference Abstract

From Data to Knowledge: A semantic knowledge graph application for curating specimen data

Peter Grobe[‡], Roman Baum[§], Philipp Bhatty[‡], Christian Köhler[‡], Sandra Meid[‡], Björn Quast[‡], Lars Vogt[§]

‡ Zoological Research Museum Alexander Koenig, Bonn, Germany
§ Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany

Corresponding author: Peter Grobe (p.grobe@leibniz-zfmk.de)

Received: 17 Jun 2019 | Published: 26 Jun 2019

Citation: Grobe P, Baum R, Bhatty P, Köhler C, Meid S, Quast B, Vogt L (2019) From Data to Knowledge: A semantic knowledge graph application for curating specimen data. Biodiversity Information Science and Standards 3: e37412. https://doi.org/10.3897/biss.3.37412

Abstract

The landscape of currently existing repositories of specimen data consists of isolated islands, with each applying its own underlying data model. Using standardized protocols such as DarwinCore or ABCD, specimen data and metadata are exchanged and published on web portals such as GBIF.

However, data models differ across repositories. This can lead to problems when comparing and integrating content from different systems. for example, in one system there is a field with the label 'determination', in another there is a field with the label 'taxonomic identification'. Both might refer to the same concepts of organism identification process (e.g., 'obi:organism identification assay'; http://purl.obolibrary.org/obo/OBI_0001624), but the intuitive meaning of the content is not clear and the understanding of the providers of the information might differ from that of the users. Without additional information, data integration across isolated repositories is thus difficult and error-prone. As a consequence, interoperability and retrievability of data across isolated repositories is difficult.

Linked Open Data (LOD) promises an improvement. URIs can be used for concepts that are ideally created and accepted by a community and that provide machine-readable meanings. LOD thereby supports transfer of data into information and then into knowledge, thus making the data FAIR (*F*indable, *A*ccessible, *I*nteroperable, *R*eusable; Wilkinson et al.

© Grobe P et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2016). Annotating specimen associated data with LOD, therefore, seems to be a promising approach to guarantee interoperability across different repositories. However, all currently used specimen collection management systems are based on relational database systems, which lack semantic transparency and thus do not provide easily accessible, machine-readable meanings for the terms used in their data models. As a consequence, transferring their data contents into an LOD framework may lead to loss or misinterpretation of information. This discrepancy between LOD and relational databases results from the lack of semantic transparency and machine-readability of data in relational databases.

Storing specimen collection data as semantic Knowledge Graphs provides semantic transparency and machine-readability of data. Semantic Knowledge Graphs are graphs that are based on the syntax of *'Subject – Property – Object'* of the Resource Description Framework (RDF). The *'Subject'* and *'Property'* position is taken by URIs and the *'Object'* position can be taken either by a URI or by a label or value. Since a given URI can take the *'Subject'* position in one RDF statement and the *'Object'* position in another RDF statement, several RDF statements can be connected to form a directed labeled graph, i.e. a semantic graph. Semantic Knowledge Graphs are graphs in which each described specimen and its parts and properties possess their own URI and thus can be individually referenced. These URIs are used to describe the respective specimen and its properties using the RDF syntax. Additional RDF statements specify the ontology class that each part and property instantiates. The reference to the URIs of the instantiated ontology classes guarantees the *F*indability, *I*nteroperability, and *R*eusability of information contained in semantic Knowledge Graphs. Specimen collection data contained in semantic Knowledge Graphs can be made *A*ccessible in a human-readable form through an interface and in a machine-readable form through a SPARQL endpoint (https://en.wikipedia.org/wiki/SPARQL). As a consequence, semantic Knowledge Graphs comply with the FAIR guiding principles. By using URIs for the semantic Knowledge Graph of each specimen in the collection, it is also available as LOD.

With semantic Morph·D·Base, we have implemented a prototype to this approach that is based on Semantic Programming. We present the prototype and discuss different aspects of how specimen collection data are handled. By using community created terminologies and standardized methods for the contents created (e.g. species identification) as well as URIs for each expression, we make the data and metadata semantically transparent and communicable.

The source code for Semantic Programming and for semantic Morph·D·Base is available from https://github.com/SemanticProgramming. The prototype of semantic Morph·D·Base can be accessed here: https://proto.morphdbase.de.

Keywords

semantic programming; FAIR data principle; knowledge graph application; collection data

Presenting author

Peter Grobe

Presented at

Biodiversity_Next 2019

Funding program

DFG LIS: Information Infrastructures for Research Data (http://gepris.dfg.de/gepris/projekt/248394582)

Grant title

eScience-Compliant Standards for Morphology

References

- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 https://doi.org/10.1038/sdata.2016.18