Let's curate together

Cornelia Fürstenau
Friedrich Schiller University Jena

Andreas Ostrowski
Friedrich Schiller University Jena

Birgitta König-Ries
Friedrich Schiller University Jena

Abstract

The strength of precise research data and concise metadata is widely recognized in all fields of science. Data curators are anxious to improve the quality of datasets to meet the evolving needs of the scientific community and the public. Furthermore, they want to train scientists to improve their skills in publishing own data in compliance with common rules such as the FAIR principles. However, this can be a very tedious process for both parties, data owner and data curators. Therefore, we want to integrate a quality feedback system into repositories to support the seamless communication between the data owner and curator and thus make improvements of datasets transparent and comprehensible for all stakeholders.

The development and implementation of these tools will be part of the further enhancement of the data management platform BEXIS2[1]. This open-source software was designed to manage research data in the fields of biodiversity and is used by different German institutions. The main objectives of BEXIS2 are to provide a supportive surrounding for data sharing and data reuse. It is designed modular and allows integration of customized modules.

The Biodiversity Exploratories Project[2] (BE) is a major biodiversity research initiative in Germany funded by the German Research Foundation since 2006. The BE sustains the scientific infrastructure and intellectual environment to address critical questions about changes in biodiversity. This project is in the migration process from the older data management software BExIS[3] to its successor BEXIS2.

Up to now, more than 100 projects uploaded data. Over the last years, the data curation has gained importance within the project. To further strengthen the curation processes, we are looking for new ways to make it more transparent, understandable and traceable for curators, data owners and to some extent also for data users. We envision the development of a set of three complementary tools to achieve this:

First, we pursue the semi-automated curation. This process includes the opportunity to specify constraints on the data when creating data structures. This information can be used for consistency checks of the data. Furthermore, a data-profiling component might generate statistics of the data and detect anomalies such as outliers or wrong units.

[1] http://bexis2.uni-jena.de
[2] http://www.biodiversity-exploratories.de/1/home/
[3] https://www.bexis.uni-jena.de

Second, as the central component, we aim to develop a curation management tool. The main objectives will be to document the curation process, e.g. status, changes and their reasons, and to develop a system indicating the quality of parts of the dataset, e.g. data structure, data or metadata. The requirements of such a system will be developed in cooperation between scientists and curators.

Third, we will create a dataset-feedback-system. This is a component to annotate and manage feedback on dataset quality issues identified during the curation process and given by dataset users. Additionally, the system can be used to compute dataset quality metrics.

We believe that together these tools will boost data quality inside our repository, facilitate data publication, and support the FAIR data initiative in science. The poster presents our conceptional framework for this new BEXIS2 module. We are looking forward to discuss our ideas and get inspired by data managers, curators, data owners, and data users. Let us get together and talk!