SciDataCon 2025

Contribution ID: **67** Type: **Presentation**

Supporting dataset curation through automation at KU Leuven

Tuesday 14 October 2025 12:36 (11 minutes)

KU Leuven RDR is the CoreTrustSeal certified institutional data repository of KU Leuven, where curation plays an important role in data FAIRification and ensuring the quality of published datasets. The curation phase is not only crucial to have some quality control on the FAIRness of the data by ensuring correct metadata input, the presence of documentation and a choice of license, but to also ensure that the researchers are fully informed and supported in their efforts to publish their data.

After the repository's launch in 2022, the monthly number of datasets published slowly increased overtime, and with that the number of dataset reviews to be carried out. As these numbers increased, it became clear that there was a need to better track the reviews and who is picking each review up, as well as a need to streamline this process in general. This would not only prevent unnecessary duplication of work, but would also potentially free up more time for support rather than the evaluation itself. To streamline the curation process, the RDR team developed an open-source review dashboard that plugs in to a Dataverse instance and automates different parts of the review process.

In the initial iteration of the dashboard, the automation focused on the administrative side of the reviews. For example, in the dashboard, reviewers can easily track who reviews what dataset, can add notes to any review and look back at the review history of said dataset. On top of that, the effort to streamline the feedback process resulted in the implementation of simple checklist in the review dashboard they can use to autogenerate feedback. This ensures uniformity in reviews, while still allowing for customizations, and prevents reviewers having to type the same feedback over and over again. This initial version of the dashboard was key to processing more datasets ready for publication and enabled reviewers to focus on the reviews themselves and not the administrative mess that previously came with it.

A second version of the review dashboard goes even a step further in its automation efforts. As the reviews were being carried out, some frequently made mistakes were flagged as having potential to be automatically found. With this idea an initial exploration began of what curation elements could all be automatically checked and how. From exploration, we found a lot of potential, such as indicating when a README file is likely missing, or when a README file is present, but empty. The list of potential automated checks was longer than expected and were easier to implement than we had anticipated. A bigger challenge, however, was to balance this automation with the human effort and input that is key in data curation. Some brainstorming on how to visualize this automation and how to always allow for human overwrites were necessary to ensure that the review supports human curation through automation and doesn't replace it.

In this presentation, we'll share our road to the creation of the review dashboard and a look at our UI, but also provide an insight into the logic of the automated checks. We hope to spark conversation on how to further support the human task of curation through tools and technology without losing the important human touch and interpretation that is so valuable to making a dataset as FAIR as possible.

Primary authors: BLOEMEN, Dieuwertje (KU Leuven); KARADENIZ, Ozgur (KU Leuven)

Presenter: BLOEMEN, Dieuwertje (KU Leuven)

Session Classification: Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI

Track Classification: SciDataCon2025 Specific Themes: Infrastructures to Support Data-Intensive Research - Local to Global