The FAIR Data Point Populator: collaborative FAIRification and population of FAIR Data Points

Daphne Wijnbergen[1,*,†], Rajaram Kaliyaperumal[1,†], Marco Roos[1] and Eleni Mina[1]

[1]*Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands*

Abstract

We created the FAIR Data Point Populator to facilitate the process of FAIRification. This tool reads metadata provided in a spreadsheet, creates RDF, and publishes these RDF documents to FAIR Data Points. We also improved interoperability and collaboration by building the tool as a GitHub workflow.

Keywords

FAIRification, collaboration, FAIR Data Point, metadata

1. Introduction

In order to allow data to be optimally reused, it is important that they are Findable, Accessible, Interoperable and Reusable (FAIR) for humans and machines. With the important role that the FAIR principles assign to metadata, a significant part of the process of increasing the level by which data follow the principles, named the FAIRification process, is dedicated to metadata. The FAIR Data Point (FDP) has been designed to serve as an example of how to publish metadata according to the FAIR principles. Although the reference implementation of the FDP provides a Web-based form for the users to enter their metadata values, many people are more comfortable using tools such as spreadsheets. To facilitate the publication of metadata in a FAIR-compliant manner, we created the FAIR Data Point Populator (FDPP), a tool that allows researchers with little or no expertise in FAIRification to create their own entry in a FDP. The tool uses spreadsheet software, GitHub repositories and GitHub workflows, which enable collaboration through their respective collaborative features. The FDPP extracts metadata from spreadsheets, converts this to RDF documents and publishes the metadata records to the target FDP. With this automation, we expect the FDPP to improve the ease of publication of metadata by non-technical users.

Although there are different tools available for metadata creation and editing, such as CEDAR and RightField, these are not integrated with FDPs. In contrast, the FDPP directly interacts with FDPs and makes the publishing of metadata to a FDP automatic. Furthermore, because the tool

is a GitHub workflow and uses a GitHub instance, it can be used without any compatibility issues or the need to install software. GitHub's features, such as version control, pull requests, reviews and comments, can be taken advantage of during FAIRification.

2. Implementation

At its core, the FDPP is a GitHub workflow that is built in Python. The user first fills in an Excel template with their metadata. This template guides less experienced users with documentation, tooltips and validation. Many users will benefit from using a tool that almost everyone is familiar with. This template can be filled by a group of people in online spreadsheet software such as Google Sheets and Microsoft 365 in order to make decisions together. The user then uploads the spreadsheet to a GitHub repository linked to the FDPP, where the administrator of that repository activates the GitHub workflow. The FDPP subsequently loads the tool from the main FDPP repository, creates RDF documents from the spreadsheet based on the FDP specification (which extends DCAT), and connects to a FDP to publish the RDF documents. The metadata is then available on the web within the FDP connected to the FDPP.

We also extended our tool to follow the metadata specifications of the European Joint Programme on Rare Diseases (EJP RD). This includes metadata for biobanks and patient registries. We created a FDP configuration that allows the FDP to validate and display metadata following the EJP RD specification. The software was tested in a workshop with data resource engineers to make their resource compliant with the specifications of the EJP RD 'Virtual Platform'.

3. Discussion

We created the FDPP, which aids FAIRification of metadata through ease of use, improved collaboration and integration with the FDP. The FDPP was tested in the rare disease community.

The FDPP lowers the barrier of entry for FAIRification, and because of that can accelerate the FAIRification of resources. Researchers only need to send in an Excel file to the administrator, or make a pull request with their metadata excel file. The administrator then needs to upload a file (or accept the pull request), check the contents, and start the FDPP workflow by clicking on the "run workflow" button within the GitHub repository.

In the future, the tool could be extended for use cases with other metadata schemas that implement FAIR principles for metadata. Providing users with a simple way to make their own templates according to their preferred schemas can be considered, and is already a feature that is offered by for instance CEDAR and Rightfield. However, this freedom can lead to more complexity for users, and templates that deviate from a chosen global standard such as DCAT.

Acknowledgments