Motivation

The SARA project (Software Archiving of Research Artefacts) aims to develop a new scientific service that allows long-term access and publishing of research data and scientific software. Its main focus is on software tools that support the processing and analysis of research data. In biological research, for example, measured data is collected and processed with the support of computers. The matching version of this software is required to faithfully reproduce the research results. Within Computer Science and Electrical Engineering, the different versions of newly developed software need to be continually stored in order to document the progress of development.

Architecture

This diagram shows the components of the SARA service and the dependencies between them. The Git Repo is a GitLab instance with which scientists work as usual and from which they can invoke the SARA service. At the user’s request, this initiates the staging of the working project in the Git Archive (a GitLab instance of the University of Konstanz, operated as a state-wide service) for long-term availability. In addition, it is also possible to add a citable bibliographical reference to the user’s Institutional Repository, so that the archived project can be found there and in Open Access platforms (e.g., Google Scholar, OpenAIR) indexing it. Two institutional repositories are currently available for user selection: KOPS (University of Konstanz) and OPARU (Ulm University), both based on the DSpace application.

All published projects receive a DOI (or a similar persistent identifier) for unambiguous and permanent referencing. Descriptive metadata is displayed on a landing page of the institutional repository. From there, users can navigate to the central Git Archive. If the publishing user has decided to publish the entire version history, the Git Archive not only allows users to download all intermediate versions, but also to explore and reconstruct the development history on-line. This architecture is based on the FORCE 11 “Software Citation Group” Software Citation Principles and other recommendations and best practices.

Metadata

We intend to use DOIs to reference the different versions of software artefacts published using the SARA service. The following metadata are captured and presented on the landing page of the Institutional Repository:

- Metadata required by DataCite: Identifier; Creator; Title; Publisher; Publication Year; Resource Type
- Other mandatory fields: Link to Git Archive; Person who Triggered the Publish Process
- The license is stored directly in the Git repo, in accordance with common practice in software engineering.

Furthermore, we investigated how metadata that are already available in Git and GitLab can be used for this purpose. We defined strategies for automatic extraction of relevant information to reduce manual entry. One of the challenges we face is the definition of author / contributor with respect to software and how to best represent them.