heiARCHIVE

COMPETENCE CENTRE FOR RESEARCH DATA

A service by URZ

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

heiARCHIVE, a long-term preservation service at Heidelberg University

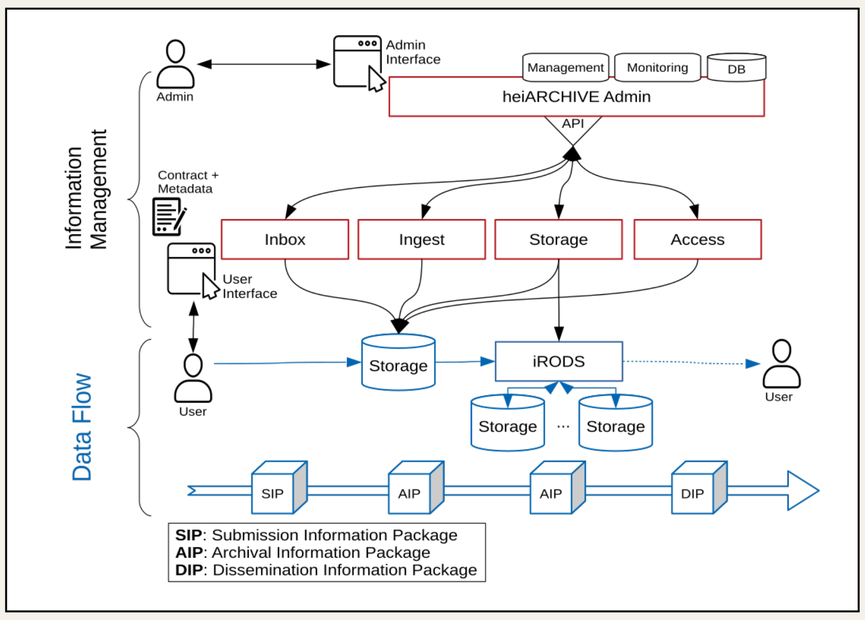
Martin Baumann[1], Florian Heß[2], Leonhard Maylein[2], Tatjana Mechler[1], Benjamin Scherbaum[1], Eric Volkmann[1]
(1) Heidelberg University Computing Centre, (2) Heidelberg University Library

A service of the Competence Centre for Research Data

The Competence Centre for Research Data is a joint facility of the university's Computing Centre (URZ) and the University Library (UB). In accordance with Heidelberg University's Research Data Policy it is **our mission to provide the best possible support for the comprehensive and coherent management of research data** for the university and its researchers. Please find further information at https://data.uni-heidelberg.de.

heiARCHIVE in a nutshell

heiARCHIVE is a new institutional service for long-term data preservation offering researchers an **easy-to-use end-user plattform** for archival of their research data. It is a **dark archive** and follows the concept of the **OAIS reference model** (Open Archival Information System). The existence of archived data can optionally be listed publicly in terms of a short description, e.g. to draw attention to the data (request and delivery is future work). heiARCHIVE is based on an in-house development and offers features like format recognition/validation and extraction of metadata from files. A storage abstraction is realized based on the open source data management software **iRODS** (https://irods.org) to manage data copies and geo replication and the **BagIt** file packaging format (RFC 8493) is used for structuring and naming directories and files. A dedicated right and role concept including a billing management is available. Through service-local identity management, also alumni can use the service and users will perspectively also be able to do authentification using their **ORCID** (https://orcid.org/).



Modular design and implementation

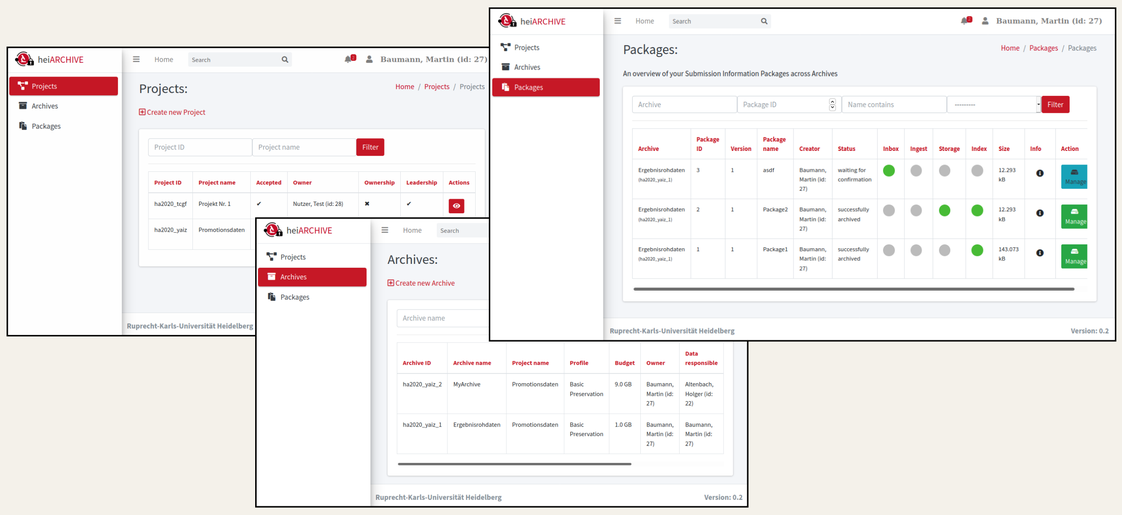
The modular design of heiARCHIVE follows the OAIS concept in implementing the data flow in terms of **SIP, AIP** and **DIP** (see figure) and introduces the **inbox** (prepare the data), **ingest** (package the data), **storage** (securely store the data) and **access** (to access the data). The different parts are conceptually separated and interact only through a dedicated **API**. The heiARCHIVE **admin** module controls the processing and state of all archive packages via the API. All parts can run on different (virtual) servers and can **scale** in compute ressources and network bandwith when adding additional (virtual) servers (helpful e.g. for intensive checksum operations). Today, data transfers from/to heiARCHIVE are realized via **SFTP**, but further protocols are intended. The implementation is done via **Python 3** and the high-level Python Web framework **Django** (https://www.djangoproject.com/).

Overall status

The main features of the software behind heiARCHIVE are implemented and the **archival and dissemination processes are running**. Extensive testing has been done to ensure the GUI and backend functions are running reliably.
Currently, the metadata model and the author's contract template are not finally defined and the access of the available tape library via iRODS is under investigation. Geo-replication will be realized before productive start.

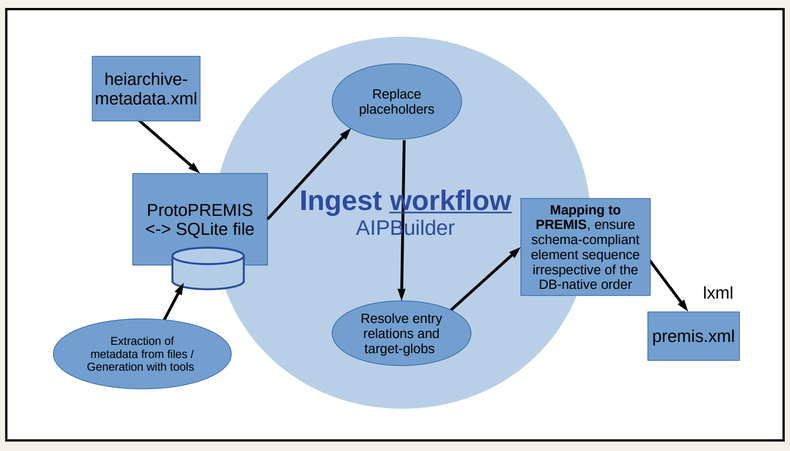
Archive- and role management

There is a hierarchical structure and a related role concept within heiARCHIVE. There highest-level structure is a **project** which can only be created by entitled persons (e.g. a professor). A project represents the link to a cost center ("Kostenstelle") and can also establish a quota.
A project can contain one or multiple **archives** each of which is related to a **data responsible person** (e.g. a PhD student). The archive defines a set of archival parameters (e.g. the archive mode to be either "long-term archiving" including format validation or simple "bitstream-preservation") that is set for all data contained within the archive. Descriptive metadata (e.g. title, short description, project context, etc.) can also be added.
The archive is the container for one or multiple **archive packages** – the structure which at the end contains the data to be archived. The data responsible person has the permission to upload data into an archive package, add metadata and start the packaging and storage process.



Metadata

For subsequent use of the preserved data and also for a description of the preservation process, a **minimal set of mandatory metadata** is stored in the heiARCHIVE database and index, but also within the AIPs. Some metadata is demanded from the user, e.g. the creator of the data and additional descriptive information. **Additional descriptive metadata** can be stored in a pre-defined location within the data package which might be considered by the indexer in the future. File formats are detected and - if needed - are validated as well.
During the ingest process, various tools are orchestrated and each yields metadata that is collected and stored in a **SQLite DB** to be finally read to write a METS/ PREMIS file in one single pass.



The **METS** standard (https://www.loc.gov/standards/mets/) is used to define a container for descriptive, administrative, and structural metadata. The **PREMIS** standard (https://www.loc.gov/standards/premis/) defines the metadata for the preservation of the data objects and their long-term usability. And, most likely, **DataCite** (https://schema.datacite.org/) will be used to represent the descriptive metadata.

https://heiarchive.uni-heidelberg.de
Website of the heiARCHIVE service.

Funded in parts by:
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST