



HAL
open science

D 4.3 - Workflow for New Services

Stefan Buddenbohm

► **To cite this version:**

Stefan Buddenbohm. D 4.3 - Workflow for New Services. [Research Report] DARIAH ERIC. 2019, pp.28. halshs-02415967

HAL Id: halshs-02415967

<https://shs.hal.science/halshs-02415967v1>

Submitted on 17 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



D4.3

Workflow for new services

DESIR

DARIAH ERIC Sustainability Refined

INFRADEV-03-2016-2017 - Individual support to ESFRI and other world-class research infrastructures

Grant Agreement no.: 731081

Date: 17-12-2019

Version: 1.0



DESIR has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731081.

| | |
|----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Grant Agreement no.: | 731081 |
| Programme: | Horizon 2020 |
| Project acronym: | DESIR |
| Project full title: | DARIAH-ERIC Sustainability Refined |
| Partners: | <p>DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES</p> <p>GEORG-AUGUST-UNIVERSITAET GOETTINGEN STIFTUNG OEFFENTLICHEN RECHTS</p> <p>UNIVERSITEIT GENT</p> <p>UNIWERSYTET WARSZAWSKI</p> <p>FACULDADE DE CIENCIAS SOCIAIS E HUMANAS DA UNIVERSIDADE NOVA DE LISBOA</p> <p>CENTAR ZA DIGITALNE HUMANISTICKE NAUKE</p> <p>GOTTFRIED WILHELM LEIBNIZ UNIVERSITAET HANNOVER</p> <p>INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE</p> <p>KING'S COLLEGE LONDON</p> <p>UNIVERSITY OF GLASGOW</p> <p>KNIHOVNA AV CR V. V. I.</p> <p>HELSINGIN YLIOPISTO</p> <p>SIB INSTITUT SUISSE DE BIOINFORMATIQUE</p> <p>UNIVERSIDAD NACIONAL DE EDUCACION A DISTANCIA</p> <p>UNIVERSITY OF HAIFA</p> <p>UNIVERSITY OF NEUCHÂTEL</p> |

| | |
|--------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| Topic: | INFRADEV-03-2016-2017 |
| Project Start Date: | 01-01-2017 |
| Project Duration: | 36 months |
| Title of the document: | Workflow for new services |
| Work Package title: | Technology |
| Estimated delivery date: | 31-12-2019 |
| Lead Beneficiary: | UGOE-SUB |
| Author(s): | Stefan Buddenbohm (buddenbohm@sub.uni-goettingen.de) |
| Quality Assessor(s): | Andrea Bertino [bertino@sub.uni-goettingen.de] Raisa Barthauer [barthauer@sub.uni-goettingen.de] Yoann Moranville [yoann.moranville@dariah.eu] |
| Keywords: | DARIAH, research infrastructure, sustainability, technology, technical reference, software quality |

Revision History

| Version | Date | Author | Beneficiary | Description |
|----------------|-------------|-------------------------------------------------------------|---------------------|----------------------------|
| 0.1 | 04.12.2019 | Stefan Buddenbohm | UGOE-SUB | Draft |
| 0.2 | 11.12.2019 | Andrea Bertino | UGOE-SUB | 1st quality assessment |
| 0.3 | 12.12.2019 | Stefan Buddenbohm | UGOE-SUB | 2nd revision after QA |
| 0.5 | 16.12.2019 | Stefan Buddenbohm Raisa Barthauer Yoann Moranville | UGOE-SUB, DARIAH | 3rd revision after QA |
| 1.0 | 17.12.2019 | Stefan Buddenbohm | UGOE-SUB | Version to be submitted |

Table of Contents

| | |
|-------------------------------------------------------------------------------------------------------------------------|-----------|
| Executive Summary | 6 |
| Introduction | 7 |
| 1. The Technical Reference | 8 |
| 2. Demonstrators for the Integration of New Services | 9 |
| 2.1 Demonstrators as tracks in the code sprints | 9 |
| 2.1.1 GROBID: Extraction of bibliographical data and citations from PDF..... | 10 |
| 2.1.2 BibSonomy: Import and export of bibliographical data and ingest in managed collections | 13 |
| 2.1.3 VisNow: Visualization of processed data with added dimensions for journals, topics, or dependency graphs | 15 |
| 2.1.4 Additional track: Securing Online Services in the DARIAH AAI using SAML/Shibboleth | 16 |
| 3. Second Code Sprint | 19 |
| 3.1 Tracks | 20 |
| 3.1.1 GROBID: Extraction of bibliographical data and citations from PDF applying GROBID..... | 20 |
| 3.1.2 BibSonomy: Automatic Import of Bibliographic Data..... | 21 |
| 3.1.3 VisNow/ViStory: Visualization of time dependent graphs of relation | 25 |
| 4. Summary of the DESIR Demonstrators | 27 |
| References | 28 |

Executive Summary

A fundamental basis of a successfully operating digital infrastructure such as DARIAH is formed by the services it provides to its users. In the particular case of the distributed setup DARIAH is using, the integration of new services requires support and guidelines that can be agreed to by all current and future service providers. Such generic guidelines can support individual research as well as new research projects just starting out, and – ideally – later enable the infrastructure to sustain their products.

By the development of three demonstrators (i.e. prototypical services) within WP4 the partners deliver an implementation of the above-mentioned guidelines and principles. The demonstrators delivered within WP4 rely on a common topic: bibliographical metadata. The demonstrators show the usage of tools for bibliographical metadata in various stages of the research process, e.g. extraction of entities, the collection and sorting of citations, visualisation of selected aspects of the data. They have been developed with the involvement of the community, applying already existing experiences and resources, and finally are open to be re-used beyond the project.

| Nature of the deliverable | | |
|---------------------------|--------|-------------------------------------------------------------------------------------|
| | R | Document, report |
| ✓ | DEM | Demonstrator, pilot, prototype |
| | DEC | Websites, patent fillings, videos, etc. |
| | OTHER | |
| Dissemination level | | |
| ✓ | P | Public |
| | CO | Confidential only for members of the consortium (including the Commission Services) |
| | EU-RES | Classified Information: RESTREINT UE (Commission Decision 2005/444/EC) |
| | EU-CON | Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC) |
| | EU-SEC | Classified Information: SECRET UE (Commission Decision 2005/444/EC) |

Disclaimer

DESIR has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731081. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Introduction

Through DESIR (DARIAH ERIC Sustainability Refined), DARIAH strives to define a roadmap for the sustainable operation of its services. This includes the administrative sustainability through a business plan as well as guidelines and best practices for the actual development and operation of the services, and finally, the deployment of three demonstrators as practical result of the work in the work package. The latter is addressed by work package 4 “Technology”, aiming to technologically enhance the DARIAH research infrastructure and its services.

Following up on previous activities, DESIR established the EURISE Network as forum to discuss and - ideally - harmonize guidelines for software and infrastructure development amongst the three ERICs CESSDA, CLARIN and DARIAH. The EURISE Network is now an ongoing activity beyond DESIR.

The second strand in this regard has been the work on three demonstrators concerning bibliographical metadata. Demonstrators in this context are prototypical services not intended for the sustainable, productive usage, even though at least one of the demonstrator will be used in production after the project and the result of another one is being integrated in the core code of the original tool.

The demonstrators concern the use of bibliographical metadata in various stages of the research process, e.g. extraction of entities, the collection and sorting of citations, visualization of selected aspect of the data. They have been developed with the involvement of the community, applying already existing experiences and resources, and finally are open to be re-used beyond the project.

| Demonstrator | Deployment | Documentation |
|--------------------|-----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Grobid | http://destracka.herokuapp.com/ | https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackA-TextMining |
| BibSonomy | https://grobid-biblio-bibsonomy.herokuapp.com | https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackB-BibliographicMetadata |
| VisNow/ ViStory | http://vistory.icm.edu.pl | https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackC-Visualization |

1. The Technical Reference

The DARIAH Technical Reference (TR) as predecessor activity to EURISE Network has been described in D4.2 Conceptual Models for the Integration of New Services and is now available through the EURISE Network's website: <https://technical-reference.readthedocs.io/en/latest/>. Initially being delivered within DESIR, the TR builds on a number of previous guidelines and best practices developed by several of the workshop attendants, in particular the CLARIAH Software Quality Guidelines (van Gompel et al 2016), the Netherlands eScience Center Guide (NLeSC 2018) and the CESSDA Software Maturity Levels (Shepherdson et al 2016). D4.2 Conceptual Models:

The purpose of the reference is to collect best practices and software development guidelines as well as quality checklists. It is a collection of best practices and guidelines for developers and maintainers of infrastructure components, as well as quality recommendations. They can be used to either gauge the quality of ongoing developments or as a starting point for new research projects, in particular when the components are planned to be attached to the infrastructures.

The TR is a reference in that it lists general requirements and considerations, but it does not always specify choices. In particular, it does not define licenses, technology stacks or hosting services. These are part of the implementation of the TR for a research infrastructure or a data centre. This approach was chosen, because it allows an increased compatibility with existing (internal) requirements on specific choices and it reduces the effect of the 'not invented here' syndrome¹.

The work of the EURISE Network is being continued beyond DESIR and open for involvement from other initiatives and institutions. The network is available through: <https://eurise-network.github.io/>

The according EURISE Network workshops are documented in detail in the WP4 section of the technical report of the DESIR project.

¹ https://en.wikipedia.org/wiki/Not_invented_here

2. Demonstrators for the Integration of New Services

The following chapter describes the development of the demonstrators along the two code sprints. Basically each code sprint was split up into tracks. The first event had four tracks: Grobid (track A), Bibsonomy (track B), VisNow (track C), and DARIAH AAI (track D). The second code sprint had three tracks: Grobid, Bibsonomy, and ViStory.

Please consult the following URLs to visit the demonstrators:

- GROBID: <http://destracka.herokuapp.com/>
- BibSonomy: <https://grobid-biblio-bibsonomy.herokuapp.com>
- VisNow/ViStory: <http://vistory.icm.edu.pl>

2.1 Demonstrators as tracks in the code sprints

Work package 4 “Technology” is tasked with utilizing the unique expertise of the three technology partners - ICM, INRIA LR3S - for the DARIAH infrastructure. Particularly the development of concepts and demonstrators for specific requirements of the DARIAH community stands in the center of this work package. Demonstrator in this context means a prototypical service or tool. The developments in WP4 centered on the topic of bibliographical metadata and have been pursued with look at the already existing DARIAH infrastructure and services landscape.

For this purpose, the partners have organized a code sprint revolving around bibliographical metadata. The code sprint took place from July 31st to August 2nd in the premises of the Institute of Library and Information Science of the Humboldt Universität zu Berlin. The event was open for everyone interested in programming for Digital Humanities use cases and was announced through DARIAH-EU and DH-affiliated channels. Although organized as part of the DESIR project, the event was branded and disseminated as DARIAH activity to gain more awareness for it and to brand it unmistakable as a Digital Humanities event. The results - and by this the work of the code sprint participants - were made available for DARIAH and will perspectival find a use within the DARIAH infrastructure. It was then planned to continue the code sprint activity with a second event in 2019.

As mentioned above the code sprint was an activity within the DESIR work package 4 “Technology”. It is embedded in the wider work plan of the WP aiming at delivering at least three demonstrators or concepts for services or applications until the project’s end.

Preliminary work before the code sprint resulted in the Gap Analysis of the DARIAH infrastructure (2017), a report investigating the DARIAH service and infrastructure landscape for gaps. Another strand of activity began with the identification of a suitable topic or grid for the code sprint. With look at the expertise of the technology partners the preparations soon

centered on the topic bibliographical data. The grid of the code sprint was split into four coding tracks, except for the track in AAI focusing on bibliographical data. The tracks and their results are described in detail below. The code sprint was opened by a keynote from Prof. Ralf Schenkel of Trier University and affiliated with the DBLP - The Digital Bibliography and Library Project focusing on computer science and related topics.



Figure 1 - Presentation of results from the GROBID track

33 participants with various backgrounds participated in the code sprint:

- 9 participants with DESIR affiliation
- 6 participants with a DARIAH affiliation
- 18 participants with neither DESIR nor DARIAH affiliation
- 17 participants with a German language background, 16 participants with a non-German language background
- 23 male participants, 10 female participants
- 31 participants with programming expertise (2 participants without from the wider project staff)

The affiliations of the participants were quite mixed with an emphasis on junior researchers and on programming affiliations.

2.1.1 GROBID: Extraction of bibliographical data and citations from PDF

As a machine learning library for extracting, parsing and re-structuring raw documents, such as PDF documents, into structured TEI-encoded ones, GROBID is a powerful tool that focuses on technical and scientific publications. For fully processing PDF documents, GROBID can manage 55 final labels used to build relatively fine-grained units ranging from traditional publication metadata to full text structures. Some of these metadata are title, author first/last/middle-name, affiliation type, detailed address, journal, volume, issue, and page.

Meanwhile, for the full text structures, it can be section title, paragraph, reference marker, head or foot note, figure captions.

With its first developments starting in 2008, GROBID has become a state-of-the-art (Lipinski: 2013; Tkaczyk: 2018) open source library for extracting metadata from technical and scientific documents in PDF format. Beyond simple bibliographical extraction tasks, the goal of the library is to reconstruct the logical structure of raw documents in order to enable large scale advanced digital library processes. For achieving this, GROBID explores a fully automated solution relying on machine learning (Linear Conditional Random Fields) models. The library is integrated today in various commercial and public scientific services such as ResearchGate, Mendeley, CERN Inspire and the HAL national publication repository in France. It is used on a daily basis by thousands of researchers and engineers. Since 2011, the library is open source under an Apache 2.0 license.

GROBID can be considered as a production-ready environment which includes a comprehensive web service API, a batch processing, a JAVA API, a generic evaluation framework, and the semi-automatic generation of training data. The GROBID Web API provides a simple and efficient way to use. For production and benchmarking, it's strongly recommended to use this web service mode on a multi-core machine and to avoid running GROBID in the batch mode.

In the scope of the code sprint workshop, track A proposed a hands-on session where users were guided through PDF data extraction and processing. The workshop was framed according to the skills available among the participants. In order to be able to follow this track, it was suggested that the participants should have some preliminary knowledge, especially in Java, Python or JavaScript and the abilities to communicate with several web services via HTTP.

The session covered the following topics (the tasks were sorted by priority, but however they were tackled depending on skills, time and interest of participants during the workshop):

- Extraction of citation data from scientific PDF documents. Required skills for these steps were Java/Python, JavaScript, HTTP, and XML/JSON.
- Visualization of extracted information using GROBID extraction services directly on the PDF documents, i.e. highlighting authors, title, tables, figures, and keywords. Required skills for these steps were Java/Python, JavaScript, HTML, XML/JSON.
- Enhancement of basic information by accessing some other external services, e.g. affiliation disambiguation, GPS coordinates concept disambiguation. Required skills for these steps were Java/Python, JavaScript, HTML, XML/JSON.

- Creation of enhanced view of PDF documents as results of combining all data extracted in previous tasks in order to produce a usable viewer. Required skills for these steps were JavaScript, HTTP.

The goal of the workshop of track A was to EXTRACT PDF documents into XML-TEI format, to ENRICH information gained from the extraction process by accessing some other web services and to VISUALISE the results collected in PDF scientific article documents.

Firstly, the participants were asked to extract the scientific PDF documents which were already prepared in five languages (English, French, German, Italian, Spanish) into TEI-XML format in order to get some important information (e.g. title, authors, abstract, keywords, tables, figures).

Based on the results in TEI-XML format, the participants were asked to visualise the extracted results in PDF documents by highlighting them i.e. to highlight the title, the authors, the abstract, the keywords in PDF documents. The participants could choose their preferred development languages, e.g. Java, Python, JavaScript for this step and further steps.

As a need to enrich the information gained from Grobid's extraction process, the participants were also asked to add some more information by accessing other external services, e.g. HAL, Entity Fishing. The last activity of track A was the development of an enhanced view of PDF documents. For this purpose, the participants were asked to develop a new tool by using their preferred development tools to produce a usable viewer. As results for track A, two prototypes have been developed, which in principle perform all steps in this track but on two different platforms, Java and Python.



Figure 2 - Introduction into VisNow by Bartosz Borucki

All codes and files of this workshop can be accessed via the GitHub repository for the tracks: <https://github.com/DESIR-CodeSprint/TrackA-TextMining>.

About 11 participants were involved in track A. They were then split into two groups concerning their basic skills, whether in Python or in Java. Since the results of this workshop were still in the prototype version, the future plan was to develop the final version of PDF document viewer. This tool will point out a number of important information in scientific PDF documents.



Figure 3 - Frank Fischer presents a demonstrator for the import of literary networks data into VisNow

2.1.2 BibSonomy: Import and export of bibliographical data and ingest in managed collections

DH researchers can benefit from a broad overview on scholarly publications relevant for their work. Thus, a bibliography of DH literature can contribute to the well-being of the discipline. For computer science, DBLP² is the de facto standard, easily allowing researchers to see who has contributed to the development of the field. Building such a great resource is a big

² <http://dblp.uni-trier.de/>

achievement. The participants of track B aimed at taking the first steps towards a DH bibliography: enabling an easy-to-use web application to import and export bibliographical metadata for the digital humanities. The goal, of course, was not to reinvent the wheel, as tools like Zotero, BibSonomy, etc. already exist. Instead, the participants focused on the simplification of data entry, e.g., by enabling import from ORCID or via drag'n'drop from PDF files (using technology developed in track A), and the use of BibSonomy³ as a backend for storing and organizing literature references. With its REST API⁴ it enabled collaborative storage and retrieval of bibliographical metadata. The choice of programming language and frameworks was adjusted with the capacity of the participants (e.g. Python, Java). Experience in web programming, particularly using web APIs and frameworks was a prerequisite.

A tool has been built for extracting bibliographical metadata from PDF files using GROBID and storing it in BibSonomy. This way bibliographical metadata can be easily added to BibSonomy with low effort. The tool comes with a user-friendly interface. The full Java code and an installation guide is published on GitHub: <https://github.com/DESIR-CodeSprint/trackB>.

The tool was created by 6 participants from different areas (mainly computer scientists). The participants split the main task into subtasks, following the Model-View-Controller pattern, to enable parallel working and to assure that every participant's expertise is best used. There were no specific plans for a possible further development in the future, but the tool could be extended, e.g. using authorization with ORCID.

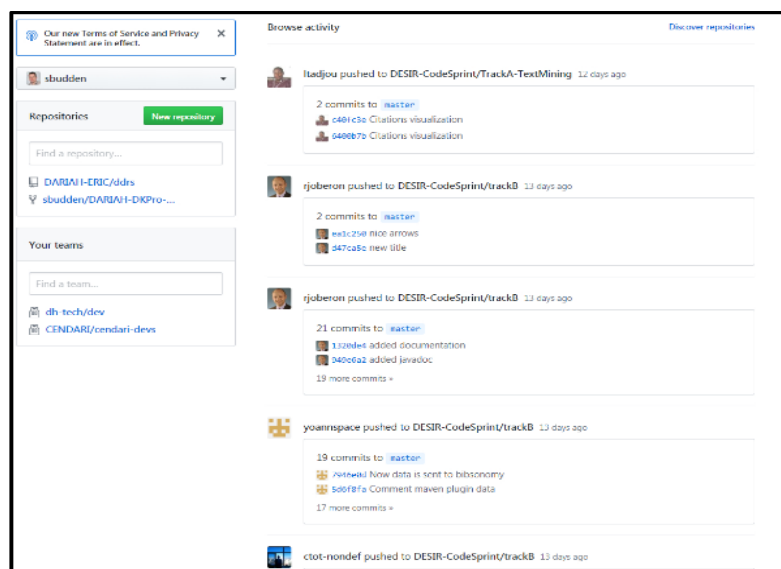


Figure 4 - All participants of the event used GitHub as platform for the exchange of results between the tracks

³ <http://dev.bibsonomy.org/>

⁴ https://bitbucket.org/bibsonomy/bibsonomy/wiki/documentation/api/REST_API

2.1.3 VisNow: Visualization of processed data with added dimensions for journals, topics, or dependency graphs

The visualization of data and results gains more and more importance as natural component of the research cycle. In DH applications most of the visualization focus is around so called information visualization - graphical approaches showing usually high-dimensional and unstructured data with structure representation, revealing hidden structure or its internal relations, usually by means of graphs, charts, maps, etc. Although a number of information visualization toolkits and services exists, many approaches and tools from scientific visualization may be applied to amplify cognition, especially for 3D or 4D interaction.

The task of this track during the code sprint was at least twofold. On one hand, to elaborate specific visualization means on the boundary of infovis and scivis for bibliographical data (e.g. author networks with additional dimensions for e.g. journals, topics or dependency graphs). On the other hand, the track was to conceptualize specific services that fit into the current DARIAH infrastructure landscape and with the preconditions provided by the other tracks in the code sprint, e.g. using data from BibSonomy. Existing building components of the generic visualization framework VisNow (<http://visnow.icm.edu.pl>) were used combined with web frameworks. The prototype web frontend for 3D graph visualization was extended by adding ego-centered view for nodes (representing authors) and adjacent edges (representing publications with other authors). The 3D interaction concepts were redesigned and example 2D maps were created. A number of expansions was implemented and tested in the 3D interaction part of the web frontend in order to work out the interaction schemes between the user and a 3D graph visualization. Data import codes were created for interaction with Bibsonomy data export files and Bibsonomy API. Modifications of backend data structuring for graph creation was tested with additional data processing and sorting layer in the backend. Additional 2D visualization was introduced on frontend side using high-level descriptive language Vega-Lite. One of the tasks covered the problem of importing literary network data into VisNow (case study "Hamlet"). Requirements for corresponding import and processing modules were defined and example visualizations prepared. The attempt was made on importing Grobid dictionaries into VisNow and prototype visualizations were created. New problems and concepts were defined on multidirectional graphs visualization. Another use case was conceptualized and tested based on the graph data from Italian music relations and geospatial information. The concept covered the relation between graph and spatial (map) visualization.



Figure 5 - Visualisation of publications data

8 participants with various backgrounds took part in track C, from computer scientists, up to digital humanities scientists. As the choice of tasks was also spreading from technical to applications, the participants were given the opportunity to either develop proof of concept technical solutions, or work on use case scenarios and practical usage.

The concepts and codes prototyped during the code sprint are laying foundations for the proof of concept services to be developed by the end of the project. The participants planned to use the outcomes as inspiration for the ongoing work. Both the functionality and concepts are projected on VisNow application and the planned services. Based on the use cases the plan was to broaden the DH planned application areas.

2.1.4 Additional track: Securing Online Services in the DARIAH AAI using SAML/Shibboleth

Researchers who want to share their online services within DARIAH can take advantage of the DARIAH Authentication and Authorization Infrastructure (AAI). The DARIAH AAI enables researchers from eduGAIN⁵ to access DARIAH services, by using the interoperable SAML

⁵ <https://technical.edugain.org/status>

standard⁶. Users can log in at their home institution, without the need to create accounts and remember passwords for the online services they want to access. Adding to this, the DARIAH AAI allows for central yet distributed management of group memberships. Thus, DARIAH online services can base their authorization decisions on these memberships.

Set into production only some weeks before the workshop, the DARIAH AAI had still lowered the barrier to connect services to DARIAH, by introducing a central AAI proxy between all DARIAH services and all eduGAIN institutions.

Key features of the AAI proxy are:

- Almost any SAML Service Provider (SP) library can be used in an application
- No registration of the SP in a federation needed anymore - just exchange SAML metadata with the proxy
- The AAI proxy ensures Identity Provider (IdP) Discovery and the connection to eduGAIN
- It supplies a service with all IdP attributes, plus information from the central DARIAH directory
- It handles user registration and terms of use approval

The proxy took over many tasks that services needed to implement previously, which now makes it much easier to connect new services.

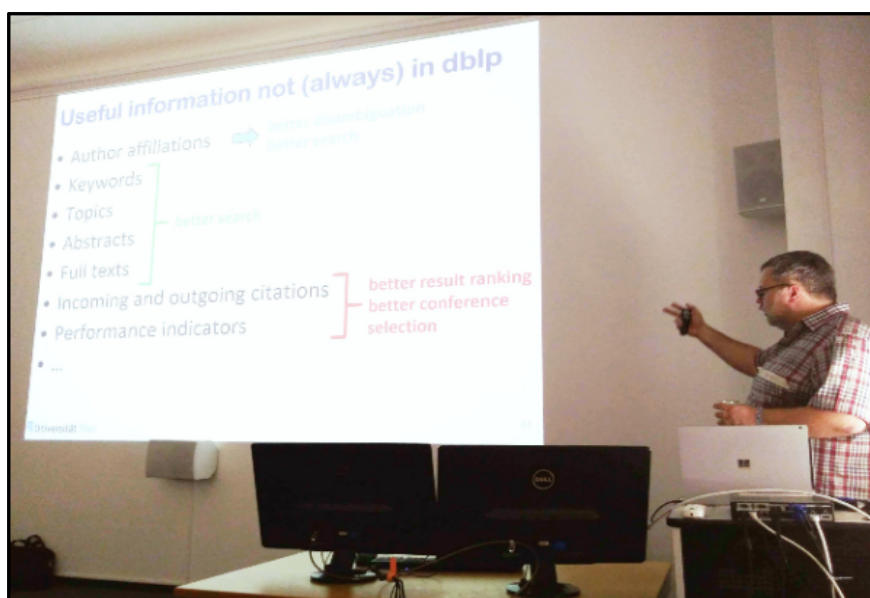


Figure 6 - Ralf Schenkel gives the keynote on DLBP

⁶ <https://www.oasis-open.org/committees/security/>

Track D introduced the DARIAH AAI including the new proxy model, and enabled its participants to install, configure and test the Shibboleth Service Provider (SP) to integrate with an online service. The goal was to make the participants familiar with the Shibboleth SP and how it integrates with their Web application. The workshop also provided for an introduction to SAML from an SP side and gave recommendations for further open-source SP implementations, and a comparison with other AAI technologies like OAuth2 and OpenID Connect.

Participants of the workshop gained a deeper understanding of the SAML standard, and on how to install and configure a Shibboleth SP to protect their online service in an interoperable way. Two test online services, and one online service that is now in a production state could be connected to the DARIAH AAI.

Documentation of the Workshop is available at <https://github.com/DESIR-CodeSprint/TrackD-AAI>, whereas an always-updated documentation of the DARIAH AAI is available at <https://wiki.de.dariah.eu/display/publicde/DARIAH+AAI+Documentation>.

This track was attended by 4 participants: one PhD student in Computer Science, and three scientific staff members, all affiliated to German research institutions. As the DARIAH AAI was running in a production mode, efforts were on the way to promote it further such that many DARIAH services can take advantage of it. The FIM4D working group (Federated Identity Management for DARIAH) is promoting this. The follow-up FIM4D Workshop called DARIAH AAI NG Service Provider Workshop took place on January 21/22, 2019, in Tübingen, Germany, see <https://wiki.de.dariah.eu/x/9nPfAw> or <https://wiki.de.dariah.eu/display/publicde/DARIAH+AAI+NG+Service+Provider+Workshop+2019>.



Figure 7 - Participants of the Code Sprint

3. Second Code Sprint

The second Code Sprint took place from September 24th to 26th 2019 in Berlin and aimed at two activities: (1) Continue the work along the three tracks of the first code sprint and - together with the undertaken work in between the two events - to finish it. As result three functional demonstrators should be available for the interested public along with the necessary documentation. (2) The second activity aimed at knowledge transfer and training of interested young researchers with a Digital Humanities affiliation.

As said above the code sprint took up again the motto of bibliographical metadata. It was not exclusively intended for participants of the first code sprint but open for everyone interested in programming for Digital Humanities use cases. The code sprint brought together interested developers and DH-affiliated people, not only from the DARIAH community. The work was again divided into three tracks which all ran in parallel and in which the demonstrators developed in the last code sprint were improved and enhanced.



Figure 8 - Participants of the code sprint on the last day

The uptake of this event was not as strong as with the first code sprint. We had 25 registrations including the project partners. The participants came from 5 countries - Germany, Poland, Czech Republic, Croatia, France - with backgrounds from the academics and libraries. After an introduction into the DARIAH and DESIR projects and a wrap up of the first code sprint, the participants introduced themselves and decided how to structure the work for the coming 2,5 days. The first afternoon was devoted to creating a common work base and the last morning to the presentation and discussion of the results.

3.1 Tracks

3.1.1 GROBID: Extraction of bibliographical data and citations from PDF applying GROBID

In this track the participants worked on an enrichment of the GROBID functionalities by adding an acknowledgment parsing service both for raw texts and pdf files. The results of the parser would be in XML/TEI format. The participants were asked to:

- Annotate a new corpus containing acknowledgment sections of ~ 3500 scientific articles in Open Access;
- Build a new model for parsing acknowledgement both in form of raw texts pdf files with Grobid (<https://github.com/kermitt2/grobid>) and DeLFT (<https://github.com/kermitt2/delft/>);
- Create an acknowledgment Web service in Grobid.
- Integrate the results of Grobid acknowledgment parser into a demonstrator of track A (<https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackA-TextMining>).



Figure 9 - Poster of the second Code Sprint



Figure 10 - Introduction into the tracks

DESIR

INFRADEV-03-2016-2017 - Individual support to ESFRI and other world-class research infrastructures, Grant Agreement no. 731081.

3.1.2 BibSonomy: Automatic Import of Bibliographic Data

The aim of this track was to extend the tool for automatic import of bibliographical metadata into BibSonomy by adding further features. In addition to the pdf file upload, two new ways of submitting data have been implemented, i.e. the option of uploading text files and the option of submitting text directly in the browser using a text field. This option was chosen to allow users to copy and paste text parts from other sources directly. Furthermore, the online version of the tool was provided with an individual user login for BibSonomy, so that users can add bibliographical items to their own BibSonomy accounts. Finally, the user interface has been improved by adding new useful features, e.g. the removal of specific items from the list of extracted bibliographical items. These improvements provide an easy-to-use way for researchers, especially from the Digital Humanities community, to store and share bibliographical data in a common place.



Figure 11 - The GROBID track

The online version of the track B tool is available at:

<https://grobid-biblio-bibsonomy.herokuapp.com>

The following figures depict the updated workflow of the tool. First, users can upload a PDF or TXT file or can copy and paste text directly into the web interface. Selecting “Submit Text” will send the content to GROBID where any bibliographical data will be extracted.

Figure 12 - Updated workflow of the tool: PDF upload interface

All bibliographical items that have been extracted by GROBID are shown to the user. The user can now modify or delete entries from the list. Any modifications are saved directly. The tool provides a “status” field where a warning for missing information is shown.

| booktitle ↑ | entryType | year | status | edit |
|-------------------------------------------------------------------------------------------------------------|-----------|------|---------------------------------------------------------|------|
| Die Infrastruktur-Angebote von DARIAH-DE und TextGrid | article | 2015 | Item missing ["editors", "tags"]. | |
| Embedded Data Manager -integriertes Forschungsdatenmanagement: Praxis, Perspektiven und Potentiale | misc | 2015 | Item missing ["editors", "tags"]. | |
| | misc | 2013 | Item missing ["title", "authors", "editors", "tags"]. | |
| Die Konzeption eines nationalen Forschungsdatenzentrums für die Archäologie und die Altertumswissenschaften | misc | 2013 | Item missing ["editors", "tags"]. | |
| Juristische Handreichung für die Geisteswissenschaften. DARIAH-DE Working Papers 12 | misc | 2015 | Item missing ["editors", "tags"]. | |

Rows per page: 5 1-5 of 20

Please log in to submit items to BibSonomy

Figure 13 - Updated workflow of the tool: entry modification after upload

After all modifications are done, the user can upload the bibliographical items to their BibSonomy account. Therefore the user must first login to BibSonomy, using their BibSonomy username and API Key.

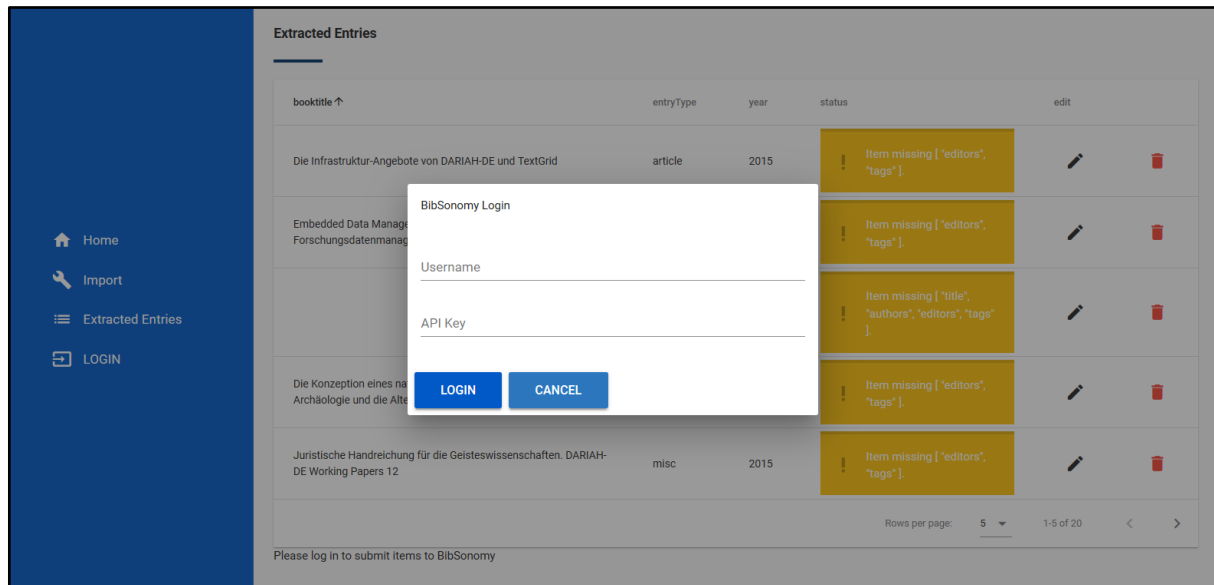


Figure 14 - Updated workflow of the tool: Login interface with API key

If the login is successful the user will be informed and the BibSonomy username appears in the upper left corner. The user can now upload all bibliographical items in the list by selecting "Submit to BibSonomy". The tool informs the user whether the submission was successful.

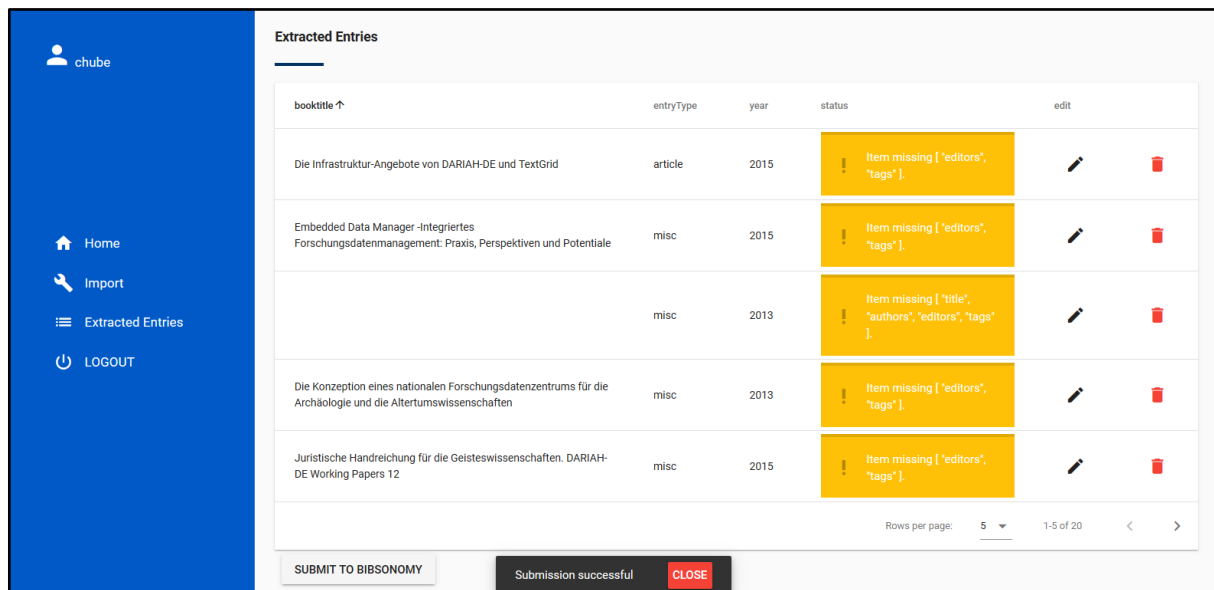


Figure 15 - Updated workflow of the tool: submission to BibSonomy

Submitted items appear in BibSonomy:



Figure 16 - Updated workflow of the tool: entries appear in the individual BibSonomy instance of the user

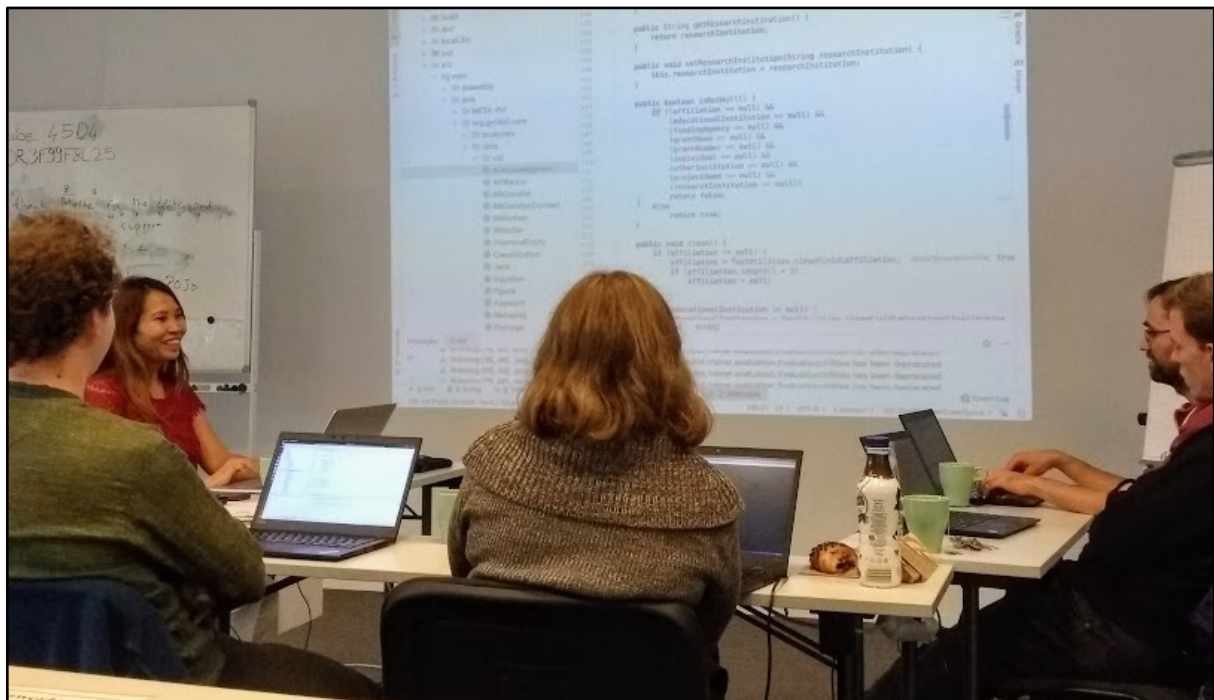


Figure 17 - Participants of the Code Sprint

DESIR

INFRADEV-03-2016-2017 - Individual support to ESFRI and other world-class research infrastructures, Grant Agreement no. 731081.



3.1.3 VisNow/ViStory: Visualization of time dependent graphs of relation

The participants of the track on *Visualization of time dependent graphs of relations* focussed on the extension of the tool developed at the last code sprint both towards new data formats and use cases, as well as new visual forms. The participants had the chance to work on the mapping of different data to the generic model of the graphs and on the translation of data formats to intermediate RDF description.

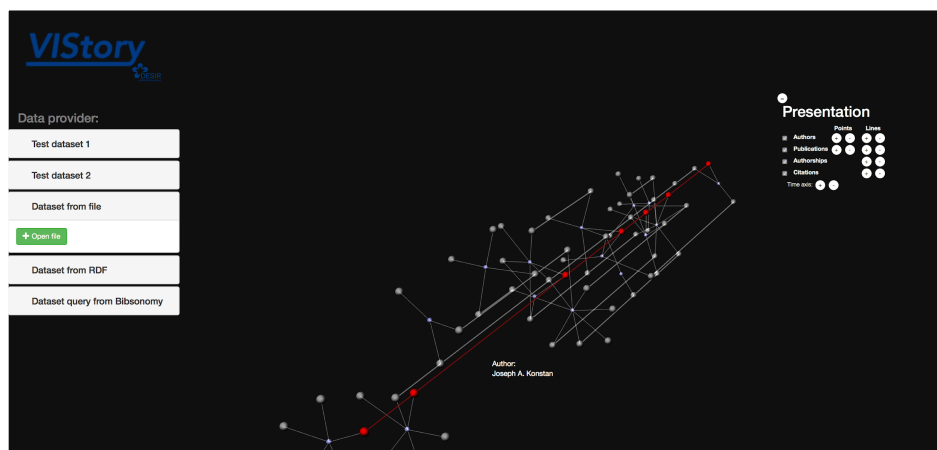


Figure 18 - User interface of ViStory

Some notes on what was actually done during the Code Sprint:

- Implementation of a function to visualize a graph showing co-authorship relations as together with the timeline. Before only the visualization of the co-authorship graph was possible.
- The frontend for the user has been improved to enhance the usability of the tool and the “readability” of the visualized data.
- Some improvement has been done on the used internal model created by ICM.
- A mapping from RDF model to the internal model has been created.
- The RDF model has been extended with regard to complexity.
- Mapping of the new model to the frontend automatically and not hard-coded anymore

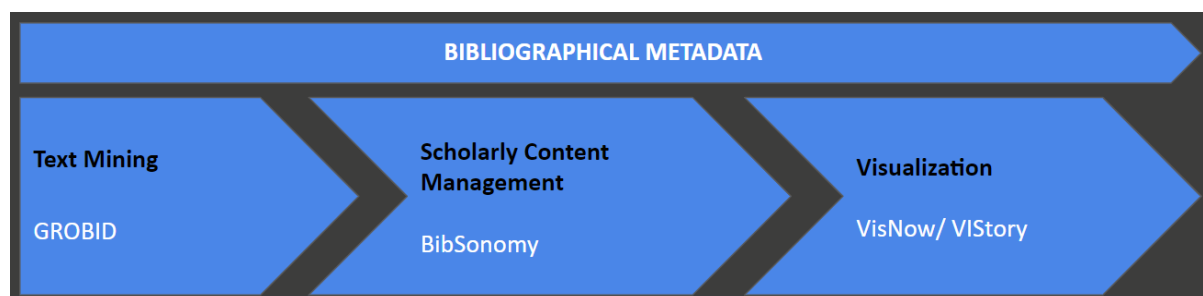


Figure 19 - Participants working on ViStory



Figure 20 - Wrap-up session

4. Summary of the DESIR Demonstrators



The work on the demonstrators developed within DESIR WP4 has been described above along the two code sprints. The code sprints can be seen as focal points of the work as they presented the work to a wider public audience and were used to work collaboratively on the services.

It is intended to deploy all three demonstrators on a common website which will be hosted and maintained by the CLARIAH-DE project. This will be finished in early 2020. The table below lists the current deployment URLs which allow the user to test the demonstrators. The documentation and the code is publicly available on the DARIAH ERIC GitHub repository. The re-use of the demonstrators by third parties is encouraged.

| Demonstrator/ Service | Deployment | Documentation |
|--------------------------|-----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Grobid | http://destracka.herokuapp.com/ | https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackA-TextMining |
| BibSonomy | https://grobid-biblio-bibsonomy.herokuapp.com | https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackB-BibliographicMetadata |
| VisNow/ViStory | http://vistory.icm.edu.pl | https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackC-Visualization |

References

- DARIAH ERIC main GitHub repository: <https://github.com/DARIAH-ERIC>
- DARIAH ERIC repository for the BibSonomy track (bibliographical metadata), maintained by Christoph Hube: <https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackB-BibliographicMetadata>
- DARIAH ERIC repository for the Grobid track (text mining), maintained by Tanti Kristanti: <https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackA-TextMining>
- DARIAH ERIC repository for the VisNow track (visualization), maintained by Bartosz Borucki: <https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackC-Visualization>
- Stefan Buddenbohm, Raisa Barthauer. 2017. D4.1 - Gap Analysis of DARIAH Research Infrastructure, <https://hal.archives-ouvertes.fr/hal-01663594>
- Maarten van Gompel, Jauco Noordzij, Reinier de Valk, Andrea Scharnhorst. 2016. CLARIAH Software Quality Guidelines, version 1.1, September 30, 2016, available at <https://github.com/CLARIAH/software-quality-guidelines>
- NLeSC, 2108, Netherlands eScience Center Guide, Available at <https://guide.esciencecenter.nl/>
- John Shepherdson, Ørnulf Risnes, Mike Priddy, Matti Heinonen, Johan Finn, Wolfgang Zenk-Möltgen. 2017. CESSDA Software Maturity Levels (v1.0) available at <https://drive.google.com/file/d/0Bwk0RK5TDo6iTWxrTXFGSG5tMzg/view>
- Markus Matoni, Stefan Schmunk, Carsten Thiel. 2017. Basic DARIAH Services and Demonstrators <https://hal.archives-ouvertes.fr/hal-01663594>
- Carsten Thiel. 2017. Workshop „Software Sustainability: Quality and Re-usability“, available at <https://dhd-blog.org/?p=8685>
- Carsten Thiel, Michelle Rodzis, Yoann Moranville. 2018. DARIAH Technical Reference, available at <https://dariah-eric.github.io/technical-reference/>