DESIR Code Sprint 1 DESIR (DARIAH) WP4 T4.4

Stefan Buddenbohm, Raisa Barthauer

WP4 T4.4 Code Sprint 1

DESIR

DARIAH ERIC Sustainability Refined

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

Date: 30-09-2018
Version: 1.0

DESIR has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731081. The EU is however not participating as a contracting authority in this procurement.

Grant Agreement no.:	731081
Programme:	Horizon 2020
Project acronym:	DESIR
Project full title:	DARIAH-ERIC Sustainability Refined
Partners:	DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES GEORG-AUGUST-UNIVERSITAET GOETTINGEN STIFTUNG OEFFENTLICHEN RECHTS UNIVERSITEIT GENT UNIWERSYTET WARSZAWSKI FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS DA UNIVERSIDADE NOVA DE LISBOA CENTAR ZA DIGITALNE HUMANISTICKE NAUKE GOTTFRIED WILHELM LEIBNIZ UNIVERSITAET HANNOVER INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE KING'S COLLEGE LONDON UNIVERSITY OF GLASGOW KNIHOVNA AV CR V. V. I. HELSINGIN YLIOPISTO SIB INSTITUT SUISSE DE BIOINFORMATIQUE UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA UNIVERSITY OF HAIFA UNIVERSITY OF NEUCHATEL
Topic:	INFRADEV-03-2016-2017
Project Start Date:	01-01-2017
Project Duration:	36 months
Title of the document:	WP4 T4.4 Code Sprint 1 (not required according to description of work)
Work Package title:	WP4 Technology
Estimated delivery date:	August 2018
Lead Beneficiary:	UGOE-SUB
Author(s):	Raisa Barthauer (barthauer@sub.uni-goettingen.de) Stefan Buddenbohm (buddenbohm@sub.uni-goettingen.de)
Quality Assessor(s):	-
Keywords:	DARIAH, research infrastructure, sustainability, technology, technical reference, software quality, code sprint

Revision history

Version	Date	Author	Beneficiary	Description
0.1	08.08.2018	Stefan Buddenbohm, Raisa Barthauer	UGOE	First Draft
0.2	25.09.2018	WP4 partners for their track results	All partners	Second Draft
1.0	26.09.2018	Stefan Buddenbohm	UGOE	Final Version

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.

Nature of the deliverable		
✓	R	Document, report
	DEM	Demonstrator, pilot, prototype
	DEC	Websites, patent filings, 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

The DESIR project is funded by the European Commission under the Horizon 2020 programme. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

DESIR WP4 Code Sprint July 31-August 2, 2018, Berlin - Report

The DESIR project sets out to strengthen the sustainability of DARIAH and firmly establish it as a long-term leader and partner within arts and humanities communities. DESIR will widen the DARIAH research infrastructure in three areas, vital for DARIAH's long-term sustainability: entity-based search, scholarly content management, visualisation and text analytic services.

General summary

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.
For this purpose the technology partners and UGOE-SUB as work package lead 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 had been announced through DARIAH-EU and DH-affiliated channels. Although organised 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 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 is planned to continue the code sprint activity with a second event in 2019.

Organisation

As mentioned above the code sprint is 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 end.
Preliminary work before the code sprint resulted in the Gap Analysis of the DARIAH infrastructure[1] (2017), a report investigating the DARIAH service and infrastructure landscape for gaps. Another strang 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

[1] https://halshs.archives-ouvertes.fr/hal-01663594

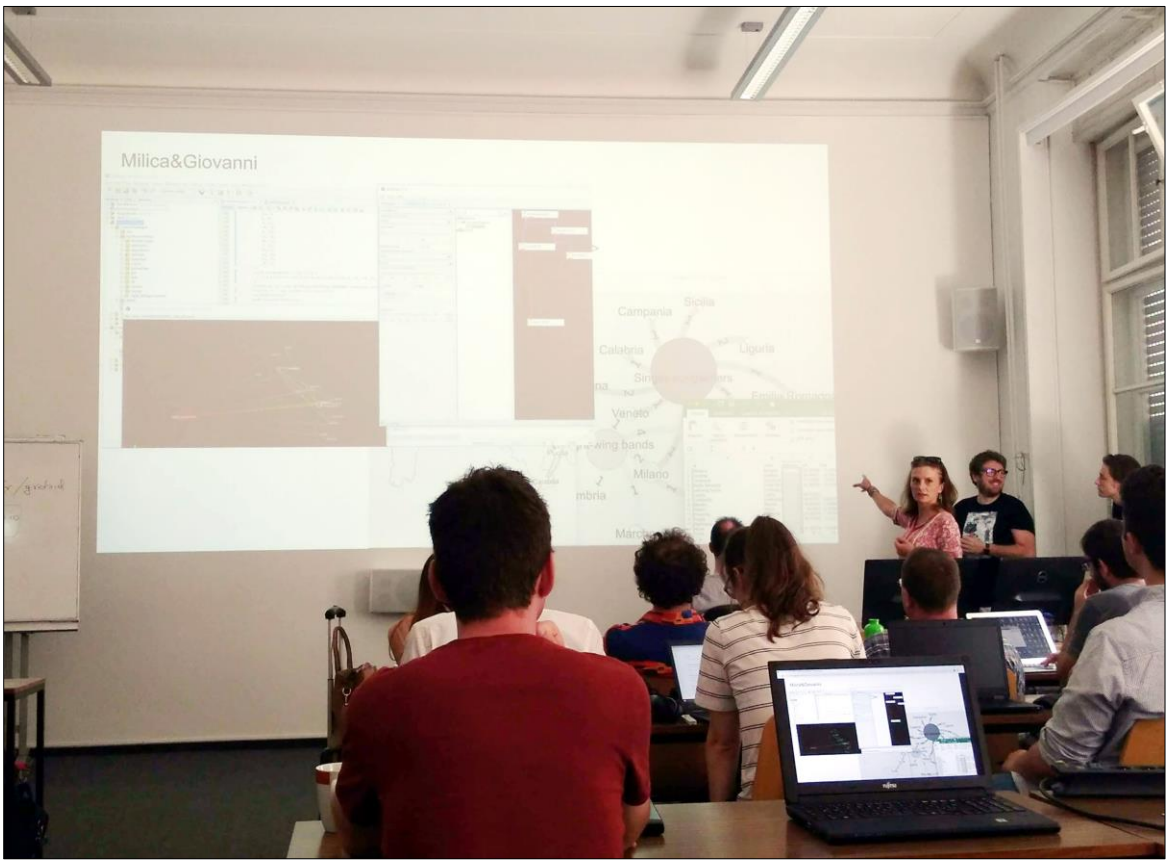
keynote[2] from Prof. Ralf Schenkel of Trier University and affiliated with the DBLP - The Digital Bibliography and Library Project focusing on computer science and related topics.

Participants

33 participants with various backgrounds participated in the code sprint. On beforehand only the group of DARIAH affiliated persons - roughly one third of the group - was known to the organisers. In detail:

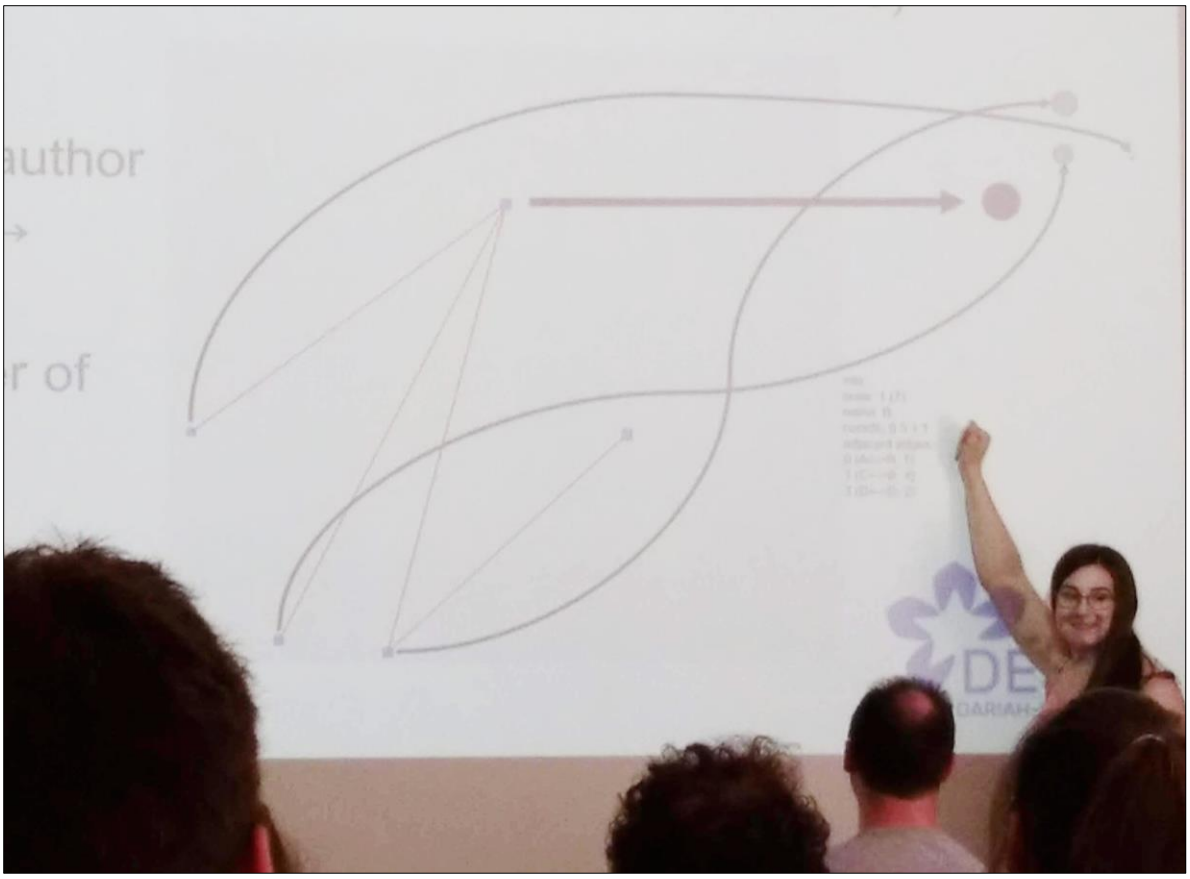
- 9 participants with DESIR affiliation
- 6 participants with a DARIAH affiliation
- 18 participants with neither DESIR or DARIAH affiliation
- 17 participants with a German background, 16 participants with a non-German background
- 23 male participants, 10 female participants
- 31 participants with 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.



Picture: DESIR

[2] https://desircodesprint.sciencesconf.org/data/pages/20180731_Desir_Codesprint_Ralf_Schenkel_1.pdf



Picture: DESIR

The tracks in detail

Track A: Extraction of bibliographical data and citations from PDF applying GROBID

As a machine learning library for extracting, parsing and re-structuring raw documents, such as PDF documents, into structured TEI-encoded ones, GROBID[3] 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 bibliographic extraction tasks, the goal of the library is to reconstruct the logical structure of raw documents in order to enable large

[3] https://github.com/kermitt2/grobid

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 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 being able to follow this track, it was suggested that participants have already had some preliminary knowledge, especially in Java, Python or JavaScript and the abilities to communicate with several web services via HTTP.
The session had 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.
- Visualisation 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 this steps were JavaScript, Http.

Results

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 5 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 some development tools they prefer, 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 asked also to add some more information by accessing other external services, e.g. HAL[4], Entity Fishing[5]. The last activity of track A was the creation of enhanced view of PDF documents. For this reason, the participants were asked to develop a new tool by using some development tools they prefer to produce a usable viewer.
As results for track A, it has been developed two prototypes which in principle performs all steps in this track but in 2 different platforms, Java and Python.
All codes and files of this workshop can be accessed via the GitHub repository for the tracks: https://github.com/DESIR-CodeSprint/TrackA-TextMining.

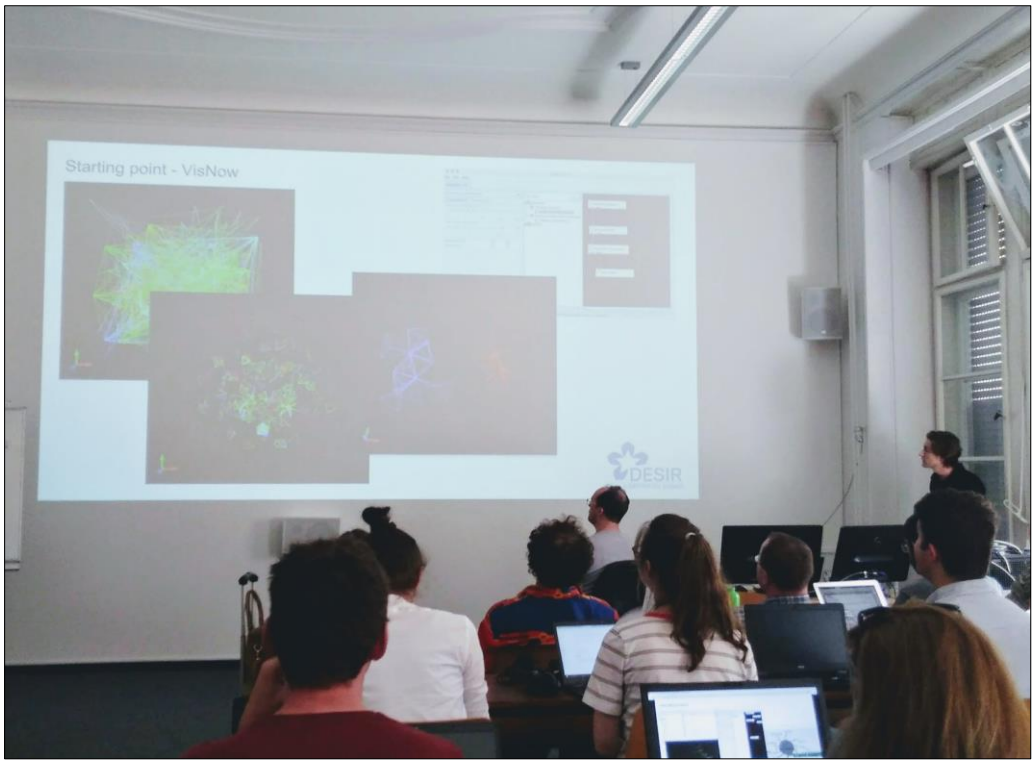
Participants

About 11 participants were involved in track A. They were then splitted into 2 groups concerning their basic skills, whether in Python or in Java.

Future Plans

Since the results of this workshop are still in the prototype version, the future plan is to develop the version final of PDF document viewer. This tool will point out a number of important information in scientific PDF documents.

[4] https://hal.archives-ouvertes.fr/
[5] https://github.com/kermitt2/entity-fishing



Picture: DESIR

Picture: DESIR

Track B: Import and export of bibliographical data from BibSonomy 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, DBLP 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 achievement and we aim at taking the first steps towards a DH bibliography: enabling an easy-to-use web application to import and export bibliographic metadata for the digital humanities. Therefore, we do not want to re-invent the wheel, as tools like Zotero, BibSonomy, etc. already exist. Instead, we focus 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 use of BibSonomy as a backend for storing and organising literature references. With its REST API it enables collaborative storage and retrieval of bibliographic metadata. The choice of programming language and frameworks is not fixed, yet, and will be decided later. Experience in web programming, particularly using web APIs and frameworks is a prerequisite.

Results

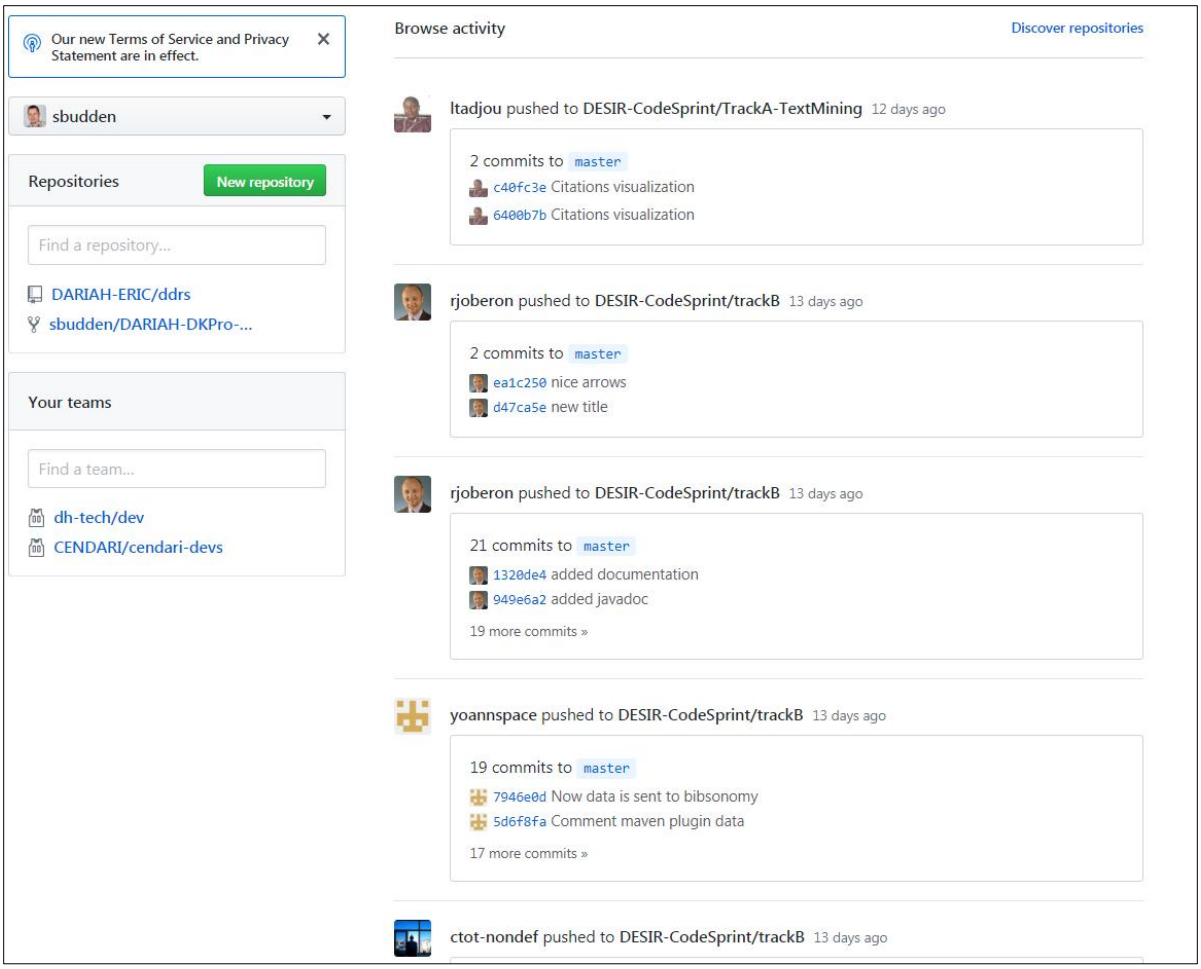
We built a tool for extracting bibliographical metadata from PDF files using GROBID and storing it in BibSonomy. This way bibliographical metadata can be easily added to BibSonomy with low effort. The tool comes with a user friendly interface. We published the full Java code and an installation guide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

Participants

The tool was created by 6 participants from different areas (mainly computer scientists). We 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.

Future Plans

There are no specific plans, but the tool could be extended, e.g. using authorization with ORCID.



Picture: DESIR

Track C: Visualisation 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.

Results

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.

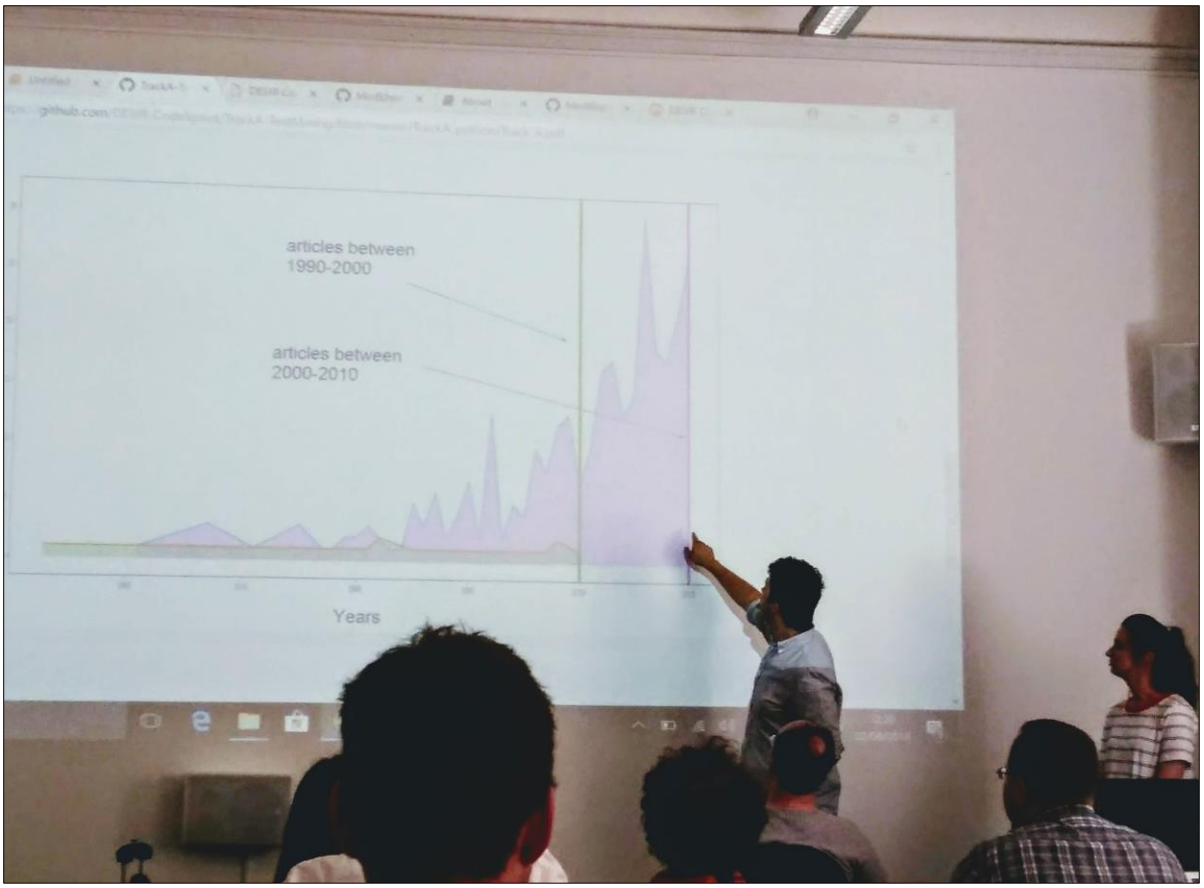
Another use case was conceptualized and tested based on the graph data from Italian music relations and geospatial information. The concept covered the relation between graph and spatial (map) visualization.

Participants

8 participants took part in Track C with various backgrounds, 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.

Future plans

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. We plan to use the outcomes as inspiration for the ongoing work. Both the functionality and concepts will be projected on VisNow application and the planned services. Based on the use cases we will broaden the DH planned application areas.



Picture: DESIR

Track D: Securing Online Services in the DARIAH AAI using SAML/Shibboleth

Researchers that 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 eduGAIN to access DARIAH services, by using the interoperable SAML standard. 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.

DARIAH AAI with an IdP-SP proxy

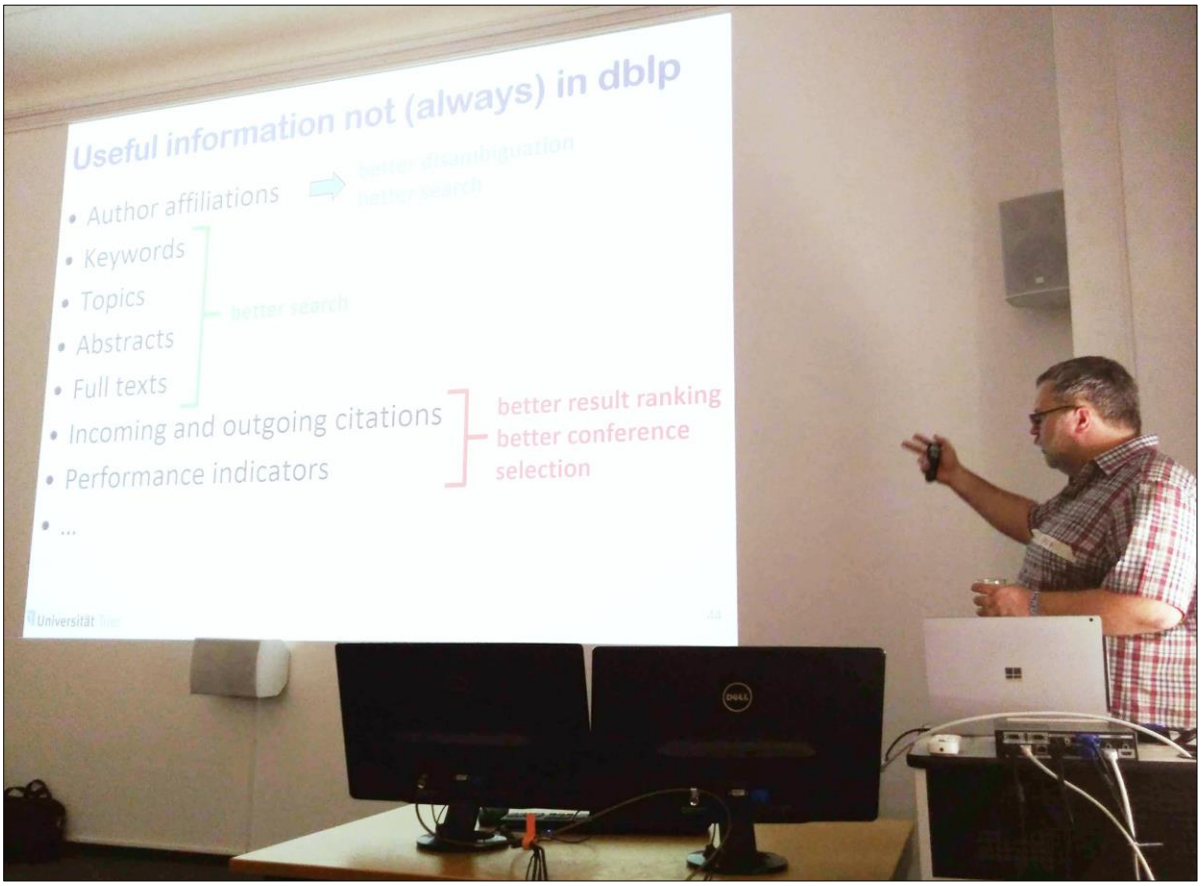
Set into production only some weeks before the workshop, the DARIAH AAI has still lowered the barrier to connect services to DARIAH, by introducing a central AAI proxy between all DARIAH services and all eduGAIN institutions.

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

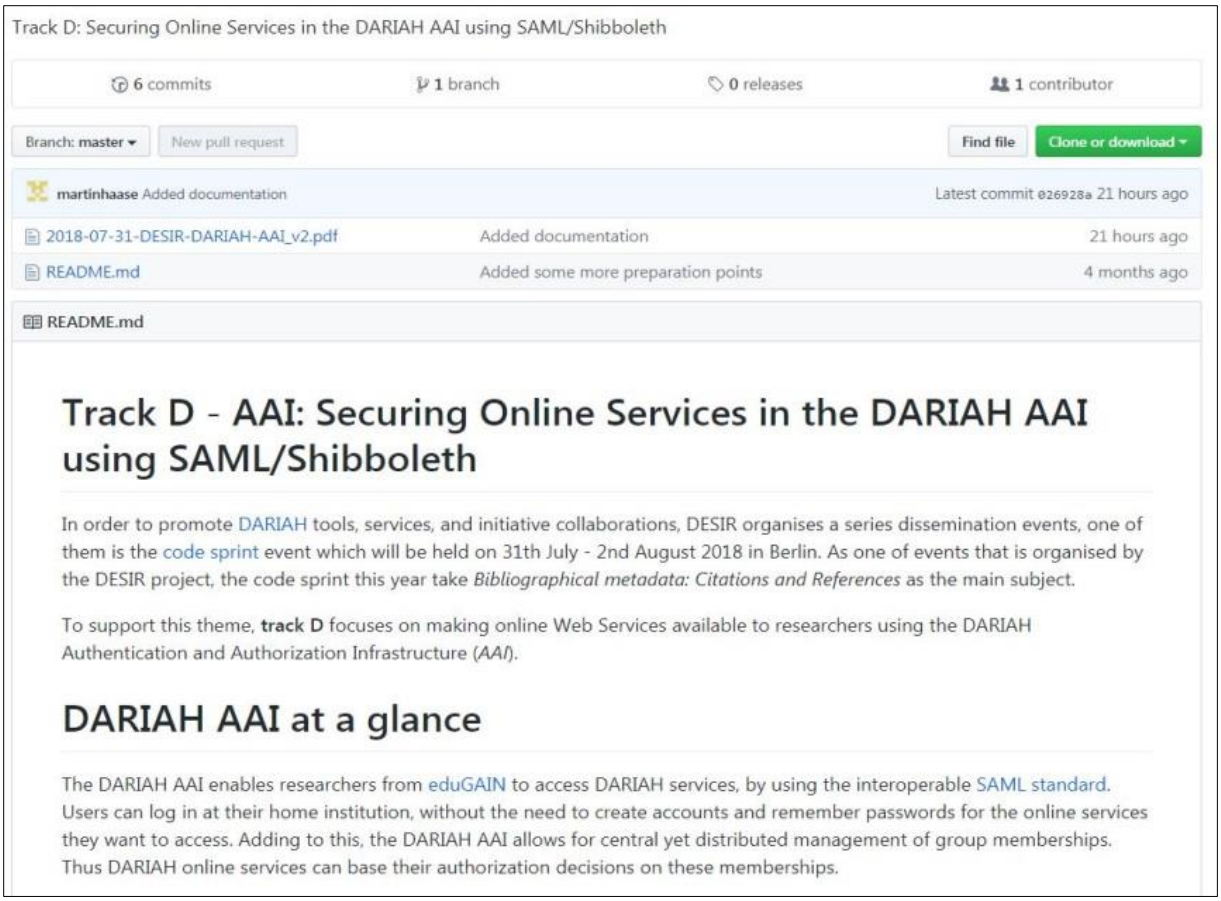
The proxy took over many tasks that services needed to implement previously, which now makes it much easier to connect new services.



Picture: DESIR

Results

This workshop 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.



Picture: DESIR

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.

Participants

The workshop was attended by 4 participants: one PhD student in Computer science, and three scientific staff members, all affiliated to German research institutions.

Future Plans

Now that the DARIAH AAI is running in a production mode, efforts are on the way to promote it further such that many DARIAH services will take advantage of it. The FIM4D working group (Federated Identity Management for DARIAH) is promoting this. The next

FIM4D Workshop called *DARIAH AAI NG Service Provider Workshop* will take 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.

Picture: DESIR