Bundesministerium
für Bildung
und Forschung

CLARIAH-DE

Repository Solutions

Technology Watch Report 1 (AP4.5)

März 2020

Version	1.0
Datum	31.03.2020
Autoren	Denis Arnold (IDS) / Bernhard Fisseni (IDS) / Felix Helfer (ULe) / Stefan Buddenbohm (SUB) / Peter Kiraly (GWDG)
Projekt	CLARIAH-DE
Förderer	Bundesministerium für Bildung und Forschung
Förderkennzeichen	01UG1910A bis I
Laufzeit	01.03.2019 bis 31.03.2021

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-98711

Table of Contents

Introduction 2
Requirements & Specifications 3
Comparative List of Repository Systems 5
3.1 DSpace 5
3.2 Fedora 6
3.3 Islandora and other Fedora branchings 7
3.4 Haplo 7
3.5 Invenio 8
3.6 MyCoRe 8
3.7 EPrints 9
3.8 Dataverse 9
Discussion 10
Sources 11
Appendix 12

1. Introduction

This is the first of three reports composing the technology watch, which is in turn the subject of CLARIAH-DE AP4.5.
The aim of all three reports is to give a detailed overview of technological developments relevant to the project and its partners, and offer recommendations concerning their adaptation within CLARIAH-DE. CLARIAH-DE is the merger of the two established German research infrastructures CLARIN-D and DARIAH-DE. An important task within this merge is the evaluation and – where possible – integration of infrastructure components or services. Parallel with the actual report an evaluation of current PID solutions is created[1].

The technology focused on in this report are *repository solutions*. With digital research infrastructures making up the core of the project, in turn, the storage, management and dissemination of research data form an essential task in this environment, and digital repositories provide the tools to fulfill it. However, the landscape of available solutions is vast, varied and constantly evolving, making a documented comparison all the more relevant for an informed overview to help selecting a viable solution.

The following section will briefly outline the identified requirements for a viable repository solution, as formulated by project partners and supporting research publications and presents various helpful specifications. Then, a comparative list of different repositories will be presented. Finally, a summarizing discussion closes the report.

All findings are the result of research undertaken by the project partners IDS Mannheim and Leipzig University, especially the IDS' reportings on the Open Repositories 2019[2] conference in Hamburg.

[1] The report on current PID solutions with focus on CLARIN and DARIAH is available here: https://zenodo.org/record/3744091
[2] http://archiv.gwin.gwiss.uni-hamburg.de/or2019/

2. Requirements & Specifications

This section outlines the identified requirements on repository systems in the context of the CLARIAH-DE project. While not all of them are seen as mandatory for a system to fulfill, they do give tangible indications on the overall viability of the different candidates.

Central to a repository system is the ingestion and storage of data and this data's medium-term provision and archiving. As text corpora make up one of the most prominent types of data managed by different partners, supporting their adequate integration into a repository system is one of the more important requirements. More specifically, corpora are often hierarchical entities, composed of subcorpora, texts, or even smaller units, as well as (possibly multiple) annotation layers. Therefore, the system should either, in the ideal case, directly support a hierarchy of entities mappable to corpora components in some fashion, or, alternatively, at least allow usage of metadata, optimally CMDI,[3] to show structures of and relationships between components.

From the perspective of partners also active in CLARIN-D, support for the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[4] is also essential, as this is a required way to provide metadata of the stored data to outside services, and is also used for service and content discovery in the CLARIN infrastructure.

Also regarding CLARIN-D partners, the delegation of users via Shibboleth and the resulting granular allocation of user privileges should ideally be supported.

As the endeavor of digital preservation is crucial for the long-term usability of the repository, appropriate supporting features are highly desirable. Preferably, the system should therefore secure data integrity via suitable checks, checksums or similar methods.

The bigger the data collection, the bigger the need for an appropriate search functionality, especially in regards to the collection's metadata. A repository system equipped with workflows that allow using the CMDI metadata for searching would thus be a considerable advantage. This however is not classified as a mandatory feature from the perspective of IDS or ULe, as the data's CMDI components are already indexed and searchable via the Virtual Language Observatory[5].

In addition, external requirements could also be considered, for example the recommendations of the Confederation of Open Access Repositories' *Next Generation Repositories* initiative[6],

[3] https://www.clarin.eu/content/component-metadata
[4] http://www.openarchives.org/OAI/openarchivesprotocol.html
[5] https://vlo.clarin.eu
[6] https://www.coar-repositories.org/news-updates/what-we-do/next-generation-repositories

which researches possible functionality improvements for repository systems. While not explicitly considered in this report, watching their continually updated requirements and recommendations is still advised, especially as they seem to be largely community driven and offer valuable insights to recent trends and new developments in the field.

A promising specification for a standard for storage of digital information in repositories, the *Oxford Common File Layout*[7], is currently under development. Although still in a beta version at the time of writing, it might be worthwhile to monitor its progress, as it proposes very sensible core principles, especially in terms of enabling re-extraction of stored data and rebuilding of the underlying repository. As it is still a work in progress however, one can naturally not expect existing systems to already adhere to its specification.

There also exist various scientific publications about comparative analyses of repository software, including [Castagné, 2013], [Pirounakis & Nikolaidou, 2009], [Pirounakis et al., 2014], [Bankier & Gleason, 2014] and [Verma & Kumar, 2018]. Unfortunately, most of those were evidently published at least six years in the past, which lessens their present significance for a field as active and prone to change as this. The only more recent work, by Verma and Kumar, restricts its comparison to only three systems (DSpace, GSDL, EPrints), of which the latter two were not deemed relevant to the context of the project. Recent work in this field would therefore be highly desirable.

An overview of established systems with regard to publication repositories can be obtained through BASE – Bielefeld Academic Search Engine[8]. A list of the most occuring systems within the TOP20 of BASE can be found in Table 1 in the Appendix (chapter 6), emphasizing the field's vast, heterogenous spread. However, it has to be noted that not all the software listed by BASE are repository systems. Instead, the list also includes non-generic, datatype-specific solutions that are not directly applicable in the context of this project's partners' use-cases (e.g. Goobi) and therefore not examined further.

Additionally, the German Registry of Research Data Repositories[9] can provide a similar overview. A list of the 20 most occuring systems within the registry can be found in Table 2 in the Appendix (chapter 6). It shows a similar spread of technology and systems.

[7] https://ocfl.io/
[8] https://www.base-search.net/about/de/about_de_sources.php
[9] https://www.re3data.org

3. Comparative List of Repository Systems

In this section, a list of repository systems will be presented and discussed, mostly regarding the previously stated requirements, if applicable.

The considered systems were all present at the Open Repositories 2019 conference[10], which at least partly suggests ongoing development, or at least an active engagement in the community by its developers. Likewise, this ensures that the selected systems have at least some ties or utilization within the scientific community, which lessens the risk of adopting a system which is ill-adapted to our purposes.

A remark: In practice a generic repository system, for instance DSpace, is possibly never used in its generic state but adapted for the specific requirements of the service provider. Usually, these are requirements coming from the nature of the content (e.g. publications, research data), the structure and appearance of the metadata describing the content (data models, interoperability, e.g. by OAI-PMH), discipline or use (archive, publications to generate spread/references). Therefore the list below focuses on the main generic solutions and leaves out specific "in-production" repositories one might expect here as well.

A user that is interested in the specific appearances of repositories therefore has to look at the contents, the metadata interoperability, and, in general, the functions and users of the repositories. Good starting points may be the aforementioned sources list of the Bielefeld Academic Search Engine (BASE), the German Registry of Research Data Repositories, or, taking a CLARIAH angle, the CLARIN Centre Registry[11] of OAI-PMH endpoints.

The following repository solutions are all available as open source software, although in most cases, setup or modifications are offered as rentable services.

3.1 DSpace

DSpace[12] is the repository system from DuraSpace, marketed as an "easy to install, out of the box" application. It is used by a large number of institutions and companies[13], making it one of the more popular repository systems on this list. Notably, DSpace' GitHub repository holds

[10] http://archiv.gwin.gwiss.uni-hamburg.de/or2019
[11] https://centres.clarin.eu/oai_pmh
[12] https://duraspace.org/dspace/
[13] https://duraspace.org/registry/?filter_10=DSpace

over 10.000 commits from 176 contributors, which speaks for a large and active community. It is written in Java.

DSpace is mainly used for document servers, which results in some disadvantages regarding the requirements of this report. Mainly, DSpace in its current version 6 does not support freely configurable content models, but instead limits usage to Dublin Core, with support for metadata beyond the 15 core properties. However, this is supposed to change with version 7, which is set to release in 2020 and already available as a beta version. This new version will also add a separation of frontend and backend, as well as a REST-API. There is also an ongoing discussion over the addition of video transcoding to further diversify supported data formats.

DSpace supports OAI-PMH, Solr-based search for metadata (and full text) and includes a checksum checker. Shibboleth integration is possible via a plugin.

3.2 Fedora

Fedora[14] is also developed by DuraSpace. Contrary to DSpace, Fedora as a framework is more of a toolbox than a ready-to-use repository system. It is implemented in Java. Its GitHub repository lists 46 contributors with over 3000 commits, also hinting at an active community. Fedora version 3 is currently in use at the IDS and Leipzig University.

At the time of writing, Fedora's latest version is 5.1.0. However, during Open Repositories 2019, the developers reported a survey result indicating that more than 80% of users still work with version 3. They cite this as an important reason to put maximal effort into the soon to be released next big iteration of the system, version 6, to convince as many users as possible to migrate.

This next version will contain several relevant changes: a specified API, support for the previously mentioned Oxford Common File Layout, a "simple search endpoint", better performance and scalability, a higher focus on digital preservation and better support for migration repositories to this version, especially coming from version 3.
Not yet planned is a core support for the OAI-PMH, which previous versions offered. However, David Wilcox suggested that a strong signal for demand from users and community members could change this. The aforementioned user concentration on version 3 again may also be a strong incentive for the developers to give them the best reasons to update.

Fedora can calculate, store and verify checksums for all managed files and can integrate with Shibboleth. Support for audio or video files does not seem to play a prominent role in Fedora, its documentation suggests using the related system *Avalon* (see below).

[14] https://duraspace.org/fedora/

3.3 Islandora and other Fedora branchings

There exist several systems based on Fedora which should be mentioned here as well.

Islandora[15] is a combination of Fedora and the content management system Drupal[16] (written in PHP). This allows the use of Drupal modules, which might be useful for e.g. setting up webshops for licensing. Up to version 7, Islandora was based on Fedora 3, but for more recent versions switched to Fedora 5 (and Drupal 8).

Islandora's way of content handling is obviously very similar to Fedora. It is a modular framework, offering customizable "solution packs" for specialized handling of different data types and models. Islandora supports Solr-based searching and a module for OAI-PMH-integration.

The Islandora website is optimistic concerning migration to the newer versions, but no experiences concerning successful migrations of live systems could be found yet. Islandora is in use at the university of Hamburg and the Max Planck Institute for Psycholinguistics in Nijmegen, amongst others, they however both have not migrated to the latest, Fedora-5-based version yet.

Another Fedora-based system is Samvera[17], which also provides a framework for building a more custom repository. As this is a resource-intensive task, several community developed frontends exist, in various stages of development: Hyrax, as a basis for using Samvera with less implementation overhead, Hyku, as an "easy to install, configure and use" solution, and Avalon, with its aforementioned focus on multimedia storage. None of these seem to offer specific benefits for our community.

3.4 Haplo

Haplo[18] is a fairly new, JavaScript based system. It has been adopted by several universities in the UK and Northern Ireland (where the company is also seated) and focuses strongly on research management and education, for example in supporting the REF[19] exercises in UK higher education.

---

[15] https://islandora.ca/
[16] https://www.drupal.org/
[17] https://samvera.org/
[18] https://www.haplo.com/
[19] https://www.ref.ac.uk/

Due to this focus, it is assumed that a possible community around this system will most likely have requirements differing from the ones presented here, even though the system, in principle, offers similar functionalities for document storage like other mentioned solutions.

Its GitHub repository does not exhibit a lot of activity, and has also only two contributors, suggesting development coming mainly (or exclusively) from the company itself for now.

3.5 Invenio

Invenio[20] is, similar to Fedora, more of a toolbox than a ready-to-use repository. It is the foundation for Zenodo and other repositories at CERN. Invenio offers a JSON-based REST-API and is optimized for large datasets and scalability. It is programmed in Python.

It supports a flexible metadata model, choosing from either an existing or custom format, and Elasticsearch based search functionality. Checksum verification and an OAI-PMH module exists, as well as third-party modules for Shibboleth authentication.

As Invenio is, among other things, also used for a video server, video transcoding is also supported.

Until the next Open Repositories, or respectively June 2020, *InvenioRDM* is supposed to be developed – a ready-to-use repository system based on Invenio.
The university of Hamburg is involved in a piloting adaption of InvenioRDM when the system is in production. Possibly, the HZSK repository will also be integrated into the general repository of the university of Hamburg. This might allow the other partners to gain some insight into the process and InvenioRDM's overall viability.

3.6 MyCoRe

MyCoRe[21] is a repository system developed by members of various German universities and libraries. It is continuously being developed, though without separate funding. It is written in Java.

MyCoRe also provides a framework instead of a ready-to-use system (although third-party applications exist, such as the *MODS Institutional Repository*). It supports a freely customizable, hierarchical data model, the OAI-PMH, metadata search and Shibboleth.

[20] https://invenio-software.org/
[21] https://www.mycore.de/

The system is mostly used for documents. Its developers claim a faster adoption of requirements for document repositories than other teams. However, support for the Oxford Common File Layout is not yet planned.

3.7 EPrints

EPrints[22] is a software package for building open access repositories. It is compliant with the OAI-PMH. However, it was not examined further, as its focus is mainly on document storage of print products, differing from the focus of the project and its partners.

3.8 Dataverse

Dataverse[23] is an open source tool originally written by Harvard University. It is an open source application for storing and sharing research data allowing proper citation with persistent identifiers (e.g. Handles, DOI) attached to the datasets. Dataverse repositories host multiple data collections. The repository has two types of data containers: first, a "dataset", which contains the data files, descriptive metadata, documentation, and any other files helping the interpretation such as source code. Second, a "dataverse", which can be described as a contextualized set of research data: it contains one or more datasets, governs the metadata schemas which can be used in the datasets, has rights management functionalities, and, moreover, might also contain other dataverses.

The data get a persistent identifier, and – in the case of Göttingen Research Online – its metadata are propagated to the central DOI database of DataCite[24], which is used by different discovery services.[25] The repository provides versioning, data citation, file previews, assignment of licenses, restricting files (and granting access). It also provides a place for organisations (departments, research groups, journals) to manage member participations via customisable roles and responsibilities. Almost all the functionalities of the service are available via APIs, so one can easily build an automated communication pipeline between different kinds of research software such as document management systems or lab notebooks and the repository. OAI-PMH is also supported.

Dataverse is currently used by more than 50 repositories worldwide having global, state-wide, or institutional scopes, but is not yet established for data from the humanities and

[22] https://www.eprints.org/uk/
[23] https://www.dataverse.org/
[24] https://datacite.org/
[25] When choosing Handle instead, metadata are exported to the ePIC database at https://www.pidconsortium.net/.

language-related research data realm.[26] Within CLARIAH-DE Dataverse is used for example by the Göttingen partner for Göttingen Research Online[27] which allows export into DARIAH-DE repository or Zenodo and may play a future role in the CLARIAH infrastructure.

4. Discussion

It is evident that for some of the systems examined in this report, big changes are imminent. Both Fedora and DSpace plan to release a major update this year, with significant changes to their respective systems. For DSpace, the switch to a more freely adaptable content model would definitely strengthen its position as a good (and already well-established) candidate. In the case of Fedora, the decision for or against OAI-PMH support will be a crucial one, which might be worthwhile to actively lobby for, as part of a "community-driven" feature request.
Furthermore, an update to Fedora as the "main branch" might also entail changes for branching systems like Islandora, at least concerning longer-term planning and development roadmaps.

An additional positive argument for Islandora itself is that it is used by three institutions with similar orientation: the Max Planck Institute for Psycholinguistics, the Meertens Instituut and the University of Hamburg (and HZSK). However, the Max Planck Institute for Psycholinguistics reports holding off on migrating to the latest Fedora-5-based version.

Dataverse is another well-established tool, with many contributors and wide usage, the latter also including two CLARIAH-DE partners. Dataverse offers a flexible data model, API functionality and OAI-PMH support. Depending on a partner's specific use case, its detailed member management could also prove to be beneficial.

Samvera, Haplo and Invenio all share the downside of very small to non-existent communities or with communities that have a different focus from ours, which lessens their recomendability. In contrast, existing expertise regarding specific systems is a significant advantage, meaning those systems which are already in use by various partners or institutes close to the project, like Fedora (used by ULe and IDS), Dataverse (used by Göttingen Research Online), or DSpace (used by various CLARIN centres). It should be noted however that the project scope involving InvenioRDM also includes establishing a community of relevant size – another reason to observe its progress and usability in the future.

[26] However, at least one CLARIN centre applies Dataverse as a repository for linguistic data, namely the Tromsø Repository of Language and Linguistics (TROLLing) in Norway: see https://hdl.handle.net/10037/17951

[27] https://data.goettingen-research-online.de/ and https://www.eresearch.uni-goettingen.de/services-and-software/gro-data/

Also important to mention is the cost of setting up a repository from one of the framework systems: even with a team of developers, this could mean a year or more of implementation work. This, in the case of a previously used system, does not include the effort of migrating existing data.

To summarize, a final recommendation of a singular system can not yet be made. Instead, development of DSpace and Fedora should be closely monitored, as both upcoming versions will bring some very relevant changes. Islandora, as another valid candidate, might also be affected by this. In addition, information about the viability concerning completed (or aborted) migration processes, especially between versions of Fedora or Islandora, could provide valuable insights. With the knowledge of these new versions' capabilities and migration viabilities, assessment of a choice between those "in-flux-systems" and other viable options – especially Dataverse – will be much more well-founded.

Development of the Oxford Common File Layout should also be closely followed, as it seems to be a promising standard, which might be especially applicable in a context of various partners and systems, and may considerably lessen vendor lock-in and migration cost in the future. Its possible adaptation into Fedora 6 is therefore especially interesting.

5. Sources

Bankier, J. G. and Gleason, K., "Institutional repository software comparison", 2014. [Online]. Available: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/news/institutional_repository_software.pdf. [Accessed: 26-Jan-2020].

Castagné, M., "Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra," 14-Aug-2013. [Online]. Available: https://open.library.ubc.ca/collections/graduateresearch/42591/items/1.0075768. [Accessed: 25-Jan-2020].

Pirounakis, G. and Nikolaidou, M., Comparing Open Source Digital Library Software, Handbook of Research on Digital Libraries: Design, Development, and Impact, IGI Global, 2009.

Pyrounakis, G., Nikolaidou, M., and Hatzopoulos, M. Building Digital Collections Using Open Source Digital Repository Software: A Comparative Study. International Journal of Digital Library Systems (IJDLS), 4(1), 10-24. doi:10.4018/ijdls.2014010102, 2014.

Verma, L. and Nishant K., "Comparative Analysis of Open Source Digital Library Softwares: A Case Study." 2018.

Stefan Buddenbohm, Nathanael Cretin, Elly Dijk, Bertrand Gaiffe, Maaike de Jong, et al.. State of the art report on open access publishing of research data in the humanities. [Other] DARIAH. 2016. ⟨halshs-01357208v3⟩

Vierkant, P., Voigt, M., Petrus, A., Zielke, D., Kindling, M., & Hartmann, T. (2015). 2015 Open Access Repository Ranking. Zenodo. http://doi.org/10.5281/zenodo.30781

6. Appendix

Table 1: Top 20 occuring software in the Bielefeld Academic Search Engine[28]. German BASE currently harvests the metadata of 165 mio. Documents (mostly publications, 60% open access) from over 8.000 sources but is not focussing on the underlying software platforms.

System	# of occurrences
Unknown	4
Goobi (not a repository platform)	3
Invenio	2
Eprints	1
freiDOK (proprietary)	1
Repox	1
PubMan	1
Fedora	1
Figshare	1
DSpace	1
Django	1
MyCoRe	1
Open Journal System	1
mediaTUM (not a repository platform)	1

[28] March 2020, https://www.base-search.net/about/de/about_sources_date.php

Table 2: Top 20 occuring software in Re3data.org[29]. The German Registry of Research Data Repositories (re3data.org) harvests 2.450 research data repositories on a global scale and offers granular metrics on these repositories and content.

System	# of occurrences
Unkown	1234
Other	475
DataVerse	88
MySQL	78
DSpace	77
CKAN	71
Fedora	37
EPrints	32
Nesstar	21
eSciDoc	3
DigitalCommons	3
dLibra	2
OPUS	1

[29] May 2020: https://www.re3data.org/metrics/software