CLARIN-D: Report for the CLARIN-D and DARIAH-DE Technical Advisory Board May 2013 – December 2014

January 2015

CLARIN-D, BMBF-FKZ: 01UG1120A-I , 01UG1420A-I

Responsible: Dirk Goldhahn

Editors: Dieter Van Uytvanck, Dirk Goldhahn, Thorsten Trippel

Contributors: Emanuel Dima, Thomas Eckart, Willem Elbers, Twan Goosen, Kees Jan van de Looij, Oliver Schonefeld, Olha Shkaravska, Thorsten Trippel, Menzo Windhouwer

Table of Contents

1 Introduction to CLARIN-D

The BMBF project CLARIN-D is developing a digital infrastructure for language-centred research in the social sciences and humanities. The main function of the CLARIN-D service centres is to provide relevant, useful data and tools in an integrated, interoperable and scalable way. CLARIN-D is rolling out this infrastructure in close collaboration with expert scholars in the humanities and social sciences, to ensure that it meets the needs of users in an orderly, accessible and user-friendly way.

CLARIN-D is a collaborative effort involving the following service centres, their host institutions, and leaders:

- Bayrisches Archiv für Sprachsignale, Ludwig-Maximilians-Universität München (PD Dr. Christoph Draxler)
- Berlin-Brandenburgische Akademie der Wissenschaften (Dr. Alexander Geyken)
- Institut für Deutsche Sprache, Mannheim (Prof. Dr. Ludwig Eichinger)
- Max Planck Institut für Psycholinguistik, Nijmegen (PD Dr. Sebastian Drude)
- Eberhard Karls Universität Tübingen, Seminar für Sprachwissenschaft (Prof. Dr. Erhard Hinrichs)
- Universität Hamburg, Zentrum für Sprachkorpora (Prof. Dr. Kristin Bührig)
- Universität Leipzig, Institut für Informatik (Prof. Dr. Gerhard Heyer)
- Universität des Saarlandes, Englische Sprach- und Übersetzungswissenschaft (Prof. Dr. Elke Teich)
- Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung (Prof. Dr. Jonas Kuhn)

CLARIN-D is operating in cooperation with German high performance computing centres to provide grid and cloud computing as well as general computing services for its users. The computing centres involved in CLARIN-D are:

- Rechenzentrum Garching
- Forschungszentrum Jülich
- GWDG Göttingen

For more detailed information about the role of the computing centres, see section 8.2.

CLARIN-D is building on the achievements of the preparatory phase of the European CLARIN initiative as well as D-SPIN - CLARIN-D's Germany-specific predecessor project. These previous projects have developed research and technical standards for the CLARIN services centres, as well as plans for the sustainable provision and long-term archiving of tools and data. Furthermore, the first results of the CLARIN-D construction phase have been reported to the Technical Advisory Board in April 2013.

CLARIN-D's work packages reflect the project's various objectives and tasks. Leadership for each work package is assigned to a specific centre, depending on the experience and expertise of particular CLARIN-D service centres. Work package 3, „Technical Infrastructure“ is led by the University of Leipzig.

2 Main Advances in the CLARIN Technical Infrastructure since April 2013

In this report, we focus on the main advances of the technical infrastructure since the last meeting of the CLARIN-D/DARIAH-DE Technical Advisory Board, which took place in April 2013 on the basis of the written report "CLARIN-D AP3 report: establishing the technical infrastructure May 2012 - April 2013".

Prior to giving a detailed account of the progress made in the various specialized technical infrastructure working groups, we would like to draw the attention to some of the milestones that have been achieved since the last report.

All CLARIN-D Centres have undergone the process of centre assessment and certification. In May 2013, all 9 CLARIN-D centres were awarded the CLARIN B centre status and received the Data Seal of Approval.

Several infrastructure registries have been implemented, such as a centre registry, a virtual collection registry and a data category registry. Together they allow for reliability, flexibility and interoperability in the distributed environment of CLARIN-D.

The Component Metadata format CMDI was submitted as an ISO standard. To support interoperability between different initiatives, it was divided into three different parts:

1. The model (ISO 24622-1:2015 Language resource management -- Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model),
2. The component description (intended to become ISO 24622-2, currently a Work Item in ISO TC 37 SC 4), and
3. Core components for language resources (intended to become ISO 24622-3).

The first part has been published as an ISO standard recently.

With regards to the Virtual Language Observatory, CLARIN's facetted browser based on metadata harvested from CLARIN centres, substantial improvements in user functionality and metadata quality have been achieved.

Finally, the CLARIN IdP has undergone substantial changes in order to improve stability before migrating it to the computing centre in Garching.

These and other recent developments will be specified in more detail in the description of the technical working groups (WGs) below. Note that rather than include long descriptive texts, we have tried to link to "living" documents or web sites (in the spirit of CLARIN) as much as possible, resulting in a more up-to-date overview.

3 WG 1: Repositories

3.1 Setup of Repositories

As one of the important pillars of the distributed CLARIN-infrastructure, each of the centres has set up a repository to store the resources they are hosting in a standardized and sustainable way. This process included the creation (and possibly conversion) of metadata and persistent identifiers.

The repository infrastructure in all centres was fully established and integrated into the CLARIN-D infrastructure, as the outcome of the centre assessment process has proven.

In August 2014 a workshop on archiving and repositories (organized by the INNET project) took place in Nijmegen, Netherlands. During this workshop several representatives of CLARIN Centres presented their repository solution and discussed the reasoning behind their choices and the details of their implementation:

http://www.clarin.eu/event/2014/innet-repository-workshop

3.2 Centre Assessment

An important part of acquiring the official CLARIN centre status is to be assessed. All 9 CLARIN-D centres have taken part in 2 assessment procedures:

- The CLARIN centre assessment, which analyses the centre's offer in terms of compliancy with the B centre requirements (the most demanding level at the moment of writing).
- By the Data Seal of Approval committee (http://www.datasealofapproval.org/), which is an independent body that checks whether a centre's data management strategy and its policy are suitable for long-term archiving.

In May 2013, all 9 CLARIN-D centres were awarded the CLARIN B centre status (http://www.clarin.eu/content/certified-centres). They also all received the Data Seal of Approval. This is major achievement and an example for other national CLARIN consortia.

Until May 2015 all CLARIN-D centres will renew their status. Changes in the assessment have been identified and consequences have been analysed by all centres. Procedures for obtaining the B centre assessment and the Data Seal of Approval are on time.

4 WG 2: Persistent Identifiers (PID)

To ensure the stability of scientific citations of language resources and the associated metadata descriptions, CLARIN relies on the use of Persistent Identifiers. By adding a level of indirection when resolving an identifier towards an URL, the long-term stability of the references can be guaranteed. A summary of the role of PIDs in CLARIN centres can be found at http://www.clarin.eu/node/3965.

All CLARIN-D centres have assigned handles to their metadata records and resources. Most of them (7 in total) are connected to the EPIC service to acquire and manage handle PIDs. The MPI and IDS are not using EPIC – they have their own prefix and handle server.

In order to stay aware of new evolutions in the field of PIDs, CLARIN has organized two workshops in the course of 2014; one documenting the use of handles in CLARIN centres and another one on the potential use of DataCite DOIs. All details about these workshops are available at:

http://www.clarin.eu/event/2014/persistent-identifier-workshop

http://www.clarin.eu/event/2014/doi-datacite-workshop

5 WG 3: Registries

5.1 Centre Registry

Accessing resources and endpoints (SRU for content searches, OAI-PMH for metadata harvesting, etc.) in a distributed setup requires an up-to-date list of addresses. Therefore CLARIN needs a machine-readable registry where such pointers can be stored and accessed.

The first versions of the CLARIN centre registry, where both technical and organizational information about centres and services is stored, grew organically from the emerging needs. In the course of 2014, the whole back-end (based on Django and SQLite) has been rewritten from scratch, including many optimizations in the data model. This has resulted in a code base which is much better documented and easier to maintain in the future. Where originally the RZG computing centre was responsible for the implementation and the hosting of the centre registry, CLARIN ERIC and CLARIN-D decided to take over the development part, as this proved to be more efficient. At the moment of writing the beta version of the new version is available at: https://centres-staging.clarin.eu:4430/

After some final tests (checking if the dependent applications can still use the updated APIs) this version will be deployed as a production service (see http://centres.clarin.eu). The centre registry is currently used by the Federated Content Search aggregator, WebLicht and the OAI-PMH harvester as authoritative information source.

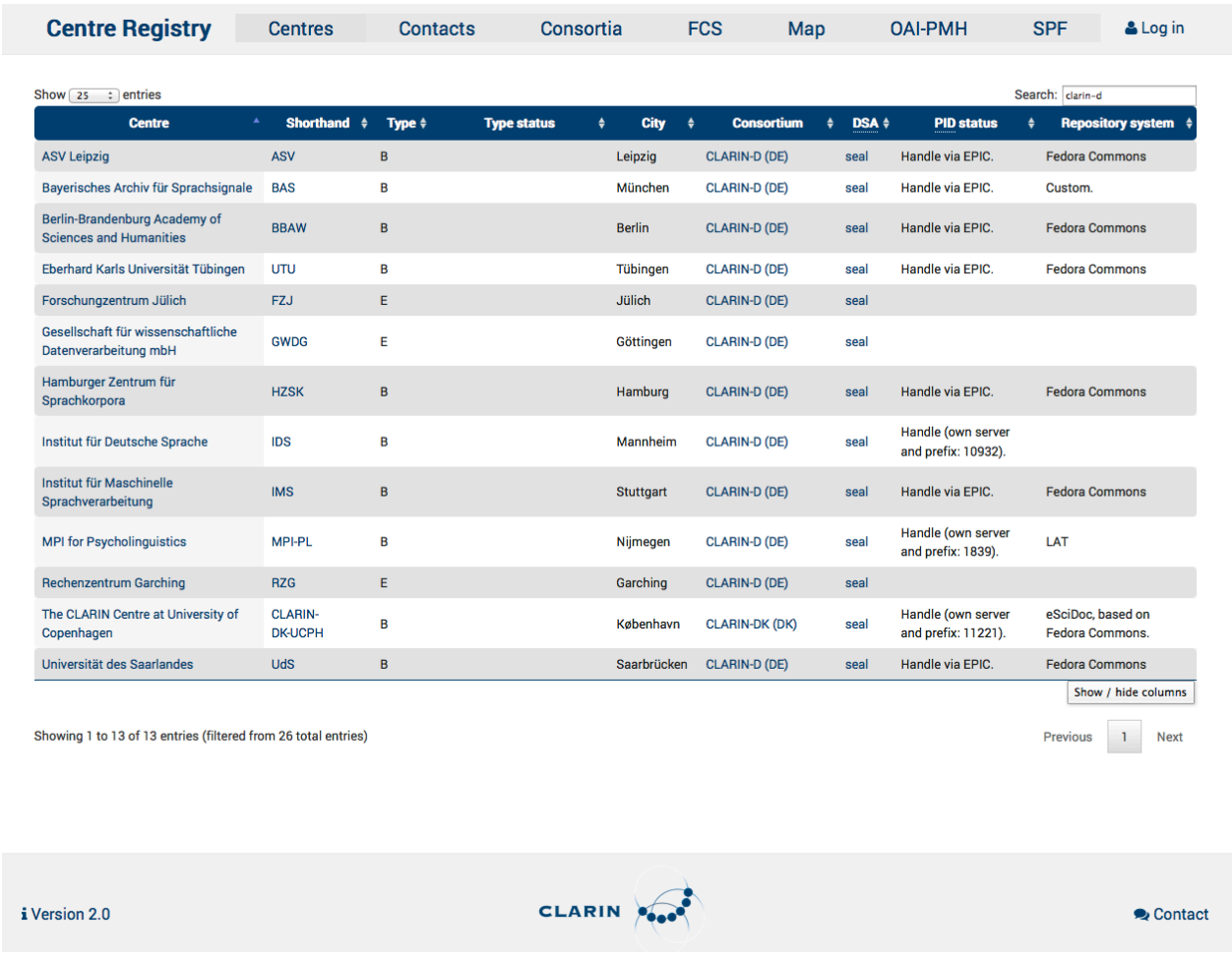

Figure: the new HTML interface of the centre registry

5.2 Virtual Collection Registry

A virtual collection is a coherent set of links to digital objects (e.g. annotated text, video) that can be easily created, accessed and cited. The links can originate from different archives, hence the term virtual. A virtual collection is suitable for manual access (using a web-browser) as well as automated processing (e.g. by a web service).

CLARIN provides a registry where scholars can create and publish their virtual collections. It is closely integrated with the infrastructure and provides persistent identifiers and federated login. The collection metadata is openly available and accessible via the Virtual Language Observatory.

The old alpha version of the Virtual Collection Registry was largely reworked and updated in cooperation with CLARIN ERIC. Important changes were the replacement of the authentication and authorization framework, use of the new EPIC API (version 2), support for content negotiation and integration of the new CLARIN layout. The exact work plan can be found at https://www.clarin.eu/node/3960, the updated web application is available at:

http://clarin.eu/vcr

5.3 Data Category Registry and Data Concept Registry

The Data Category Registry is a step in the direction of interoperability at the level of linguistic encoding (tag sets, metadata elements, etc.). The basic idea is to register all widely used concepts/terminology so that everyone can refer to them. All is based on the ISO 12620 standard, which is a generic model not restricted to linguistics.

The ISO TC37 Data Category Registry (DCR) was created in 2008 as one of the first ISO standards delivered in the form of a database (ISOcat[1]). The Max Planck Institute for Psycholinguistics (MPI) has provided development, hosting, and support services and acted as the Registration Authority (RA). The use of the DCR has grown over the years. Feedback from users, coupled with changes in ISO standardization procedures, necessitated a review of the current system and operational framework with an aim towards improving usability.

The MPI stopped being the RA and hosting provider in December 2014. After reviewing potential replacement systems, ISO TC37 selected TermWeb, from Interverbum Technology, due to its support of the required data model.

For users from the CLARIN ERIC[2], the Meertens Institute will host a new registry for CLARIN relevant concepts based on the corresponding ISOcat data categories, such as those used for the Component MetaData Infrastructure (CMDI).

The new CLARIN Concept Registry[3] (CCR) is less complex than ISOcat, and is a closed registry: only the national CCR-coordinators[4] will be able to input and edit (new) concepts. All Data Category Selections

[1] http://www.isocat.org/

[2] http://www.clarin.eu/

[3] http://www.clarin.eu/conceptregistry

[4] http://www.clarin.eu/content/concept-registry-coordinators

(and individual DCs) the national ISOcat coordinators wanted to keep are available in the new CCR as well. Also their PIDs will remain recognizable in the CCR.

For now the entries of the ISOcat Data Category Registry are still available in a static manner. All Data Category Peristent IDentifiers, e.g., http://www.isocat.org/datcat/DC-4146, remain resolvable. The public part of the ISOcat registry can be browsed using the ISOcat Guest workspace[5].

The CLARIN Concept Registry[6] is available as of January 2015, the updated Component Registry (using CCR instead of ISOcat) is expected in the course of February 2015.

The switch from Data Category Registry to Concept Registry and its migration from the MPI to the Meertens Institute are one example of a successful migration. If a migration of services is necessary it is crucial that it is well prepared. For the most crucial CLARIN services, the so-called A-services, such plans are prepared as a fallback procedure if a centre can no longer support a service. The A-services are identified by a set of criteria[7] that were established by CLARIN ERIC. Those central services are listed together with their the current status[8] and the migration plans that have been prepared.

5.4 Relation Registry

The relation registry is discontinued, as relations can now be stored in CLARIN concept registry.

6 WG 4: Metadata

The Component Metadata framework forms the base of metadata modelling in CLARIN. As the work done in this workgroup was spread across several subgroups, we mention the activities here as subsections. In general, detailed information and links to the software mentioned can be found at http://www.clarin.eu/cmdi

6.1 CMDI metadata framework

Metadata for language resources and tools exists in a multitude of formats. To overcome this dispersion CLARIN has initiated the Component MetaData Infrastructure (CMDI). The ISO-standardized CMDI model (ISO 24622-1:2015) provides a framework to describe and reuse metadata blueprints, the standardized description of the components allows the development of interchangeable metadata schemas.

In October 2013, the newly formed CMDI task force organized a workshop on the future of CMDI. The participants discussed the current state and expected developments and challenges related to the Component Metadata infrastructure. In the light of ISO standardization of the description language and the implementation of CMDI by more and more data providers, a number of shortcomings of the current implementations became obvious and feedback was provided that required some changes.

---

[5] http://www.isocat.org/rest/user/guest/workspace

[6] http://www.clarin.eu/conceptregistry

[7] http://www.clarin.eu/node/4001

[8] http://www.clarin.eu/node/4002

In February 2014, after a number of virtual meetings, the CMDI task force came together in Utrecht to decide on the global specification of CMDI 1.2, the planned successor to the current version of CMDI. By April, a full description of the changes and additions to be included in CMDI 1.2 had been completed and approved by the Centre Committee. This CMDI 1.2 will be the basis for further development of ISO 24622-2, the standardization of the component description language.

Major changes in CMDI 1.2 include the storage of lifecycle information (versioning and deprecation), support for external vocabularies, the possibility of mandatory attributes and a cleaner and more compatible XML representation (see CMDI 1.2 changes - executive summary[9]).

In the following months, most of the core 'toolkit' was developed, including scripts to convert existing CMDI documents to the new version. By the end of 2014, some work remains to be done before CMDI 1.2 can be rolled out into production environments:

- thorough testing of the toolkit
- adaptation of the infrastructure components (most importantly the Component Registry and the Virtual Language Observatory)

Centres will be able to migrate to CMDI 1.2 at their own pace, and CMDI 1.1 will keep being supported for the time being.

Publications

Twan Goosen, Menzo Windhouwer, Oddrun Ohren, Axel Herold, Thomas Eckart, Matej Ďurčo and Oliver Schonefeld: CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure (http://www.clarin.eu/sites/default/files/cac2014_submission_5_0.pdf)

6.2 Arbil

Arbil is a metadata editor that was initially designed to fill the needs of a defined set of users with a predefined workflow. Since then a great deal of additional functionality has been added, including support for CMDI.

Version 2.5 of the Arbil metadata editor was released in July 2014 (see http://www.mpi.nl/tg/j2se/jnlp/arbil/release_notes-arbil-stable_2.5.txt). This is the first multi-lingual version of Arbil, which can now be set to either English, German, Spanish or Italian. Other new features include the import and export of favourites (which was available as a plugin before and is now integrated into the application), the deletion of fields and nodes directly from the table (rather than only from the tree) and an alternative display mode for the tree ('verbatim XML'), which reflects the exact component structure of CMDI documents. Furthermore, a large number of small improvements and bug fixes are included[10].

A successor version 2.6 has been prepared. It will contain a number of internal fixes and updates, and is expected to be released early 2015.

---

[9] http://www.clarin.eu/node/3921

[10] https://tla.mpi.nl/tla-news/arbil-2-5-release-notes/

6.3 Component Registry

The component registry, browser and editor (altogether commonly referred to as 'Component Registry', http://catalog.clarin.eu/ds/ComponentRegistry) allows metadata modelers to create, edit and store CMDI profiles and makes published profiles available through a public REST interface and a user-friendly web application on top of this service.

In release 1.14.0 user groups have been added. User groups are collectives of users who share access to individual components or profiles; as such, a component or profile is still owned by a particular user, but once made available to a group, every member of that group can access, modify, publish that profile or component or move it to a different group. A simple admin interface is implemented as well. It allows to add/remove groups and add/remove group members for a given group.

Next, the backend-code has been significantly refactored which has made the code easier to debug and less error-prone. In particular, profile- and component-structures have been merged on persistence level into BaseComponent.

Simple data categories no longer show up in ISOcat search results, since the ISOcat tool is not supported since the end of 2014. It is planned to replace ISOcat search by search in CLARIN Concept Registry at the beginning of 2015.

Starting development of a new JavaScript-based front-end, which implements the user interface, is planned for the beginning of 2015. The current front-end architecture and its language Action Script / Flex have become obsolete since they cannot handle effectively the features added over the years of development of Component Registry.

6.4 OAI Harvester

The OAI harvester gathers metadata from centres all over the world and makes them available for other services such as the Virtual Language Observatory.

The new version of the OAI Harvester is now in production. Generally speaking, the provider has performed adequately. To cope with the growing number of records that need to be harvested and made available to the Virtual Language Observatory, harvesting has been distributed over time. That is: smaller endpoints are harvested together while big or slow endpoints are treated separately. Clearly, this can only be a temporary situation. Over the past months preparations have been made to update the harvester. Two major optimizations can be foreseen. The first is to use the HTTP protocol in a more efficient way. The second is to only incrementally harvest. For both these improvements code has been developed and will be integrated in the harvester in the next months.

6.5 Virtual Language Observatory

The VLO (http://www.clarin.eu/vlo) is the low-barrier end-user metadata portal, bringing together all CMDI metadata records within a faceted browser. Significant improvements that have been added in the last two years of CLARIN-D are:

- New whitelist/blacklist mechanisms were implemented that improved recall & precision of extracted facet values.

- A VLO/metadata task force/working group was established to promote user feedback (Haaf et al. 2014)
- User feedback from all CLARIN-D centres was evaluated and the VLO facet configuration was adapted in several feedback rounds.
- Feedback about CMDI profiles with incorrect concept links and incorrectly created CMDI files was given to several CLARIN centres.
- Usability and the user interface was improved (by support of tooltips, a revised resource description page, renamed facets, improved autocomplete functionality, support of new facets, first support of multilingual values etc.)
- Support of themes for the user interface was added (particularly support of a CLARIN-D theme).
- A stricter separation of code and configuration improved the configurability of the VLO.
- Optimization of configuration and deployment process at production server reduced downtimes.
- Extensive bugfixing (heap space problems, error handling, update of library dependencies etc.)
- Extension of unit test set
- Implementation and deployment of an improved web interface (Goosen & Eckart 2014)
- Various improvements in value postprocessing and normalization (for facets like language, organization, national project etc.), support of CLAVAS vocabulary for organization names
- Support of user feedback via the CLARIN-D helpdesk

For an extensive analysis of the VLO see VLO Analysis[11]. A work plan for the further development of the VLO can be found here: Workplan for VLO 3.0[12].

Publications

Susanne Haaf, Peter Fankhauser, Thorsten Trippel, Kerstin Eckart, Thomas Eckart, Hanna Hedeland, Axel Herold, Jörg Knappen, Florian Schiel, Jens Stegmann und Dieter Van Uytvanck: *CLARIN's Virtual Language Observatory (VLO) under scrutiny -- The VLO taskforce of the CLARIN-D centres*. In: *CLARIN annual conference 2014 in Soesterberg, The Netherlands*, 2014

Twan Goosen und Thomas Eckart: *Virtual Language Observatory 3.0: What's New?*. In: *CLARIN annual conference 2014 in Soesterberg, The Netherlands*, 2014

[11] http://www.clarin.eu/node/3886

[12] http://www.clarin.eu/node/3897

7 WG 5: Authentication & Authorization Infrastructure

In the CLARIN preparatory phase, the so-called Service Provider Federation (http://www.clarin.eu/spf) was initiated, cross-connecting Identity and Service Providers from Germany, the Netherlands and Finland. Although this experience was a significant step forward, it was clear that additional steps were necessary to advance the SPF to a state where it can be used as the basis for daily work. In CLARIN-D such steps forward were made, as described in detail in the subsections below.

7.1 CLARIN IdP

To provide users without a functioning IdP (or one that does not release necessary attributes) with a fallback-system to log in to CLARIN SPs, the CLARIN IdP was developed. More details about this can be found at: http://www.clarin.eu/page/3398

To improve the Identity Provider's stability, the backend was migrated in 2013 from a MySQL database to an LDAP server (OpenDJ). The latter does not lead any longer to connection problems from the side of the Shibboleth Identity Provider when the daemon is restarted.

In December 2014 work has started to migrate the CLARIN IdP to RZG and to the domain idp.clarin.eu. Currently the application has been migrated. However, the underlying user store is still running on infra.clarin.eu. Part of the work to make this IdP less dependent on external service is to replicate this user store via LDAP to the IdP server itself. This will ensure a functioning IdP even if the infra.clarin.eu server is unavailable. This work is planned to be finished in February 2015. Investigation of a redundant setup will be performed when this migration has finished.

7.2 CLARIN Discovery Service

With hundreds of Identity Providers to choose from when a user wants to log in, it is important to offer a simple and user-friendly method to select the home organization. With the deployment and configuration of a central discovery service for CLARIN, based on DiscoJuice, this issue was addressed. Instead of browsing through long lists of IdPs, the user can now filter on the fly by typing a part of the organization's name or selecting a geographically nearby IdP. More information about this can be accessed from: http://www.clarin.eu/page/3496

The discovery service has also been moved to a server (VM) running at the RZG data centre. The discovery service is now reachable from the discovery.clarin.eu domain. The underlying data source, used by the discovery service to generate the list of identity providers, is monitored via nagios. Since this service is now running at a highly available VM monitored and operated by the RZG, we are not actively looking into a redundant setup for the moment.

7.3 Web service authentication

In the Dutch BigGrid project several alternatives of accessing web services on the basis of SAML assertions have been investigated (see http://www.clarin.eu/content/web-services-and-aai). In the end using an OAuth2 Authentication Service was proven to be most promising. A first use case involving the CMDI Component Registry and the, now defunct, ISOcat Data Category Registry was proven to be successful. In 2013 the NDG OAuth2 Authentication Server was deployed to the catalog.clarin.eu Service Provider. Its setup and the changes required are now documented at https://github.com/TheLanguageArchive/ndg_oauth.

In the context of one of CLARIN-D's discipline-specific working groups another use case involving web service access to TLA-based resources from a tool developed by the University of Cologne was deployed. Here the use of the SAML/OAuth2 Bridge turned out to be relatively straightforward for both sides, e.g., the server side could use Java Spring's OAuth2 support basically out-of-the-box and the client could use common OAuth2 packages for Python.

Another use case developed in CLARIN-D is to use the SAML/OAuth2 Bridge for connecting WebLicht with the ownCloud solution for private workspaces. First experiments for this are being done in Tübingen. An interesting new aspect of this use case is the draft RFC for OAuth 2.0 Token Inspection[13], which is supported by an ownCloud app[14] but not yet by the NDG OAuth2 Authentication Server. As the Authentication Server runs behind Apache it might be possible to put a wrapper in between that translates the RFC-based requests to the propriety requests of the NDG OAuth2 Authentication Server.

Publication

Jonathan Blumtritt, Willem Elbers, Twan Goosen, Mischa Sallé and Menzo Windhouwer: User Delegation in the CLARIN Infrastructure. In: *CLARIN annual conference 2014 in Soesterberg, The Netherlands*, 2014

8 WG 6: Workspaces and Hosting

8.1 Workspaces

A *Personal Workspace* in CLARIN-D consists of online storage for individual users, which can be accessed via a programming API by CLARIN-D applications. After evaluation of several software packages, OwnCloud was deemed best suited for use as the basis software for implementing personal workspaces in CLARIN-D.

The Forschungszentrum Jülich (FZJ) hosts a test installation of OwnCloud for CLARIN-D, which allows login via Shibboleth. Although CLARIN-D applications also allow login through Shibboleth, the security architecture needs to be extended for a seamless single sign-on experience on the user side. The Shibboleth authentication allows authenticated communication between a web application (or other online resource) and a user (typically through a web browser). However, to be able to exchange data between Owncloud and CLARIN-D web services, the web applications need to act on behalf of the user. This is currently not possible using only Shibboleth protocols. The current focus within this working group has been extending authentication mechanisms used in the CLARIN-D infrastructure to allow web applications or services to access the Owncloud storage service on behalf of an authenticated Shibboleth user. For this purpose, combining the Oauth protocol with the Shibboleth authentication is being considered, and a prototype implementation of the combined architecture is under development. The authentication scheme that will be the outcome of this effort can also be used by other applications and services with similar requirements.

[13] https://tools.ietf.org/html/draft-ietf-oauth-introspection-04

[14] https://github.com/owncloud/apps/tree/master/user_oauth

8.2 Hosting

Three computing centres are partners in CLARIN-D: the Rechenzentrum Garching of the Max Planck Society and the IPP (RZG), the Forschungszentrum Jülich (FZJ) and the Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG). These three computing centres are responsible for hosting the CLARIN-D services. In close cooperation with the computing centres, the following division of hosting tasks was created:

Computing Centre	**Hosting Tasks**	**Comment**
RZG	Web services	The Stuttgart Dependency parser was implemented as a web service by the Stuttgart CLARIN-D centre and deployed at the RZG. Afterwards, the parser was applied to the TüBa-D/DC
	WebLicht failover	Virtual Machine that will provide load balancing as well as act as a failover in case of migration/maintenance of the Tübingen-hosted main server
	Centre registry	Hosted by the RZG – see section 4.2 for details
	Discovery Service	Fully migrated from MPI to RZG, to ensure higher availability
	Identity Provider	Partially migrated from MPI to RZG, further migration planned
	VLO alpha version	Test environment for new versions of the VLO
	Mind Research repository	Data repository for (largely psycholinguistic) experiment data, including statistical data bundles, see http://openscience.uni-leipzig.de/
FZJ	Workspaces	Software is installed and available as beta service. Integration with other CLARIN-D applications is in progress.
	Helpdesk	OTRS installation is online, in close cooperation with the Hamburg CLARIN-D centre
	Monitoring	Installed and configured by FZJ
	PiWik user statistics	An installation of PiWik (http://piwik.org) is planned
GWDG	PID services via EPIC consortium	EPIC API version 2 is in production

9 WG 7: Web services, General services

WebLicht is an execution environment for automatic annotation of text corpora. Linguistic tools such as tokenizers, part of speech taggers, and parsers are encapsulated as web services, which can be combined by the user into custom processing chains. The resulting annotations can then be visualized in an appropriate way, such as in a table or tree format.

The WebLicht web application and its web services are constantly being further developed and improved. There were also some new web applications and web services introduced in the third year of the project.

New web services include:

- Orthographic Canonicalizer (CAB) for orthographic normalization of historical texts
- Lexical Database service (dlexDB) to obtain statistical information for lexical types
- A morphological analyser and parsers for Dutch, English, and German
- Externally Trained Named Entity Recognizer which uses external training models

Many of the existing services have been extensively tested for reliability and scalability with the Bombard command-line tool, which simulates over 80 simultaneous users by submitting both short texts and novels to webs ervices. Bombard will be made available to the developer group in the near future. Frameworks for improving both vertical and horizontal scalability were evaluated and necessary extensions are being developed for our needs. The results will be presented to the web service developers, so that better performance of more web services can be achieved.

In addition to reliability and scalability testing, existing web services have been evaluated in terms of quality of output and many were retrained or otherwise updated to produce better output.

The most recent version of the WebLicht web application includes new features and integrates more CLARIN-D components:

- CLARIN-D HelpDesk integration for users to ask questions or give feedback
- Tündra integration to search and view parsed result sets
- Introduction of predefined tool chains which have been tested for reliability and quality
- 3 modes of operation based on user expertise

Several new general services have been added or are under construction:

- WaaS (WebLicht as a Service) can be used to invoke WebLicht chains programmatically or from the command line.
- As part of a WebAnno-WebLicht cooperation and as an effort to integrate manual and automatic linguistic tools into a common workflow, an online tool for training named entity models was created. The NER Model Trainer tool is accessed via a web interface (http://weblicht.sfs.uni-tuebingen.de/rws/service-opennlp/train-ner-model.html). The efficiency of manual annotation work can be improved by using the NER Model Trainer and the Externally Trained NER web service together.

10 WG 8: Simple store

The Simple Store in CLARIN is intended to provide a way for researchers without deep technical background and without a direct connection to a CLARIN centre to easily save their research data and reference it by a PID. The specification of the simple store does not include requirements with regard to Quality Assurance (QA) by an archive manager nor data descriptions in form of structured and detailed metadata. Access control can be simple, allowing either public access or access only by the data

depositor. Due to cooperation with EUDAT, the system to be employed for the Simple Store is EUDAT's B2SHARE (see https://b2share.eudat.eu), which is a productive service in EUDAT and fulfills CLARIN-D's needs for the Simple Store. Its functionality extends beyond these requirements, for example by providing a community based comment and evaluation system, instead of an expert QA, which might especially be suitable for the long tail of research data not in the focus of CLARIN centres.

11 WG 9: Monitoring

To achieve a high level of service it is critically important to have automatic checks (probes) in place for CLARIN repositories, web applications and web services. As specified in document CLARIND-AP3-005 (http://www.clarin.eu/specification-documents), a careful analysis showed that a standard package as Nagios (or the compatible forked version, Icinga) fulfills these monitoring needs. In cooperation with the Forschungszentrum Jülich, Nagios was installed and several plugins to monitor services were deployed. This has been extended in several ways:

- The setup of a public webpage where CLARIN users can check the status of the centres and their services (http://www.clarin-d.de/en/news/status-infrastructure.html).
- Automatic checks based on the information contained in the centre registry, like OAI-PMH and SRU/CQL endpoints.
- Checks to see if the issued handles are resolving and how long this resolution process takes.

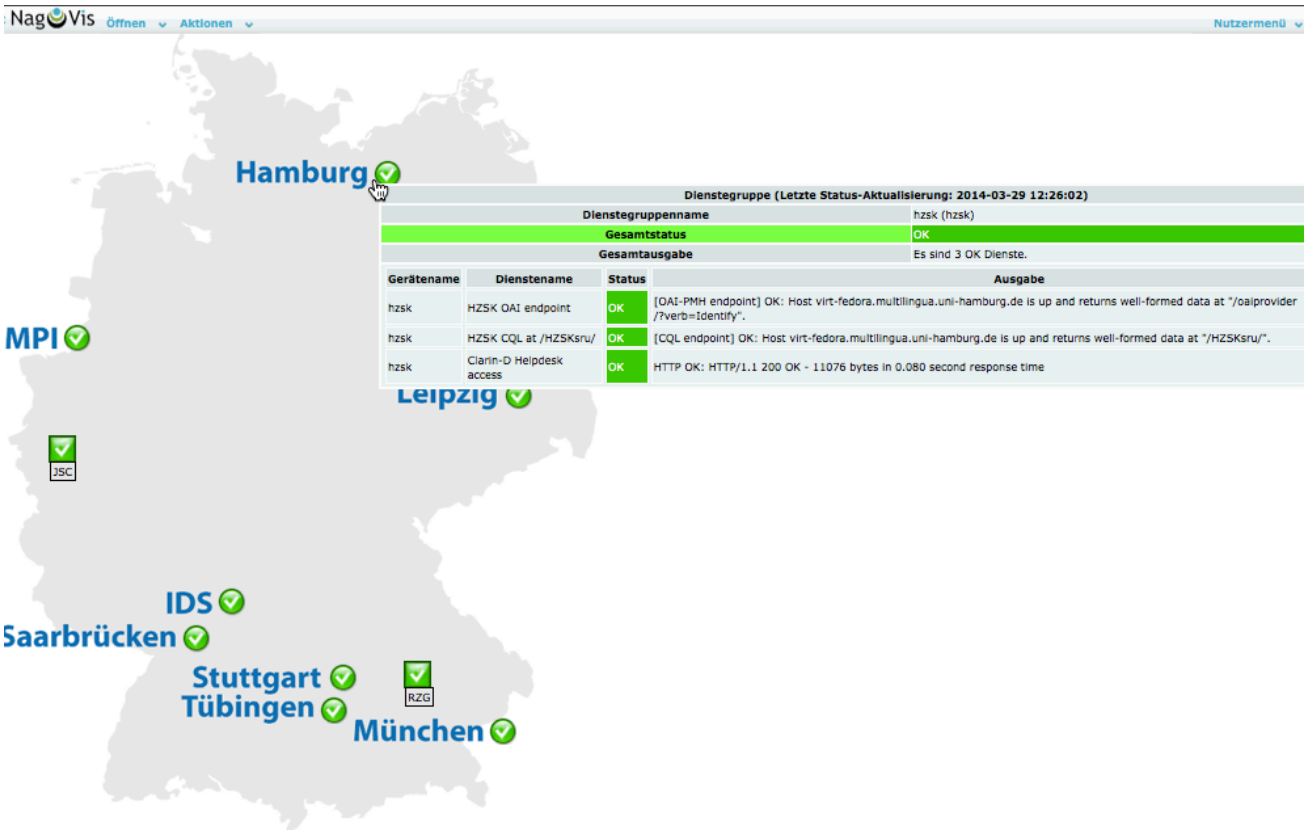

Figure: Nagios monitoring system

12 WG 10: Federated Content Search

The goal of the CLARIN Federated Content Search (CLARIN-FCS) - Core specification is to introduce an interface specification that decouples the search engine functionality from its exploitation, i.e. user-interfaces, third-party applications, and to allow services to access heterogeneous search engines in a uniform way.

All components of the CLARIN Federated Content Search (FCS) infrastructure have been further developed. The endpoints have implemented the interface specification and have added more resources. The user interface for the aggregated search (the Aggregator) is being evaluated and revised. The supporting environment has continued to develop, including:

- SRUClient, SRUServer,
- FCSSimpleEndpoint Java libraries and
- the Aggregator

Specific work done in the CLARIN-D FCS components include:

Interface Specification:

The initial FCS specification based on SRU/CQL-protocol was not very well structured and intuitive and therefore was revised to become clearer. The revised specification was approved by the CLARIN Centre Committee and now has the state of an official CLARIN specification.

Now the work is focused on extending the specification to support more advanced query scenarios, e.g. support for linguistic annotation layers like part-of-speech (POS). However, this is a challenging task, because of the heterogeneous nature of the resources, i.e. the different tag sets that are used by the centres. Furthermore, CQL's syntax will need tailoring to support querying linguistic structures and features. Currently, it is being investigated, if upgrading to SRU 2.0 is worthwhile. In addition a possible switch to the CQP query language[15] for advanced queries is investigated. On the one hand there are a number of endpoints that use CQP in their search engine and on the other hand the language is more familiar for the researches than CQL. For POS tags, the Universal Part-of-Speech tagset[16] is investigated.

CE-2014-0317 – Federated Content Search – Data Views

CE-2014-0316 – Federated Content Search – Core Specification

Aggregator:

The Aggregator web application is currently in the process of being updated and improved. The user interface has been completely redesigned and now presents a lighter look and feel. It has also been completely rewritten using modern web technologies, which offer better responsivity to the user's actions. The Aggregator backend, on which the user interface rests upon, has also been heavily modified and now makes use of the latest version of the SRUClient library, which supports the new FCS

[15]Corpus Workbench Query Language, http://cwb.sourceforge.net/documentation.php
[16] Petrov et al. (2012), "A Universal Part-of-Speech Tagset", Proceedings of LREC 2012, Istanbul, https://code.google.com/p/universal-pos-tags/ or alternatively Universal Dependencies, http://universaldependencies.github.io/docs/

specification. These changes all for a better presentation of multiple search hits found in the same text fragment and a more efficient way of harvesting collections metadata from each endpoint.

The compatibility with VLO has been achieved for most of the endpoints and their resources. Aggregator conformance with the current FCS specification was improved (e.g. added support for the querying of 'next' records). Based on a formal usability study and users' feedback, the user interface was improved and the integration with WebLicht implemented. The user can now go directly from their search results in the Aggregator to the WebLicht application and apply available linguistic tools on their search results.

http://weblicht.sfs.uni-tuebingen.de/Aggregator-testing/ (NB: this version is in an early state and thus may contain bugs or be unavailable)

Endpoints/Resources:

Work on compliance with the revised FCS specification is slightly delayed, because the work on the Aggregator is not yet finished.

However, some more resources have been made available via the current FCS specification: LINDAT-Clarin has added the HamleDT 2.0 corpus (13 languages) and the Prague Dependency Treebank (multiple versions). DANS has contributed an experimental endpoint.

Once the revised Aggregator is in beta stage, centres asked to upgrade to the revised FCS specification. This is scheduled to start mid-2015.

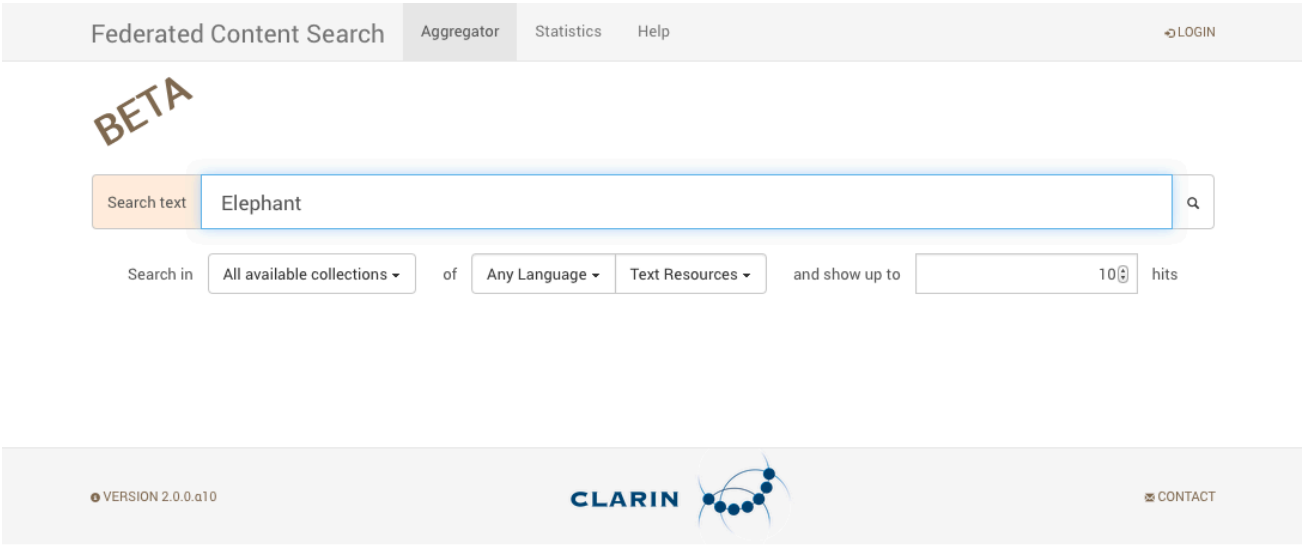

Figure: The CLARIN-D Federated Content Search Aggregator

13 WG 11: Replication

Within the context of the Clarin-D and EUDAT projects and as a proof of concept, the University of Tübingen has developed a replication software prototype to ensure long-term, reliable access to the

resources stored in its Fedora repository. Integrating with the B2SAFE (http://www.eudat.eu/b2safe) replication service offered by EUDAT, this utility is currently used for backing up the data from Tübingen to the Rechenzentrum Garching (RZG).

This work is being further developed within EUDAT into a more general and mature solution for replicating Fedora repositories, which will be available to the Clarin-D centres, improving the overall resilience, integrity and availability of Clarin-D resources.

The prototype package can only be used for backup, whereas the final package will allow access to data from its replication site, using the infrastructure provided by the EPIC PIDs.

13.1 Technical Details of the Prototype Software

The Clarin-D repository in Tübingen is backed up by a Fedora Commons system and is managed using a custom user interface written in-house. A digital object contains:

- CMDI metadata, the Clarin standard for metadata, as a special data stream
- contextual, human-targeted data such as scientific papers or other documents
- primary or processed research data

The repository also uses the PID system for referring a digital object. A PID points to the CMDI data stream of a digital object and the CMDI references all of the other data streams.

Fedora Commons can use various storage backends. The default one is Akubra, an open-source, pluggable file storage interface. Fedora and Akubra are configured in such a way that the data hosted in the repository is immutable. Modifying a data stream only creates a new version of the data; the old version is still in the repository and can be easily accessed. Due to this workflow, backing up a versioned Fedora datastream can be done safely on the file system level.

iRODS is a data grid software system with policy based data management facilities and data interfacing and sharing capabilities. iRODS is open source under a BSD license, is community-driven, has a simple installation procedure and is hardware agnostic, working on all major platforms. It is used in EUDAT for building a data federation across computing centres and, as part of the data federation, it is also the means for performing safe replication of data from one site to another.

An important choice made early in the implementation phase was to keep the existing Fedora repository independent of the replication service. This constraint directed the implementation: we decided to mount the file system directory that contains the Fedora Commons data as an external collection into the iRODS system.

Each new file being replicated is first assigned a new PID, only for the purpose of Safe Replication; this PID is stored locally. The file is then replicated via iRODS and subsequently, via the EUDAT iRODS rules, has the PID location field updated.

Using the replicas, the restoration of a corrupted Fedora repository is simple and fast. First the data is transferred from the replication site to the local site, using the common iRODS tools for data transfers. Then the Fedora restoration scripts are executed and the repository is set back online. Finally, a data integrity check is performed for assurance.

14 Frontend developments

In the past, the technical development has mainly focused on the backend of the technical infrastructure and the required APIs, protocols and frameworks, including the technical quality and robustness. A new focus has been laid on the user side of the infrastructure. Though interfaces exist to all services, these were defined primarily for the participants of the infrastructure development, sometimes leaving out important parts for the users of the Social Sciences and Humanities. Current plans are to provide interfaces with an increased usability to non-technical users, in order to provide a better user experience to those intended users. This requires a redesign of user interfaces, reducing the information load on the technical side and providing more application-oriented access to the users. A first version of a new portal allowing direct access to the services is expected at the end of March, 2015, allowing usability tests with the CLARIN user community. A launch of the website is planned for June 2015.

For existing tools and services the user focus was added by inserting helpdesk functionality. Using the OTRS helpdesk system, a ticketing system for user requests was implemented that allows the distribution of tickets to the experts via a central point of access. The helpdesk system also records the answers for future reference and for inclusion in FAQs.

Another user group is targeted by CLARIN-D by means of a repository of teaching and training material. With TeLeMaCo, a central site exists that allows registering teaching material on the use of tools, technologies and resources from CLARIN-D. Sharing this material allows the reuse of high quality teaching material and tutorials for those working on providing their own tools and services or using the ones available in CLARIN-D.

15 Summary

At the beginning of CLARIN-D's third year the continuous efforts have been clearly rewarded by the certification of all 9 German CLARIN centres. The foundations for a reliable research infrastructure are thus in place. That does not mean work has been completed — for year three and four the real challenge lay and still lie in a better integration of the existing components and improvement of the user interfaces.

Another important aspect, which needs attention, is the organizational flexibility. Within a five-year construction phase many institutional and personal factors change over time. While it is not always simple to deal with this, the migration of several A-services, such as the necessary shift from ISOcat to CLARIN Concept Registry and the migration of the AAI services to computing centres, can serve as an example. Due to the distributed centre architecture CLARIN was able to accommodate for this fundamental change with a minimal impact on the running services. This shows the robustness of CLARIN's distributed setup. At the same time the use of components such as the Centre Registry increase this robustness by allowing to access resources and service APIs in this dynamic distributed environment with high reliability. This shows the increasingly flexible nature of CLARIN's distributed setup.