CLARIN-D AP3 report: establishing the technical infrastructure

May 2012 – April 2013

March 2013

CLARIN-D, BMBF-FKZ: 01UG0801B

Responsible: Sebastian Drude

Editors: Dieter Van Uytvanck, Peter Wittenburg, Thomas Zastrow
Contributors: Willem Elbers, Twan Goosen, Herman Stehouwer, Menzo Windhouwer

1 Introduction to CLARIN-D

The BMBF project CLARIN-D is developing a digital infrastructure for language-centered 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 this infrastructure out 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. Florian Schiel)
- Berlin-Brandenburgische Akademie der Wissenschaften (Prof. Dr. Wolfgang Klein)
- 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 centers, see chapter 8.

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.

CLARIN-D's work packages reflect the project's various objectives and tasks. Leadership for each work package is assigned to a specific center, depending on the experience and expertise of particular CLARIN-D service centres. Work package 3, „Technical Infrastructure“ is led by the MPI Nijmegen. Below is a list of the specialized technical infrastructure working groups that have done the bulk of the tasks in this work package as well as their specific contributions. Note that rather than including long descriptive texts, we have tried to link to web pages that, in the spirit of CLARIN, are actively maintained and may be updated from time to time..

2 AG 1: Repositories

2.1 Setup of Repositories

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

In the course of CLARIN-D's first year, all nine center candidates have each made serious efforts to establish a repository. After a workshop devoted to this subject (http://www.clarin.eu/events/3443) and some internal discussions, most (7 out of 9) centers have selected Fedora Commons (http://www.fedora-commons.org) as a repository system, often with PrOAI (http://proai.sourceforge.net/) as a module to export the metadata via OAI-PMH.

Two centers with a longer archiving tradition – BAS and the MPI – did not choose Fedora. The MPI continues to use its self-built IMDI & Lamus repository system, extending it to support CMDI. BAS initially experimented with Fedora but decided in the end that it was too heavy weight for its purposes and set up a self-developed system.

2.2 Center Assessment

An important part of acquiring the official CLARIN center status is to be assessed. All 9 CLARIN-D centers will take part in 2 assessment procedures:

- By the CLARIN center assessment committee (see http://www.clarin.eu/sc-centers), which will analyze the center's offer in terms of compliancy with the B center requirements. All 9 centers are aiming for the B status; MPI, IDS and Tübingen University are also offering A-services in addition to that. It is the first time that this assessment takes place; hence it is expected that quite some interaction will be necessary to clarify certain points with the center representatives.
- By the Data Seal of Approval committee (http://www.datasealofapproval.org/), which is an independent body that checks if a center's data management strategy and it's policy are suitable for long-term archiving.

By the end of February 2013 all CLARIN-D centers submitted the documents needed to be assessed by both bodies. Authoring these documents was for all of the centers a good incentive to reconsider and to update their long-term archiving and service policies.

The results of the assessments are expected by May 2013.

3 AG 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.

However, this approach needs some reliable components to be in place. Therefore the PID workgroup has undertaken the following activities:

Integrate handles in the center repositories

All CLARIN-D centers have developed a PID strategy. Most of them are or will be connected to the EPIC service to acquire and manage handle PIDs. The BBAW currently has URN:NBNs in place but is about to issue handles. The MPI and IDS do not plan to use EPIC – they have their own prefix and handle server.

Setting up and testing part identifiers

For the citation of fragments of resources there is a 2-step resolving procedure, compatible with the ISO draft on part identifiers (see http://www.clarin.eu/node/269), which has been implemented by the EPIC consortium. Details about the use of part identifiers in the CLARIN-D context can be found at:

http://www.clarin.eu/faq/243

Training course on the use of EPIC services

On September 8, 2011, a tutorial about PIDs and the EPIC service was organized at the MPI for the CLARIN-D centers:

http://www.clarin.eu/events/3443

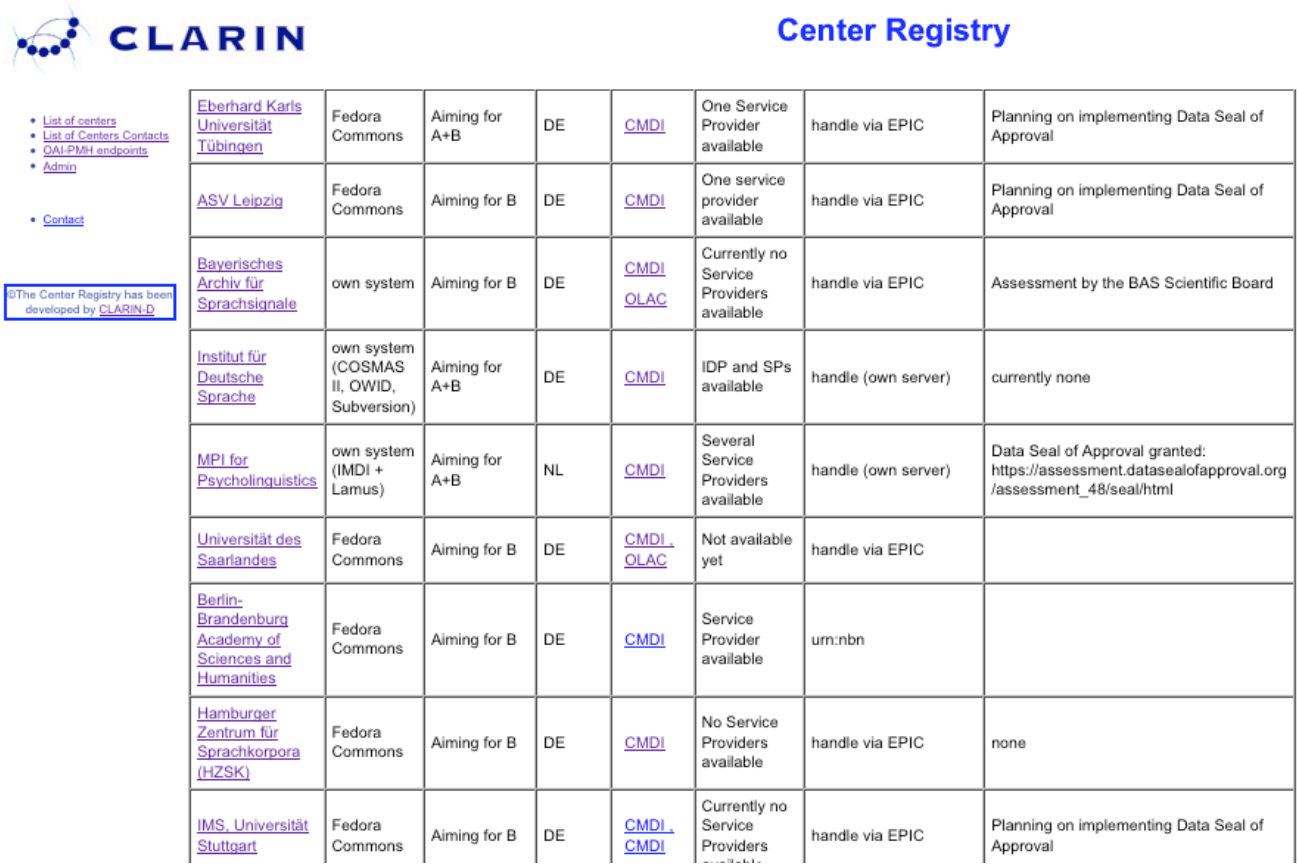
A similar workshop, in cooperation with EUdat, and – next to EPIC – also focusing on the DataCite handle service took place in June 2012:

http://www.eudat.eu/1st-training-days-presentations

4 AG 3: Registries

4.1 Center Registry

As 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, it is clear that CLARIN needs a machine-readable registry where such pointers can be stored and accessed. Additionally, the need for a list of centers, with e.g. contact information (for technical or administrative questions) resulted in the concept of a center registry. The requirements for this web-accessible center database are outlined in document CLARIND-AP3-001 (see http://www.clarin.eu/specification-documents) and were used by the RZG computing center as the basis for an implementation (see http://centerregistry-clarin.esc.rzg.mpg.de). At the moment of writing a production version of the center registry is available, both for human and programmatic access (via a web site and a REST interface, respectively). It is filled with information about the CLARIN-D centers and will be extended with the other centers in other countries.



Eberhard Karls Universität Tübingen	Fedora Commons	Aiming for A+B	DE	CMDI	One Service Provider available	handle via EPIC	Planning on implementing Data Seal of Approval
ASV Leipzig	Fedora Commons	Aiming for B	DE	CMDI	One service provider available	handle via EPIC	Planning on implementing Data Seal of Approval
Bayerisches Archiv für Sprachsignale	own system	Aiming for B	DE	CMDI OLAC	Currently no Service Providers available	handle via EPIC	Assessment by the BAS Scientific Board
Institut für Deutsche Sprache	own system (COSMAS II, OWID, Subversion)	Aiming for A+B	DE	CMDI	IDP and SPs available	handle (own server)	currently none
MPI for Psycholinguistics	own system (IMDI + Lamus)	Aiming for A+B	NL	CMDI	Several Service Providers available	handle (own server)	Data Seal of Approval granted: https://assessment.datasealofapproval.org /assessment_48/seal/html
Universität des Saarlandes	Fedora Commons	Aiming for B	DE	CMDI , OLAC	Not available yet	handle via EPIC	
Berlin-Brandenburg Academy of Sciences and Humanities	Fedora Commons	Aiming for B	DE	CMDI	Service Provider available	urn:nbn	
Hamburger Zentrum für Sprachkorpora (HZSK)	Fedora Commons	Aiming for B	DE	CMDI	No Service Providers available	handle via EPIC	none
IMS, Universität Stuttgart	Fedora Commons	Aiming for B	DE	CMDI , CMDI	Currently no Service Providers	handle via EPIC	Planning on implementing Data Seal of Approval

Figure: the HTML interface of the center registry

4.2 Data Category Registry

A stable ISOcat (http://www.isocat.org/) was achieved in the past year. Next to developments in the area of usability (users can now change their password) and performance (optimizing the database to XML mapping) the focus has been on supporting some form of standardization. The ISO standardization process has been implemented and tweaked (introduction of a timed ballot for the evaluation and validation phase). HTML rendering of data categories has been improved to function as discussion documents for the decision groups. However, none of the current Thematic Domain Groups have started actual standardization process for any data categories. But user communities, like CLARIN-NL, do actually need ways to guide their members to approved subsets of the DCR. To accommodate this ISOcat now supports views, e.g., http://www.isocat.org/interface/index.html?view=CLARIN-NL/VL. In a view, only data categories which are in a selection shared with the group are accessible. These selections shared with the group thus form the community-approved subset of the DCR. Steps are also undertaken to prepare ISOcat to connect with the CLARIN Shibboleth AAI infrastructure.

Publications

O. Crasborn, M. Windhouwer. ISOcat data categories for signed language resources. In the *9th International Gesture Workshop* (GW 2011), Athens, Greece, May 25-27, 2011

I. Schuurman, M. Windhouwer. Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have To Offer? In the *Proceedings of the 2nd Supporting Digital Humanities conference* (SDH 2011). Copenhagen, Denmark, November 17-18, 2011. (draft)

M. Windhouwer, S.E. Wright. Linking to linguistic data categories in ISOcat. In C. Chiarcos, S. Nordhoff and S. Hellmann (eds), *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata* (LDL 2012), © Springer-Verlag, pp 99-107, Frankfurt/Main, Germany, March 7-9, 2012. (draft)

Tutorials

I. Schuurman, M. Windhouwer. *CLARIN-D ISOcat tutorial*. Berlin, Germany, January 11, 2012.

I. Schuurman, M. Windhouwer. ISOcat in daily life: possible uses of a large repository of widely used linguistic concepts. Tutorial at *LREC 2012*.

4.3 Relation Registry

The alpha version of RELcat has seen the addition of various new sets of relations. These relation sets are accessible via a revised API, which (among other improvements) supports SPARQL queries on multiple sets. The new API also supports various representations including various graphical formats. The sets and API are available at http://lux13.mpi.nl/relcat/. Ongoing design efforts focus on the proper taxonomy of relation types and its intended usage.

Publications

I. Schuurman, M. Windhouwer. Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have To Offer? In the *Proceedings of the 2nd Supporting Digital Humanities conference* (SDH 2011). Copenhagen, Denmark, November 17-18, 2011. (draft)

M. Windhouwer, S.E. Wright. Linking to linguistic data categories in ISOcat. In C. Chiarcos, S. Nordhoff and S. Hellmann (eds), *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata* (LDL 2012), © Springer-Verlag, pp 99-107, Frankfurt/Main, Germany, March 7-9, 2012. (draft)

M. Windhouwer. RELcat: a Relation Registry for ISOcat data categories. Accepted for a poster and demonstration at *LREC 2012*. (draft)

4.4 Schema Registry

Initial ideas for the design of SCHEMAcat have been written down and an early alpha version provides access to some of the first schemata. The MPI handle server has also been extended to hand out PIDs for these schemata. Special attention has been paid to support for the non-XML ISO 14977:1996 EBNF notation, which is useful for defining tagsets, e.g. the Dutch CGN or (in the future) the German STTS, and annotating them with ISOcat data category PIDs.

Publications

I. Schuurman, M. Windhouwer. Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have To Offer? In the *Proceedings of the 2nd Supporting Digital Humanities conference* (SDH 2011). Copenhagen, Denmark, November 17-18, 2011. (draft)

5 AG 4: CMDI

The Component Metadata framework forms the base of metadata modeling 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

5.1 CMDI core schema

At the backend of CMDI the following changes took place:

- Addition of extra schematron checks
- All profiles and components have been checked and unused ones were removed after discussing this with the respective owners
- Replacement of some proprietary XSLT extensions by standard XSLT v2 elements
- Initial proposals have been made to implement a basic versioning mechanism for components/profiles and CMDI instances (currently at the CLARIN developers wiki, soon to be released as an official specification document)
- An initial draft (see CLARIND-AP3-007 at http://www.clarin.eu/specification-documents) was drafted regarding common issues when modeling metadata (granularity, hierarchies, cycles in metadata instances, etc.)

5.2 Arbil

A new version of the Arbil desktop metadata editor (http://tla.mpi.nl/tools/tla-tools/arbil) has been released with improved handling of CMDI and a better user experience for CMDI users (Arbil also supports the IMDI metadata standard). The changes include:

- Arbil has been made compatible with the latest version of the CMDI specification. It allows the user to read, create and edit all specified components and elements, attributes on both components and elements, and deal with controlled vocabularies, resource proxies, data categories and display priority.
- A checkbox has been added to the profiles dialog that toggles filtering of all available CMDI profiles, so that only the ones selected for metadata editing in the Component Registry are displayed
- Descriptions contained in ISOCat data category definitions are now used by Arbil as field descriptions in instances where no explicit field documentation is provided
- A wizard has been added that guides the user through the process of choosing a metadata format and selecting profiles
- Improved editing modes, local search and tree representation

5.3 Component Registry

The component registry, browser and editor (commonly referred to as 'Component Registry') 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. Over the last year, a substantial number of improvements and new features have been implemented and released to the server (http://catalog.clarin.eu/ds/ComponentRegistry/). Changes include:

- The backend has been re-implemented to use a relational database instead of file system storage
- Attributes can be added and edited on components.
- Concept links (data categories) can be assigned to components and attributes of components and elements
- Private components can be edited if they are used in other components, a warning message is shown in Flex UI
- Validation of component specifications (upon import/editing) now also supports Schematron rules in the general component schema
- Client side validation while editing has been extended
- Comments can be posted on profiles and components
- Recursion detection with respect to components takes place at every component/profile registration or update
- Users can set their display name through a web form linked from the Flex UI
- Components and Profiles can be monitored via an RSS feed

5.4 OAI Harvester

The main application[1] that is responsible for the harvesting of metadata records from the OAI providers has been significantly improved over the last year. These changes include:

- The implementation of a web-GUI to inspect the harvested metadata files for each center: http://catalog.clarin.eu/oai-harvester/
- The creation of a regularly downloadable set of metadata files (for consumption by e.g. other search engines): http://catalog.clarin.eu/oai-harvester/resultsets/
- A flexible mapping (for OLAC files) from the OAI identifier to a MdCollectionDisplayName, which results in clear collection names in the VLO.
- The code has been completely revisited and made far more robust.

[1] There are other applications which are doing this, for example the WebLicht orchestration data harvester

- The configuration of the harvester has been streamlined.

5.5 Virtual Language Observatory

The VLO is the low-barrier end-user metadata portal, bringing together all CMDI metadata records within a facet browser. Significant improvements that have been added in this first year of CLARIN-D are:

- The addition of a "national project" facet, giving the user the possibility to explore, e.g., all the resources within CLARIN-D (http://catalog.clarin.eu/ds/vlo/?fq=nationalProject:CLARIN-D)
- Links to a Language information page (e.g. http://www.clarin.eu/external/language.php?code=deu)
- The conversion script that creates CMDI instances for the records of the CLARIN LRT inventory was extended and updated.
- The backend of the VLO has been re-engineered to ensure easy deployment of new versions and easier debugging (e.g. moved from ad-hoc configuration scripts to maven and standardized configuration files).
- A link was implemented between the VLO and the Federated Content Search aggregator, making it possible to perform a content search within a set of resources that was first identified with the facet browser.
- A feedback button and webform has been setup, to stimulate users to report metadata anomalies.

An extensive description of the technology and workflow of the VLO can be found at:

http://pubman.mpdl.mpg.de/pubman/item/escidoc:1454694:6

6 AG 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.

6.1 Training

At the CLARIN-D tutorial in Nijmegen there was a special session on the use of Shibboleth for Service Providers: http://www.clarin.eu/events/3443

6.2 Testing

To check if Service Providers receive some important attributes (user name, email address, etc.) from the German Identity Providers, the CLARIN-D centers started checking mutual connections between

IdPs and SPs. This lead to the (worrisome) conclusion that a significant portion of the IdPs in the DFN-AAI federation do not release enough information to ensure a well-functioning setup. As a (hopefully temporary) alternative, the CLARIN IdP (see below) was launched. At the same time, CLARIN-D and DARIAH-DE released a call for action to the German Identity Providers and the DFN-AAI, stressing the importance of releasing personal attributes to trustworthy academic SPs:

http://www.clarin.eu/page/3500

6.3 CLARIN IdP

To provide users without a functioning IdP (or one that does not release necessary attributes, like eduPersonPrincipleName) 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

In addition to this, the requirements from some of the major resource providers (BBAW and IDS) are being analyzed to find out how strict the procedure to allow new users to the CLARIN IdP should be.

6.4 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

6.5 Web service authentication

Combining AAI and web services is, in the current context of web-based authentication, a difficult problem. In cooperation with the Dutch BigGrid project, some options in this field were and are investigated, see: http://www.clarin.eu/page/3482

At the same time, the CLARIN-D developers are preparing a more pragmatic short-term solution by constructing a trust network based on server SSL certificates (i.e. public key encryption).

7 AG 6: Workspaces and Hosting

7.1 Workspaces

A *Personal Workspace* in CLARIN-D consists of two parts:

- Online storage which is managed in a computing center and which belongs to an individual user.
- A programming API which makes it possible to embed the online storage into CLARIN-D applications.

Online storage and the API together allow a seamless flow of data between the single applications of CLARIN-D. In the first year of CLARIN-D, several workflow and software solutions were tested by the CLARIN-D partners. Finally, the decision was made to make use of the OwnCloud software (http://www.owncloud.org/) which fits the CLARIN-D requirements:

- OwnCloud is OpenSource and can be easily extended
- It offers a convenient user interface
- Integration of the online storage via WebDAV into a users local machine
- The possibility of using Shibboleth Sign On for user authentication
- Well defined and exhaustive application programming interface

The Forschungszentrum Jülich (FZJ) hosts a test installation of OwnCloud for CLARIN-D. This test installation is used to implement the integration of the OwnCloud software into the components of the CLARIN-D infrastructure, for example WebLicht, the Federated Content Search etc.

Figure: The OwnCloud web interface

8 Computing Centers

Three computing centers 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 centers are responsible for hosting the CLARIN-D services. In close cooperation with the computing centers, the following division of hosting tasks was created:

Computing Center	Hosting Tasks	Status	Comment
RZG	Web services	✔	The Stuttgart Dependenz parser was implemented as a web service by the Stuttgart CLARIN-D center and deployed at the RZG. Afterwards, the parser was applied to the TüBa-D/DC
	Center registry	✔	Implemented by the RZG – already in use for the WebLicht harvesting
FZJ	Workspaces	✔	See above: the software is in place and can be used
	Helpdesk	(✔)	Concept done by the Hamburg CLARIN-D center in cooperation with FZJ, test installation is online
	Monitoring	✔	Installed and configured by FZJ
GWDG	PID services via EPIC consortium	(✔)	EPIC API version 1 is in productive status and used, version 2 is in beta and is being tested

9 AG 7: Webservices, Generic services

9.1 Webservices

In the first year of CLARIN-D, the WebLicht web service infrastructure was further developed and reached version 2. With this version, WebLicht is deeply integrated into the CLARIN-D infrastructure:

- Every web service is described with CMDI metadata
- The web services are identified via PID
- The metadata of the web services is stored in the center's repositories and harvested by the chaining engine from Tübingen
- A web services integrates the ISOcat service into WebLicht
- Further development of the *Text Corpus Format* (TCF)

In addition, the following functionality was added to WebLicht:

- Parameterization of web services
- Integration of the CQP query engine
- New user interface

New web services were developed:

- Audio-Video web services (München / MPI)

- Dependence parser (Stuttgart / Tübingen)

9.2 Generic Services

To establish interoperability between different resources and tools, converters are necessary. Converters in CLARIN-D are implemented as web services. WebLicht contains a growing number of converters for widely used linguistic data formats. For the following data formats, converters are already available or planned to be realized in the near future:

- LAF/MAF
- TEI
- Negra Export
- TigerXML
- Paula
- Exmeralda
- TEI-Drop
- EAF
- Folker
- BPF

10 AG 8: Simple store

Within the context of CLARIN-D's first year, setting up a simple store was considered to be one of the tasks better taken up at a later time. In the second year the Simple Store concept was adopted by EUDAT (see http://www.eudat.eu/simple-store), which is currently implementing a prototype service that at the first sight seems to fulfill CLARIN-D's needs. Close coordination with EUDAT on this topic is of course foreseen.

11 AG 9: Monitoring

To achieve a high level of service it is utterly 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 is expected to be extended in 4 ways:

- The setup of a public webpage where CLARIN users can check the status of the centers and their services (to be available by June 2013).
- Automatic checks based on the information contained in the center registry, like OAI-PMH and SRU/CQL endpoints.
- Including more detailed probes for each center.

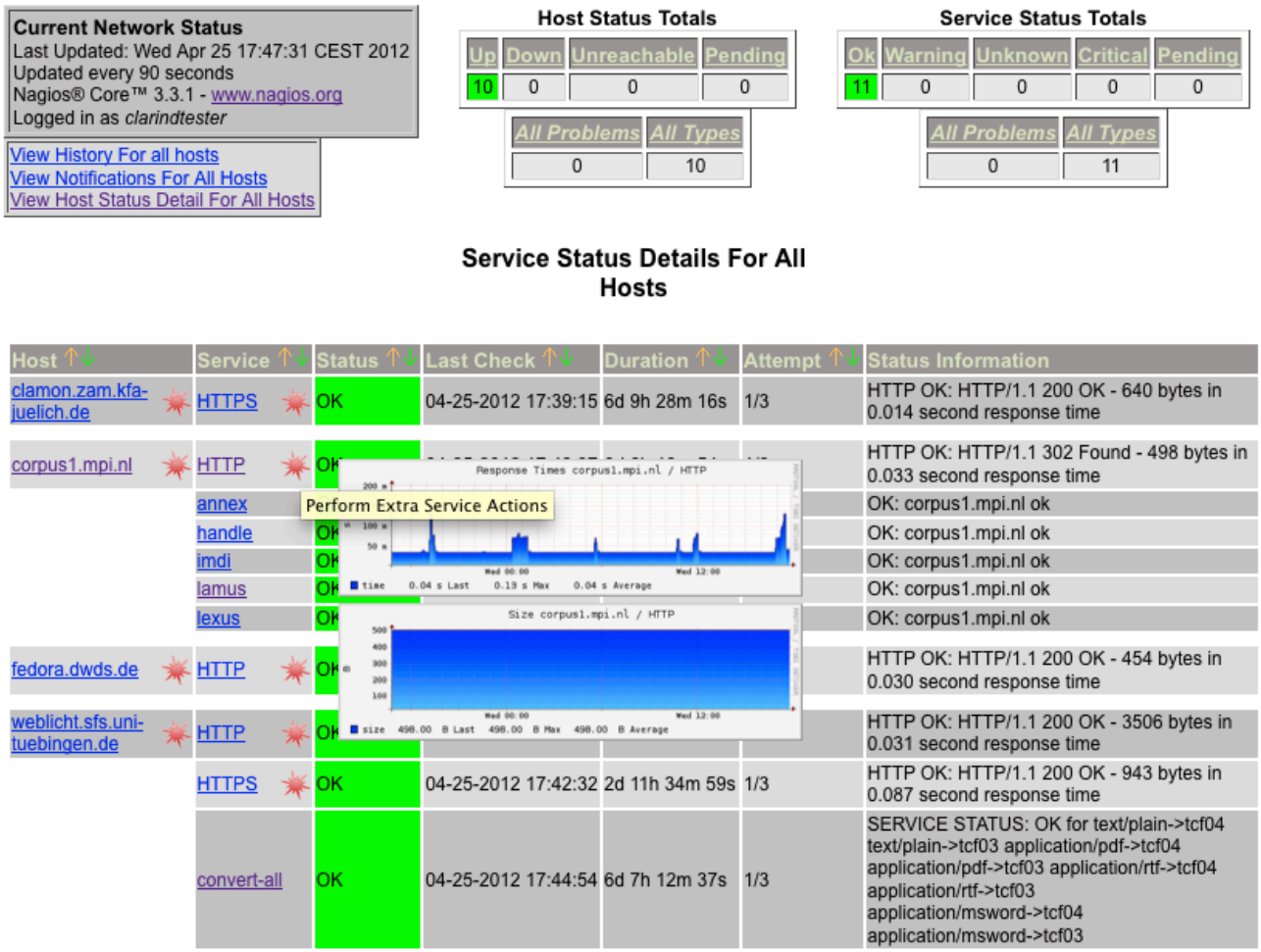


Figure: Nagios monitoring system

12 AG 10: Federated Content Search

The CLARIN-D Federated Content Search (FCS) started up this past year. We have had several rounds of (early) implementations, resulting in a clear vision of how the parts of the FCS infrastructure will interact in the future.

We defined which tasks are performed by the endpoints and which tasks by the aggregator. We also defined the interaction with the rest of the FCS infrastructure, i.e., how to search in a set of documents as defined by the VLO, the Virtual Collection Registry, or the metadata search.

Based on the vision of the overall FCS infrastructure, we defined the precise behavior of the endpoint , in the second part of the FCS document. Furthermore, demands on the whole CLARIN-D infrastructure, as related to the FCS, where defined.

To give an example of a decision with an impact on the infrastructure, we defined that the FCS endpoints must understand the MDSelfLink URI's from the CMDI metadata files that correspond to the data that they unlock through search. By "understanding" we mean that the endpoint must be able to restrict the search space to the data described by any set of its own CMDI files. Other aspects of the endpoints (e.g., announcement of resources, browsing through the results set, restricting the search, announcing available search methods) were also precisely defined in the document. Based on the

definitions of the document, most participating centers have put up an initial endpoint, which allows the process to move forward smoothly.

The definition of the acceptable return formats for the endpoints resulted in a healthy discussion. Over several meeting a definition was agreed upon that is agreeable to the stakeholders involved in the FCS process.

Finally, we mention the aggregator. Similarly to the endpoints we have defined what the aggregator should do and how it should communicate with the outside world as well as the known endpoint. A first beta version of the FCS aggregator is currently available via:

http://weblicht.sfs.uni-tuebingen.de/Aggregator/

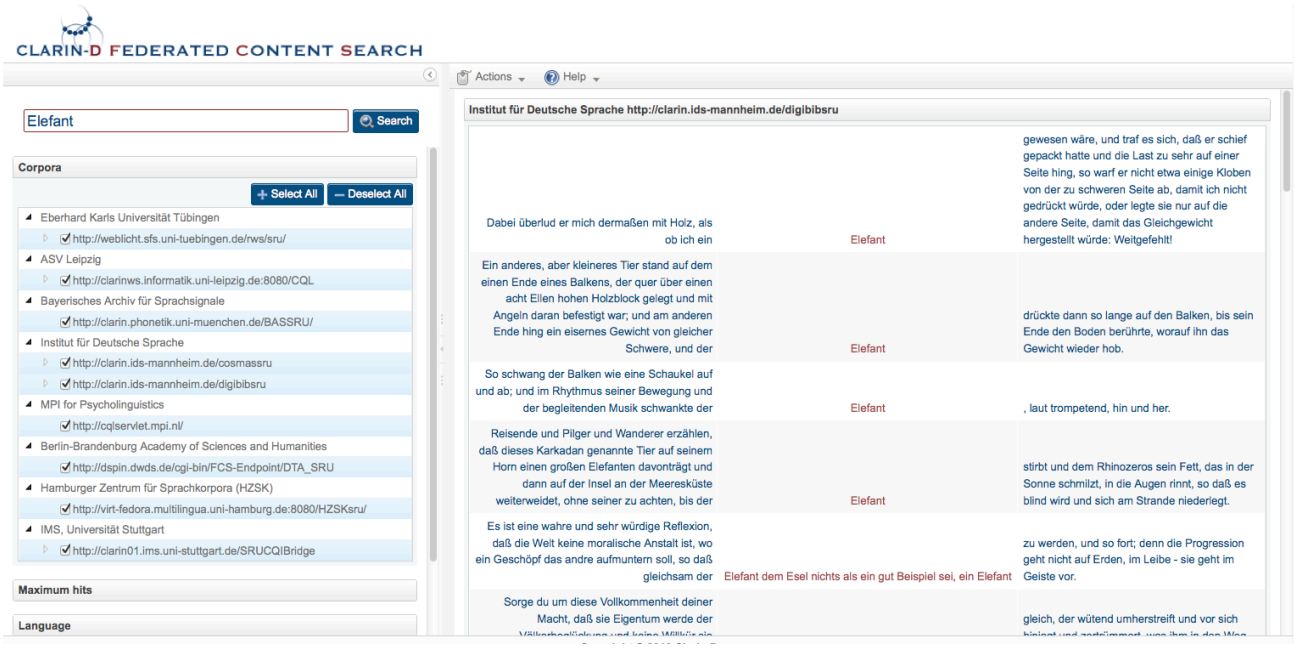


Figure: The CLARIN-D Federated Content Search Aggregator

8 CLARIN-D centers are currently offering SRU/CQL endpoints, which can be considered a success given the relatively short timeframe to set this up. The large challenge on longer term will for sure be the more sophisticated search operations and semantic harmonization (where it turns out to be feasible).

More information about this subject, including technical documentation can be found at:

http://www.clarin.eu/fcs

13 AG 11: Replication

13.1 Logical replication

Within the context of CLARIN-D, the MPI-TLA[2] has been looking for solutions to bring (1) persistency of data and (2) persistency of services. For this purpose we have been working on the REPLIX project to create a logical replication layer, based on the Integrated Rule-Oriented Data System (iRODS[3]), and a logical replication policy implemented on top of this logical replication layer.

The logical replication policy is responsible for replicating the data on a collection level, achieving (1), and for a number of additional steps, needed to achieve (2). The full logical replication policy consists of the following steps, implemented in the last year:

- Generation of the file list to replicate based on the supplied node-id (i.e. a subcollection identification code).
- Transfer the files with an rsync-based micro-service.
- Link the new data into the destination archive.
- Refresh auxiliary databases in the destination archive
- Synchronize the authorization information (AMS[4] rules).

The last step, currently being implemented:

- Administration of the PID records with the new digital object (DO) locations

An iRODS federation has been set-up between MPI-TLA in Nijmegen, the Netherlands, and RZG[5] in Garching, Germany. This iRODS federation has been used to test logical replication and is intended to run the archive backup procedure. The logical replication has also been demonstrated during two meetings, ICRI 2012 in Copenhagen and RDAP 2012 USA. For this demonstration, an iRODS federation was created between MPI-TLA and RENCI[6], demonstrating the logical synchronization across continents. The experiences and ideas behind the REPLIX project are also used as input for the EUDAT[7] project.

13.2 SAM-FS integration

The data in the MPI-TLA archive is stored in a hierarchical storage management (HSM) system, SAM-FS. Therefore, part of the data is available from online, fast cache and part of the data is migrated to offline tapes. Based on the SAM-FS rpc API we have developed a first set of iRODS micro-services that takes the online/offline state of files into account based on the following three rules:

1. Online files are replicated straight away. If the file was initially offline the file is released after the replication.

2. Offline files are staged and the replication command is queued again.

[2] https://tla.mpi.nl

[3] https://www.irods.org/

[4] Access rights Management System, https://tla.mpi.nl/tools/tla-tools/ams

[5] http://www.rzg.mpg.de/

[6] http://www.renci.org/

[7] http://www.eudat.eu

3. Files being staged are skipped and the replication command is queued again.

The iRODS delayed rule system is being tested as a queuing mechanism. The delayed rule mechanism also allows the execution of a number of rules in parallel. This is expected to increase the performance significantly. Again this has been tested in the iRODS federation setup between MPI-TLA and RZG.

More details about this topic can be found at: http://www.clarin.eu/page/3497

14 Summary

The first two years of CLARIN-D have proven to be very productive. In terms of the general infrastructure, there was the advantage that some components were already available from the CLARIN preparatory phase, so that these could be extended immediately without the need to build up the infrastructure foundations from scratch. On the other hand, CLARIN-D has initiated quite some new additions to the infrastructural landscape as well: e.g. the center registry, the monitoring setup, the use of part identifiers, a wide set of new (metadata-described) webservices, an aggregator for federated content search. These are all important building blocks that will certainly serve as inspiration for the entire European CLARIN network.

This successful start does not mean there are no challenges left for the next year. A large amount of additional software needs to be developed, which will most probably prove to be the greatest challenge. The amount of effort needed to polish user interfaces should also not be underestimated.

In short, though, the experiences, hard work and cooperative spirit of the CLARIN-D partners throughout the past years give us hope for continued success, cooperation, and innovation in the upcoming period.