<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>2</td>
<td>AG 1: Repositories</td>
</tr>
<tr>
<td>2.1</td>
<td>Setup of Repositories</td>
</tr>
<tr>
<td>2.2</td>
<td>Center Assessment</td>
</tr>
<tr>
<td>3</td>
<td>AG 2: Persistent Identifiers (PID)</td>
</tr>
<tr>
<td>4</td>
<td>AG 3: Registries</td>
</tr>
<tr>
<td>4.1</td>
<td>Center Registry</td>
</tr>
<tr>
<td>4.2</td>
<td>Data Category Registry</td>
</tr>
<tr>
<td>4.3</td>
<td>Relation Registry</td>
</tr>
<tr>
<td>4.4</td>
<td>Schema Registry</td>
</tr>
<tr>
<td>5</td>
<td>AG 4: CMDI</td>
</tr>
<tr>
<td>5.1</td>
<td>CMDI core schema</td>
</tr>
<tr>
<td>5.2</td>
<td>Arbil</td>
</tr>
<tr>
<td>5.3</td>
<td>Component Registry</td>
</tr>
<tr>
<td>5.4</td>
<td>OAI Harvester</td>
</tr>
<tr>
<td>5.5</td>
<td>Virtual Language Observatory</td>
</tr>
<tr>
<td>6</td>
<td>AG 5: Authentication &amp; Authorization Infrastructure</td>
</tr>
<tr>
<td>6.1</td>
<td>CLARIN IdP</td>
</tr>
<tr>
<td>6.2</td>
<td>CLARIN Discovery Service</td>
</tr>
<tr>
<td>6.3</td>
<td>Web service authentication</td>
</tr>
<tr>
<td>7</td>
<td>AG 6: Workspaces and Hosting</td>
</tr>
<tr>
<td>7.1</td>
<td>Workspaces</td>
</tr>
<tr>
<td>7.2</td>
<td>Hosting</td>
</tr>
<tr>
<td>8</td>
<td>AG 7: Webservices, General services</td>
</tr>
<tr>
<td>9</td>
<td>AG 8: Simple store</td>
</tr>
<tr>
<td>10</td>
<td>AG 9: Monitoring</td>
</tr>
<tr>
<td>11</td>
<td>AG 10: Federated Content Search</td>
</tr>
<tr>
<td>12</td>
<td>AG 11: Replication</td>
</tr>
<tr>
<td>12.1</td>
<td>Technical Details of the Prototype Software</td>
</tr>
<tr>
<td>13</td>
<td>Summary</td>
</tr>
</tbody>
</table>
1 Introduction

As most of the work in the technical infrastructure work package has been done within specialized working groups, we list all of these below and describe in detail which activities have been undertaken. 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.

2 AG 1: Repositories

2.1 Setup of Repositories

As one of the important pillars of the distributed CLARIN-infrastructure, each of the centers sets 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 third project year, the repository infrastructure in all centers was fully established and integrated into the CLARIN-D infrastructure, as the outcome of the center assessment process proves (see next point).

2.2 Center Assessment

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

- The CLARIN center assessment, which analyzes the center’s offer in terms of compliancy with the B center 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 center’s data management strategy and its policy are suitable for long-term archiving.

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

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.

Integration of handles in the center repositories

All CLARIN-D centers 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.
PID-related events

CLARIN-D was represented at the EPIC user forum 2013:


On 20 February 2013, a workshop about PIDs in the context of CLARIN centers was organized:

http://www.clarin.eu/node/3882

4 AG 3: Registries

4.1 Center 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. 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 were outlined earlier on 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://centres.clarin.eu). 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). The center registry is currently used by the Federated Content Search aggregator, WebLicht and the OAI-PMH harvester as authoritative information source.

<table>
<thead>
<tr>
<th>Center Name</th>
<th>Fedora Commons</th>
<th>Certification Status</th>
<th>Access</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS, Universität Stuttgart</td>
<td>Fedora Commons</td>
<td>certified B</td>
<td>Test SP available</td>
<td>handle via EPIC</td>
</tr>
<tr>
<td>Hamburger Zentrum für Sprachkorpora (HZSB)</td>
<td>Fedora Commons</td>
<td>certified B</td>
<td>Test SP available</td>
<td>handle via EPIC</td>
</tr>
<tr>
<td>Berlin-Brandenburg Academy of Sciences and Humanities</td>
<td>Fedora Commons</td>
<td>certified B</td>
<td>Service Provider available</td>
<td>handle via EPIC</td>
</tr>
<tr>
<td>Universität des Saarlandes</td>
<td>Fedora Commons</td>
<td>certified B</td>
<td>Test SP available</td>
<td>handle via EPIC</td>
</tr>
<tr>
<td>MPI for Psycholinguistics</td>
<td>LAT (own system)</td>
<td>certified B</td>
<td>Service Provider available</td>
<td>handle (own server and prefix 1839)</td>
</tr>
<tr>
<td>Institut für Deutsche Sprache</td>
<td>Fedora Commons</td>
<td>certified B</td>
<td>IDP and SPs available</td>
<td>handle (own server and prefix 1092)</td>
</tr>
<tr>
<td>Bayerisches Archiv für Sprachsignale</td>
<td>own system</td>
<td>certified B</td>
<td>Service Provider available</td>
<td>handle via EPIC</td>
</tr>
<tr>
<td>ASV Leipzig</td>
<td>Fedora Commons</td>
<td>certified B</td>
<td>One service provider available</td>
<td>handle via EPIC</td>
</tr>
<tr>
<td>Eberhard Karls Universität Tübingen</td>
<td>Fedora Commons</td>
<td>certified B</td>
<td>One service provider available</td>
<td>handle via EPIC</td>
</tr>
</tbody>
</table>

Figure: the HTML interface of the center registry
4.2 Data Category Registry

The ISOcat Data Category Registry (http://www.isocat.org) has run stably over the last year. Updates have been mostly related to small bug fixes and minor usability improvements. There has been an internal evaluation of the problems encountered and the uptake of ISOcat by both ISO TC 37 and CLARIN. This resulted in a meeting of these two major user communities at the Semantic Registries workshop, which was organized at the end of 2013 (see below). The aim was to see in how far the requirements of both communities still overlap. The result is that the successor of ISOcat will have a simpler data model focused on concept specifications, and also a workflow that is more geared towards community agreement than official ISO standardization.

Publications


Posters


Workshops


4.3 Relation Registry

Some additional relation sets have become available in the alpha version of the RELcat Relation Registry (http://www.isocat.org/relcat-test/). Investigation and first trials have been done to switch from the Open Anzo triple store to the open source version of OpenLink Virtuoso Universal Server. However, further development has stalled in favor of SCHEMAcat (see below), with the aim to bootstrap RELcat from the SCHEMACat code base.
Publications


Presentations


4.4 Schema Registry

The SCHEMAcat Schema Registry (http://www.isocat.org/schemacat-test/) has made much progress and is implementation wise nearing its first beta release. Its implementation is based on BaseX and NetKernel 5 as respectively the database and the application server at the backend, and a HTML 5/JavaScript frontend using AngularJS and Bootstrap CSS. Features of the current version include:

- Access via Shibboleth, i.e., recognizes a new user, asks for additional information and creates a local user account
- Viewing and managing metadata, e.g., description, status, license, for a schema
- Downloading and viewing a schema, e.g., rendering using syntax colorization
- Sharing a schema, i.e., provides an URL to be used in the development phase
- Optimized access to schemas, i.e., extensive internal caching and proper HTTP caching directives to foster fast response times
- Full user interface addressability through URL’s i.e. bookmarks can be created for any point of the application e.g. show a certain schema, schema license, ...

Features currently lacking for a first release:

- Starting with a complete new schema, i.e., upload a new schema
- Publishing a schema, i.e., assigning it a handle and freezing it
- Searching through the available schemas

As stated above the developed codebase and expertise will be used to bootstrap the development of RELcat.

Publications


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/cmddi

5.1 CMDI core schema

At the backend of CMDI the following changes took place:

• The general component schema allows only one root on component and profile specifications
• Schematron rules have been added to prevent illegal combinations of CardinalityMin and CardinalityMax

The CMDI taskforce has formed a number of working groups, working towards a specification for a new version (1.2) of CMDI (see https://www.clarin.eu/node/3883). The specification is expected to be completed in the next few months, followed by a period of implementation. The new version of CMDI will bring major improvements and new features with respect to lifecycle management, vocabularies and schema sanity as well as a number of small improvements and fixes:

• 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

Initially Arbil was 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. We are now focusing on usability needs of a broader range of users by working towards a greatly simplified user interface.

In 2013, a new minor version (2.4) of Arbil was released and another (2.5) was developed and will be released as a stable version soon. These new versions provide a number of additions and improvements, particularly in the domains of usability and CMDI support. In parallel, development efforts have been put into:

1) the development of a plugin-architecture allowing for the development of additional functionality by third parties and enhancing the sharing of capacities and features between Arbil and other applications;

2) the development of a database-driven backend which will allow for fast and powerful search functionality, both locally and remote as well as a web based search portal that makes use of these technologies;
3) designing and preparing for implementation of a branch of Arbil that will allow the user to work on existing file structures, creating and editing metadata on the fly.

More details can be found in the release notes available at

http://tla.mpi.nl/tools/tla-tools/arbil/release-notes

5.3 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.

Development of user group features started in 1.14 but were not delivered conclusively. 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 serious bug was uncovered in the late testing phase where access checks are systematically ignoring the new access paths implemented by groups, so we had to delay 1.14. We started rewriting the back end (services and persistence) for 1.15 in order to reduce complexity and simplify maintenance. Midway the test results for 1.14 came in, at that point it was decided that the 1.14 release would use parts of the 1.15 code base. Also planned, but not started yet for 1.15 is a rewrite of the flash front end in order to reduce the substantial code complexity and reduce code duplication.

Development activities were temporarily paused in December 2013 due to an internal reorganization of the development team and personnel changes at The Language Archive. Continuation of the development of the component registry is one of the priorities for the next project year.

5.4 OAI Harvester

A new, largely rewritten version of the OAI-PMH harvester has been completely decoupled from the MPI’s archive structure. Support for configurable workflows has been added with the intention of making the software component generally usable for anyone interested in metadata interoperability, both within and outside CLARIN. However, there are also some improvements more specific to CLARIN: integration with the center registry and support for centers that use multiple metadata prefixes for CMDI metadata.

More details and the source code can be found at:

https://github.com/TheLanguageArchive/oai-harvest-manager

5.5 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 facet browser. Significant improvements that have been added in this year of CLARIN-D are:
• New whitelist/blacklist mechanisms were implemented that improved recall & precision of extracted facet values.
• A VLO/metadata task force was established to promote user feedback.
• User feedback from 8 CLARIN-D centers was evaluated and the VLO facet configuration was adapted in several feedback rounds.
• Feedback about CMDI profiles with incorrect conceptlinks 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 etc.)
• Support of themes was added (particularly support of a CLARIN-D theme).
• A stricter separation of code and configuration improved the configurability of the VLO.
• Optimisation of configuration and deployment process at production server reduced downtimes.
• extensive bugfixing (heap space problems, error handling etc.)
• first alpha version of an improved web interface

6 AG 5: Authentication & Authorization Infrastructure

In the CLARIN preparatory phase, the so-called Service Provider Federation ([http://www.clarin.eu/spf](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.

6.1 CLARIN IdP

To provide users without a functioning IdP (or one that does not release necessary attributes, like eduPersonPrincipalName) 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](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. Next step is to investigate a redundant setup.

6.2 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](http://www.clarin.eu/page/3496)

Here too, a redundant setup to avoid the risks related to a single point of failure, would make sense and will be considered.
6.3 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 ISOcat Data Category Registry was proven to be successful. Last year the NDG OAuth2 Authentication Server was deployed to the catalog.clarin.eu Service Provider.

In the context of F-AG 3 a new use case involving web service access to TLA-based resources from a tool developed by the University of Cologne was to be deployed. Although the use case is not fully into production yet, 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.

At the moment of writing the OAuth2 Authentication Server is still suffering from some configuration problems; as soon as these have been overcome the installation will be documented and other interested users in the CLARIN-D community will be informed about the details.

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 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.

The focus in the 3rd project year went to:

• connecting the OwnCloud web application to Shibboleth (successfully tested)
• connecting the OwnCloud webdav-access to the CLARIN IdP via LDAP (successfully tested)
• development of a plugin that allows CLARIN-D services to write to the workspaces (successfully tested but the use of self-developed plugins still needs to be further discussed with the hosting computing center)

Figure: The OwnCloud web interface

7.2 Hosting

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:

<table>
<thead>
<tr>
<th>Computing Center</th>
<th>Hosting Tasks</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>RZG</td>
<td>Web services</td>
<td>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</td>
</tr>
<tr>
<td></td>
<td>Center registry</td>
<td>Implemented by the RZG – already in use for the WebLicht harvesting</td>
</tr>
<tr>
<td>FZJ</td>
<td>Workspaces</td>
<td>See above: the software is in place and can be used</td>
</tr>
<tr>
<td></td>
<td>Helpdesk</td>
<td>OTRS installation is online, in close cooperation with the Hamburg CLARIN-D center</td>
</tr>
<tr>
<td></td>
<td>Monitoring</td>
<td>Installed and configured by FZJ</td>
</tr>
<tr>
<td>GWDG</td>
<td>PID services via EPIC consortium</td>
<td>EPIC API version 1 is in production but will be phased out in 2014 in favor of version 2 (in production too)</td>
</tr>
</tbody>
</table>
8  AG 7: Webservices, General services

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

New webservices include:

- Orthographic Canonicalizer (CAB) for orthographic normalization of historical texts
- Lexical Database service (dlexDB) to obtain statistical information for lexical types
- A morphological analyzer and parsers for Dutch, English, and German
- Externally Trained Named EntityRecognizer 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 webservices. 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 webservice developers, so that better performance of more webservices can be achieved.

In addition to reliability and scalability testing, existing webservices 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 from the command line and will be rolled out in the near future.

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 webservice together.

9  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 (now called B2SHARE - see https://b2share.eudat.eu), 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. Among other things, this is ensured by the fact that UTU, MPI and the CLARIN computing centers are also participating in EUDAT.

10 AG 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 centers and their services (http://www.clarin-d.de/en/news/status-infrastructure.html).
- Automatic checks based on the information contained in the center 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

11 AG 10: Federated Content Search

All components of the CLARIN-D 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 is being evaluated and updated. The supporting environment has continued to develop, including:

- SRUClient and SRUServer Java libraries
- SRU/CQL server endpoint reference implementation
- SRUQIBridge - implementation of SRU end point front end for a CQI Server
- a CLARIN-FCS endpoint conformance tester web application (http://clarin.ids-mannheim.de/srutest/app/)

As endpoints began implementing the interface, some flaws in the specification became clear. To address those, the FCS task force has prepared the new version of the FCS interface specification.

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

**Interface Specification:**

The initial FCS specification based on SRU/CQL-protocol was revised to make it better structured and more intuitive, to provide support for more advanced linguistic queries and to provision for future extensions in terms of query and resources support. To give an example, a revised specification provides full support for boolean queries and for representing search results of such queries in a basic records type. The draft of the proposed specification is scheduled to be submitted to the CLARIN Center Committee for approval.

**Aggregator:**

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/](http://weblicht.sfs.uni-tuebingen.de/Aggregator/)

**Endpoints/Resources:**

Endpoints have improved their level of compliance with the current FCS specifications. A number of new resources and endpoints have joined FCS, among them: 7 speech corpora from Bayerisches Archiv für Sprachsignale, Cookbook Corpus and GRUG Treebank from Saarbrücken, Corpora of Berlin-Brandenburg Academy of Sciences and Humanities (DTA Corpus, Dingler Online, C4 Corpus), HZSK Hamburg Maptask Corpus, CoNLL 2009 Shared Task - Czech Data corpus and Hindi Web Texts from LINDAT-Clar in. Work for compliance with the new FCS specification is scheduled to start in the mid-2014.

**12 AG 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](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 centers, 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.

12.1 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 centers 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.
13 Summary

In CLARIN-D’s third year, the continuous efforts have been clearly rewarded by the certification of all 9 German CLARIN centers. The foundations for a reliable research infrastructure are thus in place. That does not mean work has been completed — the real challenge now lies in a better integration of the existing components and polishing of the user interfaces. Another important aspect for the upcoming years, which will need some 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 cooperative spirit that has been demonstrated over the last years seems to indicate CLARIN-D will find it’s way in a constantly changing environment.