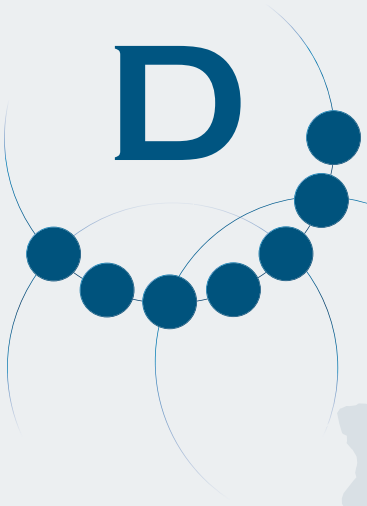

D-SPIN

D-SPIN Report R 2.3

Integration of German Resources into the Registry, Selection of Typical Workflows, and Implementation of Typical Basic Services

March 2011

D-SPIN, BMBF-FKZ: 01UG0801

Deliverable R-2.3: Integration of German Resources into the Registry, Selection of Typical Workflows, and Implementation of Typical Basic Services

© All rights reserved by MPI on behalf of D-SPIN

Editors: Peter Wittenburg, Thomas Zastrow, Volker Böhlke, Lothar Lemnitzer

Contents

1 Background

CLARIN/D-SPIN is finishing its preparatory phase and preparing its construction phase. In the construction phase ready-made services for users need to become available in user-friendly ways. In the preparatory phase we focused on infrastructure components and first service prototypes that can show the potential of a research infrastructure. Yet we did not focus on optimizations and user friendliness, although some new tools such as WebLicht and ARBIL exhibit already much professionalism. In this report we will summarize the achievements in the direction of the state of integration and user oriented services and we will refer to infrastructure components that enable services. In the conclusions we will line out what needs to be done in the construction phase.

This report is based on a couple of reports that have already been created in CLARIN and/or D-SPIN. In particular the following reports need to be mentioned:

Number	Date	Title	Description
CLARIN D2R-5a/b	January 2010	Registry Infrastructure	A description of the Metadata Infrastructure including a description of CMDI
CLARIN D5R-3	December 2009	Linguistic Processing Chains as Web Services: Initial linguistic considerations	Overview about existing Web Services and Workflow environments being used in the linguistic domain
CLARIN D5R-2	June 2009	Usage and Workflow Scenarios	Elaborate descriptions of a large set of use cases and typical research workflow descriptions
D-SPIN R2.2b	February 2010	The German Resource Landscape and a Portal	Brief overview about the setup of the Virtual Language Observatory portal and the representation of German resources being harvested
D-SPIN R3.3	March 2010	Case Studies - Intermediate Report	Describing the needs of humanities researchers with respect to linguistic tools

These reports describe in detail

- The metadata and registry infrastructure
- Typical linguistic processing chains
- Typical use cases of humanities researchers from which the need for certain services can be extracted
- The data resource landscape in Germany.

2 Integration of German Resources into the Registry

2.1 Registry

D-SPIN as well as CLARIN decided to

- Describe all resources and tools/services according to the CMDI scheme (Component Metadata Infrastructure);
- Allow everyone to harvest the metadata records via OAI-PMH;
- Allow people creating derived portals and sub-registries that are optimized towards special purposes.

This decision was discussed again and confirmed at the D-SPIN Workshop with the topic "Metadata and Registries for Web Services" which took place in Nijmegen on January20, 2011. The following diagram indicates the principles schematically.

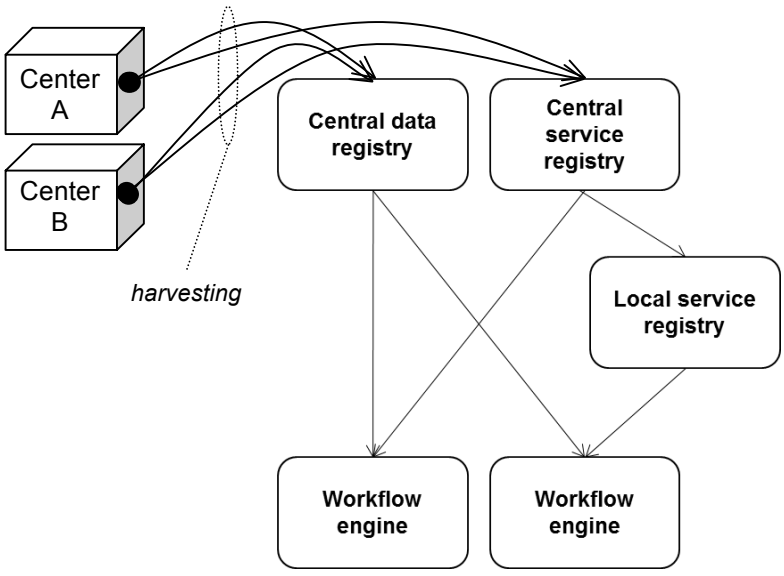

Fig 1:Individual centers that offer CMDI metadata descriptions will be harvested via OAI-PMH so that there is one „central" portal including all registered data and tools. Local specialized services may handle a selection of metadata descriptions with for example reduced information optimized for specific purposes. This architecture will guarantee that there will be no further fragmented islands. Workflow engines can operate with one of these registries for example.

Due to the enormous time pressure which we imposed on us to come up with visible results in the preparation period a temporary landscape emerged which does not represent the state as described in the diagram. The following registries were set up or are in use:

- IMDI and OLAC based data providers are harvested by MPI for some years to be integrated into the global IMDI registry.
- TEI based and other metadata descriptions have been transformed either to OLAC, IMDI or CMDI and are harvested by MPI and are either integrated into the IMDI registry or directly into the CMDI registry.
- A LRT Inventory was created supporting a restricted element set to motivate data/service providers to quickly enter preliminary metadata descriptions and thus make them visible.
- A Web Service registry was setup for the WebLicht related work to achieve fast progress in allowing users to create workflow chains.
- All registries except for the WebLicht Registry are harvested by VLO (Virtual Language Observatory) including a semantic mapping to the agreed ISOcat defined metadata elements

and vocabularies. Thus for data resources VLO has a wide coverage, for services it is possible that there are gaps and overlaps between VLO and the WebLicht registries.

It is agreed that in the construction phase this fragmented situation needs to be overcome in the way indicated above with highest priority. The basis for setting up all registries and creating metadata descriptions is

- The usage of the categories which have been defined in the metadata profile in ISOcat which is currently in a formal revision process
- The CMDI component model and the component specifications as specified in D2R-5a Registry Infrastructure - v1 which also is now suggested as a new item for standardization in ISO TC37/SC4
- The usage of OAI-PMH for harvesting
- The usage of persistent identifiers as reference mechanism

Currently, all semantic mapping is done via implicit methods. This will also be changed in the construction phase when the relation registry has been implemented. This allows every researcher to define their own preferred mapping schemes and to share it with others.

2.2 Integration of German Data Resources

Based on the described mechanisms above we can report about some figures extracted from the VLO. We should note here that there is no facet in the VLO search portal[1] that allows calculating all resources from German centres directly. There is a facet "language" which is of relevance for researchers and where one can select "German", but this gives information about all resources in the German language stored somewhere on a centre. In addition it does not contain all non-German resources stored in D-SPIN centres. If selecting "country = German" this informs about the resources being recorded/created in Germany which also does not give the intended answer directly (see appendix A). To derive all resources from German centres we need to select "centres" and filter out those from the German centres, which currently requires manual work due to the missing centre vocabulary[2].

2.2.1 Data Resources

Totals

- 24388 language resources have been registered and harvested from German centres.
- 8291 resources include the German language but these have been harvested from centres also in non-German countries.
- The granularity of the resources being described is heterogeneous, i.e. some have a metadata record for one large corpus, and others describe every single resource[3].

Most of the metadata providers are not from German centres. The CLARIN LRT inventory includes many resources from other European providers, the IMDI registry and the OLAC registries include data from providers worldwide, ECHO and DBD data are European resp. Dutch contributions, etc.

[1] The reason is that metadata sets used until now did not support an element such as "organization address" and that the
[2] There are still different spellings for individual institutions and typing errors making automatic processing impossible.
[3] Yet there is no guideline which needs to be changed in the construction phase. However, we realized that for some centers a high granularity will require major changes in their repository setup, i.e. we expect a time consuming adaptation process.

Organization
The 24388 data resources originate from about 300 German organizations amongst which are mainly research institutions and publishers (for details see appendix A). We are missing a curation step on organization names to compensate for typing errors and name variants. This curation in principle needs to be done by the data creators, but CLARIN will setup an ontology to make fast steps ahead.

We assume that there are still many resources out there that are neither visible nor accessible. This puts an emphasis on the need of an archiving campaign in the construction phase.

2.3 Integration of German Tools

As indicated above we need to distinguish at this moment the VLO registry and the special one created for WebLicht.

2.3.1 Tools/Services in VLO

According to the above-described mechanisms we can report that **German centres registered 65 tools and services** in VLO (for details see appendix B). Amongst these are German institutions, publishers and companies that are not centres in the CLARIN sense, but active technology providers that want to make their tools visible through CLARIN channels. We believe that the coverage is already fairly good with one exception: many search and access tools have been created specifically for one specific corpus. These tools cannot be maintained over time and are so specific that it would not make sense to register them.

Again it is not straightforward to determine the nature of the tools and services provided by German institutes. However, we can assume that the distribution will not be so different than for the whole set of tools which is listed below (in percentage). The majority is for written language and classical NLP tasks, but there are also quite a number of annotation tools and tools for speech/multimedia/multimodality tasks.

- Written Language (40)
- NLP Development Aid (16)
- Annotation Tools (14)
- Other (11)
- Spoken Language (8)
- Multimedia (3)
- Multimodality (3)
- Evaluation Tools (2)
- Single task tool (2)
- Multiple task tool (1)

Yet we do not have an overview, which of these tools are desktop applications, web applications and web services. In the construction phase we will look at many of these tools to check whether curation and CLARIN integration will make sense.

2.3.2 Services in WebLicht Registry

The D-SPIN WebLicht registry stores and provides metadata on web services compatible to the temporary WebLicht-Infrastructure approach. In addition to basic metadata, like information on the creator/maintainer of a service etc., orchestration metadata[4] is stored. This orchestration metadata

[4] The set is included in the Athens Core metadata set currently registered in ISOcat.

consists of strongly typed interface descriptions, and technical information (url etc.) on how to invoke a service. Orchestration metadata is used by the D-SPIN chaining algorithm[5] to support the semiautomatic specification of workflows and is used by the workflow invoker that is part of the back-end of the WebLicht web-interface[6]. The registry itself can be accessed via REST web service interfaces. Human users are supported in the usage of the registry by the D-SPIN registry management tool, which allows easy specification and modification of the metadata stored in the registry.

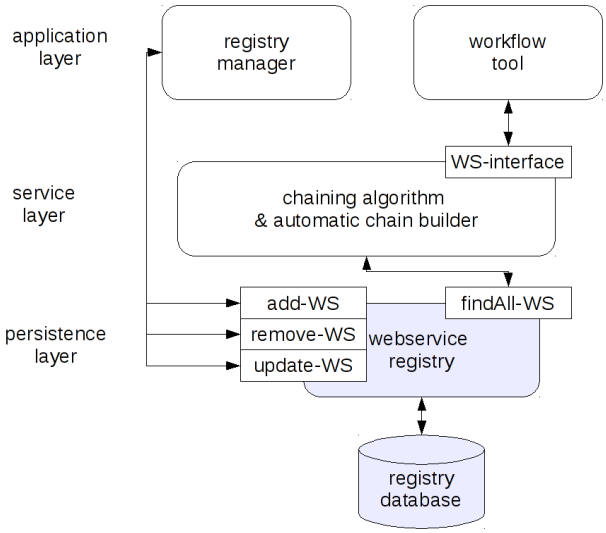

Fig 2: WebLicht infrastructure - overview

Currently (March 2011) there are 134 services registered to the D-SPIN WebLicht registry. These services are maintained by 8 D-SPIN/CLARIN partners:

- ASV - University of Leipzig (30)
- BBAW Berlin (10)
- IDS Mannheim (1)
- IMS, ILR - University of Stuttgart (32)
- SfS - University of Tübingen (40)
- GL - University of Helsinki (3)
- Institute of Computer Science - Polish Academy of Sciences (3)
- RACAI - Romanian Academy (15)

The services provide access to a variety of textual resources and tools in the following 11 languages: German, English, Italian, French, Spanish, Finish, Romanian, Polish, Slovenian, Hungarian and Czech.

There are 25 services providing access to text or text related (statistical information; co-occurrences, frequencies e.g.) resources like the 2 DeReKo Goethe Corpus[7], the C4 Corpus[8] and multiple corpora

5 Volker Boehlke: A Generic Chaining Algorithm for NLP Webservices. Vortrag und Publikation: Web Services and Processing Pipelines in HLT Workshop, LREC 2010; Valletta, Malta; Mai 2010

6 https://weblicht.sfs.uni-tuebingen.de/

7 http://www.ids-mannheim.de/cosmas2/projekt/referenz/korpora1.html?sigle=GOE

of the wortschatz project[9]. The matching counterparts are 109 services wrapping linguistic tools. The tool wrappers may be further categorized in the following way:

- Converters (30) from/into the following formats: Word, RTF, PlainText, Negra, TEI[10], MAF[11], D-SPIN TextCorpus (TCF[12]) and D-SPIN Lexicon
- Tokenizers (27)
- POS-taggers (20)
- Parsers (7)
- Named entity recognizers (4)
- Chunkers (3)

Additionally there are several tool-wrapper services such as a service maintained by the partner in Tübingen that creates a KML[13] file for Google Earth[14] from all geographic locations found in a given text.

3 Selection of typical Workflows and their Implementation as Web Services

Many workflow scenarios have been listed and implemented in the work reported in D5R-2 to which German researchers contributed actively. In the area of textual analysis typical canonical workflows are the most frequent reoccurring patterns. For audio/video analysis for example we cannot refer to so clearly defined patterns.

3.1 Workflows and Services for Text-Resources

For an in-depth analysis of texts, it is useful to add linguistic information to the original text(s), which can be done in the form of linguistic annotations. This means that in several steps

- The original text is enriched with linguistic information (inline annotations) or
- The linguistic information is separated from the original text and stored in individual layers (stand-off annotations).

In both strategies, the linguistic information is extracted from the original text or is derived from already existing annotations. This process, where every service builds up upon its predecessors, can be organized as a workflow. With the help of the technical metadata stored in a registry (see chapter 2.3.2), a workflow engine can create such workflows (semi-) automatically.

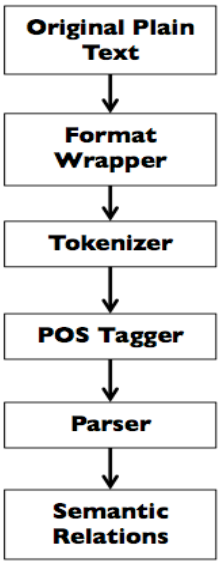

Fig 3: A typical workflow for annotating text resources

A typical workflow for text resources looks like denoted in Figure 3. Starting with the plain text itself, it is first wrapped into a data format that is

8 http://chtk.unibas.ch/korpus-c4
9 http://wortschatz.uni-leipzig.de/ and http://corpora.informatik.uni-leipzig.de/
10 http://www.tei-c.org/index.xml
11 ISO/DIS 24611. 2008. Language resource management - Morpho-syntactic annotation framework (MAF).
12 Heid, U.; Schmid, H.; Eckart, K.; Hinrich E. (2010). *A corpus representation format for linguistic web services: the D-SPIN Text Corpus Format and its relationship with ISO standards*. LREC 2010.
13 http://code.google.com/intl/de-DE/apis/kml/
14 http://www.google.com/earth/index.html

compatible with the following services. After that, linguistic annotations are added step by step (here: tokenization, part-of-speech tagging, parse trees and semantic relations). Every service of the workflow can be replaced by a different service of the same type. This makes it possible to apply for example several different tokenizers to one and the same text and then compare the results. For being compatible with each other, it is necessary that all services are using the same standardized data format for exchanging and transporting the data through the workflow from one service to the next one or to apply converters.

3.2 Workflows and Services for Audio/Video-Resources

Typical workflows in the area of audio/video processing for multimodal and sign language work have been investigated recently by analysing the use cases in DR5-2, by visiting a number of institutes and departments, by having asked researchers to contribute in workshops[15] and by talking to researchers at meetings. We cannot claim that a systematic investigation has been carried out yet which partly is due to the fact that the group of multimodal researchers is rather fragmented and that it was not possible yet to organize a joint meeting of them neither in the CLARIN nor in the D-SPIN realm[16]. The following researchers and research groups were included in such discussions during the last year: group of Asli Ozyurek, group of Onno Crasborn (U Nijmegen), Hedda Lausberg (U Cologne), Steve Levinson, Nick Enfield, Jeremy Hammond, Marc Dingemanse, Tyko Dirksmeyer, Gunter Senft (MPI), group of Irene Mittelberg (RWTH Aachen), the group of Stefan Kopp (U Bielefeld), the group of Costanza Navarretta (U Copenhagen), Marianne Gulberg (U Lund) and Lorenza Mondada (U Lyon2).

The following topics turned out to be most important for the kind of research these researchers are doing:

- Availability of an efficient annotation and manipulation framework from which a rich set of functionalities can be accessed
- Availability of a rich set of services that offer smart, robust and scalable audio/video recognizers to create automatic annotations
- High degree of usability to allow users to search for annotation patterns, to browse through them, to visualize and manipulate them
- Pure increase of efficiency is an important issue, but equally important is the capability of supporting theorization, i.e. researchers are not interested in a black box solution solving a certain problem

In collaboration with U Nijmegen and two Fraunhofer institutes (IAIS, HHI) a first set of services has been provided and can be accessed from the extended ELAN workbench[17]. The available **audio detectors** are: silence detector, vowel/pitch pattern detector, speech/non-speech detector, utterance segmentation, gender detector, speaker clustering and identification and forced language dependent alignment. In preparation are: a speech recognition service, language independent aligner and a query-by-example annotator. The available **video detectors** are: shot and sub-shot boundary detection, homogeneous segment detection, key frame extraction, camera motion detector, skin colour estimator, hand and head tracker. In preparation are: specific gesture detectors, stroke

[15] http://www.mpi.nl/research/research-projects/the-language-archive/events
[16] Such a workshop is planned for 2011.
[17] http://www.mpi.nl/avatech; http://www.mpi.nl/tools/elan

detector. In the coming construction phase more institutes will be approached to provide smart detectors[18].

All detectors are described by CMDI type of metadata and a protocol has been defined that allows invoking them remotely on a server. Yet this is only implemented for a server on a local area network, since we need to provide an efficient way of accessing the data streams. Audio/video streams cannot be transferred fast enough via the web during workflow execution, thus we need to solve three problems: (1) how to take care that the services are being deployed on servers which allow fast access to the data, (2) how to take care that intermediate data which can be new time series with high data volumes such as video motion vectors is generated in a workspace with high throughput and (3) how to transfer data to be visualized to the user which also can be lengthy time series. These problems can only be solved in collaboration with a computing centre which will be done in the construction phase, in particular, if users do not only work interactively on one data stream, but if they want to execute the chosen workflow on a whole series of streams.

With respect to typical workflows we are in a phase of testing, evaluation and improvement. We can however describe a couple of re-occurring workflows:

- Use of a detector -> inspection of results -> modification of parameters -> re-use of the detector -> manual correction of annotations
- Use of a segmentation service -> apply specific detectors to specific segments
- Use a basic detector such as hand detector -> apply a cascaded hand movement detector
- Use audio detectors -> use video detectors -> use a pattern detector that combines different types of annotations to new annotations

In general we can assume that workflows do not include so many steps as in classical NLP applications. The emphasis is on usability due to the probabilistic and erroneous nature of audio and video pattern recognition and the need of manual adjustment after almost every step.

4. Implementation of typical Basic Services

In the context of workflows, applications and web services D-SPIN/CLARIN has already created a number of basic services giving access to collected information. However we have to admit that some are in a prototypical state and need improvements in the construction phase. We differentiate between protocols, infrastructure, metadata services and content services some of which are generic and others being specific for text, audio and/or video resources.

4.1 Protocols

For basic services we currently use the following standard protocols in D-SPIN/CLARIN:

- OAI-PMH for metadata harvesting
- SRU/CQL for distributed searches
- REST for invoking remote web services (for example in WebLicht)
- SOAP/WSDL for invoking remote web services (for example for ISOcat)

OAI-PMH, REST and SOAP/WSDL are well known and don't have to be described in this document.

[18] Discussions with other well-known groups in Germany (U Bielefeld, RWTH Aachen) and in Europe (LIMSI, U Barcelona) have been started to extend the domain of services.

SRU/CQL (Search/Retrieval via URL/Contextual Query Language) is a protocol that was invented in the library world for distributed searching. The typical situation in CLARIN will be that there is a „centralized“ metadata portal that can accept all kinds of metadata queries, but that the content (the data resources) will remain at different centres. Thus the only way of carrying out joint content queries on a virtual collection is to spread the query to different local search engines and to retrieve the responses via a standardized protocol such as SRU/CQL. More details about SRU/CQl are explained in appendix C.

4.2 Infrastructure Services

With respect to generic infrastructure services CLARIN has been active in setting up a few for general use. Here we want to refer only to two examples (for more we refer to the documents about the D-SPIN/CLARIN infrastructure):

- A persistent identifier service provided by EPIC and based on Handles has been setup to register and resolve PIDs. Registration can either be done manually or automatically by using an API. Resolving is done via the standard Handle facilities. Registration can be done by any accepted centre that can indicate that it takes care of the persistence of its resources. This service is of great importance in the world of web services where annotations are created in workflow chains executed in workspaces on some compute centre for example and where it is important to maintain clear identities.
- An infrastructure for distributed authentication has been setup based on SAML assertions and using Shibboleth software that allows protecting the access to any type of web-resources. This facility is also important for accessing workspaces which act as a container for executing workflows.

4.3 Metadata

The basic services supporting a joint metadata domain have already been mentioned (for details see CLARIN D2R-5a/b) so that we only need to summarize them:

- ISOcat concept registry to store all metadata concepts
- CMDI component editor
- CMDI metadata description editor
- Harvesting with the help of OAI-PMH
- Semantic gateway with hardcoded mapping relations[19]
- Virtual Language Observatory with different access possibilities:
  - Standard catalogue function
  - Standard search function
  - Faceted browser supporting a few standard facets

All these services are operational.

4.4 Content Services

4.4.1 Basis-Services for Text-Resources

As described in chapter 3.1, a typical workflow for text resources consists of multiple services, which are applied to a plain text and these are mediated by two basic services: (1) an orchestration framework

[19] This will need to be made explicit in the construction phase so that researchers can use different relation sets.

allowing users to create and manipulate workflows and (2) a workflow execution engine. In some cases the workflow execution engines can execute workflows specified in a generic language such as BPEL. For both tasks there exist general-purpose frameworks such as jBPM to graphically create workflows and ODE, which is executing BPEL scripts. Several others have been developed. On the other hand there is the experience that at least orchestration environments should be specific for a domain to make the user interface simple and to support community traditions and semantics. This and reasons of simplicity and flexibility convinced the D-SPIN team to focus on an orchestration engine which covers a domain specific orchestration environment and a specific execution framework.

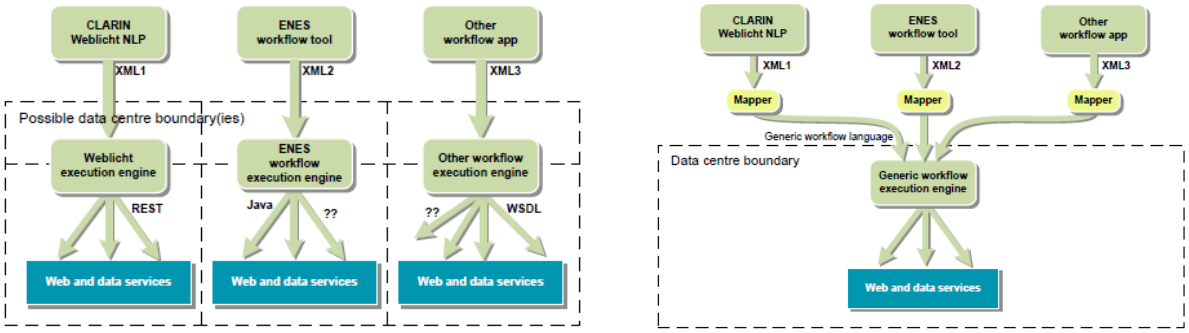

Fig4: In the left figure it is indicated that every domain (linguistics, climate research, etc) has defined its own complete workflow orchestration and execution environment. When this needs to be installed on a large computer center it would mean that a variety of software tools would have to be supported and maintained. A much more economic solution would combine the advantages by using a a domain specific orchestration tool with which the user is being confronted and a generic execution engine allowing proper support.

D-SPIN/CLARIN is aware of the need to come to economic solutions if we want to allow many users to create and execute workflows. This can only be done by deploying the services on large compute centres. For the computer centres it would hardly be possible to support and maintain a large number of different execution engines provided by different communities. Therefore, in a different context D-SPIN/CLARIN already agreed with the computer centres in Jülich and Garching that a scenario as indicated in the right half of figure 4 would be optimal, where the domain specific orchestration tools create the workflows in a generic specification language so that they can be executed by one generic execution engine. This will be investigated and implemented in the coming years.

Part of the basic services offered in the WebLicht Framework are of course all types of converters that allow including different types of data resources. These basic services all together made it possible to have included so many services and resource types as indicated in chapter 2.3.2.

4.4.2 The C4 Corpus

The C4-Corpus is a common effort of the Berlin-Brandenburg Academy of Sciences, the Austrian Academy of Sciences, the University of Basel (Schweizer Textkorpus) and the Free University of Bozen (Korpus Südtirol). The history of the C4-Korpus reaches back to the year 2006, when the four partners formally agreed to construct a shared corpus. The common goal was to provide a reference corpus for the study of regional varieties of standard German. Each partner is supposed to supply 20 million tokens from their corpus materials. The C4 corpus is not yet completed. The corpus currently consists of 20 Million tokens of the German and Swiss Corpus each, 4.1 million tokens from the Austrian corpus and 1.7 tokens from the corpus of South Tyrol. However, it can already be queried

e.g. through the DWDS-website of the Berlin-Brandenburg Academy of Sciences (www.dwds.de) and by a portal which is offered by the University of Basel http://chtk.unibas.ch/korpus-c4/search).

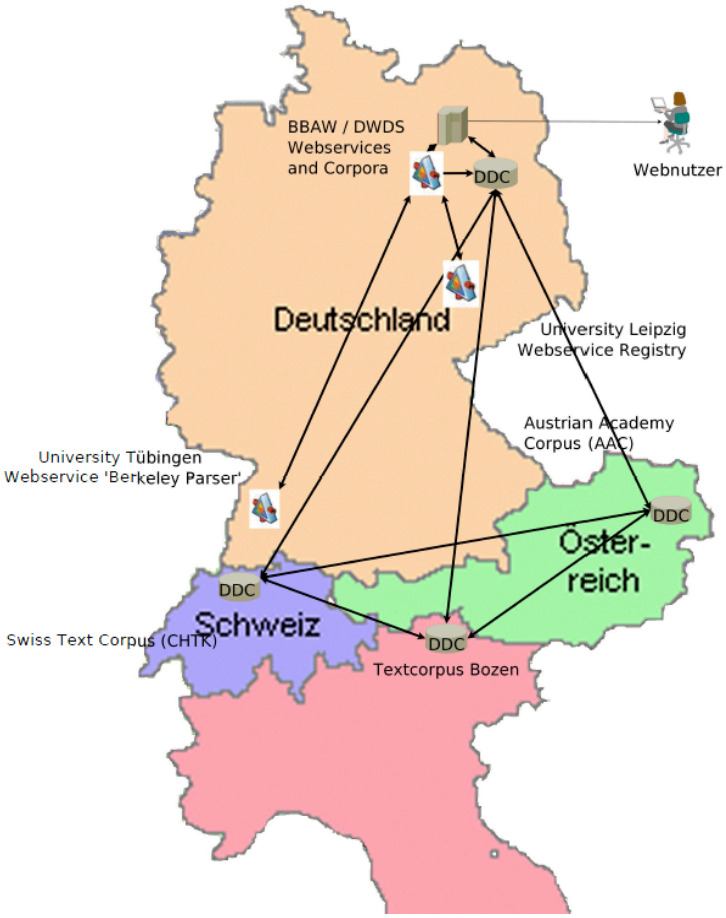

Fig 5: Distribution of the C4-Korpus and related services

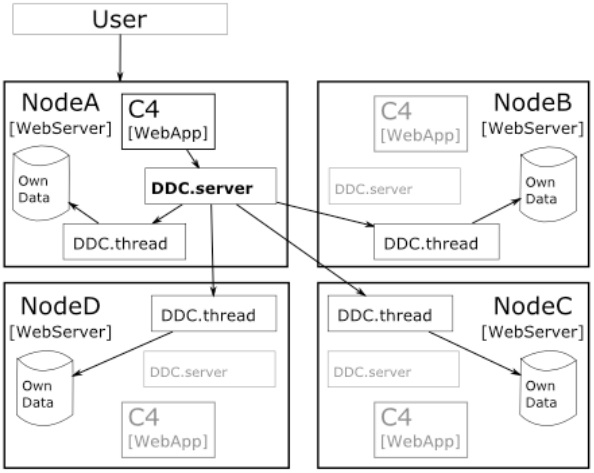

Fig. 6: Interaction and data flow

This collection of texts is virtual in a sense that there is no ONE place (server etc.) where all the texts reside. Indeed the four parts of this corpus reside on the servers of the donating institutions.

Nevertheless they can be considered, and, even more important, queried as a whole (see fig. 6). Thus it is a virtual corpus users can use.

The heart of this virtual corpus is the search engine DDC ("DWDS Dialing Concordance"). This search engine[20] has been implemented for the DWDS project and has been provided since then as a web service by the DWDS team. The interaction and data flow is presented in figure 6.

In CLARIN/D-SPIN this virtual corpus was connected to the WebLicht workflow framework to execute workflows on C4 texts.

4.4.3 Basis-Services for Audio/Video-Resources

Due to the special nature of dealing with audio, video and other time series streams there is still no usable alternative to a desktop solution for the accurate creation and analysis of annotations, this research domain needs to rely on powerful desktop applications from which all types of services can be accessed. The setup is similar to the solution that has been chosen for example for Google Earth or complex Flash applications where major processing is also being done on the local computer. This is the reason why the widely used ELAN tool has been extended to access external services via APIs such as ISOcat for making use of registered concepts when creating annotation schemes and external resources by using URLs or PIDs pointing to different repositories.

So ELAN can be seen as an application that will emerge to a Virtual Research Environment and thus forming a basic service for annotating and analysing annotated streams. There is a web-based cousin of ELAN called ANNEX[21], which currently is used as a visualization tool. Due to a lack of alternatives this web-based tool is linked with the MPI archive setup. In the construction phase of CLARIN such a visualization tool should become available as an independent tool that can be invoked via an API.

As explained above it is now possible to execute for example remotely installed audio and video detectors. An API has been defined which can easily be ported to a REST interface for example to make it a true web service. Summarizing we can say that in the multimedia domain applications such as ELAN are "basic" services.

Another basic service is a search solution for annotations stored on distributed repositories as indicated above. CLARIN/D-SPIN is working on this and will make use of the SRU/CQL protocol as indicated above. Yet there are not so many centres that have an integrated archive for all their multimodal annotations and that support a search engine such as TROVA[22] at MPI. Making this accessible via an API supporting SRU/CQL is work to be carried out in the construction phase.

3 Conclusions

Since this is the last report from AP 2 in the preparatory phase we take the opportunity to outline future activities in the construction phase on the basis of what has already been achieved. We differentiate between four dimensions: User-oriented activities, data resource issues, tools/services issues and infrastructure issues.

[20] Sokirko, Alexey (2003). *DDC – A search engine for linguistically annotated corpora*. In: Proceedings of Dialogue 2003, Protvino, Russia, June 2003

[21] http://www.lat-mpi.eu/tools/annex

[22] TROVA is integral part of both ANNEX and ELAN.

User Oriented Activities

- Based on the early use case studies and what has been achieved so far we need to take the time to re-adjust all our efforts in infrastructure building, i.e. we need to analyse the documented use cases again, collect new ones, discuss again with community members and extract the type of services that are required and specify the type of infrastructure services again;
- We need to look for more integrated scenarios that have the potential to overcome the traditional boundaries for example between audio/video processing on the one hand and NLP processing on the other - more examples could be mentioned;

Data Resource Issues

- Obviously we need to invest time to carry out a digitization/archiving campaign to find out which other valuable resources are out there and to upload them into the centre repositories in a curated form with high-quality metadata;
- Obviously we need to curate the current metadata by providing background ontologies, since we recognized that some data providers are not in the state to modify the original sources easily;
- In addition the quality of the metadata description needs to be adapted to the expectations required by state-of-the-art processing;
- A granularity guideline needs to be worked out that specifies a proper setup of a repository so that individual resources from a collection can be addressed;
- Time needs to be spent to increase the quality of the resources by relating the used categories with the ones in ISOcat and to provide explicit schemas that are compliant with standards or best practices;

Tools/Services Issues

- In the construction phase we need to address issues such as usability, user friendliness and code consolidation and maintenance more than before;
- We need to focus more on jointly built and thus non fragmented Virtual Research Environments that give the researcher access to a wide variety of functionality overcoming current boundaries;
- This and extended functionality needs to be designed based on a new analysis of user wishes; we assume that small workflows and manual interventions will need to be considered since for many daily research problems standard procedures will not work (but this needs to be verified);
- In this respect it is important to make audio/video/time series processing components available as web services as well and to solve the data exchange and/or deployment problem;
- Studies to make use of a generic execution framework should be carried out and all services should be deployed at computer centres to effectively allow any researcher to execute services;
- Available visualization tools need to become available as services and new ones need to be provided to offer data in attractive ways;

Infrastructure Issues

- The fragmented registry and metadata situation needs to be overcome as described above; this needs to be paralleled with a quality check of all registered metadata and registries;
- ISOcat and CMDI as agreed must be the basis, since this is the widely accepted standard for integration;

Obviously in the preparatory phase we were under an enormous time pressure to deliver prototypes, which necessarily resulted in some ad hoc decisions. In the following construction time we need to take the time also for re-adjustments of plans, for overcoming the ad hoc decisions, improving usability and focus more on end-services.

Compared to many other research infrastructures D-SPIN/CLARIN tackled a number of so-called hard problems and came up with prototypical solutions. For us it is good to see that by having done so D-SPIN/CLARIN is very well respected in the overall discussions about research infrastructures.

Appendix A: Resources from German Institutions

Abteilung Computerlinguistik der Uni. Erlangen (1)
Abteilung für Computerlinguistik der Universität Erlangen-Nürnberg (1)
Akademie (1)
Akademie Verlag (2)
Akademie der Wissenschaften (1)
Akademie der Wissenschaften und der Literatur (1)
Akademie-Verlag (2)
Akademie-Verlag. (1)
Akademische Buchdruckerei (1)
Alfred Hölder (1)
Arbeitsgruppe Kognitionsforschung, FB2, Universität Paderborn (1)
Arnoldus-Druckerei (1)
BIS Verlag (1)
Berlin (2)
Berlin : Akademie Verlag (1)
Berlin : Akademie-Verlag (11)
Berlin : Asienkunde (1)
Berlin : Buchhandlung der Berliner ev. Missionsgesellschaft (1)
Berlin : D. Reimer (9)
Berlin : D. Reimer (E. Vohsen) (1)
Berlin : D. Reimer/E. Vohsen (1)
Berlin : De Gruyter (2)
Berlin : Deitrich Reimer (1)
Berlin : Dietrich Reimer (Ernst Vohsen) (1)
Berlin : Dietrich Reimer Verlag (1)
Berlin : G. Reimer (2)
Berlin : J.J. Augustin (1)
Berlin : Langenscheidt (1)
Berlin : Langenscheidt KG (1)
Berlin : Mouton de Gruyter (1)
Berlin : Mouton/de Gruyter (1)
Berlin : Reuther &Reichard (1)
Berlin : Springer-Verlag (1)
Berlin : W. de Gruyter (1)
Berlin : Walter De Gruyter (1)
Berlin ; New York : M. de Gruyter (2)
Berlin ; New York : Mouton (1)
Berlin ; New York : Mouton Publishers (2)
Berlin ; New York : Mouton de Gruyter (12)
Berlin ; New York : Springer-Verlag (1)
Berlin ; New York : W. de Gruyter (1)
Berlin ; New York : Walter de Gruyter (2)
Berlin New York : Mouton de Gruyter (4)
Berlin, Germany : Akademie-Verlag (1)
Berlin, Germany : Verlag Von Dietrich Reimer (1)
Berlin-Brandenburg Academy of Sciences (9)
Berlin-Schoneberg : Langenscheidt ; London : Methuen (1)
Berlin; New York : M. de Gruyter (1)
Bertelsmann (1)
Bielefeld, Germany : University of Bielefeld (1)
Braumüller (1)
C. Klincksieck (2)
Carl Gerold's Sohn (1)
Carl Winter (1)
Carl Winter Universitatsverlag (1)
Carl Winter Universitätsverlag (1)
D. Reidel (2)

D. Reimer (2)
D. Reimer (Ernst Vohsen) (1)
DOBES (15)
Darmstadt : Wissenschaftliche Buchgesellschaft (1)
Deutsche Akademie der Wissenschaften (1)
Deutsche Morgenlandische Gesellschaft (1)
Deutsche Morgenländische Gesellschaft (DMG) (1)
Dissertations Druck (1)
Dr. Ludwig Reichert Verlag (1)
Dresden-Leibzig : Alexander Köhler (1)
Druck von Junge & Sohn (1)
Duden-Verlag (1)
Dusseldorf : Padagogischer Verlag Schwann (1)
Dusseldorf : s.n (1)
Düsseldorf : Pädagogischer Verlag Schwann (1)
Düsseldorf : Schwann (2)
Düsseldorf : s.n (1)
Eisenbrauns (23)
Enzyklopädie (6)
Enzyklopädie. (1)
Erfurt University (1)
FA Brockhaus (1)
Fischer Verlag (1)
Frankfurt am Main New York : P. Lang (1)
Frankfurt on the main : Vittorio Klostermann (1)
Frankfurt-on-Main : Peter Lang Pub (1)
Franz Steiner Verlag GmbH (1)
Freie Universität Berlin (182)
Freie Universität Berlin, Department of Neurology (315)
German Research Center for Artificial Intelligence - DFKI- (2)
German Research Foundation (DFG); Freie Universität Berlin (715)
Giessen : Justus-Liebig Universitat (1)
Gottingen, West Germany : Vanderhoeck et Ruprecht (1)
Gottschalk. (1)
Grossen-Linden : Hoffmann (1)
Gunter Narr (7)
Gunter Narr Verlag (7)
Gunter Narr. (1)
Göschen (1)
Göttigen : Vandenhoeck & Ruprecht (1)
Göttingen : Vandenhoeck & Ruprecht (3)
Haag and Herchen Verlag (1)
Haag und Herchen (1)
Halle : Buchhandlung des Waisenhauses (1)
Halle : M. Niemeyer (1)
Halle, Germany : Max Niemeyer (1)
Halle/Saale : M. Niemeyer (1)
Hamburg : Augustin (1)
Hamburg : Buske (2)
Hamburg : Cram, De Gruyter (1)
Hamburg : Cram, De Gruyter & Co (1)
Hamburg : Deutsches Institut für Afrika-Forschung (1)
Hamburg : Deutsches Institut für Afrika-forschung (1)
Hamburg : Geographischen Gesellschaft (1)
Hamburg : H. Buske (6)
Hamburg : Helmut Buske (2)
Hamburg : L. Friederichsen & Co (2)
Hamburg : L. Friederichsen & Co. (L. & R. Friederichsen) (1)
Hamburg : L. Friederichsen & co (2)
Hamburg : O. Meissner (1)

Hamburg : [Glückstadt, J. J. Augustin] (1)
Hamburg, Germany : Deutsches Institut für Afrika-Forschung (1)
Heidelberg : C. Winter (2)
Heidelberg : C. Winter's Universitätsbuchhandlung (1)
Heidelberg : Groos (1)
Heidelberg : J. Groos (2)
Heidelberg : Julius Groos (2)
Heidelberg : South Asia Institute Kathmandu Tribhuvan University (1)
Heidelberg : Sprachwissenschaftlichen Institut und Seminaren Linguistische Berichte (1)
Heidelberg, Germany : Julius Groos (1)
Heineman (1)
Helmut Buske (11)
Helmut Buske Verlag (20)
Helmut Buske Verlag Hamburg (1)
Hermann (1)
Hermann Böhlaus Nachf (1)
Hildesheim : G. Olms (1)
Humanethologisches Filmarchiv der Max-Planck-Gesellschaft und Humanwissenschaftliches Zentrum der Ludwig-Maximilian Universität München (427)
Humboldt Universität zu Berlin (2)
Humboldt-Universität zu Berlin (141)
Hölder (1)
Im Selbstverlag (1)
Innsbruck : University of Innsbruck (1)
Innsbrucker Beiträge zur Sprachwissenschaft (1)
Institut fur Afrikanistik, Universität Köln (1)
Institut fur Afrikanistik, Universität zu Köln (2)
Institut für Afrikanistik (1)
Institut für Afrikanistik und Ägyptologie der Universität Wien (4)
Institut für Afrikanistik und Äthiopistik der Universität Hamburg (1)
Institut für Afrikanistik, Universität zu Köln (2)
Institut für Deutsche Sprache (13)
Institut für Sprachwissenschaft der Universität Innsbruck (1)
Institut für Sprachwissenschaft, Universität zu Köln (2)
Institute for Natural Language Processing - IMS- , University of Stuttgart (2)
Institute of Indian Studies, University of Groningen (1)
Institute of Indology and Tamil Studies, Cologne University (1)
Jena : Friedrich-Schiller-Universität Jena (1)
Kaiserliche Akademie der Wissenschaften (3)
Kiel : [s.n.] (1)
Kiel University (1)
Klett Verlag (1)
Koln, Germany : Hermann Bohlaus Nachf (1)
Koln, Germany : Institut für Afrikanistik (1)
Kommissionsverlag von G. Reimer (1)
Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften (1)
Köln : Institut für Afrikanistik, Universität zu Köln (1)
Köln : Institut für Linguistik, Abteilung Allgemeine Sprachwissenschaft, Universität zu Köln (2)
Köln : Institut für Sprachwissenschaft der Universität zu Köln (11)
Köln : Institut für Sprachwissenschaft, Universität zu Köln (4)
Köln : Köppe (1)
Köln : R. Köppe (1)
Köln : Rüdiger Köppe Verlag (1)
Köln : Universität zu Köln (1)
Köln, Germany : Institut für Afrikanistik, Universität zu Köln (1)
Köln, Germany : Rüdiger Köppe (1)
Köppe (14)
Köppe Verlag (2)
L. Friederichsen (2)
L. Friederichsen & Co. (1)

LIT Verlag (5)
Langenscheidt (3)
Langenscheidt Enzyklopädie (4)
Langenscheidt. (1)
Leipzig : B.G. Teubner (1)
Leipzig : J. C. Hinrichs (2)
Leipzig : Otto Harrassowitz (1)
Leipzig : VEB Verlag Enzyklopädie (1)
Leipzig : Verlag Enzyklopädie (3)
Leipzig; Baltimore : J.C. Hinrichs'sche Buchhandlung ; Baltimore : Johns Hopkins Press (1)
MPI-EVA (137)
Magdeburg-Stendal University of Applied Sciences (10)
Mainz : Akademie der Wissenschaften und der Literatur ; Wiesbaden : Steiner (1)
Mainz Germany : Methods and Materials Development Center of the International Inlingua Schools of Languages (1)
Max Niemeyer (2)
Max Niemeyer Verlag (1)
Max Planck Digital Library (http://mpdl.mpg.de/) (2561)
Max Planck Digital Library, http://wals.info/author (1)
Max Planck Gesellschaft (2)
Max Planck Institut fuer Psycholinguistik, Nijmegen, Nl. (2)
Max Planck Institute for Psycholinguistics (Nijmegen, Netherlands) (7)
Max Planck Institute for Evolutionary Anthropology (16)
Max Planck Institute for Evolutionary Anthropology, Department of Linguistics (929)
Max Planck Institute for Psycholinguistics (13495)
Max Planck Institute for Psycholinguistics, Nijmegen, Nl (3)
Max Planck Institute for Psycholinguistics; Académie Marquisienne (Tuhuna 'Eo 'Enata) (81)
Max Planck Institute for Psycholinguistics; Académie marquisienne (Tuhuna 'Eo Enata) (7)
Max Planck Institute for Psycholinguistics; Centre Territorial de Recherche et de Documentation Pédagogiques (CTRDP) (5)
Max Planck Institute for Psycholinguistics; Centre Territorial de Recherche et de Documentation Pédagogiques (CTRDP); Académie marquisienne (Tuhuna 'Eo Enata) (129)
Max Planck Institute for Psycholinguistics; Max Planck Institute for Evolutionary Anthropology (5)
Max Planck Institute for Psycholinguistics; Volkswagen Foundation (103)
Max Planck Institute for Psycholinguisticsc (1)
Max-Planck-Institüt für Psycholinguïstik (21)
Mechitharisten-Buchdruckerei (1)
Mitteilungen der Ausland-Hochschule Universität Berlin (1)
München : C.H. Beck (1)
München : Hueber (1)
München : LINCOM EUROPA (1)
München : Lincom Europa (2)
München : R. Oldenbourg (1)
München : Weltforum Verlag (1)
München ; Newcastle : Lincom Europa (1)
München [i.e.] Unterschleissheim Newcastle : LINCOM Europa (1)
München, Germany : LINCOM-Europa (1)
München] : M. Hueber (1)
Münster, Germany : Aschendorff (1)
N. G. Elwert Verlag (1)
Neumünster : Wachholtz (1)
Regensburg University (164)
Reichert (2)
Reichert Verlag (1)
Ruhr-University Bochum, Department of Linguistics (730)
Ruhr-Universität Bochum (1)
Rüdiger Köppe (18)
Rüdiger Köppe Verlag (12)
Saarland University, Computational Linguistics (1)
SaarlandUniversity, Computational Linguistics (1)
Schriftenreihe des Kölner Instituts für Afrikanistik (1)

Schuster (1)
Seminar für Deutsche Philologie; Georg August-Universität Göttingen (1)
Sprachwissenschaftliches Institut, Universität Bochum (1)
Springer (1)
Springer Verlag (1)
Springer Verlag. (1)
Stuttgart: United Bible Societies (1)
Stuttgart: W. Spemann (1)
Stuttgart, GERMANY : Ernst Klett (1)
Stuttgart, Germany : Deutsche Bibelgesellschaft : United Bible Societies (1)
Technische Universität, Chemnitz , Universität Bayreuth (1)
Teubner (1)
Teubner Verlag (1)
Trier, Germany : Linguistic Agency, University of Trier (1)
Tübingen: Gunter Narr (1)
Tübingen: Max Niemeyer Verlag (1)
Tübingen: G. Narr (2)
Tübingen: G. Narr Verlag (1)
Tübingen: Gunter Narr Verlag (5)
Tübingen: J.C.B. Mohr (Paul Siebeck) ; Louisville : Westminster/John Knox Press (1)
Tübingen: M. Niemeyer (1)
Tübingen: Max Niemeyer (1)
Tübingen: Max Niemeyer Verlag (3)
Tübingen: Maz Niemeyer Verlag (1)
Tübingen: Narr (3)
Tübingen: Narr ; Philadelphia, Pa. : Distributed by John Benjamins North America (1)
Tübingen: Stauffenberg Verlag (1)
Tübingen: TBL-Verlag Narr (1)
Tübingen: Universität Tübigen (1)
Tübingen, Germany: Max Niemeyer (1)
University of Bremen (1)
University of Cologne (1074)
University of Cologne, Department of Linguistics (207)
University of Cologne, Institute for African Studies (92)
University of Erfurt (6)
University of Hamburg (145)
University of Hamburg, Archaeological Institute, Mesoamerican Studies (26)
University of Hamburg, Institut für Finnougristik/Uralistik (3)
University of Heidelberg, Heidelberger Akademie der Wissenschaften (1)
University of Kiel (315)
University of Kiel; Association culturelle Te Reo o te Tuamotu (199)
University of Kiel; Max Planck Institute for Psycholinguistics (37)
University of Konstanz (1)
University of Leiden (86)
University of Leipzig (1265)
University of Potsdam (16)
Universität Bamberg, World Language Documentation Centre (1)
Universität Bayreuth, Afrikanistik (1)
Universität Hamburg (10)
Universität Köln (1)
Universität München (2)
Universität Osnabrück (1)
Universität Regensburg (1)
Universität Tübingen, Seminar für Sprachwissenschaft (6)
Universität Zürich (2)
Universität Zürich, Seminar für Allgemeine Sprachwissenschaft (1)
Universität zu Koln (8)
Universität zu Köln, Allgemeine Sprachwissenschaft (61)
Universität zu Köln, Institut für Afrikanistik (1)
Universität zu Köln, Institut für Sprachwissenschaft (2)

Universitätsverlag (1)
Universitätsverlag C. Winter (2)
Universitätsverlag Heidelberg (1)
VEB Verlag Enzyklopaedie (1)
VEB Verlag Enzyklopädie (1)
VGH Wissenschaftsverlag (1)
Verlag Brockhaus (1)
Verlag Dietrich Reimer (Ernst Vohsen) A. G. (1)
Verlag Enzyklopädie (14)
Verlag Enzyklopädie / Max Hueber (1)
Verlag J. J. Augustin (2)
Verlag der Akademie der Wissenschaften (1)
Verlag für Kultur und Wissenschaft (1)
Volk und Wissen (1)
Volk und Wissen Verlag (1)
Volkswagen Foundation (39)
Westdeutscher Verlag (3)
Westfälische Vereinsdruckerei (1)
Westfälische Wilhelms-Universität Münster, Department for General Linguistics (3)
Wiesbaden (1)
Wiesbaden : Akademie der Wissenschaften und der Literatur (1)
Wiesbaden : Dr. Ludwig Reichert Verlag (1)
Wiesbaden : Franz Steiner Verlag (1)
Wiesbaden : Harrasowitz (1)
Wiesbaden : Harrassowitz (3)
Wiesbaden : O. Harassowitz (1)
Wiesbaden : O. Harrassowitz (5)
Wiesbaden : Otto Harrassowitz (1)
Winter (3)
Winter Verlag (1)
Winter. (1)
Wissenschaft und Technik Verlag (1)
Zentral-Antiquariat der DDR (1)
Zentrum für Allgemeine Sprachwissenschaft, Berlin (33)
(München) : Hueber (1)
[Berlin : Akademie-Verlag (1)
[Berlin : s. n.] (1)
[Berlin] : Langenscheidt (1)
[Heidelberg, etc.] : Julius Groos (1)
[München : Huber (1)

Appendix B: Tools from German Institutions

Abteilung Computerlinguistik der Uni. Erlangen (1)
DFKI GmbH (6)
DFKI Language Technology Lab (1)
DFKI, Saarbrücken University (1)
FernUniversität Hagen (1)
Free University Berlin (1)
German Research Center for Artificial Intelligence (DFKI) (2)
Heinrich-Heine-Universität Düsseldorf / Seminar für Allgemeine Sprachwissenschaft (1)
IMS, University of Stuttgart (8)
Institut für Kommunikationsforschung und Phonetik, University of Bonn (1)
Institut für Deutsche Sprache, Mannheim (1)
Institut für Phonetik und Sprachverarbeitung, München (1)
Institute for Communications Research and Phonetics, University of Bonn (1)
Institute for Natural Language Processing (IMS), University of Stuttgart (2)
Institute for Natural Language Processing, University of Stuttgart (5)
Institute of Phonetics and Speech Processing, LMU Munich (1)
Knowbotic Systems GmbH & Co. KG, Frankfurt am Main, Germany (1)
LT-Lab, DFKI GmbH, Germany (1)
Lehrstuhl für Informatik VI, RWTH Aachen - University of Technology (1)
Max Planck Institute for Psycholinguistics (13)
SFB 'Multilingualism', IDS Mannheim (1)
Saarland University, Computational Linguistics (2)
Saarland University, Computational Linguistics Department (1)
SaarlandUniversity, Computational Linguistics (1)
TEMIS Deutschland GmbH / H.A.S.E. GmbH (1)
University of Tübingen (1)
University of Bielefeld and Lomonossov University (1)
University of Bremen (1)
University of Potsdam, Department of Linguistics (1)
University of the Saarland (1)
Universität Bielefeld (1)
Universität Erlangen-Nürnberg (1)
Universität Stuttgart (1)
Universität Hamburg (1)

Appendix 3: SRU/CQL

Within the CLARIN context Search/Retrieval via URL (SRU)/ Contextual Query Language (CQL)[23] as a candidate combining both metadata and content search. This document will not describe the full functionality of SRU/CQL, but be limited to the agreed functionality for two distributed search scenarios in CLARIN/D-SPIN: (1) The C4 corpus project brings together German resources from BBAW, U Vienna, U Basel and IDS and is intended to allow joint searches on all of them. (2) The multimedia/multimodal search case is intended to bring together data sets from a number of institutes including the D-SPIN partners: MPI, IDS, U Hamburg, BAS. We will not describe the details of these projects but describe briefly how SRU/CQL is used in these projects in the first phase.

The queries will be limited to a simple Google like search for strings contained in the annotations of the resources.

In the following we will

- Introduce the SRU packaging format
- The CQL language focusing on the intended search cases
- And a bit about CQL syntax (for more info we refer to the web-information)

SRU

Search/Retrieval via URL (SRU)[24] is a standard XML-focused search protocol for Internet search queries, utilizing CQL (Contextual Query Language), a standard syntax for representing queries. The main operation in SRU is 'searchAndRetrieve' allowing a client to submit a query and retrieve the results for matching records.

The mandatory request parameters are:

Name	Mandatory/Optional	Description
operation	Mandatory	The string: 'searchRetrieve'.
version	Mandatory	The version of the request, and a statement by the client that it wants the response to be less than, or preferably equal to, that version.
query	Mandatory	Contains a query expressed in CQL to be processed by the server. See 2.1 CQL

An example of SRU query is given here:
http://z3950.loc.gov:7090/voyager?version=1.1&operation=searchRetrieve &query=dinosaur

The response to a searchRetrieve operation is an XML document containing, amongst others, the result set information. The main response parameters for SRU are listed below:

[23] SRU is a standard that is already widely used in the library domain for distributed search, thus it makes sense to adopt SRU as well.
[24] http://www.loc.gov/standards/sru/specs/search-retrieve.html

Name	Type	Mandatory/Optional	Description
version	xsd:string	Mandatory	The version of the response. This MUST be less than or equal to the version requested by the client
numberOfRecords	xsd:integer	Mandatory	The number of records matched by the query. If the query fails this MUST be 0.
Records	sequence of <record>	Optional	A sequence of records matched by the query

All other response parameters are optional in SRU.

For the return of the hits a structure was chosen which returns the typical words in context information (concordance) including eventually three tiers. For every tier there needs to be an indication at which positions the string has been found. For every hit there needs to be a reference to the media resource or fragment or to a player with a corresponding identifier. The figure shown below indicates how metadata search results and content search results are combined:

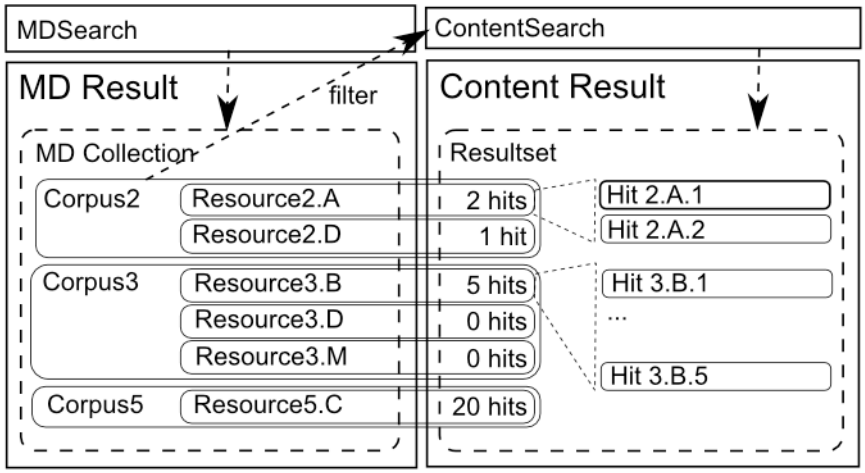

Figure 1: Metadata and content search result sets. The left side indicates the resources in which hits were found, they can come from different collections/corpora and probably there are some hits in a given collection. Of course as hits the URLs or PIDs of the resources are returned. For resource 2.A obviously from collection 2 two hits may be the result. In the content part we expect then a concordance style of result to be returned (where possible). This is shown in the following table in more detail. For each hit we expect a reference, a start and end specification (for videos this obviously are frame numbers) and where possible a KWIC concordance string to be returned.

1	reference (Resource2.A)	start	End	kwic string
		start	End	kwic string
2	reference (Resource2.D)	start	End	kwic string

3	reference (Resource3.B)	start	End	kwic string
		start	End	kwic string
		start	End	kwic string
		…	…	…
4	reference (Resource5.C)	start	End	kwic string
		…	…	…

Where:

Source reference refers to the persistent identifier of the resource or the metadata data record (up for discussion)
Start refers to the fragment identifier of the start of resource fragment
End refers to the fragment identifier of the end of the resource fragment
Kwic string provides a representation of the resource fragment itself.

This information needs to be packaged into the SRU *record* specification[25]. In SRU each record is however structured according to the following mandatory elements shown in the table below. At the time of writing however the structure of the XML fragment describing the individual record data of the result set (<recordData>) are still to be specified.

Name	Type	Mandatory/Optional	Description
recordSchema	xsd:string	Mandatory	The URI identifier of the XML schema in which the record is encoded. Although the request may use the server's assigned short name, the response must always be the full URI.
recordPacking	xsd:string	Mandatory	The packing used in recordData, as requested by the client or the default..
recordData	<stringOrXmlFragment>	Mandatory	The record itself, either as a string or embedded XML.

CQL

The Contextual Query Language (CQL) is a formal language for representing queries targeted at information systems. In CLARIN queries are combined metadata and content search queries, i.e. a single query is constructed that will allow searches through both CLARIN's central metadata catalogue and individual search engines. The execution method is performed in two steps; first the query is sent to the metadata search engine which will extract all relevant metadata search criteria and return metadata records of resources that match these. Next, the query, modified with the result set, is delivered to those content search engines capable of executing the query on the filtered resources.

[25] http://www.loc.gov/standards/sru/specs/search-retrieve.html

An example of a proposed CLARIN specific CQL[26] query is shown below.

Query = cmd.organization isA "University" and cmd.title any "Liebe" and word="Herz" prox/unit=word/distance>2 word="zerreißen"

This query combines a number of CQL extension features proposed by the team. In the query shown above both metadata as well as content search parameters are passed. Organization (cmd.organization isA "University") is typically handled at the metadata level in the distributed search. Title (cmd.title any "Liebe") certainly indicates an overlap between metadata search and content search since we can assume that certain content search engines also include metadata terms to perform the search. Thus "Liebe" will be dealt with at external metadata level and might also be applied by the individual search engines. Some content search engines are able to use these metadata characteristics to improve results. The terms "Herz" and "zerreißen" are used to search on annotations. They can be found at word level and both terms should be found with a distance of 2 words between them.

The prefixes *cmd* and *css* indicate context sets that define the scope of the semantics. If no context set is defined, as for *word*, the default CQL context set is to be used. For CLARIN two context sets are defined to differentiate between metadata and content search:

cmd Component Metadata - proposed CLARIN's adopted context set in SRU for metadata
ccs CLARIN Content Search - proposed CLARIN's adopted context set in SRU for content

Given the limited amount of time we therefore suggest limiting the content search clause for the multimedia/multimodal demonstrator to allow users to search for the following simple content patterns:

- Case 1: search for a string in all annotation tiers; these strings can contain words or tags, i.e. strings appearing somewhere in the annotations where the usual separators are used (space, ".", ",", "!", "?" etc.) to identify units. The specified string can appear as a substring somewhere.
  pattern1 = "string" allowed are all UNICODE characters
  Example:
  "Liebe"

- Case 2: search for two strings in all tiers that are separated by a number of words where the usual separators are used (space, ".", ",", "!", "?" etc.) to identify units and where the strings are found on the same tier.
  pattern1 = "string" allowed are all UNICODE characters
  pattern2 = "string" allowed are all UNICODE characters
  distance = "1|2|3|4|5" "1" means adjacent where pattern1 is the first
  Example:
  "Liebe" prox/unit=word/distance=1 "Herz"

- Case 3: search for three strings in all tiers where the usual separators are used (space, ".", ",", "!", "?" etc.) to identify units and where all patterns belong to the same moment in time; thus

[26] http://www.loc.gov/standards/sru/specs/cql.html

in multimodal annotations it could be what someone is speaking and gesturing at the same moment in time[27] and for texts it will be for example a “word” with its “POS” and “morphology” annotation.

pattern1 = “string” allowed are all UNICODE characters
pattern2 = “string” allowed are all UNICODE characters
pattern3 = “string” allowed are all UNICODE characters
distance = “0” “0” means the same time period
css.morphology means to find the substring on the morphology tier that presupposes that we all use the correct tier labelling (therefore we probably need to do a mapping)

Example:

ccs.word = "Liebe" prox/unit=annotation/distance=0 ccs.POS = "V"
prox/unit=annotation/distance = 0 css.morphology = "lieb"

For all strings submitted 10 alternative strings can be specified to cope with naming differences within and across languages. These alternatives can be generated by the user, by looking at ISOcat or by looking into an online lexicon such as LEO.

[27] For reasons of simplicity an overlap also counts as same time period.