D-SPIN Report R5.1: Guidelines for the Standard-Conformant Adaptation and Creation of Language Resources

June 2009

D-SPIN, BMBF-FKZ: 01UG0801A

Deliverable: R5.1: Guidelines for the standard-conformant adaptation and creation of language resources

Responsible: Erhard Hinrichs

Editors: Erhard Hinrichs, Lothar Lemnitzer, Kathrin Beck, Iris Vogel, Daan Broeder, and Peter Wittenburg

1 Scope of the Document

This document provides guidelines for the standard compliant adaptation and creation of language resources. It is addressed at providers of such resources who want their resources to be integrated into the CLARIN/D-SPIN LRT federation.

The document builds on work that has been performed mainly by CLARIN/D-SPIN WP 2, "Technical Infrastructure" and WP5 "Language Resources". In particular, it builds on the following CLARIN deliverables:

- D2R-2a Federation Foundation – LRT
- D2R-2b Federation Foundation - PIDs
- D2R-4 Registry Requirements
- D2R-6a Web Services and Workflow Requirements
- "Standards for LRT" – CLARIN Technical Report by Nuria Bel et al.
- "CLARIN Metadata Infrastructure" – CLARIN Technical Report by Daan Broeder

As well as on the following D-SPIN reports:

- Helmut Schmid: "The technical details of the D-SPIN Architecture" (Internal technical report 2009)

This document will be discussed in the relevant D-SPIN working groups and boards and will be subject to regular adaptations reflecting the progress of the D-SPIN project.

2 Introduction

D-SPIN is the German counterpart of the CLARIN project. These guidelines will therefore refer to CLARIN specifications whenever this is appropriate and also include some of the key documents of CLARIN as appendices. Integration of language resources into the LRT federation has to be kept in line with the CLARIN efforts.

Resource providers are the primary target group of these guidelines. In the course of time, we expect many German resource providers to join the D-SPIN efforts in building an LRT federation and to provide their resources. To this end, D-SPIN organized a German LRT-Summit in Mannheim in May 2009, which brought together the major resource providers in Germany. These resource providers as well as other potential participants need to understand what is required for a resource to be CLARIN/D-SPIN conformant as well as which support services can be offered by CLARIN and D-SPIN to the resource providers to meet these requirements.

Work on standards, e.g. for major types of resources, for their metadata and encoding formats, as well as for tools and web services is currently going on in CLARIN and D-SPIN. There is good progress in all fields. Nevertheless, most standards and specifications are not finalized yet. Since work towards stable specifications for all aspects of the LRT federation is still ongoing, these guidelines should be considered to be preliminary and subject to further adaptations as the projects move on.

3 Purpose of this Document

The aims of this document are, with respect to metadata:

- To explain the building blocks of the CLARIN metadata sets for language resources and tools
- To describe the relation of the CLARIN metadata sets to existing metadata schemata, e.g. Dublin Core, OLAC, IMDI and TEI
- To give recommendations to those who want to integrate their language resources into the CLARIN/D-SPIN LRT federation and therefore want to make their metadata conformant with the CLARIN metadata sets.

With respect to the encoding and annotation of resources:

- To explain standards for data formats of various resource types (i.e. text, audio, pictures etc.)
- To describe the position of the CLARIN federation towards existing standards
- To give recommendations to those who want to integrate their language resources into the CLARIN/S-SPIN LRT federation and therefore want to make keep their data conformant with the CLARIN supported standards

More generally:

- To describe in which ways CLARIN and D-SPIN can give support to resource providers with respect to standards.

Note that we are still in an early phase of the preparatory phase of both CLARIN and D-SPIN. The goals of this phase are, among others, to specify stable specifications for various aspects of the infrastructure, to establish a prototype with limited scope and to assess the costs and efforts for the implementation of an encompassing infrastructure. However, we consider it reasonable to inform the LRT community at an early stage and to engage them in discussion about standards for LRT. The LRT community should be made aware of the infrastructure building activities, be able to provide feedback and to look at these efforts from their particular perspective e.g. as language resource providers.

4 The CLARIN Metadata Infrastructure

Metadata are an essential part of all language resources[1]. They are used to describe various aspects of these resources and to classify these resources according to many criteria. Stored in catalogues, they will enable potential users to find exactly those resources that they need for their research, for reference or for any other purpose. Thus, metadata increase the visibility of the resources and guarantee their widest and appropriate use.

In the area of metadata for language resources there are many and often competing metadata schemes. Existing metadata schemes are often too inflexible to accommodate the wide range of LRT resources that need to be modeled. Secondly, the semantics of metadata attributes and values is often too vague to achieve the goal of semantic interoperability between metadata categories.

The CLARIN LRT federation, of which the D-SPIN project is a partner, will try to address these problems by (a) taking a modular approach to metadata categories and (b) establishing links to data category and concept registries with clear definition of the categories. We will explain both aspects in more detail in the following sections.

4.1 Generation of Metadata Descriptions: Elements, Components, and Profiles

CLARIN/D-SPIN conformant metadata will be modular structures built from *metadata elements*. The metadata elements are the building blocks for *components* while components will be combined into *metadata profiles*.

Definition 1

A **Metadata element** (or **Metadata descriptor**) is an atomic part of a metadata description - a combination of a name and value that together with the other metadata elements form the metadata descriptions describing the associated resource. In a metadata schema, a metadata element is characterized by a name and a domain of values. The domain can be a finite vocabulary or a regular expression constraining the value. A CLARIN metadata element also provides a link to a concept in a registry that could provide vocabularies and constraints.

Definition 2

A **CLARIN Metadata component** is an aggregation of metadata elements and components aimed at describing a specific aspect of a resource.

Definition 3

A **CLARIN Metadata profile** is specification of an aggregation of metadata components that can be (re-) used to create metadata descriptions. The profile is not different from a component itself except that it is used to describe all relevant aspects of a single resource or resource collection. The profile contains a specification of components together with specifications of cardinality, mandatory presence, default values and guidance for applications. The profile can be exported into a suitable XML schema.

[1] In this document the term "resources" includes data resources as well as tools, web applications and web services.

Typical metadata elements are: AUTHOR, CHARACTER-ENCODING, DATE. While the domain of the AUTHOR element can be any string, the domain of the ENCODING element can only be one of a fixed set of encoding schemes, e.g. ISO-8859-1 or UTF-8 if the element refers to the character encoding of texts. The value of the DATE element can be restricted to a pattern according to international standards. CLARIN/D-SPIN will provide definitions along the names and domains of the data elements that are supported. With this information the meaning of each metadata element should be well defined and transparent to everybody who wants to use this element. Commonly used metadata elements and their corresponding concepts will be registered in a concept registry by the responsible CLARIN staff. The main candidate for such a registry is ISOcat. However, users who do not find the metadata element that they need to properly describe their resource can make a suggestion to extend the concept registry with this metadata element.

The CLARIN and D-SPIN staff will take care of building **components** from these elements, tailored to explicit needs of the resource providers. These components are handy aggregations of metadata elements which describe aspects of these resources, e.g. location and project description, and which are expected to be of use to many resource providers. The same holds for **metadata profiles.** These aggregations of metadata elements and components describe a resource in sufficient breadth and depth. They are supposed to be used "off the shelf" for major resource types. Resource providers can of course customize these profiles to their needs. They are nevertheless recommended to use CLARIN conformant metadata elements wherever possible.

Definition 4

A **CLARIN XML metadata schema** is an XML-Schema (presented as a Document Type Definition, in the XML Schema language, in RELAX NG etc.) that formally defines a metadata description as built up from metadata components.

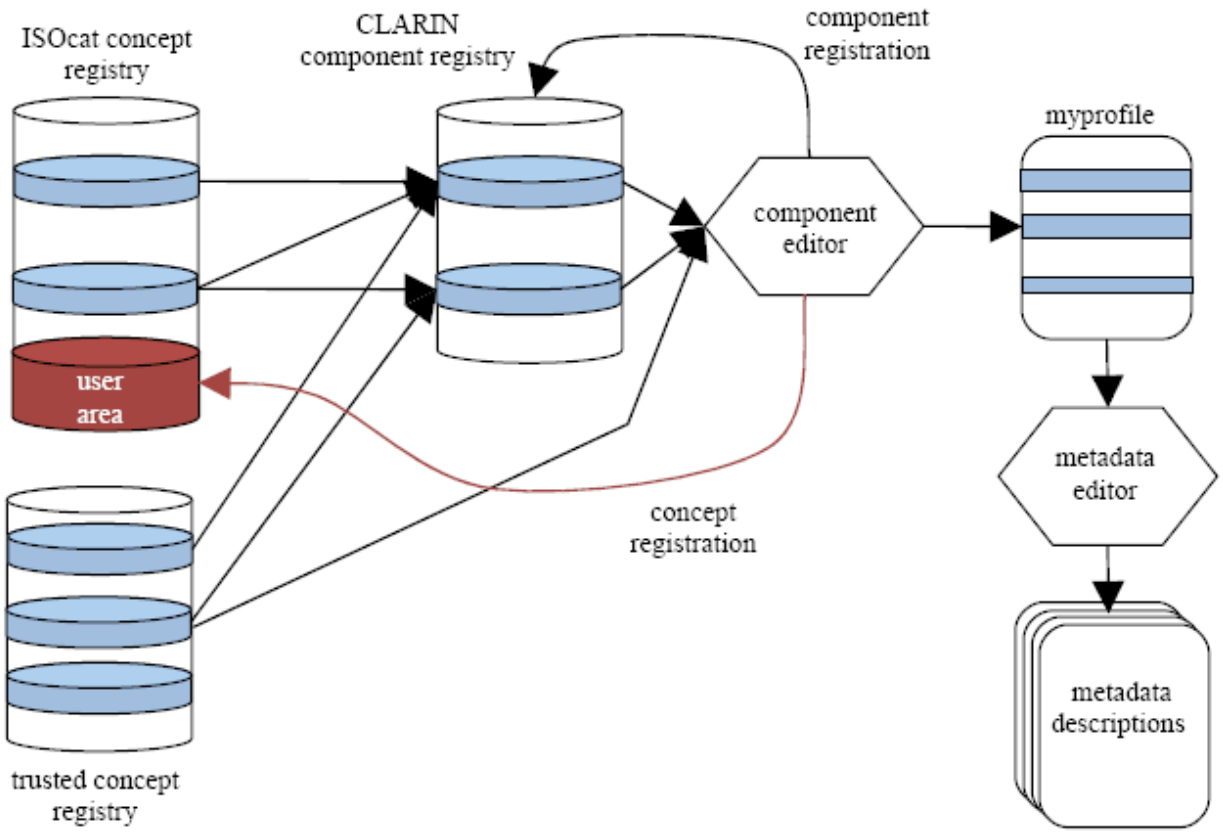
Metadata components and profiles will be provided in the form of XML-Schemata. Schemata are grammars that describe the set of well-formed instances. In other words, document grammars are a helpful tool for the validation of a concrete metadata set or profile against a specification.

Figure 1 depicts a typical use case of the CLARIN metadata infrastructure:

A user, i.e. a resource provider, will create a **metadata description**. Users should be able to use a **metadata editor** for this purpose, although they do not need to use one. Such an editor will be provided, in the near future, by the CLARIN project. It will support users in the process of manually generating and maintaining metadata records.

Metadata descriptions should be aggregated from components that are available in the CLARIN component registry. Of course users will also be able to build their custom components and profiles from the set of available metadata elements. Metadata elements can be fetched from the ISOcat concept registry (this is the preferred way) or from any other "trusted concept registry". **Profiles** are aggregations of metadata elements and components. If an appropriate profile is available, this profile can be taken as the blueprint for the metadata description. If this is not the case, the user will be able to create their profile ("myprofile"), to store it for future use and to generate an XML-Schema from this profile. This schema can then be used to validate a metadata description, which has been drafted with a general-purpose XML editor, e.g. Oxygen.

So far we have described how resource providers wanting to create new metadata descriptions should proceed.



Recommendation 1

For these resource providers, **it is recommended** that they (a) use the metadata elements and components which are registered in accepted category registries such as DC, ISOcat and TEI, that they (b) provide metadata elements, their names, domains and definitions, for inclusion into the registries for metadata elements and components which are not yet available and that they (c) use the tools provided by CLARIN, in particular the metadata editor and component editor.

4.2 Using Existing Metadata Descriptions

Many resource providers will already have their resources described using existing metadata schemes, e.g. Dublin Core, OLAC, IMDI or the TEI header, to mention just a few.

These metadata descriptions will be supported by CLARIN/D-SPIN. In the following we will briefly describe the standards which will be supported and afterwards outline what this support will look like.

4.2.1 Dublin Core Metadata Initiative (DCMI)

The Dublin Core Metadata Initiative started by defining a restrictive set of elements with semantically broad categories. The original 13 core elements were later increased to 15: *Title*,

Creator, *Subject*, *Description*, *Publisher*, *Contributor*, *Date*, *Type*, *Format*, *Identifier*, *Source*, *Language*, *Relation*, *Coverage*, and *Rights*. Later on this was extended this by a second level, the qualified Dublin Core set with three more elements and rather precise semantics. For further details, cf. CLARIN Metadata deliverable, section 4.2.3.

Within the CLARIN metadata infrastructure, it might be useful to generate a DC-compatible metadata description from other metadata descriptions. These minimal metadata records can then be shared with initiatives and repositories that are more strongly oriented to the world of digital libraries. Sharing the metadata of our resources will enhance their visibility.

CLARIN will provide support for the conversion of CLARIN-conformant metadata sets to DUBLIN Core.

4.2.2 Open Language Archiving Community (OLAC) Metadata Set

The OLAC Metadata Set is the set of metadata elements that members of the Open Language Archiving Community have agreed to use for discovering language resources that come from various archives. The metadata set consists of all the elements of the Dublin Core Metadata Set. To this core set, OLACMS adds a set of refinements and qualifications that are designed for describing fundamental properties of language resources, such as subject language, language data type, and software functionality. The OLACMS Standard uses XML to represent metadata descriptions. For further details, cf. CLARIN Metadata deliverable, section 4.2.7.

The OLAC Metadata Set will be supported as a pivot format in the CLARIN LRT domain. In practice this means that services like the harvesting and cataloguing of metadata will be based on that format. CLARIN will provide support for the conversion of CLARIN-conformant metadata sets to OLAC.

4.2.3 Text Encoding Initiative (TEI) – The Metadata Header

The Text Encoding Initiative has developed a widely used standard for the markup of electronic resources in the humanities domain (ranging from corpora like the British National Corpus to poems and dramas). Metadata can be embedded in the header of a TEI-file, which generally contains fields that correspond to a bibliographic record (e.g. title, distributor). As from the most recent version of TEI (called "P5") its schema can be customized by creating a file that contains prose descriptions and a formal specification of the newly introduced elements.

TEI header elements are widely known also in the LRT domain and are used in a number of projects to characterize resources. It seems that TEI has received a new momentum. CLARIN/D-SPIN will support TEI-header elements. It is not possible to include the full richness of TEI headers, but TEI has decided to make their header categories referable so that they can be used in simpler frameworks.

4.2.4 ISLE Metadata Initiative (IMDI)

The IMDI Framework [IMDI] offers, on top of to a suitable set of XML-based metadata descriptors for language resources, a set of tools and an infrastructure to use these. IMDI focuses on the description of annotated multimedia/multi-modal resources. Despite this focus it covers a number of other aspects: (a) It provides a schema for lexica with multimedia extensions; (b) It has a special schema for corpora; (c) It has a special profile to allow the integration of the TEI header elements of the Dutch spoken corpus; (d) It has a special profile for Sign Language. For further details, cf.

CLARIN Metadata deliverable, section 4.2.5. IMDI will be supported by the CLARIN Metadata Infrastructure.

Besides these Metadata description formats, which are widely known and used in the Language Resource domain, there are some other formats which we will not mention here. For details of these formats, cf. CLARIN Metadata deliverable, section 4.2.

Recommendation 2

Resource providers who already have produced metadata for these resources **are advised**

1. to register these resources to the CLARIN ad-hoc repository (cf. http://www.clarin.eu/summary-wp5-language-resources-and-technology-overview);
2. to inform the CLARIN/D-SPIN metadata colleagues which metadata standard they are using;
3. to provide to the CLARIN/D-SPIN metadata colleagues a specimen of their metadata descriptions;
4. to help the CLARIN/D-SPIN metadata colleagues to identify matching CLARIN metadata elements and components where they are available and, if this is not the case, to register the elements and components;
5. to assist the CLARIN/D-SPIN metadata specialists in writing converters from their metadata description to the CLARIN pivot formats (DC and OLAC).

Towards the end of the preparatory phase we hope to have taken into account most of the widely used metadata standards for language resources and tools and to have provided ways for these metadata descriptions to be integrated into the CLARIN Metadata Infrastructure.

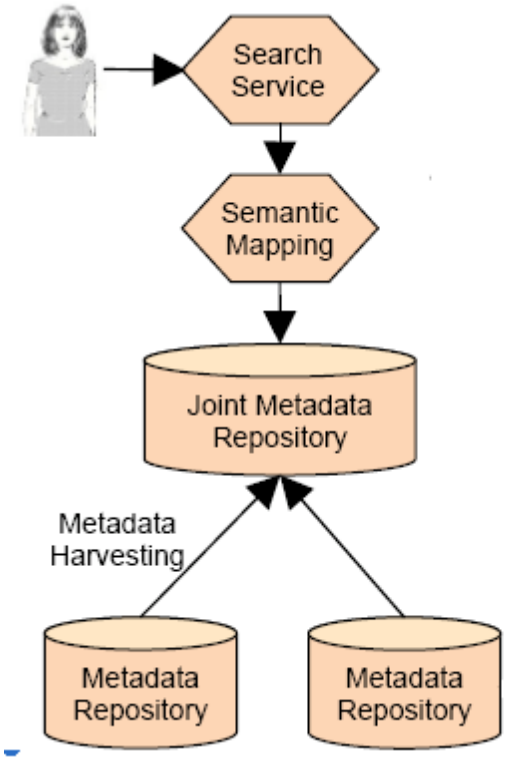
In the next section we will describe what will be done with these metadata descriptions within the metadata infrastructure: harvesting and providing access via catalogues that can be queried and browsed.

4.2.5 DFKI Tool Registry

At DFKI the tool registry has been maintained. Tools can be described with the help of a number of elements. These were taken as basis for the CLARIN tool inventory and also found their way into the CLARIN metadata category set which is now subject of registration in ISOcat.

4.3 Harvesting and Cataloguing of Metadata for LRT

Available metadata descriptions will be harvested and stored in metadata repositories at service provider sites offering portals. Such metadata repositories will offer a variety of services such as a metadata search and browsing service to meet the wishes of the user community (see figure 2) that needs to easily find relevant resources and tools in the emerging big market place.



The wish to have a central catalog covering metadata from different repositories has led to the emergence of the OAI-PMH protocol (which stands for "Open Access Initiative – Protocol for Harvesting Metadata") as a de-facto standard for gathering metadata – a process that is called metadata harvesting. In the OAI-PMH model the world is divided into data providers who offer metadata for harvesting and service providers that harvest the metadata and offer a service to the users, for instance a central metadata catalog.

The OAI-PMH protocol requires that the metadata be offered in DC format next to any other format. This allows that metadata from different disciplines can be harvested and put into one catalog although presumably much information was lost by mapping all metadata to the DC set. When all harvested data providers have agreed to also provide metadata of another set than DC, it is of course possible to create a more useful catalog. This will be a strategy for the CLARIN/D-SPIN communities.

The strategy of gathering all metadata into a single catalog can be augmented by having the service providers transform the offered metadata into a standard set other than DC such as for example compliant with the modular CLARIN component set. The service provider would need to know different mappings for all the different offered sets.

In all the above scenarios the catalog is located at one site and metadata searching is done within that catalog. However, different sites can create portals and act as metadata service providers. In fact we assume that a whole new business will emerge, since many researchers want to be guided. A different scenario is that of "federated search", where a special web portal offers a web interface for metadata search and has knowledge about all relevant repositories that offer a search service. The portal also knows how to access the search service and how to formulate queries and how to map these queries to the specific metadata sets used at the different repositories. The portal then converts a query by the user into the appropriate format for every relevant search service and propagates it.

The portal is also responsible to solve the ranking problem by merging the results of the different service providers in a meaningful way, which can be difficult. For metadata processing such as search in general federated search is not seen as an efficient solution and due to the openness of metadata there is no IPR issue requiring federation.

Recommendation 3

Providers of metadata for LRT resources are recommended to prepare and store these metadata on their servers so that they can be harvested and catalogued by the emerging CLARIN and D-SPIN service providers in the supported formats. Advice on how to prepare and store the metadata will be given in separate guidelines and in tutorial events to which the service providers will be invited. Metadata needs to be open to promote the development of suitable portals.

5 CLARIN/D-SPIN Conformant Encoding and Annotation Standards

Following a CLARIN-supported standard in the encoding and/or annotation of a language resource has the following advantages:

1. Data elements of various resources are compatible with one another. For example, the headwords of several lexical resources can be extracted and merged into one list if the notion of "headword" is well defined and according to a standard. The "Lexical Markup Framework" (LMF) is a good point of reference. It is also recommendable to stick to ISOcat categories whenever possible. Note that sticking to standards and well-defined categories does not solve all compatibility problems.
2. Resources can more easily interact with processing tools if the tools stick to supported standards in their input and output formats.

These features of tools and resources are aspects of what is called interoperability. There is a CLARIN working group (Working Group 5.6) that takes care of all aspects of interoperability between language resources and tools. We recommend that you as a provider of language resources and tools join this working group via the CLARIN website (http://www.clarin.eu).

In the following, we will describe some of the relevant standards and specifications that have already been established as (ISO) standards or are in the process of becoming standards. We will point to the authoritative documents that describe this standard. We will summarize the current position of CLARIN with respect to these standards and give recommendations to those who consider using these standards.

5.1 UNICODE (also: ISO 10646)

Status: Formally a version of the Unicode Standard is defined by an edition of the book *The Unicode Standard*, together with the online Unicode Standard Annexes and the Unicode Character Database, which updates and extends the book's normative specifications and informative content. The current version is 5.1.0.

Canonical document/website: http://www.unicode.org

UNICODE and its derivatives UTF-8 and UTF-16 are mature and widely applied standards for the encoding of characters. It is highly recommended to encode textual resources in one of the character encoding schemes that are derived from the UNICODE character set. The encoding should be mentioned explicitly in the metadata of a resource. Tools should be enabled to process texts with UNICODE encoding. Where this is not possible with acceptable effort, the supported character encoding schemes for input/output should be made explicit in the metadata and/or documentation of the tool.

5.2 Extensible Markup Language (XML)

Status: Version 1.1 is a recommendation of W3C

Canonical document/website: http://www.w3.org/XML/ and http://www.w3.org/TR/xml11/

Since its publication by the W3C in 1998, the XML recommendation has become one of the most widely disseminated specifications for representing semi-structured information. Its wide distribution and use has lead to the availability of a large range of tools, commercial as well as open source and freely available ones, and of accompanying recommendations for the manipulation of XML documents (e.g. XML Stylesheet Language Transformation, XSLT) or for their embedding in distributed applications (e.g. SOAP). Still, being a meta-language allowing one to define specific document models (by means of DTDs, RelaxNG schemata or W3C schemata), widely agreed generic models for some of the major linguistic resource types are still a desideratum.

CLARIN and D-SPIN fully endorse XML as the reference syntax for any representation, exchange or archival of linguistic information. The projects will support activities to come to generic models expressed as XML schemata for the major linguistic resource types. This is however not a statement about internal processing formats.

5.3 Data Category Registry (DCR)

Status: Standard - ISO 12620

Canonical document/website: http://www.isocat.org/

The ISO DCR is based on 12620, which is compliant with ISO 11179, a big initiative that crosses the borders of academic disciplines. Currently, categories resulting from decades of linguistic discussion (EAGLES, ISLE/MILE, IMDI) are entered into the implementation of ISO DCR called ISOcat. It can be expected that many researchers will not accept its category definitions. Two ways are suggested to make progress regardless of this: (a) Sub-communities are enabled to add their categories into a separate profile in ISOcat and it is the task of the researchers to establish relations between the different categories where this is possible. (b) These sub-committees can also add entries to the user space in ISOcat or create their own instance and register it. Then it is a matter of trust of other researchers in the persistence of the registry and in the stability of the definitions whether they want to use them. Again it will be required that these relations are interoperable with the ones in ISOcat. It is obvious though that in some/many cases a mapping between categories will not be possible.

The ISO DCR is a cornerstone of the standardization and interoperability framework. ISOcat will become available this summer. It is the only suggestion for achieving semantic interoperability at the level of linguistic categories, that linguist from all over the world agreed upon. CLARIN and D-SPIN will nevertheless promote the work with ISO DCR and ISOcat. There is no practicable alternative and we need to go a step ahead in standardization. We therefore recommend everyone to make themselves familiar with the ISO DCR and ISOcat (see the link above) and to make suggestions to their improvement.

5.4 Text Encoding Initiative (TEI) – Guidelines for Text Encoding

The Text Encoding Initiative (TEI) is a consortium that collectively develops and maintains a standard for the representation of texts in digital form.

Status: The chief deliverable of TEI is a set of guidelines that specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. The latest version of these guidelines is called “P5”. There are XML schemata for document types or “tag-sets”. The

TEI also provides a tool called "ROMA" (cf. http://www.tei-c.org/Roma/) which allows users of the TEI Guidelines to generate a schema accommodating their needs.

Canonical document/website: http://www.tei-c.org/ and the Guidelines, version 5 at: http://www.tei-c.org/Guidelines/P5/

TEI will be supported by CLARIN. It is however, necessary to provide a schema as an accompanying description of a language resource. We recommend sticking as close as possible to the tags and tag sets which are specified in the guidelines. The semantics of the tags used should be made explicit in order to enable interoperability between resources. Ideally, the categories should be registered in the Date Category Registry to make their semantics explicit.

5.5 Corpus Encoding Standard ((X)CES)

Status: XCES instantiates the EAGLES Corpus Encoding Standard (CES) DTDs for linguistic corpora, developed by the Department of Computer Science, Vassar College, and Equipe Langue et Dialogue, LORIA/CNRS. XCES is the XML version of this "standard". Indeed, it is not a real standard since it is not endorsed by any standardization organization, but it is a specification that is widely used in the community. There are ongoing activities to make XCES compliant with TEI P5. The current version of XCES is 1.0.4.

Canonical document/website: http://www.xces.org/

As a widely used specification, XCES will be supported by CLARIN/D-SPIN.

5.6 Lexical Markup Framework (LMF)

Status: LMF is the ISO standard for Natural Language Processing (NLP) lexicons and Machine Readable Dictionaries (MRD). The ISO code number for LMF is ISO-24613: 2008.

Canonical document/website: http://www.lexicalmarkupframework.org/, the LMF specification with examples. The latest version is available through this website. The current version is version 16.

The development of the Lexical Markup Framework responds to the fact that lexicon developers all come up with different structures and lexical attributes. LMF can be seen as a flexible framework that allows lexicon builders to create lexica of different complexity where the individual attributes are linked to a registered reference category.

LMF has been widely standardized and first tools are supporting this standard. CLARIN and D-SPIN should promote the usage of LMF, its thorough testing and if required its further standardization process. It could play a role as pivot model for lexicon interoperability, i.e. existing converters should be made available as reusable services. Note that the primary goal of LMF is to cover machine readable and computational lexical resources. Digital versions of human-readable dictionaries are covered by the TEI tag set for dictionaries (cf. chapter 12 of the TEI "Guidelines for Text Encoding") and by ISO Standard 1951 – Presentati*on/representation of entries in dictionaries – Requirements, recommendations and information.*

Still, LMF is fairly new. No sufficiently realistic tests have been carried out to make it a well-proven standard. However, its existence can be used to push forward all aspects that have to do with format interoperability for lexica. Some linguists say that LMF is too little constrained, i.e. that researchers could create any structures. ISO addressed this issue and created reference structures for NLP types

of lexica. It will take a while until there will be such reference1) structures for other subdomains. We highly recommend the tutorial style introduction into issues of converting existing formats for lexical resources into LMF by Gil Francopoulo ("Extended examples of lexicons using LMF", cf. http://lirics.loria.fr/doc_pub/ExtendedExamplesOfLexiconsUsingLMF29August05.pdf).

5.7 Linguistic Annotation Framework (LAF)

Within ISO committee TC 37/SC 4 a set of new more generic standards (i.e. Frameworks) are being worked out. It seems that this area is still much under development. We document the current state of this work shortly in this and the following sections.

Status: ISO proposal for standardization (ISO/DIS 24612)

Canonical document/website: http://www.tc37sc4.org/new_doc/

ISO_TC_37-4_N076_Proposal_of_NP_linguistic_annotation_fram_A1_A6.pdf

LAF provides a generic framework for representing annotated resources as graphs and nodes and links associated to feature structures (conformant to IO 24610). It is particularly useful when integrating heterogeneous resources within one single repository. Moreover, LAF ensures a coherence scheme across all other ISO/TC 37/SC4 projects. The specifications are at a very abstract level, so that LAF can only be seen as a set of very basic and general guidelines.

5.8 Morpho-syntactic Annotation Framework (MAF)

Status: Draft International Standard ISO/DIS 24611

Canonical document/website: http://atoll.inria.fr/~clerger/MAF/html/index.html;

http://www.tc37sc4.org/new_doc/ISO_TC_37-4_N225_CD_MAF.pdf

MAF offers a model as well as a format for the representation of morpho-syntactic annotation on a two-tier principle (token – word form). It provides means of representing complex annotation cases (ambiguities, multiple segmentations) as a well as a tag-set definition framework based on feature structure libraries. However, MAF does not standardize any specific tagsets, leaving this to specific projects.

5.9 Syntactic Annotation Framework (SynAF)

Status: SynAF has already been accepted at the ISO Level as a Draft International Standard (DIS 24615).

Canonical document/website: http://www.tc37sc4.org/document.php?p=tc37sc4_list_total.txt&search_text=SynAF&project_category=on

This standard provides a reference format for the representation of syntactic annotations. It provides a generic model for representing both constituent and dependency based syntactic annotation and has been inspired by initiatives like Tiger. The document is still a draft and cannot be applied in its current state.

5.10 Semantic Annotation Framework (SemAF)

Status: Draft International Standards (DIS 24617-1, Part 1)

Canonical document/website: http://www.tc37sc4.org/new_doc/

ISO_TC37_SC4_N269_rev06_WG2_WD_24617-1_%20SemAF-Time.pdf

This proposed international standard provides a formal specification language called ISO-TimeML for temporal information markup and includes a specific set of guidelines for such a markup. Adopting XML as its formal language, the SemAF/Time annotation standard provides a formalized markup language called ISO-TimeML with a systematic way to extract and to represent temporal information, as well as to facilitate the exchange of temporal information, both between operational language processing systems and between different temporal representation schemes.

Recommendations 4

If you want to integrate an existing resource into the CLARIN/D-SPIN infrastructure, there is an ad-hoc repository on the CLARIN website where you can register and describe your resource(s) and tools (http://www.clarin.eu/summary-wp5-language-resources-and-technology-overview). Currently, there is only a minimal set of descriptive categories. In the future you will be able to provide more information, including a link to a schema or a document grammar that describes the encoding and the logical structure or your resource. For tools, you will be able to specify the input requirements and output format(s) of the processed or, in the case of lexical lookup, the consulted data. You will be able to provide a specimen of your resource that helps to understand the structure of it. We will keep you informed about the further development of this ad-hoc repository. If your resource(s) is/are in a format that is currently not supported by CLARIN, you should consult the leader of WP5 of the D-SPIN project and discuss the possibilities of converting your resource into a CLARIN-supported format.

If you want to build a new language resource, you should make yourself familiar with one or more of the above mentioned standards and specifications and check whether they are adequate for your resource. Most of the "frameworks" are either rather abstract and need concrete instantiations or they are still lacking sufficient testing. Your assistance in applying and testing these standards on your resource(s) is very welcome. If you want to (a) build a corpus-like resource, you should have a look at XCES or TEI; (b) if you want to build a linguistically annotated corpus, you should have a look at the Linguistic Annotation Framework (LAF) or at XCES; (c) if you want to build a lexical resource, you should have a look at the Lexical Markup Framework. In either case, you should check whether the data categories you are going to use are registered in the ISO Data Category Registry.

6 CLARIN/D-SPIN Conformant Tools and (Web) Services

In D-SPIN, a group of developers has started (a) to provide language technology tools as web services and (b) to combine these services into processing chains or workflows. One aim of this exercise is to prove that tools and services can be combined into processing chains that implement complex operations if some requirements are met. Most of the requirements are of an architectural nature. They are described in detail in the paper by Helmut Schmid (cf. section 1 of this document).

In the following, we distinguish between two types of language resource related web services:

1. Tools and services that extract bits of information from e.g. textual or lexical resources and present them to the user (a human or a piece of software). We call this type of service a data service.
2. Tools and services that enrich a resource, e.g. a text, by adding information to it. We call this type of service an annotation service.

Only services of the second type have been in the focus of the D-SPIN project; services of the first type will be implemented in the near future.

You can join this effort with your tools and resources. The framework is open to your participation; however, a few requirements have to be met:

- Your resources must be open and freely available. No efforts are currently made to restrict access to the resources that are integrated into the framework. Legal issues of this kind are subject to further decisions in D-SPIN.
- Your resources have to match the input specification(s) of the tools that extract or add information. Currently, only the input format for text is specified, further specifications will follow. There is also a chance to develop a converter from the format of your resource to the current D-SPIN format if the format of your resource sticks to one of the widely used standard formats which have been described above.
- Similarly, the input/output specification of your tool(s) must fit the specifications of the other tools and, in general, be able to cope with the D-SPIN data formats. Again, there is a chance to develop wrappers for your tools to make them compatible with the D-SPIN web services infrastructure.

There are still a number of issues that need more discussion, among them are:

- It has been analyzed which metadata need to be used for web services and the current CLARIN metadata category set covers the major elements. Indeed at an expert meeting in Tübingen the issue of profile matching was discussed, i.e. can we compare the resource and service metadata by automatic means to find out whether a service/tool can be applied to the given resource? Yet we need more practical tests to see where we need to amend the set.
- Preliminary results of processing chains need to be stored in workspaces in order to be available for further processing or carry out certain steps again with different parameter sets. They need to be treated the same way as data that is valuable for other users and that the creators want to share (e.g. by metadata descriptions, persistent identifiers etc.). Only then it

will be a simple operation for the creator to check it in a visible repository. However, also in this respect we need more experience and interaction with the grid community for example that tackled these issues already.

- Metadata, provenance data and persistent identifiers will be generated by standardized CLARIN wrappers that will be provided by CLARIN WP2 and assigned to the results of a processing chain. Despite that the requirements specifications have been formulated, we need practical experience to test all aspects including the workspace issue.
- Currently it is assumed that by using the service bus construction and including the mentioned wrapper functionality at a central place we will come to a manageable infrastructure for web services. This idea can be combined with the workspaces idea, i.e. to have a number of centres that act as execution spaces for web services. These centres could be chosen selectively to also maintain the service bus architecture, i.e. install new versions of the wrapper functionality etc.
- Yet there are not fully satisfying workflows tools that meet all requirements of the LRT world. We need to study in more detail which functionality we can get by professional tools such as described in the CLARIN requirements document and start negotiating with experts about extensions. We cannot expect CLARIN to build a full-fledged workflow system from scratch of course.

Some of these and other, related questions have been addressed by CLARIN Work Package 2. The recommendations of this Work package will be used as guidelines for the further specifications in D-SPIN.

Recommendation 5

It is the right time to join the Web services implementation group if you are (a) an implementer or (b) a tool developer who wants his tool(s) to be integrated into a larger structure or (c) a resource provider who wants to explore the potential of combining resources with tools. Your participation is welcome!

7 Outlook

Several activities have already been planned for the second year of D-SPIN, that directly concern the issues of LRT standards:

1. D-SPIN members will participate and contribute to a special session on LRT standards during the CLARIN consortium meeting in Barcelona (May 11th – 13th 2009).

2. D-SPIN/CLARIN workshop on standards for finite state technologies will be hosted by the University of Stuttgart and by the University of Tübingen in Freudenstadt (June 4th – 6th 2009).

3. A special meeting concerning LRT standards will be jointly organized by D-SPIN and CLARIN at the BBAW Berlin (July 11th 2009).