Standardisation Action Plan for Clarin

State: Proposal to CLARIN community

Nuria Bel, Jonas Beskow, Lou Boves, Gerhard Budin, Nicoletta Calzolari, Khalid Choukri, Erhard Hinrichs, Steven Krauwer, Lothar Lemnitzer, Stelios Piperidis, Adam Przepiorkowski, Laurent Romary, Florian Schiel, Helmut Schmidt, Hans Uszkoreit, Peter Wittenburg

August 2009

Summary

This document describes a proposal for a Standardisation Action Plan (SAP) for the Clarin initiative in close synchronization with other relevant initiatives such as Flarenet, ELRA, ISO and TEI. While Flarenet is oriented towards a broader scope since it is also addressing standards that are typically used in industry, CLARIN wants to be more focussed in its statements to the research domain. Due to the overlap it is agreed that the Flarenet and CLARIN documents on standards need to be closely synchronized. This note covers standards that are generic (XML, UNICODE) as well as standards that are domain specific where naturally the LRT community has much more influence.

This Standardization Action Plan wants to give an orientation for all practical work in CLARIN to achieve a harmonized domain of language resources and technology stepwise and therefore its core message is to overcome fragmentation. To meet these goals it wants to keep its message as simple as possible. A web-site will be established that will contain more information about examples, guidelines, explanations, tools, converters and training events such as summer schools.

The organization of the document is as follows:

- Chapter 1: Introduction to the topic.
- Chapter 2: Recommended standards that CLARIN should endorse page 4
- Chapter 3: Standards that are emerging and relevant for CLARIN page 8
- Chapter 4: General guidelines that need to be followed page 12
- Chapter 5: Reference to community practices page 14
- Chapter 6: References

This document tries to be short and will give comments, recommendations and discuss open issues for each of the standards.

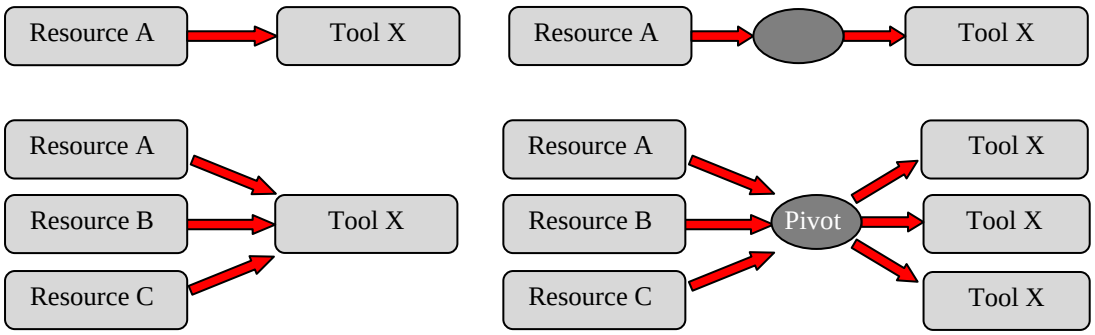
User Guidance

- If you want to get clear recommendations of what is agreed upon in CLARIN, just got to chapters 2 for concrete standards and 4 for general guidelines.
- If you want to know which standards are currently being worked on and what CLARIN should test go to chapter 3.

It should be noted that the standardization process is open to experts, i.e. they are welcome to participate in the ISO discussions at national and international level.

1. Introduction

The major objective of the CLARIN research infrastructure initiative is to create an integrated and interoperable domain of language resources and tools. While for the integration part CLARIN has widely sorted out the baselines (single sign-on, persistent identifiers, metadata, etc.) and is now working on their implementation, the interoperability part is as difficult as expected. It is obvious that this is a good moment to intensify the discussions and to synchronize the opinions even beyond the boundaries of CLARIN.



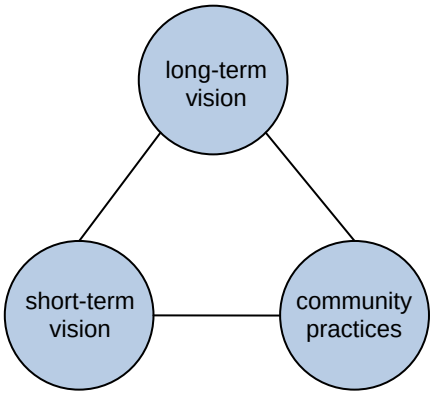
When CLARIN wants to enable humanities researchers to combine language resources and tools into new collections and workflows it is a prerequisite obstacles on the way to full interoperability are removed stepwise. In general we can describe the problem as an import/export problem as indicated in the two figures above. In the first example (upper two drawings), we have the simple case that a conversion step may solve the problem that the expected format and encoding does not fit with the one offered by the resource. The second example indicates the complexity of the problem when several resources and tools are involved, a situation which we envisage to occur quite often in CLARIN. Theoretically, it is obvious that ideally the introduction of a widely agreed pivot format and encoding convention will be optimal. The introduction of such formats and conventions needs to be based on a standardisation process to involve many people and to avoid to loss of investments in time and effort.

Before discussing standards and their potential relevance we want to make a few clarifications of issues which frequently led to misinterpretations:

- The problem of interoperability only emerges when linguists are ready to offer their resources and tools to other researchers. As long as they produce their resources and tools for individual usage interoperability and therefore the need to adhere to standards or best practices is of little relevance.
- No linguist should be required to read long documents about standards; it is primarily the task of the tool, service and converter developers to provide frameworks that help the researcher and that hide complex formalisms as much as possible.
- There are established communities that use certain formats and encoding conventions - no one is arguing that these established procedures need to be changed in the near future.
- There is no need to have one single format or encoding convention for a certain level of linguistic description, in most cases we will expect a few of them. The point of concern is that the investments for maintaining resources and making them interoperable in an open CLARIN scenario need to be limited.
- Of course we can expect that increasingly often tool builders will adapt to standards when they are available and show a chance of broad acceptance. Again users should not be affected in their productivity.
- Standards for interoperability need to be viewed under pragmatic aspects. In the above mentioned case the issue is to solve cross-resource and technology problems, but not to re-invent linguistic theory. In some cases of transformation we will not be able to solve this without losing essential information. In other cases we will be able

to create abstractions that allow us to more easily map between a variety of descriptive systems.

- Members of the LRT community assume different roles: (1) they are researchers and in this role they do not like to be bound to strict standards and (2) increasingly many act as service providers (language documentation, NLP, lexica, etc.), i.e. they create data and tools that are useful for others - often without accepting this role as service provider explicitly.



Therefore CLARIN needs to find a proper balance between three major aspects:

- Standardization and harmonization as a long-term vision for cost reduction when creating interoperability.
- Short-term needs with sufficient coverage to solve the short-term goals when testing and implementing interoperability at this moment.
- No productivity decrease for sub-communities and little affect on linguists who do not want to act as "service providers".

For the linguistic community standards are fairly new and there is still a reluctance to speak about standardization, since it may hamper scientific progress and flexibility. However, we are all used to de-facto or proprietary standards such as the MS Word formats and we have seen that it partially allows us to exchange documents easily. Finally everyone will benefit from the introduction of standards in areas that are mature enough.

CLARIN WP5 will set up a web-page where

- examples, use cases and motivations will be given for the various standards recommended
- tools will be mentioned that support the standards
- people can comment on standards and in doing so influence further developments.

2. Recommended standards

In this chapter those standards are listed that CLARIN members should endorse, disseminate and implement.

Recommendation: CLARIN will establish a list of standardisation experts that can act as liaison to provide information as well as feedback to the relevant standardisation bodies (ISO, TEI, W3C). For ISOcat profiles the appropriate boards have been determined including many experts from CLARIN.

2.1 Unicode - ISO 10646

Comments: ISO 10646 and its industry counterpart UNICODE are now widely agreed, in particular in the form of the UTF-8 encoding scheme. It is now supported by all relevant software vendors.

Recommendations: CLARIN community will apply ISO 10646/UNICODE in all resources and tools.

Open Issues: There are still characters out there which linguists are confronted with that have not yet been integrated in UNICODE such as Cuneiform characters and where special arrangements are required. Also there are known problems of different operating systems to handle the complete UNICODE set correctly. However, increasingly more characters are captured. The linguistic community is represented in the UNICODE boards. CLARIN would communicate with the UNICODE boards if this is required by its members.

2.2 Country codes - ISO 3166

Comments: ISO 3166 provides 2 and 3 letter country codes and is related to a maintenance agency since 1974. It is widely disseminated across all types of IT applications.

Recommendations: CLARIN community will apply ISO 3166 in all resources and tools.

Open Issues: Hardly any. One has to be careful with the code for United Kingdom ('GB')

2.3 Language codes - ISO 639-1/2/3

Comments: There is a history of language codes starting with the 2-letter code 639-1, the 3 letter code 639-2 and recently the Ethnologue code set has been adopted by ISO as 639-3, since it covers about 6000 languages. Thus the latter somehow covers the variety that linguists need. Yet there are many languages not covered and many of the definitions of Ethnologue are heavily debated. A new set of standards is in preparation, where 639-4 defines the principles for language naming, 639-5 the language families and 639-6 finally will extend to dialects. A harmonization between all sub-standards is being worked on currently.

Recommendations: CLARIN will adopt ISO 639-3 as basis for all its resources and tools and collaborate with ISO TC 37/SC 2 to work on the new family of standards 639-4/5/6. For the missing languages and dialects it will offer a registry to make the codes re-usable.

Open Issues: With respect to language codes a few issues can be identified: (1) The standards do not cover all languages and in particular dialects, i.e. there must be an extension mechanism and a registry for these extensions. (2) Language names are a matter of political debate for various reasons, therefore only a broad process organized by ISO and relying on the community of experts will lead to a widely accepted standard. (3) The language family codes in the new standard proposal cannot be seen as stable. CLARIN should play an active role as mediator between informed communities and push forward the new family 639-4/5/6.

2.4 Codes for the representation of names of scripts - ISO 15924

Comments: ISO 15924 provides codes for the representation of scripts for written languages. Like the 639 series, it is maintained by a Registration Authority (the Unicode consortium) and

is thus updated on a regular basis. The current set of codes is also freely accessible from the Unicode web site[1].

Recommendations: CLARIN community will apply ISO 15924 when needed in all resources and tools.

Open Issues: Hardly any. Missing scripts have to be reported to the registration authority.

2.5 XML

Comments: Since its publication by the W3C in 1998, the XML recommendation has become one of the most widely disseminated syntax for representing semi-structured information. Its fame has lead to the availability of a large range of tools and accompanying recommendation for the manipulation of XML documents (e.g. XSLT) or their embedding in distributed applications (e.g. SOAP).

Recommendations: CLARIN fully endorses XML as the reference syntax for any representation, exchange or archival of linguistic information. It will support activities to come to generic schemas for the major linguistic resource types and to define a strategy for providing better semantic interoperability. This does not make statements about internal processing formats, which could make use for example of relational databases for fast operations.

Open Issues: Being a meta-language allowing one to define specific document models (by means of DTDs, RelaxNG schemas or W3C schemas), we lack widely agreed generic models for some of the major linguistic resource types and it does not provide means to control the semantics of XML components. CLARIN will set up a web-page where it will be explained to users in simple terms why it makes sense to represent their textual data in XML.

2.6 ISO 24610-1:2006 Language resource management -- Feature structures -- Part 1: Feature structure representation (FSR)

Comments: The FSR standard has been established jointly between ISO and the TEI to provide a reference XML vocabulary for the representation of feature structures. It can be embedded as a module in other applications and covers a wide range of functionalities.

Recommendations: CLARIN community will apply ISO 24610 in all resources and tools, whenever feature structures are embedded in other formats.

Open Issues: Work is ongoing to have the feature structure description module also adopted by ISO.

2.7 TEI for the representation of primary sources

Comments: The TEI guidelines provide a modular framework and vocabularies to represent textual content across a variety of possible genres (prose, drama, dictionaries, transcription of speech, etc.).

Recommendations: CLARIN will recommend that all source documents that require more than plain text format (e.g. representation of division and paragraph level) will use an agreed upon minimal subset of the TEI guidelines where suitable.

Open Issues: TEI offers very flexible mechanisms which in practise leads to the situation that there is a large variety of simplified subsets. TEI will adapt so that the vocabulary can be re-used in various frameworks for semantic interoperability reasons. TEI offers tools such as ODD, ROMA to create customizations. CLARIN needs to set up a practical guide to TEI with simple examples to guide the non-professional user.

2.8 Knowledge Engineering

Comments: In the area of knowledge engineering quite a number of frameworks have been defined in particular by W3C such as RDF (Resource Description Framework), RDF-S

[1] http://unicode.org/iso15924/codelists.html

(Schema extension), SKOS (Simple knowledge Organization System) and OWL (Web ontology language coming along in four different flavours). They are all based on XML syntax and address certain needs to deal with concepts and relations between them. RDF is a simple schema that allows users to define their concepts and the relation between them in term of triples. RDF-S is a first simple extension to RDF to allow users to specify a domain vocabulary and their ontological relations. SKOS is a framework with simplified logic that allows users to represent for example hierarchical concept systems such as thesauri. OWL builds on RDF and RDF-S and adds more vocabulary for describing more complex ontologies. Due to its inherent complexity it comes with different flavours that address different needs.

Recommendations: CLARIN recommends to make use of the W3C standards wherever knowledge needs to be represented in flexible formats. Various frameworks recommended in CLARIN should provide an export into these formats making use of RDF and OWL.

Open Issues: The structure of complex data types with implicit relation types such as lexica can be defined by an XML schema or as a set of RDF triples where structure is flattened, but relations are made explicit. Dependent on the intentions and the nature of the processing steps involved the user may want to chose the one or the other representation. When automatic reasoning is intended making all relations explicit has advantages. For other types of operations the compact representation as complex structure has advantages, but the tools need to know how to interpret the elements. The semantic web community has widely agreed to use the W3C recommendations, i.e. interoperability requires their usage.

2.9 Audio/Speech Standards

Comments: The best way to digitize sound waves is to use a direct digital representation of the analogue waveform which is called linear PCM (Pulse Code Modulation), however increasingly often sound material is born digital already. Consumer products come with small recorders that do compression such as MP3 and ATRAC (MiniDisk) which carry out a reduction of components our human perception is not aware of as is said. Since these compression schemes are lossy and since we cannot know where the sound recordings will be used for in future it is strongly recommended to use linear PCM techniques.

In certain research areas phonetic transcriptions are required for further speech processing - here the Alphabet of the International Phonetic Association (IPA) is used. A frequently used scheme is to use SAM-PA and X-SAM-PA for this purpose which specifies IPA characters in terms of ASCII characters. With the appearance and wide support of UNICODE this has become widely obsolete.

Recommendations: For audio recordings CLARIN recommends to make recordings in the best possible quality and not use compressed formats. In general linear PCM with 44/48 kHz sample frequency and 16 bit resolution will be sufficient to represent speech. For specific type of purposes 96 kHz and 24 bit resolution would be better due to its better time resolution and its higher dynamic range.

For representing phonemes the international practice is to use the IPA Alphabet which is included in the UNICODE standard.

Open Issues: Some phonetic researchers want to describe special phonetic characteristics not yet included in IPA. However, here we suggest to try to convince first the IPA board which will contact the UNICODE board. It is known that SAM-PA and X-SAM-PA specifications are not without errors. For other areas such as intonation annotation certain schemes are used, but they are heavily debated, so that CLARIN cannot make recommendations.

2.10 Video/Multimodality Standards

Comments: Video digitization is a highly dynamic field because on the one hand the interest in higher resolution schemes is obvious and on the other hand the data rates need to be kept manageable, i.e. heavy compressions is applied. Currently H.264 based variants are replacing old codes for representing video in consumer electronics and for web streaming due to their improved quality/data-rate ratio compared to MPEG1 and MPEG2. In general video data is born digital and compressed. For archiving purposes the motion film industry has

decided to go with MJPEG2000 lossless compression which is defined for various resolution schemes. But the amount of data cannot be dealt with in normal applications, i.e. as working format codecs such as H.264 or lossy MJPEG2000 will be chosen.

Multimodality analysis is applied to a wide range of different modalities such as eye tracking, gesture, hand motion, body motion, facial expressions, haptics etc. For most of these channels there are no standardized or widely agreed encoding systems. For some as for example facial expressions, hand shapes etc there are suggestions that are widely used. CLARIN cannot make recommendations at this moment.

Recommendations: For video recordings CLARIN recommends to use MJPEG2000 lossless as backend format, although most data is already generated in compressed form. For handling and processing video data in general MPEG2 or even better H.264 (included in MPEG4 in general) are recommended.

Open Issues: The usage of standards in the video area is widely dependent on the available equipment and software. Only now lossless schemes such as MJPEG2000 seem to be manageable for archiving purpose. For low price recordings and for the daily work codecs such as H.264 will be used. There is software to convert formats, but users need to be aware of concatenation effects which may appear when applying series of transformations. Highly compressing codecs apply heavy reductions, i.e. it will depend on the intentions which technique will be applied.

3. Ongoing standardisation projects

In addition to the strong recommendations expressed above, CLARIN needs to actively track a number of ongoing standardisation activities at the two major levels: (1) linguistic structures/formats and (2) linguistic encoding. CLARIN as an infrastructure project has the duty to evaluate, test and comment these proposals in close relation with the relevant standardisation bodies. When necessary, CLARIN may take the lead in initiating new standardisation activities when a clear gap in coverage is identified.

3.1 Standards for Semantic Interoperability

Currently the work on the ISO data category registry - its model and its implementation - are at the core of all standardization efforts in ISO. The categories to be included are those that result from years of discussion in the discipline, from widely used practices and from new initiatives such as for example TimeML.

ISO DCR and ISOcat

Comments: The ISO DCR is based on 12620 which in itself is compliant with ISO 11179 which is a big initiative crossing multiple disciplines. Currently, categories resulting from decades of linguistic discussion (EAGLES, ISLE/MILE, IMDI) are entered into the implementation of ISO DCR called ISOcat. Of course, we can assume that many sub-communities will not use these category definitions for their daily work. Two ways are suggested to make progress nevertheless: (1) 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 semantically possible. (2) They 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 the stability of the definitions whether they want to use them. Again to achieve interoperability relations to the ones in ISOcat would be required. It is obvious that in some/many cases a mapping between categories will not be possible.

Recommendations: The ISO DCR is a fact and ISOcat will become available this summer. It is the only suggestion for achieving semantic interoperability at the level of linguistic categories where linguists from all over the world agreed upon. Yet not all sub-communities have participated in these discussions and may have objections to participate. CLARIN should nevertheless promote the work with ISO DCR and ISOcat, since it is at the core of various standardization and harmonization activities. Currently there is no practicable alternative and we need to shift the borders that will still be there. CLARIN recommends to include other widely used tag sets to make them re-usable for others and to relate them to other categories in the DCR.

Open Issues: Of course quite a number of problems where already mentioned which need to be taken up in future steps. We just want to mention three of them: (1) The 12620 model is restricted if one compares this for example with other mechanisms such as Framenet or unrestricted RDF based suggestions. However, we also need to take care of feasibility, i.e. it must be possible in a limited amount of time to add a large number of relevant linguistic categories including its most relevant features. (2) The model does not allow to enter relations between categories. This is seen as a strength, since relations are very often dependent on the concrete usage intentions. Where relations are agreed amongst linguists or where relations are part of definitions it is suggested to define them outside of the DCR in relation registries which have not been defined yet. (3) In many languages or in certain contexts the usage of categories needs to be constrained. The DCR does not offer any means to enter them. Again it is suggested that the schemas that refer to a category include constraints, meaning that every schema instance needs to define them properly. There are other open issues such as the semantic granularity of the categories. This again is widely dependent on the application and the community will need to gather more experience to improve the representations. CLARIN needs to create a web-page with examples, guidelines, process descriptions and help facilities.

TimeML[2]

Comments: TimeML is now part of the ISO standardization effort, within TC 37/SC 4, and enlarges the representational capabilities of the original TimeML scheme by offering a metamodel and a formal semantics associated with the scheme. ISO –TimeML offers a format for the annotation of temporal entities, namely: temporal expressions, eventualities (both events and states), signals, such as temporal prepositions and conjuncts, and, finally, a set of relations between these entities, namely temporal relations, aspectual or phrasal relations and subordinating relations which should facilitate the development of reasoning algorithms. TimeML is designed to address four problems in event and temporal expression markup: (i) time stamping of events (identifying an event and anchoring it in time); (ii) ordering events with respect to one another (lexical vs. discourse ordering); (iii) reasoning with contextually underspecified temporal expressions (temporal expressions such as 'last week' and 'two weeks before'); (iv) reasoning about events. TimeML tags have improved the representational capabilities of previous annotations schemes for event annotation and temporal expressions (e.g. TIDES TIMEX2 tag).

Recommendations: CLARIN endorses ISO TimeML (ISO DIS 24617-1 Semantic Annotation Framework – Part 1: Time and events) as the pivot format for the annotation of eventualities and temporal expressions, promotes its use and the development of tools which support this format. TimeML has been integrated with OWL Time (DAML Time).

Open Issues: TimeML is now quite a stable markup language. A simplified version is currently employed for the data set of the 2010 SemEval task 13 (TempEval-2) which will provide annotated data for five languages: English, Italian, Spanish, Chinese and Korean. Annotation in other languages should be promoted and also the fixing of some minor shortcomings of the TimeML scheme, e.g. as far as the annotation of events spanning over multiple tokens (i.e. multiword expressions) is concerned.

3.2 Standards for Structural Interoperability

Within ISO committee TC 37/SC 4 a set of new more generic standards are being worked out. It seems that this area is still much under development, since again and again sub-communities are working on new proposals with certain optimization criteria in mind. These attempts increase the fragmentation which CLARIN wants to overcome.

ISO/DIS 24611 Morpho-syntactic Annotation Framework

Comments: 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. The suggestion has been worked out by looking at various examples from diverse languages. Nevertheless, more testing is required to stabilize the standard. MAF is a structural framework that needs to be filled with morpho-syntactic tags that should be taken from a recognized category registry. Well-known registries are ISOcat and TEI, although many tag sets in use are not registered yet.

Recommendations: CLARIN endorses MAF as the pivot format for the exchange of morpho-syntactic information and encourages CLARIN partners to identify possible mappings with their own formats and tools. Above all existing tag-sets should be progressively defined and disseminated according to the MAF guidelines.

Open Issues: MAF does not standardise any specific tag sets, leaving this to specific projects. But it requires to make use of registered tag sets or at least to refer to them to achieve semantic interoperability at the tag level. CLARIN should promote the adaptation of tools to support MAF.

[2] This section was contributed by Tommaso Caselli.

ISO/CD 24615 Syntactic Annotation Framework (SynAF)

Comments: SynAF provides a generic model for representing both constituent and dependency based syntactic annotation and has been inspired by initiatives like Tiger which is very close to SynAF.

Recommendations: CLARIN should identify a task force to help finalise and test the SynAF proposal, so that it could be endorsed before the end of the project.

Open Issues: The document is still a draft and cannot be applied in its current state. Various best practices such as the TIGER format, the various Treebank formats and the Prague-Dependency format should be compared with SynAF to validate its representational power.

ISO 24613:2008 Lexical Markup Framework (LMF)

Comments: The development of the Lexical Markup Framework was driven by the fact that lexicon developers all come up with different structures and their lexical attributes being embedded in various contexts. LMF can be seen as a flexible framework that allows researchers to build lexica of different complexity where the individual attributes need to point to a registered reference category. TEI describing mainly printed dictionaries can be represented in LMF indicating a certain overlap. Since LMF is a flexible framework CLARIN needs to come up with example lexica for different sub-communities. Currently, only examples for NLP lexica have been worked out.

Recommendations: LMF has been widely standardized and first tools are supporting this standard. CLARIN should promote the usage of LMF, its thorough testing and if required its further standardization process. It will play a role as pivot model for lexicon interoperability, i.e. existing converters should be made available as re-usable services.

Open Issues: LMF is fairly new and yet not enough realistic tests have been carried out to speak about 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 not strict enough, i.e. researchers could create any structures. ISO addressed this issue and created reference structures for NLP type of lexica for example. It will take a while until there will be such reference structures for other sub-domains. CLARIN should promote the creation of such reference structures that can be re-used for similar intentions. LMF has already been tested to represent Wordnets for example, although the connection between ontologies and lexica is still an issue under debate. Also CLARIN should promote the adaptation of tools and the creation of converters to this format.

3.3 Mechanisms for resource management

TEI/ODD

Comments: One of the TEI modules offers a fully fledged language for the specification and documentation of XML applications (named ODD and based on RelaxNG fragments). This format is used for the specification of the TEI itself as well as for the management of some ISO documents. From an ODD specification, one can generate HTML, MSWord or PDF documentations, as well as DTDs, RelaxNG and W3C schemas. The flexible metadata infrastructure of CLARIN will be based on XML components and the infrastructure will have an ODD generation for documentation purposes.

Recommendations: When CLARIN partners prefer to define their own formats for whatever level of annotation, CLARIN recommends that ODD be used systematically to ensure a proper documentation and dissemination of the schemas.

Open Issues: Ongoing work intends to extend the capacities of ODD to design families of related schemas. Also this framework needs to be applied independent of the TEI mechanisms to understand its representational power.

ISO/CD 24619 Persistent identification and access in language technology application (Citer)

Comments: Citer provides a core set of recommendations for the unique identification and referencing of language resources. It is based on the knowledge that references must address resources and even resource fragments in a persistent way. The Citer document follows widely the requirements that have been worked out in CLARIN.

Recommendations: CLARIN should establish a task force for evaluating the possible adoption of Citer in all its technical activities. A Citer compliant service has already been set up which is available for all CLARIN members for testing.

Open Issues: The service offered is based on the Handle System. National libraries are making use of the URN:NBN scheme, however no services are known that allow researchers to register and resolve millions of urn-based references and that can be used by researchers on the basis of a feasible cost model. ELRA makes use of the Library of France service, but registers at the catalogue level.

ISO/DIS 24612 Linguistic annotation framework (LAF)

Comments: LAF provides a generic framework for representing annotated resources as graphs and nodes and links associated to feature structures (conformant to ISO 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/SC 4 projects. While MAF, LMF etc are addressing the linguists building resources, LAF is addressing the data modelling experts.

Recommendations: Whatever formats CLARIN is applying it should devote some time to check compatibility with LAF. If problems are seen with the current specification these should be communicated to the ISO representatives within Clarin.

Open Issues: The specifications are at a very abstract level, so that LAF can only be seen as a set of very basic and general guidelines addressing specialists and not the linguist.

4. General Guidelines for CLARIN

CLARIN should follow a number of abstract principles which are partly already addressed by concrete standards mentioned beforehand. Nevertheless we should mention them explicitly, since they should be seen as general guidelines for all development and construction work.

4.1 Models, Schemas, Categories

Complex linguistic resource types such as lexica should be described as abstract models by using a proper model description notation. It may have various instantiations and there should be one in the form of a schema specifying an XML format. All linguistic elements that are used in such XML schemas should be registered in a recognized and persistent data category registry.

4.2 Data Category Registry

It is a trend across many disciplines to register domain categories/concepts in community specific, persistent and open accessible registries so that they can be re-used across applications or that they can be used for referencing purposes. CLARIN should promote the registration of all categories that have a broad usage in specific sub-disciplines, push ahead the development of typical profiles and schemas to guide the naive user and motivate tool builders to interact with the programming interface.

4.3 Atomic Objects and Standoff

Complex objects combining various basic data types such as text and images are optimal for visualization, but difficult for processing. Therefore, CLARIN supports the notion of "atomic objects" wherever possible and widely agreed. In particular this is true for metadata (in the general sense of additional data) about resources and the resources themselves. Without being complete a few examples can be mentioned:

- Keyword type of metadata (descriptive metadata) is accepted as representing possibly large resources (multimedia files or representations of large collections) in many types of operations and in contrary to many resources descriptive metadata is open. For these reasons metadata describing resources and tools need to be accessible as separate objects. This does not mean that they cannot exist as TEI headers for example within specific resources, but in such cases they need to be extracted as separately harvestable resources as well.
- (Multilevel) annotations on resources are often stored as stand-alone resources to not modify the original resource. Yet there are many tools that require inline annotations. This is not in line with the general principles on which a large resource infrastructure such as CLARIN has to build. Metadata is the glue that provides the relations between these closely related objects.
- In workflows resources are the result of a number of sub-sequent processing steps and of course a provenance file is required to inform users and machines how a certain resource has been created. This is yet another type of information that should adhere to a certain schema. Therefore it needs to be separated from other information.

4.4 Persistent Identifiers (PID)

CLARIN should require for all its resources and services that they are associated with persistent und unique identifiers that can be resolved to valid paths. Resource and service providers should register their components and enter the corresponding PIDs in the metadata records. Currently there are a few widely used schemas for PIDs such as URNs, Handles, DOIs[3], ARKs etc. Yet there are only very few services that offer an open registration and a robust and reliable resolving mechanism. CLARIN will not require the usage of a specific system, but it requires a resource/service registration, a performant and available resolution to a valid path and a persistency of the PID.

[3] Actually DOIs are Handles, but the IDF issuing DOIs combines DOIs with a specific business model.

4.5 Component based Metadata

Although not all ingredients of the component metadata model have been worked out, CLARIN needs to adopt the usage of this component model as defined by its requirement specification document to describe its resources and services/tools. On purpose it makes use of accepted categories as registered by Dublin Core, IMDI, OLAC, ELRA and TEI. A short-term solution based on IMDI and OLAC has been defined to allow us to start harvesting and using existing portals right now. CLARIN WP2 needs to take care that this data will be transformed to the component model.

4.6 Web Services and Resources

From careful investigations of a variety of metadata sets including UDDI and ebXML and the analysis of the future workflow chains scenario it became obvious that data resources and services/tools need to be described with the same type of semantics. The reasons are that the primary users will be the same, i.e. for users it will be easier to search for suitable objects and that it will be easier to apply advanced techniques such as automatic profile matching. As an example we can imagine the situation that a user has selected a certain (set of) resource(s) and knows the kind of function he wants to execute. By just specifying the resources and the function the workflow tool should check by profile matching (comparing the metadata descriptions of resources and services/tools) which would best help him to solve his task. A in-depth discussion about descriptive categories has been carried out that would be suitable for this kind of automatic metadata processing.

It would be up to the service providers maintaining a certain portal to offer different interfaces for resources and services/tools. In the metadata registry based on the component model there is no need to separate the descriptions, i.e. they could exist in the same metadata repository.

4.7 Internal Use vs. Exchange

For various reasons repositories and tools use optimized internal formats. A repository could for example decide to store all language resources in one big relational database. CLARIN will make statements about such decisions as long as basic requirements such as availability of proper harvestable metadata, association with a PID at "object" level etc are met, but only give advice. In the same way a tool may chose to gather a set of XML resources, convert them into an optimal search index with whatever technology as long as the objects themselves are still available in a format compliant to a schema for example. This all is internal use or optimization that may be implemented by not making use of standards.

Whenever data exchange or data preservation is required standards become essential, i.e. all data providers need to have a clear strategy of how to generate standard compliant formats if they internally use proprietary mechanisms. Many people store metadata for example in large index files without considering the relevance for long term preservation for example. It is important also for this data that it is at least exported regularly into an XML format.

4.8 Data Preservation

Long-term data preservation is one of the great challenges research is faced with. with respect to data we can distinguish between pure bit-stream preservation and maintaining interpretability of data. With respect to bit-stream preservation services will be given by some CLARIN centres and CLARIN made already statements about associating persistent identifiers to resources that are amended by MD5 checksums for example. These will allow repositories to assess authenticity of data. With respect to interpretability the best we can do is to convince researchers and repositories to rely on data resources that are in standardized formats, since there we can expect that the formats will have a long lifetime and that there will be converters when new standards will come up. It needs to be the community that will test the quality of transformations since often they are associated with transformation errors or reductions. It is important that the type of transformation is being described in the metadata descriptions.

Therefore, CLARIN will push forward standards as described in this document and support all initiatives that will provide a more systematic approach to bit-stream preservation in Europe.

5. Community Practices

The standardization attempts of the communities participating in the ISO discussions so far are paralleled by best practices of sub-communities and new attempts to come up with suitable solutions for problems to be solved. Whatever CLARIN does must ber related to these practices and where possible we have to find ways to integrate the resources and tools/services of these sub-communities.

There is a wide range of such practices in sub-communities that need to be taken care of. Here we can give only a limited account of such sub-disciplines, i.e. the list is not at all exhaustive. But CLARIN should make statements which practices will be incorporated in the sense that converters etc will be made available, probably in a stepwise fashion. Communities with activities such as word sense disambiguation, named entity recognition, semantic role analysis, co-referencing, discourse annotation, language documentation, sign language research, child language research, machine translation, terminology, typology and many more need to be covered, but this document is not the place to list their practices in detail.

5.1 Formats

With respect to formats a huge number of different suggestions have been made by some of those communities that are mentioned above. A few are indicated here as examples:

- CHAT (CHILDES) is a format for text corpora and lexica in the field of child language research existing already for many years and being used as interchange format amongst researchers world wide. It is not formally specified as a schema, but a set of widely used tools work on the resources and but robust converters to EAF and LMF for example exist.
- SHOEBOX/TOOLBOX is an interesting tool set widely used in the field of field-linguistic research and language documentation to create text corpora and lexica. The tool is also existing for many years already and works with plain text databases. It is not formally specified as a schema, but conversions to for example EAF and LMF exist if users adhered to the guidelines. Since the tool does not enforce hard constraints mostly some curation effort is needed to get the data into proper XML/UNICODE.
- EAF is a format based on a flexible XML schema applied by researchers world wide that do audio, video and time series annotation at multiple levels. The format emerged during the discussions about Liberman&Bird's Annotation Graph paper and is flexible in so far as it can have as many levels and annotation types as needed. XSLT conversions can be applied to transform to other XML formats. The ELAN tool generating EAF is one of the most widely used media annotation tool.
- EXMERALDA is a similar format compared to EAF and there are converters between them.
- MATE is another multilevel annotation format comparable with EAF and EXMERALDA.
- XCES is an XML based corpus format that is widely used to create text corpora with multilevel annotations on the texts. It is a subset of the TEI specifications to make processing feasible.
- TIGER is a flexible format to represent syntactical annotations on texts. It was the basis for designing SynAF (see before).
- PentreeBank is also a format to represent syntactical annotations on texts. Also this format should be covered by SynAF.
- PAULA is a format that is realized as a set of relational database tables and that is meant to bring together various formats such as EAF, EXMERALDA, XCES etc into one database to make it searchable with the help of one index.
- TMX is an exchange format widely used in the area of machine translation experts.
- others to come

5.2 Tag Sets

With respect to the encoding of linguistic phenomena we can identify a variety of encoding systems or habitudes, most of them are not described in a formal way, but exist as individual selections. As discussed above it should be the intention to cover them in ISOcat if they have a broad enough community and exhibited some stability over time. Again we can only mention a few of them.

- EGLAES/ISLE/MILE produced a set of categories for the description of various phenomena in text corpora, lexica and metadata. The main ideas have now been transferred to ISOcat.
- GOLD is an ontology covering the essential categories and their relations used for morphology encoding that was created within the E-Meld project. It is intended to bring their category definitions into ISOcat.
- Typological Database System specifies an ontology covering categories and their relations as they appear in a number of typology databases that have been analyzed. It is intended to bring their category definitions into ISOcat.
- STTS is a list of morphological categories as they are typical for German. It is intended to bring their category definitions into ISOcat.
- others to come

6. References to standards and Best-Practices