The Standards' Landscape Towards an Interoperability Framework

The FLaReNet proposal

Building on the CLARIN Standardisation Action Plan

July 2011

Monica Monachini, Valeria Quochi, Nicoletta Calzolari, Núria Bel, Gerhard Budin, Tommaso Caselli, Khalid Choukri, Gil Francopoulo, Erhard Hinrichs, Steven Krauwer, Lothar Lemnitzer, Joseph Mariani, Jan Odijk, Stelios Piperidis, Adam Przepiorkowski, Laurent Romary, Helmut Schmidt, Hans Uszkoreit, Peter Wittenburg[1]

[1] This group of experts was involved in the making of the document at different stages both in CLARIN and FLaReNet.

Table of Content

Summary

This document proposes an overview of the current scene towards an Interoperability Framework and acts as a reference point for the current standards that the community fosters and encourages to adopt/improve. This initiative is in close synchronization with other relevant initiatives such as CLARIN, ELRA, ISO and TEI and META-Share.

The document builds on the CLARIN Standardisation Action Plan and adapts and extends it to the needs of the broader LT Community, beyond the SSH research areas including the industry.

The main goal of this document is to give a practical orientation for various LT players, both commercial and academic; the main message being that a harmonized domain of language resources and technology can be achieved stepwise, but that an effort to adopt standards is necessary to overcome fragmentation.

NB: This is to be intended by no means as a static, closed document, rather a dynamic one which needs to be constantly/periodically revised and updated by the community itself.

Structure of the document

Drawing on the results of a previous report drafted by the CLARIN project together with FLaReNet, META-NET and ELRA, the "CLARIN Standardisation Action Plan" (Bel et al. 2009) has been revised and updated with relevant standards for the broader LT community, also addressing those that are typically used in industry, at different levels of granularity[2]. It is meant as a general reference guide for the whole community and is particularly useful for LT organizations such as META-SHARE as it provides concrete indications about standards and best practices that are important for given tasks or media in LT.

Both current standards and on-going promising standardisation efforts are listed, so that the community can monitor and actively contribute to them. These standards are at different stages of development: some are already very well known and widely used, others more LR-specific standards, and especially those developed in the framework of the ISO Technical Committee devoted to LR management are in the process of development or are being revised.

LR standards become increasingly relevant for all industry branches where LRs are being produced and used, information technology, automation/robotics, telecommunications, data mining, information retrieval, and for all sectors supported by information technologies: eCommerce, eHealth, eLearning, eGovernment, eEnvironment.

A number of standards exist, creating a potentially useful framework, ready for adoption. Currently, relatively small sets of basic standards (defined as foundational standards) can be identified that have gained wide consensus. These are not necessarily specific to language resources, but provide a minimum basis for interoperability. On top of these come standards that specifically address language resource management and representation that should also be considered as foundational. They are increasingly recognized as fundamental for real-world interoperability and exchange. A set of other standards focusing on specific aspects of linguistic and terminological representation are also currently in place and officially established, resulting from years of work and discussions among groups of experts of various areas of language technology. Most of the more linguistically oriented standards are also periodically under revision in an attempt to make them ever more comprehensive as new

[2] Input was collected also from LRE Map, Multilingual Web, the FLaReNet Forums, LREC Workshops, ISO and W3C.

technologies appear and new languages are being considered. Considerable effort seems to be still needed for their promotion and to spread awareness to a wider community.

The standards related to terminology management and translational technologies are probably the most widespread and consolidated, in part because of the real market behind the translation industry. Finally, the current situation witnesses a stream of on-going standardisation projects and initiatives focused mainly on recent mature areas of linguistic analysis and on emerging technologies such as semantic annotation which includes temporal and space annotation. These are initiatives the community needs to monitor closely and actively participate in.

Along with the standards mentioned above, in specific communities there are established practices that can be considered de-facto standards. For these a number of tools exist that facilitate the usage of the resources, e.g. WordNet, PennTreeBank, etc. As these need not to change, at least not in the near future, it is recommended *the development of mappers/converters from these best practices/common formats to the other endorsed/official standards*.

1. Introduction

De-facto or proprietary standards are being adopted in the community. The agreed orientation is toward the introduction of standards in those area that that are mature enough that will produce benefit for the field and for industry, e.g. massive data use increase, combination of language resource and tools into new collection and workflows, infrastructure development.

The FLaReNet community comprises a wide range of standardisation experts that can act as liaison to provide information as well as feedback to the relevant standardisation bodies (ISO, TEI, W3C) in different appropriate boards.

As general considerations regarding standards the community should take into account that:

- Standards should be relatively easy to apply, users should not be required to read long specification documents; instead, there should be tools, services and converters available that facilitate end users in using the standards by hiding the complex formalisms.

- 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.

- Mappers/converters form the well-established practices to the proposed standards should be provided

- 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.

2. Recommended standards

2.1. Basic Standards

In this chapter section FLaReNet provides a list of basic standards which receive large consensus in the community and are to be seen as almost obligatory.

2.1.1. Text

2.1.1.1. Unicode - ISO 10646

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

It is imperative to use character encoding standards to really support multilingualism. The community is thus strongly recommended to apply ISO 10646/UNICODE in all resources and tools.

There are still characters out there where linguists are confronted with that have not yet been integrated in UNICODE such as Cuneiform characters and where special arrangements are required. However, increasingly more characters are captured. The linguistic community is represented in the UNICODE boards.

Related to Unicode are ISO standards on language identifiers and scripts.

2.1.1.2. Country codes - ISO 3166

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.

The community will apply ISO 3166 in all resources and tools.

2.1.1.3. ISO 639 series of the International Standard on Language Coding

The ISO 639 series currently consists of 6 parts:

ISO 639-1:2002 Codes for the representation of names of languages – Part 1: Alpha-2 code, developed by ISO/TC 37/SC 2; confirmed 2007-12.

ISO 639-2:1998 Codes for the representation of names of languages – Part 2: Alpha-3 code, developed by a joint committee of ISO/TC 46/SC 4 and ISO/TC 37/SC 2; confirmed 2008-12.

ISO 639-3:2007 Codes for the representation of names of languages – Part 3: Alpha-3 code for comprehensive coverage of languages, developed by ISO/TC 37/SC 2; published 2007-02.

ISO 639-4:2010 Codes for the representation of names of languages – Part 4: General principles of coding of the representation of names of languages and related entities, and application guidelines, developed by ISO/TC 37/SC 2; published 2010-07.

ISO 639-5:2008 Codes for the representation of names of languages – Part 5: Alpha-3 code for language families and groups, developed by ISO/TC 37/SC 2; published 2008-05.

ISO 639-6:2009 Codes for the representation of names of languages – Part 6: Alpha-4 code for comprehensive coverage of language variants, being developed by ISO/TC 37/SC 2; published 2009-11.

2.1.1.4. ISO 639 Governance Structure, current transition, and future development

ISO 639 Registration Authorities (RAs)

Parts 1,2,3,5, and 6 are maintained by registration authorities:

- ISO 639-1 Registration Authority: Infoterm
- ISO 639-2, ISO 639-5 Registration Authority: Library of Congress
- ISO 639-3 Registration Authority: SIL International
- ISO 639-6 Registration Authority: GeoLang

Web sites and information policy

The Library of Congress hosts the home page of the ISO 639 RAs-JAC: http://www.loc.gov/standards/iso639-2/. The ISO 639-1, ISO 639-2 code tables are available on that site, by approval from the ISO Central Secretariat. ISO 639-5 code tables are available at: http://www.loc.gov/standards/iso639-5/

SIL International hosts a web site containing all ISO 639-1, ISO 639-2, and ISO 639-3 code tables, as well as the home page of ISO 639-3: http://www.sil.org/iso639-3/.

GeoLang hosts a web site containing the ISO 639-6 code table: http://www.geolang.com/iso639-6/.

Since the various parts of ISO 639 are continuously updated, external users are encouraged to visit the web sites for up-to-date information about language identifiers.

ISO 639 RAs Joint Advisory Committee

ISO 639 RAs-JAC has been functioning since 1999, consisting of one representative of each of the ISO 639 RAs, three voting members nominated by ISO/TC 37, three voting members nominated by ISO/TC 46. In addition, up to five technical experts functioning as non-voting observers may participate.

Role and operation of ISO 639 RAs-JAC

ISO 639 RAs-JAC was established to advise the RAs to guide the application of the coding rules as laid down in the various parts of ISO 639. Details on the working principles of ISO RAs-JAC and further information are available on the web site, in particular http://www.loc.gov/standards/iso639-2/iso639jac_n3r.html.

Secretariat of ISO 639 RAs-JAC

ISO 639 RAs-JAC has a Secretary at Standards Norway: Mr. Håvard Hjulstad.

Changes in and additions to the ISO 639 code tables

The ISO 639 RAs-JAC has worked with the identification and removal of inconsistencies in the code tables of ISO 639. No additional items have been encoded for ISO 639-1, ISO 639-2, or ISO 639-5, but a number of changes to language names have been approved and one item has been deprecated. See http://www.loc.gov/ standards/iso639-2/php/code_changes.php for complete documentation of changes to ISO 639-1 and ISO 639-2.

In addition, a number of new items have been approved following the procedures for addition of new items to ISO 639-3. These additions are documented separately by ISO 639-3 RA. See http://www.sil.org/iso639-3/changes.asp for documentation of updates.

Further development of the ISO 639 series

Through the finalization of ISO 639-6 the ISO 639 series of International Standards provides mechanisms to encode living and historical languages, language groups, and language variants. The intention is to develop the ISO 639 code tables to provide language encoding solutions for the needs of language and content industries as well as for documentation and information centres and providers.

A migration process has been initiated for the ISO 639 database to be technically and methodologically related to ISOCAT (ISO 12620). The governante structure is also being expanded to include various communities of linguists and other experts on linguistic diversity and language documentation, actively involving CLARIN, FlareNet, Meta-Net, UNESCO, and related scientific communities.

The lists of language identifiers that are standardized in ISO 639-1 (alpha-2

It has been decided to transform current part 4 of ISO 639 into ISO 639 as the only standard, while the contents of parts 1,2,3,5, and 6, i.e. the language code lists themselves from now on are part of the ISO 639 database for widespread use.

2.1.1.5. Codes for the representation of names of scripts - ISO 15924

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].

The community will apply ISO 15924 when needed in all resources and tools.

Missing scripts have to be reported to the registration authority.

2.1.1.6. XML

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). XML is at the hearth of XQuery, XML Schema, XSLT, XPATH, DTD, Relax-NG, Schematron, I18N. These standards allow to define schemas to manage Language Resources or their embedding in distributed applications and web standards (e.g. SOAP).

The community 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.

Being a meta-language allowing one to define specific document models (by means of DTDs, RelaxNG schemas or W3C schemas), it does not provide means to control the semantics of XML components.

Microsoft has increasingly adopted XML in their products, XML is taking off the ground also in IT sectors until now reluctant to adopt it. This is good for those LR standards that offer open exchange formats and conversion routines for and to XML. Important for XML is the use of UTF-8 UTF-16.

2.1.2. Audio

The rapid development of telecommunications industry and of multimedia industry favoured the taking off of audio/video standards . The International Telecommunications Union (ITU) has also been active in this field for many years and has produced important standards on the technical level of speech processing (ITU-T Speech and audio coding standardisation)..

2.1.2.1. PCM

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.

2.1.2.2. MP3 and ATRAC

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.

2.1.2.3. SAMPA

In certain research areas phonetic transcriptions are required for further speech processing - here the International Phonetic Alphabet is used. A frequently used scheme was to use SAMPA for this purpose which specifies IPA characters in terms of ASCII characters.

For audio recordings it is recommended 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 (International Phonetic Alphabet) which is included in the UNICODE standard.

There is the need to describe special phonetic characteristics not yet included in IPA.

2.1.3. Video/Multimodality

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.

2.1.3.1. H.264

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.

2.1.3.2. MJPEG2000

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 will be chosen.

For video recordings it it is recommended 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.

It seems that H.264 is better because usable in real applications see Benoit, Martin, Pelachaud, Shomaker, Suhm “Audio-visual and multimodal Speech Systems”].

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.

- avi
- dv
- mpeg-ps (vob)
- mpeg-ts
- mov
- mp4
- avchd

Adobe Flash Video can be used for video embedded in documents, such as manuals.. The data in a file container may use any one of a variety of codecs.

- mpeg1
- mpeg2
- mpeg4
- DivX
- Mjpeg
- Dv
- Xvid

2.1.4. Web services

There a set of W3C specifications and guidelines, at different levels of stability, for Web technology standards. Manipulating data with XML requires sometimes integrity, authentication and privacy: XML signature, encryption, and xkms can help creating a secure environment for XML. The Web of Services is based on technologies such as HTTP, XML, SOAP, WSDL, and others. In the last few years, REST (Representational State Transfer) is gaining popularity.

2.1.4.1. HTTP

HTTP is the core protocol for exchanging information on the Web. HTTP has been coordinated by the Internet Engineering Task Force (IETF) and the World Wide Web Consortium (W3C). HTTP protocol functions in a client-server computing model. The version in use today, HTTP/1.1, was defined in June 1999.

2.1.4.2. SOAP REST

REST is gaining a momentum as a simpler, more straightforward way to access web-based services. REST services are much simpler to consume than SOAP based services and much more flexible for rapid integration into Web application.

It seems that SOAP is being preferred for services within the enterprise whereas REST is being preferred for services that are exposed as public APIs.

2.1.4.3. WSDL

WSDL Web Services Description Language Version 2.0, provides the reference model and an XML format for describing Web services.

2.1.4.4. XML Encryption

XML Encryption specifies a process for encrypting data and representing the result in XML. The data may be arbitrary data (including an XML document), an XML element, or XML element content. The result of encrypting data is an XML Encryption element which contains or references the cipher data.

2.1.4.5. XML Signature

XML Signatures provide integrity, message authentication, and/or signer authentication services for data of any type.

2.2. Language Resource Specific Standards

2.2.1. Language Resource management

2.1.6.2. ISO 24610-1:2006 -- Feature structures -- Feature structure representation (FSR)

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.

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

This standard is currently under revision by the ISO TC 37 committee. Work is ongoing to have the feature structure description module adopted by ISO.

2.2.2. Representation of Lexical Resources

2.2.2.1. ISO 24613:2008 Lexical Markup Framework (LMF)

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 there is the need to come up with example lexica for different sub-communities. Currently, only examples for NLP lexica have been worked out.

LMF has been widely standardized and first tools are supporting this standard. The usage of LMF should be promoted, 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.

LMF is fairly new and it is not possible 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. The creation of such reference structures that can be re-used for similar intentions should be promoted. LMF has already been tested to represent Wordnets for example, although the connection between ontologies and lexica is still an issue under debate. The adaptation of tools and the creation of converters to this format should be promoted.

2.2.3. Representation of primary sources

2.1.6.3. TEI

The TEI guidelines provide a modular framework of LR-specific standards mainly for corpus representation, markup and annotation.

TEI Guidelines are well established the Humanties more than in the industrial context. The community 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.

TEI offers very flexible mechanisms which in practice 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.

TEI invested considerably in tutorials, facilities and initiatives to facilitate its adoption and use., which resulted in a wide use in the humanities community. An activity which should be taken as a positive example.for encouraging the use of standards and common formats.

2.1.6.4. XCES

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.

2.1.6.5. TEI/ODD

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.

TEI/ODD is recommended by CLARIN. It will be used systematically to ensure a proper documentation and dissemination of the schemas.

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.

2.2.4. Representation of annotated text

2.2.4.1. ISO/DIS 24611 Morpho-syntactic Annotation Framework (MAF)

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.

MAF is endorsed as the pivot format for the exchange of morpho-syntactic information and encourages the community 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.

MAF does not standardise any specific tagsets, 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. The community should promote the adaptation of tools to support MAF.

2.2.4.2. ISO/CD 24615:2010 Syntactic Annotation Framework (SynAF)

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.

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.

2.2.4.3. ISO 24617-1:2009 Semantic annotation framework (SemAF) -- Part 1: Time and events (ISO-TimeML)

TimeML offers a format for the annotation of temporal entities, namely: temporal expressions, eventualities (i.e. 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 phasal 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 scheme for event annotation and temporal expressions (e.g. TIDES TIMEX2 tag).

TimeML should be the pivot format for temporal annotation. TimeML has been integrated with OWL Time (DAML Time). TimeML and ISO TimeML should be endorsed and its use and the development of tools which support these formats should be promoted.

TimeML is now part of an ISO standardization effort within TC 37/SC 4, Semantic Annotation Framework (SemAF) ISO 24617-1. ISO – TimeML enlarges the representational capabilities of the original TimeML scheme by offering a metamodel and a formal semantics associated with the scheme. ISO – 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. It could be useful also to fix some minor shortcomings of the TimeML scheme as far as the annotation of events spanning over multiple tokens (i.e. multiword expressions) should be performed.

2.2.5. Knowledge Representation – W3C Semantic Web

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 (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.

2.2.5.1. RDF

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.

2.2.5.2. SKOS

SKOS is a framework with simplified logic that allows users to represent for example hierarchical concept systems such as thesauri.

2.2.5.3. OWL

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.

The recommendation is to make use of the W3C standards wherever knowledge needs to be represented in flexible formats. Various frameworks should provide an export into these formats making use of RDF and OWL.

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.

2.2.6. Terminologies and Translation

The standards in this sections are mostly and commonly used by the localization and translation industry as well as by public translation and terminology units and organisations

The ISO standards are developed and maintained by ISO/TC 37, the Technical Committee on Terminology and Other Language and Content Resources. This TC has 4 sub-committees

- SC 1: Principles and methods
- SC 2: Terminographical and lexicographical working methods, covering: Layout, lexicography, pragmatic applications; language codes, and translation management
- SC 3: Systems to manage terminology, knowledge and content, covering: Computer assisted terminology management
- SC 4: Language resource management, covering Natural language processing and other language resources
- SC 5 (growing out of SC 2 working group 6) is being established in autumn 2011 and focuses focus on standards for translation, interpreting, multilingual technical communication, and localization, from the perspectives of service requirements, processes, specifications for quality metrics and assessment, etc.

LISA (Localisation Industry Standards Association) was founded in 1990 and ceased to exist in 2011. It created several standards widely used by the translation and localization industry. Stewardship for these standards has been taken over by ETSI in summer 2011. Several relevant standards have been developed and are maintained by OASIS..

2.2.6.1. ISO 1987 (Part 1 and 2) Terminology work – Vocabulary

ISO 1087 contains the central terms and concepts of professional domain of terminology, covering terminology theory, terminology management, terminology work.

Note 1: originally there were 2 parts of this standard, the second part focusing on computer-based terminology work. This part has been withdrawn and the current revision of ISO 1087 part 1 will lead to a new version integrating all aspects in a single standard document.

Note 2: This standard is a meta-document laying down the language of the profession of terminologists and providing a consistent conceptual basis for all other ISO TC 37 standards.

2.2.6.2. ISO 704 Terminology work – Vocabulary

The current version of ISO 704 is from 2009 and contains the basic principles and working methods for terminology work, in particular principles for coining new terms and evaluating existing terms, creating concept systems, and writing definitions.

2.2.6.3. ISO 12620 Terminology and other language and content resources -- Specification and management of a Data Category Registry for LRs

This standard was published in 2009. Its main purpose is to (1) specify data categories with a consistent semantics to be shared by all people building up or running a terminology database or any language resource in XML, for the Semantic Web, and many other application scenarios and to (2) lay down the rules, procedures, governance, and requirements for an online Data Category Registry for language resources (on the web known as ISOCat) for communities such as translation, localization, domain-specific or multi-domain terminology management, information systems, knowledge management, corpus linguistics, etc.

2.2.6.4. ISO 26162:2010 Systems to manage terminology, knowledge and content – Design, implementation and maintenance

This standard specifies principles and requirements for appropriately designing, implementing and maintaining a terminology management system. It deals with principles of data modeling relevant to terminology databases and functionalities of terminology management systems usually expected by different user communities.

2.2.6.5. ISO 16642:2003 Computer applications in terminology – Terminological Markup Framework

TMF is the overarching foundational standard for all forms of terminology markup. TBX (see 2.2.6.6) is one TMF-conformant markup language for terminological data. TMF exists in XML, UML and more recently in RDF. It is useful for meta-data modeling activities, for concrete database implementations TBX is recommended.

2.2.6.6. ISO 30042:2008 Systems to manage terminology, knowledge and content – TermBase eXchange (TBX)

Term Base eXchange (TBX) is the open, XML-based standard for exchanging structured terminological data that has been approved as an international standard by LISA and ISO. In 2011 LISA ceased to exist and stewardship of all LISA standards was taken over by ETSI.

By providing a universal markup framework, TBX enables companies to take control of their terminology and to share it more easily with those who need it, such as business partners and language service providers.

TBX as one of the concrete manifestations of TMF. TBX is widely recognized and supported by tool providers, large companies and public organisations.

2.2.6.7. TBX-Basic terminology-related standard

TBX-Basic is a lighter version of TBX, particularly suited to small or medium sized language industries. It is also suited for any language application that requires a lightweight approach to terminology management, such as some controlled authoring applications.

TBX-Basic is a TBX-compliant terminology markup language that allows a limited set of data categories. It is intended for terminology resources that are commonly developed to support translation and localization processes. The purpose of TBX-Basic is to formalize the translation and localization industry's needs for terminology markup in an XML standard, in order to increase the ability to exchange terminology resources between users and to use these resources in various computerized environments.

2.2.6.8. TMX (Translation Memory eXchange)

TMX (Translation Memory eXchange) is the vendor-neutral open XML standard for the exchange of Translation Memory (TM) data created by Computer Aided Translation (CAT) and localization tools. The purpose of TMX is to allow easier exchange of translation memory data between tools and/or translation vendors with little or no loss of critical data during the process. In existence since 1998, TMX is a certifiable standard format. TMX was developed and maintained by OSCAR (Open Standards for Container/Content Allowing Re-use), a LISA Special Interest Group, in 2011 taken over by ETSI.

2.2.6.9. SRX (Segmentation Rules eXchange)

Segmentation Rules eXchange (SRX) is the vendor-neutral standard for describing how translation and other language-processing tools segment text for processing. It allows Translation Memory (TM) and other linguistic tools to describe the language-specific processes by which text is broken into segments (usually sentences or paragraphs) for further processing. It was developed when it was realized that TMX leverage was sometimes lower than expected because different tools segmented text in different ways, preventing a direct correlation between results between the tools. When implemented with TMX, SRX allows the transmission of the segmentation rules that were used when a TM was created so that tools can improve the leverage achieved when deploying TM data. SRX can also be used by any tool that segments text to improve integration with other processes.

SRX version 2.0 was officially accepted as an OSCAR standard in April 2008, in 2011 taken over by ETSI and is a new work item in ISO/TC 37/SC 3.

2.2.6.10. GMX (Global Information Management Metrics Exchange)

Global information Management Metrics Exchange is a three-part standard from LISA OSCAR that focuses on translation metrics. GMX/V defines what constitutes word and character counts, and allows for the exchange of metrics information within an XML vocabulary. GMX/V defines a canonical form for counting words and characters in a transparent and unambiguous way. The two associated standards, yet to be defined, will be GMX/C for complexity and GMX/Q for quality. Once the three GMX standards are available, they will provide a comprehensive way of defining a given localization task.

2.2.6.11. XLIFF XML Localization Interchange File Format

The XML Localization Interchange File Format is an OASIS standard for the exchange of data for translation. Rather than having to send full unprotected electronic documents for localization, with the inevitable problems of data and file corruption, XLIFF provides a loss-less way of round tripping text to be translated. Language Service Providers, rather than having to acquire/write filters for different file formats or XML vocabularies, have merely to be able to process XLIFF files, which can include translation memory matching, terminology, etc.

2.2.6.12. OAXAL: Open Architecture for XML Authoring and Localization Reference Model

As one of the most recent and comprehensive initiatives, OAXAL is made up of a number of core standards from W3C, OASIS and LISA (from 2011 ETSI).

The approach is a promising reference architecture for localization management: OAXAL – Open Architecture for XML Authoring and Localization Reference Model. It is not only a reference model, but also a newly founded OASIS reference architecture technical committee under the same name.

It is interesting for several reasons: (i) it is not limited to localization in the narrower sense, but open to all authoring and publishing processes in dynamic workflows in multilingual communication environments; (ii) it is consistently based upon and oriented towards open standards; (iii) it is component-based and flexible, yet integrative in nature; (iv) it is practice-oriented and driven by practical needs arising from real-life localization industry workflows

The OAXAL reference model is open and will further be extended. TBX, for instance, will be added to the model (see chapter 2.2.6.6 for details on TBX). OAXAL is certainly useful already from a conceptual and strategic point of view, as it invites decision makers in industry not to take a look at each individual standard in an isolated way but rather to look at the whole model from a workflow and integration perspective. Then one can decide which building blocks or components are actually relevant for a particular implementation and application scenario.

2.2.6.13. Translation services

The European standard EN 15038 Translation – Service Requirements (published in 2006) focuses on translation service provider quality management aspects. It replaced several previously existing national standards in EU member countries. Outside Europe there are national translation service requirements standards in the USA, Canada, China, Russia, and other countries. ISO TC 37 SC 5 (previously SC 2/WG 6) has started a world-wide effort to create an ISO standard for this purpose and as started other standards projects for translation quality metrics and assessment, requirements for community interpreting, etc.

3. Ongoing/Upcoming Standards

This sections reports those standards that are still *in fieri*, or very recently approved and that appear to be interesting and promising for the development of the LT field.

FLaReNet thus recommends the community to monitor their evolution and possibly to actively contribute to their definition and adoption.

3.1. Annotated Text

3.1.1. ISO/DIS 24612 Linguistic annotation framework (LAF)

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 modeling experts.

It should be devoted some time to check compatibility with LAF. If problems are seen with the current specification these should be communicated to the ISO representatives.

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.

3.1.2. ISO/DIS 24617-2 Semantic Annotation Framework (SemAF) - Part 2: Dialogue Acts (SemAF-DA)

The ISO standard for Dialogue Acts (ISO-DiAML) is an abstract meta-model for the annotation of dialogue corpora, following up on the EU-supported project LIRICS (Linguistic Infrastructure for the Interoperable Resources and Systems) developed in collaboration with TC 37/SC 4 ad-hoc Thematic group 3, Semantic content. DiAML is still under development and has been accepted for Draft International Standard balloting. The standard has been designed in accordance with the ISO Linguistic Annotation Framework (LAF, ISO 24612, 2009).

3.3.3. ISO/AWI 24617-3 Semantic Annotation Framework - Part 3: Named Entities (SemAF-NE)

The main area of application of NEs, and thus of this standard, is information management systems in the context of the Semantic Web, and in particular technologies such as: question answering, automatic or semi-automatic construction of ontologies, information structure computation, document comparison, machine translation, message identification for automatic filtering, classification,…

The aim of the SemAF-NE standard is to specify a consensual model, or annotation scheme, for the annotation of named entities (NEs) in texts and speech contents.

The specification proposal thus aims at allowing comparison and merging of different pre-existing annotations, providing a “best practice” for new annotations that will thus be natively interoperable with each other and with pre-existing annotations; permitting integration of NEs into other annotation schemes like TimeML and ISO-Space; facilitating the development and provision of common tools.

3.1.4. ISO/AWI 24617-4 Semantic Annotation Framework - Part 4: Semantic Roles (SemAF-SRL)

The ISO standard for semantic role annotation aims at defining a annotation scheme for Semantic Roles (SRs), which are receiving increasing interest in the information processing community because they make explicit key conceptual relations between a verb and its arguments.

The current proposal is informed by the various semantic role frameworks being used to support data annotation, such as FrameNet, Verbnet, PropBank and LIRICS, which have been found to bear strong underlying compatibilities. The documentation explains such compatibilities and gives a loose mapping between definitions of individual semantic roles from the different frameworks. The general goal is to provide language neutral semantic representations for semantic roles, a pivot representation to facilitate mapping between different formalisms, and guidelines for creating new resources for languages that would be immediately interoperable with each other and with pre-existing resources.

The specification is envisaged to be used in two different situations: 1) in annotations where semantic roles are recorded in annotated corpora; 2) as a dynamic structure produced by automatic systems.

3.3.5. ISO/AWI 24617-5 Semantic Annotation Framework - Part 5: Discourse Structure (SemAF-DS)

The ISO Standard for Discourse Structure representation is based upon LAF and targets the description of how a discourse is organized in terms of its semantic and pragmatic content. It mainly addresses the semantic structure of discourse, whose modality may be text, audio, video, hypertext, games, …. It is meant toa basis for annotation, production, translation of various types of documents as discourse structures can be found not only in linguistic content, but also in non-linguistic content such as (possibly silent) video.

The goal is to define a scheme that can provide a common, language-neutral pivot for the interoperation among diverse formats of discourse structures of various types of documents, linguistic or not. The standard scheme proposed specifies the organization of discourse structures consisting of eventualities and the discourse relations among them. Discourse relations have traditionally been assumed to carry both semantic and presentational (syntactic and pragmatic) information, but this standard simplifies them and minimizes the set of discourse relations by attributing presentational information to other parts of discourse structures.

3.3.6. ISO 24616:2011 -The Multilingual Information Framework (MLIF)

In addition to further annotation standards, MLIF is integrative and workflow oriented in nature, referring to a number of existing standards in language industry. MLIF provides a generic platform for modeling and managing multilingual information in various domains: e.g. localization, translation, multimedia annotation. As most ISO standards, it provides a metamodel and a set of generic data categories for various application domains. MLIF also provides strategies for the interoperability and/or linking of models including, but not limited to, XLIFF (cross ref) and TMX (cross ref).

MLIF has been recently approved as an official standard, albeit not published yet. Potentially, this standard can become useful for multilingual information workflows, but as of today still lacks large adoption and dissemination.

3.3.7. W3C Emotion Markup Language (EML)

An Emotion Markup Language (EML or EmotionML) is defined by the W3C Emotion Incubator Group (EmoXG) as a general-purpose emotion annotation and representation language, which should be usable in a large variety of technological contexts where emotions need to be represented.

Cf. EARL: HUMAINE Emotion Annotation and Representation Language (EARL).

3.3. Video/Multimodality

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. It is hard to make strong recommendations at this moment. But, a series of interesting standardisation proposals and useful formats can be mentioned.

3.3.1. EMMA: Extensible MultiModal Annotation markup language

EMMA, Extensible MultiModal Annotation markup language, is a standard framework for multimodality developed by the W3C from 2009. Its main elements (with several working groups dedicated to developing and maintaining these standard recommendations) are:

- “Extended Hypertext Markup Language (XHTML)10.an XML version of HTML for presenting visual information on screens
- Speech Synthesis Markup Language (SSML)11.an XML-based language used to render text as speech
- Scalar Vector Graphics 1.2 (SVG)12.an XML-based language for writing two-dimensional vector and mixed vector/raster graphics
- Synchronized Multimedia Integration Language 2.0 (SMIL)13.an XML-based language for writing interactive multimedia presentations” (Larson 2005)

3.3.2. Ink Markup Language (InkML)

The Ink Markup Language (W3C new candidate) serves as the data format for representing ink entered with an electronic pen or stylus. The markup allows for the input and processing of handwriting, gestures, sketches, music and other notational languages in applications. It provides a common format for the exchange of ink data between components such as handwriting and gesture recognizers, signature verifiers, and other ink-aware modules.

3.3.4. Multimodal Interaction Framework

The W3C Multimodal Interaction Framework describes input and output modes widely used today and can be extended to include additional modes of user input and output as they become available.

3.4. Web services

3.4.1. OASIS WS-I

OASIS WS-I (http://www.oasis-ws-i.org/) comprises a diverse community of Web services leaders from a wide range of organizations around the world. OASIS WS-I Technical Committees maintain Profiles and supporting Testing Tools based on Best Practices for the selected groups of Web services standards. Profiles are guidelines based on Best Practices for the selected groups of Web services standards to assist the Web services community in developing and deploying interoperable Web services.

The key WS-I Profiles are:

- **Basic Profile 1.0 and 1.1,** which establishes core Web services specifications (SOAP, WSDL, UDDI, XML Schema, HTTPS) that should be used together to develop interoperable Web services.

- **Attachments Profile 1.0**, which complements the Basic Profile 1.1 to add support for interoperable SOAP Messages with attachments-based Web services.

- **Simple SOAP Binding Profile**, which consists of those Basic Profile 1.0 requirements related to the serialization of the envelope and its representation in the message, incorporating any errata to date.

- **Basic Security Profile 1.0**, an interoperability Profile that addresses transport security, SOAP message security and other security considerations, and composes with other WS-I Profiles. It references existing specifications and standards, including the OASIS Web Services Security 1.0 and SOAP Message Security 1.0 specifications, and provides clarification and guidance designed to promote interoperability of Web services created according to those specifications.

3.4.2 W3C Web Services Activity

The W3C Web Services Activity (http://www.w3.org/standards/webofservices/ and http://www.w3.org/2002/ws/) is designing the infrastructure, defining the architecture and creating the core technologies for Web services. The SOAP 1.2 XML-based messaging framework became a W3C Recommendation in June 2003 and the SOAP Message Transmission Optimization Mechanism (MTOM) in January 2005. The W3C Web of Services has different working groups dealing with Protocols, Service description, Security and Internationalization.

Currently, the W3C Web Services Activity has seven “Candidate Recommendations”, briefly described below.

3.4.2.1. Web Services Enumeration (WS-Enumeration)

This specification describes a general SOAP-based protocol for enumerating a sequence of XML elements from a SOAP enabled information source.

3.4.2.2. Web Services Eventing (WS-Eventing)

This specification describes a protocol that allows Web services to subscribe to or accept subscriptions for notification messages.

3.4.2.3. Web Services Fragment (WS-Fragment)

This specification extends the WS-Transfer specification to enable clients to retrieve and manipulate parts or fragments of a WS-Transfer enabled resource without needing to include the entire XML representation in a message exchange.

3.4.2.4. Web Services Metadata Exchange (WS-MetadataExchange)

This specification defines how metadata associated with a Web service endpoint can be represented as resources, how metadata can be embedded in endpoint references, how metadata could be retrieved from a metadata resource, and how metadata associated with implicit features can be advertised.

3.4.2.5. Web Services Transfer (WS-Transfer)

This specification describes a general SOAP-based protocol for accessing XML representations of Web service-based resources.

3.4.2.6. Web Services Event Descriptions (WS-EventDescriptions)

This specification describes a mechanism by which an endpoint can advertise the structure and contents of the events it might generate.

3.4.2.7. Web Services SOAP Assertions (WS-SOAPAssertions)

This specification defines two WS-Policy assertions that can be used to advertise the requirement to use a certain version of SOAP in message exchanges.

3.6. Language Resource Management

3.6.1. ISO/FDIS 24619 -- Persistent identification and sustainable access (PISA)

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.

A task force should be established 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.

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.

3.6.2. NWIP_ISO - DFO10328 Component based Metadata

Although not all ingredients of the component metadata model have been worked out, a new work item has been proposed within ISO. There is the need 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.

3.7. Terminologies/Translation

3.7.1. SAE J2450

Ongoing standards development work in ISO TC 37 SC 2/ WG 6 now also focuses on translation metrics similarly to the domain-specific standard from the automotive industry, Translation Quality Metric. It provides a method for assigning weighted rating values to errors in translation processes.

3.8. Standards for semantic interoperability

3.8.1. ISO DCR and ISOcat

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.

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. The work with ISO DCR and ISOcat should be promoted, since it is at the core of various standardization and harmonization activities. Currently CLARIN recommends it and also asks to include other widely used tag sets to make them re-usable for others and to relate them to other categories in the DCR.

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.