Interoperability Framework: The FLaReNet action plan proposal

Nicoletta Calzolari, Monica Monachini, Valeria Quochi
Istituto di Linguistica Computazionale "A. Zampolli"
Consiglio Nazionale delle Ricerche
Via Moruzzi 1, Pisa, Italy
name.surname@ilc.cnr.it

Abstract

Standards are fundamental to ex-change, preserve, maintain and integrate data and language resources, and as an essential basis of any language resource infrastructure. This paper promotes an Interoperability Framework as a dynamic environment of standards and guidelines, also intended to support the provision of language-(web)service interoperability. In the past two decades, the need to define common practices and formats for linguistic resources has been increasingly recognized and sought. Today open, collaborative, shared data is at the core of a sound language strategy, and standardisation is actively on the move. This paper first describes the current landscape of standards, and presents the major barriers to their adoption; then, it describes those scenarios that critically involve the use of standards and provide a strong motivation for their adoption; lastly, a series of actions and steps needed to operationalise standards and achieve a full interoperability for Language Resources and Technologies are proposed.

1 Interoperability and Interoperability Framework

Today open, collaborative, shared data is at the core of a sound language strategy. Standards are fundamental to exchange, preserve, maintain and integrate data and Language Resources (LRs), to achieve interoperability in general; and they are an essential basis of any language resource infrastructure.

In the past, we used the notion of "reusability" that today has evolved into "interoperability". *Interoperability* means the ability of information and communication systems to exchange data and to enable the sharing of information and knowledge. Interoperability was declared one of the major priorities for the LT field at the first FLaReNet Forum in Vienna (Calzolari et al. 2009)

An *Interoperability Framework* can be defined as a dynamic environment of language (and other) standards and guidelines, where different standards are coherently related to one another and guidelines clearly describe how the specifications may be applied to various types of resources. Such a framework must be dynamic in several ways. First, as it is not feasible to define one single standard that can cover all the various linguistic representation levels and applications, a series of specific standards should continue to exist, but they should form a coherent system (i.e. coherence among the various standard specifications must be ensured so that they can "speak" to each other). Then, standards themselves must be conceived as dynamic, because they need to follow and adapt to new technologies and domains of application. As the Language Technology (LT) field is expanding, standards need to be periodically revised, updated and integrated in order to keep the pace of technological advancements.

An Interoperability framework is also intended to support the provision of language services interoperability.

Enterprises nowadays seem to need such a language strategy, and to be key players they must rely on interoperability, otherwise they are out of business. A recent report by TAUS (TAUS/LISA 2011) states that: "The lack of interoperability costs the translation industry a fortune", where the highest price is paid mainly for adjusting data formats.

2 The "History" of Standards

In the past two decades, because of the robustness and industrial applicability of some NLP technology, the need to define common practices and formats for linguistic data resources has been increasingly understood and sought. Language data resources, in fact, serve LT development in various ways. They are

- the data which is passed and exchanged among software components or applications;

Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm, pages 41–49,
Chiang Mai, Thailand, November 12, 2011.

- the lexical, terminological and semantic resources needed to perform various tasks such as information extraction, machine translation (MT), question answering;
- the primary source for statistical language modelling, fundamental for example in statistical machine translation (SMT), or automatic speech recognition, and many other applications.

Several projects laid the foundations for standardisation of resource representation and annotation, e.g. the Expert Advisory Group on Language Engineering Standards (EAGLES 1996) within which also the Corpus Encoding Standard (CES and XCES, Ide 1998) was developed, and the International Standard for Language Engineering (ISLE, Calzolari et al. 2002). With these projects Europe in the '90s was at the forefront in establishing standards for LT.

All these efforts bring us to the current landscape where actual standardisation is on the move. Consensus has begun to emerge, and in certain areas stable standards have already been defined. However, for many areas work is still ongoing either because "the emergence of a solid body of web-based standards have dramatically impacted and re-defined many of our ideas about the ways in which resources will be stored and accessed over the past several years" (Ide and Romary 2007), or because there are new emerging technologies, such as multimodal ones, that have specific requirements not covered by existing formats and established practices.

We therefore observe a continuum of standardisation initiatives at various stages of consolidation and the rising on new proposals, as the various areas of LTs become mature. Also, while some standards are "official", that is designed and promoted by standardisation bodies - i.e. ISO, W3C and LISA - others emerged bottom-up. These are the so-called *de-facto* standards or *best practices*: formats and representation frameworks that have gained community consensus and are widely used: e.g. WordNet (Fellbaum 1998), PennTreeBank (Marcus et al. 1993), CoNLL[1] (Nivre et al. 2007).

2.1 The FLaReNet Landscape

Drawing on a previous report drafted by the CLARIN[2] project (Bel et al. 2009), together with FLaReNet, META-SHARE and ELRA the original document has been revised and updated with standards relevant for the broader LT community, also addressing those that are typically used in industry, at different levels of granularity. "The Standards' Landscape Towards an Interoperability Framework"[3] (Bel et al., to appear) thus lists both current standards and on-going promising standardisation efforts so that the community can monitor and actively contribute to them. This document is conceived like a "live" document to be adopted and updated by the community (e.g. in future projects and networks), so as not to restart similar efforts over and over.

It is meant to be a general reference guide for the whole community and particularly useful for LT initiatives such as the META-SHARE infrastructure, as it provides concrete indications about standards and best practices that are important for given tasks or media in LT. These standards are at different stages of development: some are already very well known and widely used, others more LR-specific standards, especially those developed in the framework of the ISO Technical Committee 37 devoted to LR management, are in the process of development or are being revised.

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: e.g. Unicode-UTF8 for character encoding, ISO639 for language codes, W3C-XML for textual data, PCM, MP3, ATRAC, for audio, etc.

On top of these we find standards specifically addressing LR management and representation that should also be considered as foundational - ISO 24610-1:2006 - Feature structure representation, TEI, and LMF for lexical resources (Francopoulo et al. 2006, 2008). 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, such as TMF (ISO 2002) for terminology, SynAF (Declerk, 2006) and MAF (Clément and de la Clérgerie, 2005) for morphological and syntactic annotation. These result from years of work and discussions among groups of

[1] http://ilk.uvt.nl/conll/#dataformat
[2] www.clarin.eu
[3] This document also collects input also from the LRE Map, Multilingual Web, the FLaReNet fora, LREC Workshops, ISO and W3C.

experts from various areas of language technology and are thought to be comprehensive enough to allow for the representation of most current annotations. Most of them address syntactic interoperability by providing pivot formats (e.g. LAF/GrAF, Ide and Suderman 2007), while today there is a greater need for semantic interoperability, which is still an actual challenge. Most of the more linguistically oriented standards are also periodically under revision in an attempt to make them ever more comprehensive as new technologies appear and new languages are being considered. Effort is still needed for their promotion and to spread awareness to a wider community.

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: we speak of TMF, TMX, TBX [4], XLIFF, an OASIS standard for the exchange of data for translation. A recent effort is the reference architecture OAXAL (Zydron, 2008), a standard component stack, made up of a number of core standards from W3C, OASIS and LISA.

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 (ISO24617 – 1-6), emotion, i.e. EML (W3C, 2007) and multimodal annotation, i.e. EMMA (W3C, 2009). 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, e.g. WordNet and the PennTreeBank. For these a number of tools exist that facilitate their usage. 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.

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

Concluding, we can safely state that today a number of standards exists that create a potentially useful framework, ready for adoption, and that efforts now should to spread their application.

2.2 Barriers and major problems

While the current picture of LTs presents a great potential for real interoperability, some problems or barriers have emerged that hamper the broad usability of the current standards framework.
The key issue is not so much a lack of standards, but, in particular for LT-specific standards, a *lack of (open) tools for an easy use* of them. This certainly is a major factor that hampers a broad standard usage. Another barrier is the lack of reference implementations and documentation, possibly open source, to enable others to understand what was done and how. A major problem has to do with lack of developer- and user education and culture in using standards. There is resilient tradition to use idiosyncratic schemes, which causes incompatibility of formats (even for minor differences), thus hampering the possibility of merging annotations or using them together. This in turn prevents easy reuse of available data.

Within ISO, general interest standards (like country codes) are free. But others are not, and this should be avoided. In fact, this may be one other major factor preventing a wider adoption of standards. There are now attempts – i.e. in ISO TC 37 – to overcome this situation by allowing direct application of standards free of charge through implemented databases with free access such as the new ISO 12620 ISOcat.

In W3C, full documentation of standards is free, so it is easy for W3C documents to be spread and largely applied. However, participation in the definition and decision-making process is costly.

Standards need to be built by consensus; therefore their creation is a slow process[5].

3 Motivations for standards

There are various scenarios that critically involve the need for standards and provide a strong motivation for their adoption and for investment in the development of the missing ones. This section briefly introduces some of them.

[4] First developed in the SALT project, TMF and TBX are now ISO standards

[5] This is also in line with all other recommendations from FLaReNet and also fits well to the strategies of METANET.

[1]. ***Use of the same tools on different data; use of different tools on the same data***. Interoperability among software components is today a major issue. In architectures for knowledge mining, or for the creation of new resources, where the same data have to be used, enriched and queried by (chains of) different tools, common formats become crucial for their success, as for instance in the KYOTO project[6] where the KAF has been defined and adopted as a common representation format for textual data and related linguistic annotations (Bosma et al. 2009). Moreover, the use of different tools on the same data is relevant for testing and comparing tools, and also in collaborative situations to process the same corpus for different purposes.

[2]. ***Creation of workflows - Web Service Interoperability***. In cases where workflows need to be built chaining together tools not originally built to work in pipeline/together, standards will ensure their execution. As of today, in most cases workflows can be run with tools that were already designed to work together, or with the use of format converters. This is a major obstacle esp. in the context of web-based platforms for distributed language services. Experiences such as PANACEA (Bel 2010, Toral et al. 2011) show that using a common standardised format facilitates integration. If tools were built/modified to work directly on common/standard formats, workflows might be simpler, easier to design and quicker to run. While this is not possible at present, when the advantages are shown, new tools could naturally go in this direction. Workflow management should be generalised to cover both local processing and web service interfaces.

[3]. ***Integration/Interlinking of resources***. This has recently become an important trend also for companies that wish to provide composite services. In order to exploit the wealth of manual annotations already existing and developed within the years mostly by academic institutions, (legacy) resources must be integrated and interlinked. This is needed for example also for generating new training data, or for re-purposing already existing one. In order to achieve this goal broadly, we need not only standard formats but also common methodologies and best practices for resource management and update. The experience of linking Propbank and PennTreebank in SemLink[7] teaches us that changes/updates in one resource cause many problems to their mapping, resulting in a lot of manual work to be done. Data lifecycle issues thus enter into play here.

[4]. ***Mashing-up***. Also for the mash-up movement, i.e. web applications that allow developers with relatively little technical skills to combine, quickly and easily, existing content (geographic data, pictures, videos, news …) and functionalities in new ways from different sources, standards are obviously critical to easily integrate data.

[5]. ***Documentation and metadata***. At a different level, documentation and adequate use of metadata is what makes resources (re-) usable in the first place. Standardising documentation in the form of standard templates would facilitate developers and users. Consensus on basic sets of Metadata agreed in the community is also of utmost importance for an easy identification and tracking of resources independently from their physical location. This is critical in the emerging infrastructures and there is a big interest and a movement towards metadata standardisation, not only in Europe and the USA, but also e.g. in Australia.

[6]. ***Validation of language resources***. In order to be able to establish a certified quality validation of LRs (an issue that is coming-up more and more often) conformity to an accepted standard is a requirement.

[7]. ***Evaluation campaigns: shared tasks***. If we want to evaluate and compare the results of different methods, approaches, or technologies, it is important to have data encoded and annotated according to a common format that different groups need to be able to process and use. Here standards clearly play a fundamental role. In fact, many de-facto standards find their origins in evaluation campaigns or shared tasks and then become commonly used in the related sub-community (e.g. CoNLL). Therefore, it must be recognised that such initiatives play an important role also in introducing/disseminating the use of standards

[8]. ***Collaborative creation of resources***. Collaborative ways of creating or updating/enhancing LRs represent a recent trend in

[6] www.kyoto-project.eu

[7] http://verbs.colorado.edu/semlink/

the LT community. To fully exploit the potential of web-based collaboration, again common formats and annotation schemes have to be employed, so that distributed annotation, editing and data aggregation tools can be easily developed.

[9]. ***Preservation***. As IT evolves, both data resources and tools need to be ported to new systems, encodings etc. Storing data and developing tools according to widely accepted or official standards should thus facilitate their portability and help avoiding mismatches. Also, standards would make preservation easier as they would allow resource structures and content to be accessible (and understandable) also in time.

4 Strategies and recommendations

This section leads to the identification of a number of strategies and actions recommended by FLaReNet for achieving full interoperability in the Language Resource/Technology sector.

4.1 Address Semantic/content interoperability

Until now we have mostly tackled the problem of syntactic interoperability, i.e. the ability of systems to process exchanged data either directly or via conversion. Pivot formats, such as GrAF, attempt to solve syntactic interoperability, enabling merging and easy transduction among formats. Semantic interoperability, i.e. ability of systems to interpret exchanged linguistic information in meaningful consistent ways (e.g. through reference to a common set of categories), still remains unattained, as it is much more difficult. Linguistic characteristics of different languages, as well as different linguistic theoretical approaches play a big role in this. Interoperability of content is however desperately needed in the current landscape (e.g. in the scenarios [1], [2], [3], [4], [8] above). A good and practical rule (already recognised as a basic principle in EAGLES) is to define the standard as the lowest common denominator, at the maximal level of granularity. But, to arrive at this point large confrontations among experts are required. A recent effort in this direction is represented by ISOCAT (Kemps-Snijders et al. 2009), but more initiatives should be brought forward in order to maximise and accelerate the process.

4.2 Push Linked Data and Open Data

Interoperability through Linked Data could mean to be able to link our objects of linguistic/semantic knowledge with corresponding knowledge in other fields, and therefore to converge both within the field and outside with other fields. This would be very beneficial in scenatrios like [3] and [4]. To achieve maximum results, data needs to be open as much as possible, or the potential exploitation advantages will be lost. We must therefore closely monitor and participate to the Linked Open Data[8] initiative, connected to issue of semantic interoperability, to understand and enhance the potentialities for our field.

4.3 Develop tools that enable the use of standards

In order to increase the availability of sharable/exchangeable data, we must foster the development and availability of tools that enable an easy use of standards.

4.4 Incentivise web services platforms

Web service platforms (as in scenario [2]) certainly offer an optimal test case for interoperability and possibly a good showcase to demonstrate the need and advantages of the adoption of standards. Such platforms need both syntactic and semantic interoperability and thus can also function as an evaluation ground for interoperability issues. Projects like Language Grid, U-Compare and PANACEA could thus be seen as models for platforms providing LT services. A possible concrete action in this direction could be to compel players to deploy results of (publicly funded) projects as web-services that can be used, tested and called by others.

Cloud-based service architectures could also be leveraged as enablers for LT development.

4.5 Experiment with collaborative and crowdsourcing platforms.

The use of the collaborative paradigm to create language resources (in [8]) may become a means to encourage or even compel standardisation and - as a consequence - to share all the more the burden and cost of resource creation. Also crowdsourcing for shared resources is somehow linked to interoperability, requiring commonly accepted specifications. Collaborative develop-

[8] http://linkeddata.org/

ment of resources would create a new culture of joint research.

4.6 Establish a collaborative multilingual annotation plan

A collaborative approach to the creation of multilingual, possibly parallel, annotated data would also help maximise the visibility, use and reuse of resources, while at the same time encouraging exploratory diversity. A huge multilingual annotation pool, where everyone can deposit data annotated at every possible different linguistic level for the same resources, or for diverse resources, should be defined as a specific type of collaborative initiative for resource annotation (Calzolari 2010). This could create a fruitful (community driven) exchange between most used annotation schemes and establishment of best practices. Such an initiative would also be extremely beneficial for infrastructures like META-SHARE.

4.7 Support evaluation and validation campaigns

As mentioned in [7], evaluation campaigns help in standardisation. The lack of a European evaluation body that coordinates and prioritises evaluation efforts is an issue that finally hampers interoperability. Shared tasks should therefore become more prominent as loci where interoperability is foregrounded, where standards are pushed forth and thus the occasion to make progress in standardising not only resources but components as well. The possibility of having official validators[9] for compliance to basic linguistic standards can/should also be investigated. This could be used to provide the community, through with validation services for the resources to be shared.

4.8 Set up Interoperability Challenges

Along with the previous proposal, the idea of organising interoperability challenges, discussed by Nancy Ide and James Pustejovsky at a SILT Workshop (April 2011), should be enforced and supported, as an international initiative to evaluate and possibly measure interoperability. This could speed up the dissemination of standards and drive interoperability forward[10]. The NLP community should be involved and an overall challenge should be defined that explicitly require the use and integration of multiple data formats, annotation schemes, and processing modules, so that players will be highly motivated to adapt, adopt, use standards and common format and could start seeing the advantages they offer.

4.9 Standards should be open and simple

As a basic rule standards should be open, simple, and relatively non-invasive to facilitate their adoption. For example, people should continue to be allowed to program/mark-up as they wish, but there should be well-formed points of contact that act as bridge between data and code that the community needs to come up with.

4.10 Maintain a repository of standards and best practices

Information on standards is essential. A repository of standards and best practices must be created and kept alive. A preparatory initiative was started within FLaReNet[11], but dedicated effort must be devoted to create and support a repository of standards and best practices so that it assumes also the effect of a cultural initiative. A repository of standards could obviously be linked to a repository of open data compliant with the them. This would maximise the benefits.

4.11 Organise awareness initiatives

Awareness about the existing standards and the motivations behind them is one of the key factor for enlarging their adoption. Educational programs should therefore be launched to explain, promote and disseminate standards especially to students and young researchers (e.g. through tutorials at conferences, summer schools, seminars…). Steps could be taken to include standardisation in regular university curricula. Also, effective ways to demonstrate the return of investment (ROI) of interoperability must be sought. Adapting one's tools and resources to standardised common formats in fact requires some investments that players may not be willing to make unless the clearly see advantages.

4.12 Set up a Standard Watch

At present, no mechanism is available to watch when a discipline deserves standardisation. We should create a permanent Observatory or Standard Watch. TAUS for example has announced an Interoperability Watchdog initiative that goes in the right direction. Examples of deficiencies

[9] http://validator.oaipmh.com/ or the OLAC validator http://www.language-archives.org/tools/xsv/

[10] https://sites.google.com/site/siltforum/files

[11] http://www.flarenet.eu/?q=FLaReNet_Repository_of_Standards_and_Guidelines

from the European side are the lack of support to official standardisation initiatives for important topics such as Space and Lexicon-Ontology, which have also an economic potential. As standardisation is a slow process and the ROI is not immediate, funding agencies should be more present in the initiatives. .

4.13 Establish a Quality Certificate

Work is needed towards the definition and establishment of some kind of Quality seal, on the model of the "Data Seal of Approval"[12], to be endorsed by the community. The Data Seal of Approval is a quality sign for resources (data) that provides a certification for data repositories to keep data visible and accessible and to ensure long term preservation. Similarly, efforts should be made to encompass not only data for archiving, but also for dynamic exchange and also for software components. For example, there is a requirement for CLARIN centres to comply with certain standards. This is linked to the concept of "preservation" and sustainability. Infrastructures like META-SHARE could introduce some mechanisms (possibly socially based) for assigning quality scores to resources and tools, also evaluating them for compliance to standards/best practices. Systems of "penalties" could be devised, as well, for not complying data resources.

4.14 Link up to web content-related standards

Collaboration and synergies must be enforced with ISO, W3C and other multilingual web content-related standards, which in the case of LT can be seen as more basic levels of representation that can to ensure the (potential) integration of LT/NLP technologies into present and future web content products. Multilinguality should be incorporated in standards, e.g. ISO standards should be instantiated/generalised for as many languages as possible, which does not always happen at present. A recommendation to standardisation bodies must be to test/apply standards multilingually.

4.15 International collaboration

In particular for standards it is important that initiatives are taken at a truly international level. This means going beyond European initiatives.

[12] http://www.datasealofapproval.org/

5 Conclusions: Operationalising Standards

A recurrent request from industrials in many recent meetings, such as the META-NET Vision Groups and the META-Council, is: "give me the standards and give me open data".

The major recommended step for an interoperability framework is operationalising standards, in the sense of making standards finally "operational" and come up with operational recommendations. Standards must be usable and actually used; otherwise they are of no relevance.

A step forward in this direction is to make standards open. However, there is no single definition of the term. The minimum requirements for open standards are availability and accessibility for all, detailed documentation and possibility to be implemented without restrictions. Publicly available standards with public specifications in fact promote their usage and adoption (Perens, 2010; Krechmer, 2006).

The basic pre-conditions to operationalise the standards and the essential steps to be taken need to be outlined. Some of these steps and conditions are summarised below.

5.1 Technical conditions

Common metadata. This is a commonly recognised pre-condition, in all the most important infrastructural initiatives: ELRA, LDC, CLARIN, META-SHARE.

Explicit semantics of metadata. Explicit semantics of annotation metadata/data categories is essential. A mechanism to be used can be ISOCat: even if there are still many problems, it is at the moment the only available instrument that allows the definition of data categories at a persistent web location and to reference them from any annotation scheme.

High level metadata is not the only set of values that are recorded in ISOCat. Until now, the linguistic categories within ISOCat have been mostly recorded from the EAGLES, MULTEXT-East and LIRICS projects (e.g. morpho-syntax, extended also to Semitic, Asian and African languages), and terminology starting from LISA and ISO-12620 sets of values. Recently ISOCat is enriched by the CLARIN project with the need of Social Sciences and Humanities (SSH) in mind. These metadata however are not enough for NLP. This gap (from SSH to NLP) is currently filled in META-SHARE and an effort must be done to involve a broad community of resource developers/users.

Creation of data category selections for the major standards/best practices would increase convergence towards common data categories. This would help taking a step towards semantic interoperability. Funding agencies could encourage entering data categories and selections in ISOCat, which could become a useful instrument if broadly used.

Tools that facilitate the use of standards. It is of utmost importance to develop (online) tools that hide the complexities of standard formats and allow for easy usage of standards and easy exportation/mapping to the standards. The development of converters from/to the major standards/best practices/common formats to other endorsed/official standards is thus recommended. This is true in particular for infrastructures like META-SHARE where best practices should be promoted also through tools.

5.2 Infrastructural conditions

A common (virtual) repository as an easy way to find the most appropriate standards. An international joint effort should take care of the indexing of different standards and best practices, to ease their finding and to keep track of the status and different versions and their history. This is critical for infrastructures like META-SHARE that should also be able to recommend standards and best practices for the resources made visible through them, in particular for the new ones.

Common templates for documentation. Currently, resource and tool documentation is often not adequate, ranging from too poor to too heavy. Nevertheless, documentation of resources is essential for reaching common understanding and practically for exchange and re-use. Therefore, a consensual set of templates for resource documentation should be devised and disseminated, with actions to facilitate their adoption.

Provide a framework that facilitates testing. Test scenarios to verify compliance are needed.

An interoperability framework for/of web services. Operationalising standards could also mean that they should be based on an interoperability framework for/of web services. We should therefore deploy linguistic services based on standards. A key point here is workflows (see the success of the KYOTO project).

"**Meta-interoperability**" among standards. We should also speak about "meta-interoperability" among standards and understand what it means operationally. Standards must constitute a coherent framework, i.e. they must be able to speak with each other. This refers to the LR specific ecology framework (as an integrated system).

5.3 Social and cultural conditions

Involvement of the community. The community as a whole must be involved in standardisation processes. It is recommended that researchers, groups and companies involved or interested in resource development/annotation/validation actively contribute to the definition of LT standards. Initiatives must be defined to change the community mentality into a "social network" for scientific collaboration, as community-level active participation is critical for attaining true interoperability. In fact, the wider the participation to such initiatives, the more robust and valid the standards would be. One possible way of making work on standardisation appealing could be to establish a framework for the citation of resources, like for publications, and measure their impact factor.

Dissemination (but not forcing). The potential and advantages of standardisation must be disseminated, standards pushed, incentives to the use of standards possibly devised, but people must not be obliged to conform. Standards must not be seen as an overhead, but people should feel that they want to use standards because it's in their own interest.

Interoperability as valid research area. Community mentality should be changed also to accept interoperability and standardisation as academically valid research areas.

Link to sustainability. In general, a virtuous circle must be established between standard-definition, adoption, feedback, and their interoperability.

Acknowledgements

This work has been supported by the FP7 FLaReNet project (ECP-2007-LANG-617001) and META-NET (FP7-ICT-4 – 249119: T4ME-NET).

We thank the whole FLaReNet community, which through participation to the meetings and discussions, considerably contributed to shaping the results and ideas reported in this paper.

References

Bel, N. et al. 2009. *CLARIN Standardisation Action Plan.* CLARIN http://www.clarin.eu/node/2841

Bel, N. et al. to appear. *The Standards' Landscape Towards an Interoperability Framework*

Bel, N. 2010. Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies: PANACEA. In *Proceedings of the 26th Annual Congress of the Spanish Society for Natural Language Processing (SEPLN)*, Valencia.

Bosma, W., et al. 2009. KAF: a generic semantic annotation format. In: *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon* (GL 2009). Pisa.

Calzolari, N., et al. 2002. "Broadening the scope of the EAGLES/ISLE lexical standardization initiative". In *Proceedings of the 3rd workshop on Asian language resources and international standardization (COLING '02)*, vol. 12. pages 1-8, Taipei.

Calzolari N., et al. (eds.). 2009. *Shaping the Future of the Multilingual Digital Europe*, 1st FLaReNet Forum, Vienna.

Calzolari, N. 2010. Invited presentation at the COLING 2010 Panel. Beijing.

Declerk, T. 2006. SynAF: Towards a Standard for Syntactic Annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 229-232, Genoa.

EAGLES 1996

http://www.ilc.cnr.it/EAGLES96/home.html

Fellbaum C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Francopoulo, G., et al. 2006. Lexical markup framework (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06*), pages 233-236, Genoa.

Francopoulo G., et al. 2009. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*. 43(1): 57-70.

Ide, N. 1998. "Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora." In *Proceedings of the First International Language Resources and Evaluation Conference (LREC'98)*, pages 463-470, Granada.

Ide, N and L. Romary. 2007 Towards International Standards for Language Resources. In Dybkjaer, L., Hemsen, H., Minker, W. (eds.), *Evaluation of Text and Speech Systems*, Springer, Dordrecht, 263-84.

Ide, N. and K, Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop at ACL 2007*, pages 1-8, Prague.

International Organization for Standardisation. 2002. *ISO:16642-2002. Terminological Markup Framework*. http://www.loria.fr/projets/TMF/

International Organization for Standardization. 2008. *ISO DIS 24611 Language Resource Management - Morpho-syntactic Annotation Framework (MAF)*. ISO/TC 37/SC4/WG 2.

International Organization for Standardization. 2008. *ISO DIS 24611- (1,2,3,4,5,6) Language Resource Management - Semantic annotation framework (SemAF)*. ISO/TC 37/SC4/WG 2.

Kemps-Snijders M., et al 2009. "ISOcat: Remodeling Metadata for Language Resources". *International Journal of Metadata, Semantics and Ontologies (IJMSO)*, 4(4): 261-276.

Krechmer K. 2006, Open Standards Requirements. *The International Journal of IT Standards and Standardization Research*, 4(1): 43-61.

Lionel C. and Éric de la Clergerie. 2005. Maf: a morphosyntactic annotation frame work. In *Proceedings of the 2nd Language and Technology Conference* (LTC'05), pages 90-94, Poznan.

Marcus, M. P., B. Santorini, M.A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19 (2): 313-330.

Nivre, J. et al. 2007. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915-932, Prague.

TAUS. 2011. *Annual Plan 2011: Translation Innovation Think Tank Interoperability Watchdog.* http://www.translationautomation.com/images/stories/pdf/taus-annual-plan-2011-extended.pdf

TAUS (2011) *Report on a TAUS research about translation interoperability*. February 25, 2011. http://www.translationautomation.com

Toral A., et al. 2011. "Towards a user-friendly webservice architecture for statistical machine translation in the PANACEA project". In: M. L. Forcada, H. Depraetere, V. Vandeghinste (eds.) *Proceedings of the 15th EAMT 2011*, pages 63-70, Leuven.

W3C 2007 EML: Emotion Incubator Group, W3C Incubator Group Report, 10 July 2007.

W3C 2009 EMMA: Extensible MultiModal Annotation markup language, W3C Recommendation, 10 February 2009.

Zydron A. 2008. OAXAL. What Is It and Why Should I Care? *Globalization Insider*. http://www.lisa.org/globalizationinsider/2008.