Linking Northern European Infrastructures for Improving the Accessibility and Documentation of Complex Resources

Gyri Smørdal Losnegaard[1],*Gunn Inger Lyse*[1], *Anje Müller Gjesdal*[1],
Koenraad De Smedt[1], *Paul Meurer*[2], *Victoria Rosén*[1,2]

(1) University of Bergen, Norway
(2) Uni Research, Norway

gyri.losnegaard@lle.uib.no, gunn.lyse@lle.uib.no, anje.gjesdal@uib.no,
desmedt@uib.no, paul.meurer@uni.no, victoria@uib.no

ABSTRACT

This paper describes our integration efforts in two Northern European language infrastructures. Specifically, this work has been a collaboration between the META-NORD team at the University of Bergen and the INESS project, a large treebanking infrastructure project in Norway, in developing and documenting two complex resources, as well as making these accessible to the R&D community.

KEYWORDS: metadata, IPR, Treebanks, research infrastructure, Northern Europe, INESS, META-NORD, CLARIN.

Gyri Losnegaard, Gunn Inger Lyse, Anje Müller Gjesdal, Koenraad De Smedt, Paul Meurer and Victoria Rosén 2013. Linking Northern European infrastructures for improving the accessibility and documentation of complex resources. *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013.* NEALT Proceedings Series 20 / Linköping Electronic Conference Proceedings 89: 44–59.

1 Introduction

Several large-scale infrastructures are currently under development across Europe for the distribution of research results, data and tools in the Humanities and Social Sciences. The various initiatives differ in the disciplines that they cover and the scope of their goals, but they have the common aim of fostering the reuse and sustainability of resources and tools. Such initiatives require a considerable effort to harmonize metadata schemes, adhere to standards and solve intellectual property rights (IPR) issues (Hinrichs et al., 2010; Duin et al., 2010; Gavrilidou et al., 2011, 2012). Moreover, since different infrastructures co-exist at different levels, infrastructure initiatives will increasingly need to focus on establishing best practice criteria to facilitate the linking of infrastructures and ensure their interoperability.

This paper describes our integration efforts in two Nordic and Baltic language infrastructures. Specifically, this work has been a collaboration in Norway between the META-NORD and INESS projects in developing and documenting two complex resources, as well as making these accessible to the R&D community.

META-NORD (Vasiļjevs et al., 2012) (2011–2013) has been a CIP ICT-PSP (Information and Communication Technologies Policy Support Programme) project aimed at creating an open infrastructure to promote the accessibility and reuse of language resources and technologies (LRT). Its consortium includes organizations from all the Nordic and Baltic countries. Among its main results has been the documentation, rights clearance, licensing and sharing of many language resources via the META-SHARE[1] catalogue and repository, thereby making LRT more readily available to R&D.

INESS (Rosén et al., 2012) (2010–2016) is an ongoing project at the University of Bergen (Norway) and Uni Computing (a division of Uni Research, also in Bergen), aimed at establishing an Infrastructure for the Exploration of Syntax and Semantics. It is funded by the Research Council of Norway and the University of Bergen. One of its activities is the implementation and operation of a comprehensive open treebanking environment in which a large number of treebanks can be hosted and made accessible through advanced web interfaces for search and visualization[2]. The other is the development of a large parsebank for Norwegian with a wide coverage LFG grammar and lexicon.

INESS has cooperated extensively with META-NORD throughout the project lifetime of the latter. The main field of cooperation has been to collect and develop treebanks, to make these more accessible in standardized ways, to document them through metadata and to link them through alignment in parallel treebanks, as will be explained in more detail below. The results of these activities are also being integrated in CLARINO (the Norwegian part of the CLARIN network) and in *Språkbanken*, a language technology resource collection for Norwegian, hosted at the National Library of Norway.

Among the challenges faced was resolving the sometimes conflicting requirements for creating and integrating treebanks in the INESS treebanking infrastructure, on the one hand, and documenting them with metadata in META-SHARE, on the other. We have made some initial efforts towards consolidating the metadata creation and description between the two infrastructures. While the integration of existing resources is an essential part of building infrastructures, we

[1] http:/meta-share.tilde.lv

[2] The different aspects of the INESS treebanking infrastructure, from visualization via interactive annotation of treebanks to treebank search, are described in detail in Rosén et al. (2012).

argue that infrastructure initiatives will increasingly need to focus on establishing best practice criteria to be applied at the data creation stage. This will in turn facilitate the linking of infrastructures and ensure their interoperability.

The rest of this paper is structured as follows: in section 2 we describe the integration and linking of treebanks in the INESS infrastructure. Section 3 addresses the challenges encountered in their documentation: metadata compilation (section 3.1), IPR clearance (section 3.2) and metadata creation in META-SHARE (section 3.4), the latter exposing special challenges in the description of complex resources. We present our work integrating the two infrastructures in section 4, before we finally, in section 5, provide suggestions for best practices in terms of standardization of formats, metadata, IPR and integration between infrastructures.

2 Creating, integrating and linking treebanks in the INESS infrastructure

In the cooperative effort between INESS and META-NORD, two parallel treebanks were constructed: the Sofie Parallel Treebank and the Acquis Parallel Treebank.

The Norwegian novel *Sofies verden* (*Sophie's World*) (Gaarder, 1991) was chosen as a suitable basis for parallel treebanking because it is linguistically rich and professionally translated into many languages, and because some monolingual treebanks already existed for text selections from this material in some languages in the META-NORD area. Existing treebanks for this material had been made in the context of the Nordic Treebank Network (NTN), funded by the Nordic Language Technology Program (2001–2005). Annotation files for Danish, Estonian, German, Icelandic and Swedish were obtained via Tekstlaboratoriet (the Text Laboratory) at the University of Oslo, and a treebank for the English version was obtained from the SMULTRON parallel treebank (Stockholm MULtilingual TReebank)[3] (Adesam, 2012). These treebanks were documented, processed and supplemented with new treebanks for the Norwegian, Georgian and Finnish versions.

The Sofie treebanks made available through INESS and META-NORD show considerable diversity with respect to both the language families that are covered and the linguistic formalisms that are represented. The Sofie Danish Treebank is a dependency treebank, semi-automatically annotated according to the guidelines used to create the Danish Dependency Treebank and automatically converted to TIGER-XML by the DTAG program. The Sofie Estonian Treebank is a constraint grammar (CG) treebank, automatically parsed with a CG parser assigning syntactic function labels and enhanced with manually added constituencies. The Sofie Icelandic Treebank is a constituency treebank which was manually annotated by the late Gunnar Hrafn Hrafnbjargarson. The Sofie Swedish Treebank is a dependency treebank, automatically created with the Maltparser tool. The Sofie German Treebank was annotated with the Annotate tool, followed by an automatic deepening of the flat syntax trees. The Sofie Finnish Treebank is a manually annotated dependency-CG treebank created by the UHEL FinnTreeBank team for FinnTreeBank and META-NORD. The Sofie Norwegian Treebank was automatically parsed with an LFG grammar developed in the NorGram and INESS projects, producing c-structures and f-structures; the analyses were manually (interactively) disambiguated by the use of discriminants (Rosén et al., 2009, 2007). The Sofie Georgian Treebank was similarly processed, but with a Georgian grammar developed by Paul Meurer. The Norwegian and Georgian treebanks are downloadable in Negra/Tiger XML format.

Furthermore, small pilot treebanks were constructed for the JRC Acquis Multilingual Parallel

[3] http://www.cl.uzh.ch/research/paralleltreebanks_en.html

Corpus of EU/EEA law texts,[4] which provides materials from a different genre. The standardized and uniform structure of the corpus and its texts facilitated the selection of a document of appropriate length which was available in all the relevant META-NORD EU-languages, and for which translations existed also for the non-EU languages Icelandic and Norwegian. Dependency annotations were produced for the Danish, Estonian, Finnish, and Swedish Acquis texts, and constituency annotations for Icelandic. INESS provided annotations of the Norwegian and English versions of the selected Acquis document, which were parsed with LFG grammars and manually disambiguated.

These existing and new treebanks were combined and integrated into INESS, providing both long-term physical storage and a platform for research and development of treebanks. INESS supports most of the standard input formats (TigerXML, CoNLL-X, CG3-dependency, Penn Treebank II bracketing or XLE prolog) and with the exception of one treebank that had to be converted from non-standard notation to one of the standard input formats, the integration was seamless. Monolingual annotations for each of the collections were then aligned at sentence level and their alignment was made downloadable in XML stand-off format. The parallel treebanks were made searchable, and individual sentences from the treebanks are visualized side by side, as illustrated in Figure 1.

Besides the Sofie and Acquis monolingual treebanks, which provided the basis for parallel treebanks, the INESS project has also made several other freestanding monolingual treebanks based on different sources available. Some of these were selected for documentation in cooperation with the META-NORD project. These include treebanks for Finnish, Icelandic and Norwegian in the linguistic area of META-NORD as well as for a number of other languages inside and outside the linguistic area of META-NET (including smaller languages in the META-NORD geographical area such as Faroese and Northern Sami).

3 Resource documentation

Adequate documentation is essential both in order to create trustworthy metadata and to resolve IPR issues, and it presupposes correct and reliable information about formats, IPR and resource creation. An important part of resource documentation consists of obtaining this information and clearing the rights for the resource so that it can be used for the intended purposes. In order to exploit the expertise and standards being developed within the META[5] network and to avoid a duplication of efforts, it was decided to delegate metadata and IPR issues in INESS to META-NORD. The parallel treebanks, as well as each monolingual treebank, were documented with structured metadata using META-SHARE.

Although the existing treebanks were originally created in NTN, an advanced research network, their documentation and IPR clearance proved especially challenging, as described in sections 3.1 and 3.2. Moreover, the complexity of the parallel treebanks brought about special documentation requirements which META-SHARE did not allow for in a straightforward way (3.3), forcing us to come up with some expedient solutions (3.4).

3.1 Metadata compilation

For the treebanks developed in NTN, substantial efforts were invested in recovering the information required to make treebanks available for download, directly or indirectly, through

[4] http://langtech.jrc.it/JRC-Acquis.html

[5] http://www.meta-net.eu/

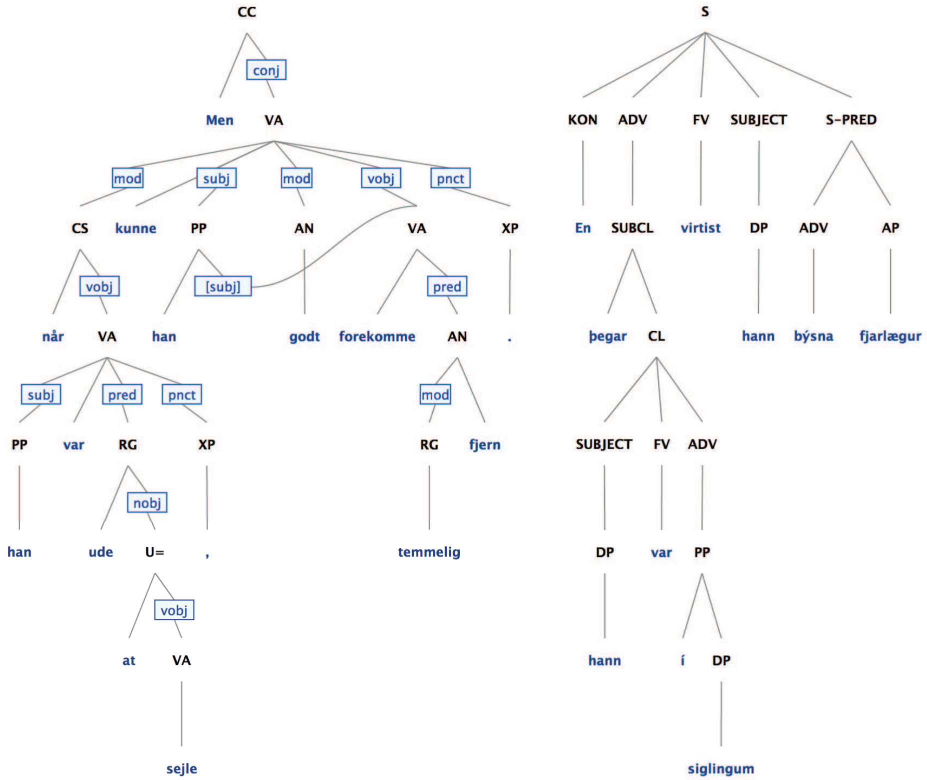

Figure 1: Visualization of sentences from the Parallel Sofie Treebank: Danish and Icelandic

META-SHARE. A major challenge was presented by the fact that NTN project results and documentation were no longer maintained and partly inaccessible. Some information was available on the NTN webpages and as metadata encoded in the XML header of some of the annotation files. However, a large part of the information necessary to create adequate metadata descriptions and to ensure rights clearance had to be recovered otherwise. This was done partly by approaching NTN network participants, and partly by studying the encoding of the annotation files. By searching for tag sets, annotation features, etc. on the web, we identified the treebank types (the formalisms used), input formats and in some cases also the origins of the annotations. Our NTN contacts then verified our educated guesses and supplied them with additional information. Some of the documentation which was missing from the webpages due to inactive URLs was uploaded to the Copenhagen Dependency Treebank's Google Code repository, including a few HTML pages documenting the tools developed in NTN and the common representation formats used in that project. The recovered information was harmonized with META-SHARE and included in the metadata records for the monolingual treebanks, and will be further maintained via the INESS webpages.

For new treebanks, the following metadata were collected from the developers and harmonized with the META-SHARE schema for text corpora description:

- Annotation mode (automatic, semi-automatic, manual)
- Grammar/parser (type and/or name of tool)
- Grammar origin/creator (project, person(s), name(s) of annotator(s))
- Grammar type/formalism (constituency, dependency, LFG, etc.)
- Output format (Tiger XML, CoNLL, etc.)
- Tagset (documentation URL, taglist, etc.)
- Terms of use and license information

3.2 IPR clearance

Rights clearance is an attempt to balance interests. On the one hand copyrighted material must be protected. On the other hand it should be possible to access material for innovative work and to allow uses of copyrighted material that may be beneficial for society. A well-developed and easy-to-grasp legal system that protects tools and resources in a general and transparent way is an incentive for people to create, share and use tools and resources. Infrastructures provide an invaluable framework for establishing best practices for clearing rights using standardized agreements.

Fixed licenses are available for instance through Creative Commons.[6] Moreover, a set of fixed licenses specifically made for the sharing of LRT is available through META-SHARE, including also standardized depositor's agreements that regulate the rights and obligations that hold between the copyright owner and the distributor of a resource. Similarly, the CLARIN infrastructure is developing editable templates for end-user licenses as well as for depositor's agreements. CLARIN licenses, however, were still under development when the work reported here was initiated.

[6] http://creativecommons.org/

In our experience, the copyright holders as well as the researchers negotiating the user terms find standardized legal texts or templates reassuring.

Rights clearance is time-consuming work, and in order to facilitate the long-term reuse of resources, it is essential that rights clearance is done with a long-term perspective in mind. Thus, extensive work has been done on rights clearance both for source texts and for the grammatical annotations of the treebanks. Rights were negotiated separately for source texts and annotations. It was endeavored, to the extent possible, to resolve IPR issues uniformly, using common or similar agreements for resources with a common or similar origin.

For the Sofie treebanks developed under NTN, rights had been cleared for the original source text and its translations in an exemplary way, but only for use in the context of that project. These agreements illustrate a limitation that must be avoided in a long-term infrastructure: in order to secure maximal reuse, permissions must be granted to user groups that are minimally restricted and for as general purposes as possible. For instance, the NTN rights clearance only allowed the acting research group to create one specific *derivative*, namely a corpus to be browsable (but not downloadable) online under certain restrictions.[7] In the context of META-NORD the rights clearance for this material therefore had to be renegotiated to allow the distribution of the treebanks for general use in language technology R&D.

A depositor's agreement was signed with Aschehoug, the publisher of the Norwegian original of *Sophie's World*. Aschehoug also wrote a recommendation letter to the publishers of the translations, which were subsequently contacted. Signed depositor's agreements have so far been obtained for the Swedish, Estonian, Danish, Icelandic, German and Georgian translations, while the English version was already freely available through SMULTRON. For the Finnish translation, unfortunately, the translator who holds the rights to the text did not give permission to distribute the treebank. The depositor's agreements used for the Sofie materials are based on the standard META-SHARE template, and have restrictions on the redistribution of the texts while allowing the use of the texts for R&D purposes in language technology, the most important purpose of META-NORD.

While rights clearance for source texts often requires a certain amount of negotiation since the rights holders did not originally intend the texts to be used for R&D purposes, clearing rights for linguistic annotations is in principle easier, since these annotations are designed specifically for such purposes. The main challenge with linguistic annotations is thus not the rights clearance itself, but the identification of the rights holders in cases where this has not been properly documented. Of the annotations created in NTN, the webpages made no mention of IPR ownership or licenses, and only a few treebanks had creation data encoded in the annotation file itself, so that it remained a challenge to identify the creator(s) and rights holders of several annotations. The solution, reached in agreement with NTN network coordinator Joakim Nivre and project co-workers Mathias Buch-Kromann and Kadri Muischnek (the creators of the Swedish, Danish and Estonian annotations respectively), was for the network coordinator to sign a common depositor's agreement on behalf of the annotation group and for all annotations created within the project. The approach of using one common agreement for all annotations developed under NTN was also adopted for the new annotations developed in META-NORD, with a few exceptions for treebanks made by third-party collaborators. Treebanks borrowed from unrelated projects were released under the conditions specified for that project. New annotations developed for META-NORD by third parties were released under the

[7] A derivative is a product that contains a substantial, or significant, part of an original resource.

same license as treebanks developed by META-NORD partners, but with individual depositor's agreements since the IPR holders were not project members.

In the META-NORD parallel treebanks, treebank alignments constitute pairwise stand-off layers of annotation. These were created in and by INESS and quality assured by META-NORD partners, and the rights to these annotations remain with the project consortium. The alignments are, however, independent annotations, and must also be supplied with a license formalizing the terms of use.

Our experiences with individual rights clearance for each layer of the treebank (source text, linguistic annotation and alignment) clearly demonstrate that the IPR aspect of complex resources is, indeed, very complex. In the process of resolving IPR issues, questions and doubts constantly arose as to whether our approaches were good enough, or whether our solutions were legally sound. Should, for example, the grammatical annotation in a treebank and its source text be considered as separate with respect to licensing? Consider Figure 1, in which the words are 'leaves' in the grammatical sentence analyses. Can the grammatical annotation, being tightly intertwined with the source text, be considered as a resource completely detached from the text that it describes, or is an annotation the *combination* of source text and linguistic marking? The answer to such a question has legal implications, since the linguistic annotations could conceivably be shared for further research under a fairly open license whereas the source text *qua* source text may remain licensed under considerably stricter terms of use. The user perspective is also an important consideration, because it might not be ideal to confront the user of the treebank with several licenses, one for each monolingual treebank and one for each annotation layer.

In project internal discussions it was tentatively concluded that it is possible to provide separate terms of use of the source text and the annotations as long as the user is explicitly told which conditions hold for which part. Annotations with for example a CC-BY license[8] can then be used freely (with attribution) even when the source text is more restricted, as long as the user has been made aware that if the source text is extracted from the annotation, the source text license applies. In other words, it is not *always* the case that the annotation will be restricted by the license of the text.

The discussion originally revolved around the Acquis treebanks, whose source texts are from the Acquis Communautaire[9] and are in the public domain, with no rights holder or restrictions of use. For the META-NORD Acquis treebanks, however, texts for the relevant languages were selected from the JRC-Acquis multilingual corpus[10], where Acquis documents have been aligned at document level. This slightly complicated the picture, since this aligned corpus applies specific terms of use which are stricter than CC-BY. It was made explicit in the META-SHARE description of these treebanks that the Acquis documents are in the Acquis Communautaire, which is available via the EUR-lex webpage, and that the same documents are available in the JRC Acquis corpus under specific conditions.

It is sometimes extremely time-consuming, if not impossible, to establish contact with all copyright owners. Standard corpora collected from many relatively small text excerpts are typical

[8] Public Creative Commons license with attribution, allowing the user to share and to modify the resource, also commercially, provided that the creator and/or licensor of the original resource is attributed; see `http://creativecommons.org/licenses/by/3.0/`.

[9] The total body of European Union (EU) law applicable in the the EU Member States, distributed via EUR-Lex (`http://eur-lex.europa.eu/`).

[10] `http://ipsc.jrc.ec.europa.eu/index.php?id=198`

examples. Fair use of quoted text fragments may sometimes be invoked in such situations. However, if the author of a text does not want the text to be distributed, for whatever reasons, that decision should be respected. In some cases, the original permission from each text contributor is recorded, but only concerned the intended use within the project creating the corpus, not its reuse; such short-sighted arrangements make it necessary to renegotiate the terms of use for new research within a research infrastructure.

The complexity of factors involved in the annotation of (possibly) copyrighted text remains a challenge which calls for juridically skilled scrutiny. We propose as the safest solution to apply one overall license to each treebank as a whole, i.e., to release a treebank under the license with the most restricted conditions of use. For the Sofie treebanks, for instance, even though the linguistic annotations were cleared for a CC-BY license, the user terms of the source texts were applied, restricting the use of these treebanks to language technology R&D purposes. The Acquis treebanks, however, based on source texts that are in the public domain, could be licensed under an open source license (CC-BY) as agreed with the creators of the linguistic annotations.

Despite the ethical and legal considerations, the decision to use only one license per treebank was primarily made out of consideration for the user, who cannot be expected to have expert knowledge about IPR and licenses. A typical treebank user will probably not be interested in the different levels and details of licensing, and will be inconvenienced by having to relate to more than one set of user terms which are often hard to interpret.

An important lesson was learned from a setback experienced when one of the source text rights holders refused to release the annotated source text, even if only for R&D purposes. The problem was identified only *after* a new annotation had been developed, and this demonstrates the importance of establishing work order routines in treebank development. Clearing rights for the source text should ideally be done as a preliminary step, before annotation. This case also suggests the advantage of having a ready-made, consistent and convincing line of arguments for use in the negotiation process. Establishing good routines for treebank development will at worst increase the chances for the resource in question to be released under a restricted license; at best it will allow for unrestricted, attributed distribution.

3.3 The description of complex resources in metadata

The description of complex resources is a general challenge that must be dealt with sooner rather than later in the development of LR infrastructures if we are to avoid a proliferation of ad hoc, nonstandard approaches towards handling them. The concept of a *complex resource* may have different interpretations, but at a very general level we will here define a resource as complex if it has several components, if it is multilingual, or if several tools or methods have been applied in the process of creating it.

Parallel treebanks are complex in at least two different respects. First, they are composed of several monolingual treebanks, which makes them diverse in terms of 'linguality' (i.e. a multilingual resource consisting of monolingual ones) and potentially also in terms of provenance, annotation type, IPR and licensing, etc. Second, monolingual treebanks are complex in their own right, having both a text component and one or more layers of annotation. This feature makes them complex in terms of metadata since it should be possible to describe a variable number of components and layers systematically, and to express clearly how the components and layers relate to each other.

The treebanks in question present complexities on all these levels, bringing forward specific requirements for their description in META-SHARE. Within the META-SHARE framework, monolingual treebanks can currently only be described satisfactorily at an appropriate level of detail if described with an individual metadata record for each monolingual treebank. As part of a parallel resource, the individual metadata descriptions must not only account for the range of resource specific features such as type, format, creation details and contact information, they must also represent relations to the other treebanks that constitute a parallel *collection*. META-SHARE allows the definition of relations in metadata, but since there are no standardized relations with fixed meanings, relations in META-SHARE are only meaningful to human users. It is not currently possible to filter or extract resources belonging to a certain collection.

The information common for all components of the complex resource must also be described, ideally without repeating this information for each individual component treebank. In the following section we describe how we ensured that the documentation requirements were met.

3.4 Metadata creation

The implementation of a metadata schema for the description of complex resources in META-SHARE was envisioned, but not accomplished, during the course of the META-NORD project. As suggested in Lyse et al. (2012), for instance, a schema for complex resources should make it possible to search and retrieve all parts of the complex resource, or to retrieve only the subpart that a user is looking for. It was therefore necessary to find an adequate way of representing such resources in a preliminary way until the provision of a more satisfactory solution becomes available.

META-SHARE represents a language resource as a metadata *record*, with mandatory and optional features for different types of resources. The mandatory set of features constitutes the *minimal description* of a resource. A treebank, which is a syntactically annotated corpus, is classified as a “TextCorpus”. A parallel treebank is a set of individual, monolingual treebanks based on texts that stand in a translational relation to each other. Of these, some or all may have been aligned, in our case at sentence level. The representation of parallel treebanks should ideally meet the requirements sketched out in the previous section: each monolingual treebank must be described at an appropriate level of detail, documenting all individual features, while at the same time preserving information about relations holding between the individual treebanks as well as information common for these treebanks, without unnecessary duplication.

Several approaches to the representation of parallel treebanks in META-SHARE were considered. The solution first proposed by the META-SHARE developers was to create one metadata record for the entire parallel treebank, using the feature “sizePerLanguage” to specify the number of sentences for each language. This would have been an acceptable option if all the treebanks had been developed within the same project, if they were of the same type, if they had the same input formats, annotation mode, IPR holder and so on. This was clearly not the case for our treebanks. Another, similar option would be to create one overall record and to add separate “corpusTextInfo” sections for each language module. The “lingualityInfo” feature, which indicates whether the resource is mono- or multilingual, and the “languageInfo” feature, specifying the relevant language, are both described in this section. It would thus be possible to specify language and ‘linguality’ for each component treebank, as well as other annotation features such as creator, type and format. However, the metadata about provenance, IPR holder, distribution, etc., can only be described in the section describing the overall resource. If the component treebanks, as in our case, have been created in different projects, have different source text and annotation rights holders and so on, this information cannot be structured in one metadata record with several "corpusTextInfo" sections. As a consequence, neither of these solutions allows for the level of detail required to describe each treebank properly. Equally important, there is no way of showing that the combination of the component treebanks is multilingual and aligned, and that the treebanks in effect constitute one, multilingual resource.

We thus opted for a resource description with one multilingual parallel 'mother' metadata record, and one record for each of its monolingual components. The metadata records were linked using a "relation" feature: the monolingual treebanks were related to the 'mother' resource with a "partOf" relation, and to their sister resources with an "alignedWith" relation. Our parallel treebanks are now represented as multilingual text corpora which list their language components both in the "sizeInfo" part and in the "relations" part.

4 Metadata links between infrastructures

While META-SHARE collects metadata for a large number of language resources and tools, the INESS system also needs to maintain metadata as documentation of its own resources. These metadata are used for presenting documentation about each treebank to the user, and can also be used for selecting treebanks based on desired features, e.g. language, license, provenance, etc. Importantly, this includes terms of use and licensing information which must be presented to the user and in many cases must be accepted by the user. In terms of usability it should also be made maximally explicit that a parallel treebank is not necessarily a uniform resource, but rather a collection of resources of potentially different provenance, type, and quality, and that aligned treebanks may not be directly comparable.

For reasons of efficiency and consistency, it is therefore important that the metadata in both infrastructures are not created and maintained separately, but are harvested and synchronized. Several solutions were considered. Considering the shortcomings in the META-SHARE schemas described in section 3.4, the ideal solution would be to define parallel treebanks in a proper way in the INESS system and to export relevant metadata to META-SHARE. However, this would necessitate a new metadata editor interface on the INESS side, as well as suitable import mechanisms on the META-SHARE side. An easier solution was adopted, consisting of the export of META-SHARE metadata to INESS. These metadata are further maintained on the INESS server, where they can be edited using any XML editor, and, after validation, uploaded again to a META-SHARE node using a simple http-based communication protocol developed at Tilde.

For licensing purposes, it is also important to implement a trusted authentication and authorization interface. Software was created on the INESS side to allow federated login via Feide, the Norwegian federation of academic ID providers. This authentication solution will be further tested and extended to eduGAIN.[11]

5 Conclusions and suggestions for best practices

5.1 Documentation and metadata

Documentation implies the provision of information on representation, provenance and IPR in order to create trustworthy metadata. Different projects produce different data and obviously

[11] http://www.geant.net/service/edugain/pages/home.aspx

have different documentation needs; it does not appear realistic to aim for fixed, predefined metadata solutions that can accommodate any documentation need. Still, there is clearly a need for some level of standardization, and we see a great potential for infrastructure initiatives to actively influence the documentation of future projects. Specifically, initiatives such as CLARIN should provide documentation templates that clearly define the minimal sets of documentation needed. META-SHARE and CMDI(Broeder et al., 2012b,a)offer interesting opportunities here for establishing metadata profiles for different kinds of resources; CLARIN, for example, is currently going through CMDI profiles that have already been created to describe existing resources in order to identify 'families' of metadata profiles.

There should also be clearly defined documentation guidelines regarding *where* to put metadata. It is often the case that a resource is represented by a collection of files. Consider for instance the case of stand-off annotation. Placing information in every document header ensures that a future user looking for information can trust which information applies for the part of a resource represented within a given document. On the other hand this may result in the same information being repeated in several files and thus being redundant. Moreover, in case a resource is upgraded or modified it may be time consuming to update documentation properly in different files or via different channels, unless sychronization is properly automatized. For the time being, we suggest as a general guideline that structured and centralized information must be provided whenever possible, but that annotation files should include, as an absolute minimum, information about *resource creator*, *creation date* and, if relevant, *originating project*. These metadata are invariable and will not become outdated, and they will ensure the future identification of the rights holder in case an annotation should become "orphaned" (i.e., separated from its repository and metadata).

With respect to the metadata schemas, we propose a new schema supporting at least the following requirements for parallel treebanks (possibly also covering other complex resources):

- There must be one metadata record for the resource as a whole, as well as individual, nested descriptions of the monolingual treebank components.

- For each monolingual component (i.e., treebank), individual descriptions of a variable number of layers (i.e., source text and any annotations) must be allowed.

- A description of the validation of each monolingual treebank must be supported, in terms of documenting the number of acceptable analyses, unacceptable analyses, unparsed sentences, etc. (as well as which sentences or parse units this information holds for).

5.2 IPR

Few researchers without legal training are happy to deal with IPR issues without assistance. The development of standardized templates and fixed law texts, such as those developed in CLARIN and META-SHARE, are therefore indispensable. Along with the dissemination of standard depositor's agreements, licenses and the establishment of a basic legal vocabulary, routines and guidelines should be established to enable research seniors and juniors to easily clear the rights for new research material. In the context of META-NORD we tested META-SHARE and Creative Commons licenses as well as the preliminary CLARIN license templates, but our experience is that sufficient guidelines are currently still missing. Among other things, neither META-SHARE nor CLARIN could provide assistance or guidelines to foresee the complex IPR problems encountered in connection with the rights clearance for treebanking. A virtual legal help desk for the CLARIN community similar to the UK JISC Legal Guidance for ICT Use in Education, Research and External Engagement[12] would be a welcome resource for researchers and deposit centers working with language data. A similar virtual competence center and additional training activities are currently planned in the DASISH project.[13]

Among the guidelines we propose is that any efforts toward the creation and distribution of resources should begin with rights clearance of the source texts for the envisaged purpose, audience and distribution scenario before investing any time in annotation and metadata creation. Moreover, license templates as well as fixed licences offer a number of different options, such as prohibiting the distribution of the original resource or allowing derivatives. In hindsight, some of these options turn out to be more decisive than others, if the use and reuse of resources within a long-term infrastructure is truly the aim. Based on our experience it should be prioritized whenever possible to make sure that the licensor accepts derivatives (i.e. allowing modifications of the original resource).

The importance of allowing derivatives may be illustrated by a straightforward scenario for treebank-based research, namely to try out a new parser on material that has previously been analysed with another parser. Unless derivatives are allowed for that source material, the new research product cannot be shared or redeposited for further research unless the new researcher takes the same (typically time-consuming) rights clearance round that was made for clearing the right to distribute the original material. Even though the emerging infrastructure initiatives hopefully will lessen the burden of clearing rights through guidelines, standards and templates, it is a fact that the source texts used for treebanking usually come from some third party that only makes the original text available for research out of goodwill (and not because treebanking research offers the prospect of profit for the source text owner). Under such circumstances repeated rights clearance requests from future researchers may not be welcomed.

5.3 Usability

While META-SHARE is a valuable infrastructure for the availability, description and exchange of resources, it has certain conspicuous shortcomings. First, its level of user-friendliness is still not well tuned towards inexperienced users. For resource owners without previous knowledge about metadata or IPR to be able to register their resource in META-SHARE, guidelines and ideally also tutorials for IPR clearance and licensing as well as metadata description are absolutely necessary. Furthermore, the metadata must be persistent and stable; it must be guaranteed that all metadata is backed up and that the updating of metadata in all META-SHARE nodes is automatic and robust.

On a more general level we insist that resource creators always document their newly created resource, and that they conform to the minimal metadata schemas developed within a collaborative, large-scale infrastructure such as META-SHARE. Integrating metadata creation as part of the resource development routine will, importantly, ensure proper documentation on resource ownership. It will hopefully also force the resource developer to keep best practices with respect to standards and IPR in mind. Drawing on our experience from treebank development in META-NORD and INESS, we claim that resources are not effectively reusable unless they

[12] http://www.jisclegal.ac.uk/
[13] http://dasish.eu

are supplied with an absolute minimum of metadata, as described in section 3.1, and until rights are cleared with an eye towards the long-term perspective, as described in section 5.2. It would, in many cases, require less work to create a new resource than to reuse a poorly documented, existing resource. Adhering to best practices in documentation and IPR clearance is thus a crucial first step towards actual usability, and hence reusability, of language resources.

5.4 Outlook on interoperability

In this paper we have presented cooperative work in two significant infrastructure projects. We have discussed several specific issues including interoperability challenges. In fact, we see interoperability in general as the greatest future challenge for cooperation between infrastructure projects. While META-SHARE has developed a specific metadata editing tool supporting fixed schemas, CLARIN has opted for CMDI as its metadata format. In order to preserve and integrate the metadata created in META-SHARE, it seems that further work on interoperability, specifically between META-SHARE and CLARIN, should have high priority. There are ongoing experiments with mapping META-SHARE schema elements to CMDI and relevant harvesting options. These initiatives hold promise that cooperation between projects, linking of infrastructures and promotion of interoperability will increasingly occupy the agenda of the research community.

References

Adesam, Y. (2012). *The Multilingual Forest: Investigating High-quality Parallel Corpus Development*. PhD thesis, Stockholm University, Stockholm, Sweden.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., and Windhouwer, M. (2012a). Standardizing a component metadata infrastructure. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1387–1390, Istanbul, Turkey. European Language Resources Association (ELRA).

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., and Trippel, T. (2012b). CMDI: a component metadata infrastructure. In Arranz, V., Broeder, D., Gaiffe, B., Gavrilidou, M., Monachini, M., and Trippel, T., editors, *Proceedings of the Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*, pages 1–4, Istanbul, Turkey. European Language Resources Association (ELRA).

Duin, P., Durco, M., Olsson, L.-J., Schonefeld, O., and Windhouwer, M. (2010). Registry Infrastructure – v2. Deliverable d2r-5b, CLARIN.

Gaarder, J. (1991). *Sofies verden: roman om filosofiens historie*. Aschehoug, Oslo, Norway.

Gavrilidou, M., Labropoulou, P., Despiri, E., Giannopoulou, I., Hamon, O., and Arranz, V. (2012). The META-SHARE metadata schema: Principles, features, implementation and conversion from other schemas. In Arranz, V., Broeder, D., Gaiffe, B., Gavrilidou, M., Monachini, M., and Trippel, T., editors, *Proceedings of the Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*, Istanbul, Turkey. European Language Resources Association (ELRA).

Gavrilidou, M., Labropoulou, P., Piperidis, S., Monachini, M., Frontini, F., Francopoulo, G., Arranz, V., and Mapelli, V. (2011). A metadata schema for the description of language resources (LRs). In Calzolari, N., Ishida, T., Piperidis, S., and Sornlertlamvanich, V., editors, *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 84–92, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Hinrichs, E., Vogel, I., Bański, P., Beck, K., Budin, G., Caselli, T., Eckart, K., Elenius, K., Faaß, G., Gavrilidou, M., Henrich, V., Quochi, V., Lemnitzer, L., Maier, W., Monachini, M., Odijk, J., Ogrodniczuk, M., Osenova, P., Pajas, P., Piasecki, M., Przepiórkowski, A., Van Uytvanck, D., Schmidt, T., Schuurman, I., Simov, K., Soria, C., Skadina, I., Stepanek, J., Stranak, P., Trilsbeek, P., and Trippel, T. (2010). Interoperability and Standards. Deliverable d5.c-3, CLARIN.

Lyse, G. I., Escartín, C. P., and De Smedt, K. (2012). Applying Current Metadata Initiatives: The META-NORD Experience. In *Proceedings of the Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*, pages 20–27.

Rosén, V., De Smedt, K., Meurer, P., and Dyvik, H. (2012). An open infrastructure for advanced treebanking. In Hajič, J., De Smedt, K., Tadić, M., and Branco, A., editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

Rosén, V., Meurer, P., and De Smedt, K. (2007). Designing and implementing discriminants for LFG grammars. In King, T. H. and Butt, M., editors, *The Proceedings of the LFG '07 Conference*, pages 397–417. CSLI Publications, Stanford.

Rosén, V., Meurer, P., and De Smedt, K. (2009). LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Van Eynde, F., Frank, A., van Noord, G., and De Smedt, K., editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht. LOT.

Vasiļjevs, A., Forsberg, M., Gornostay, T., Haltrup Hansen, D., Jóhannsdóttir, K., Lyse, G., Lindén, K., Offersgaard, L., Olsen, S., Pedersen, B., Rögnvaldsson, E., Skadiņa, I., De Smedt, K., Oksanen, V., and Rozis, R. (2012). Creation of an open shared language resource repository in the Nordic and Baltic countries. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12)*, pages 1076–1083, Istanbul, Turkey. European Language Resources Association (ELRA).