META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools

Christian Federmann[1], Ioanna Giannopoulou[4], Christian Girardi[3], Olivier Hamon[4], Dimitris Mavroeidis[2], Salvatore Minutoli[5], Marc Schröder[1]

[1] Language Technology Lab (DFKI), Saarbrücken, Germany
[2] R.C. "Athena" (ILSP), Athens, Greece
[3] Fondazione Bruno Kessler (FBK), Trento, Italy
[4] Evaluations and Language resources Distribution Agency (ELDA), Paris, France
[5] Consiglio Nazionale Ricerche (CNR), Pisa, Italy
cfedermann@dfki.de, ioanna@elda.org, cgirardi@fbk.eu, hamon@elda.org, dmavroeidis@ilsp.gr, Salvatore.Minutoli@iit.cnr.it, schroed@dfki.de

Abstract

We describe META-SHARE which aims at providing an open, distributed, secure, and interoperable infrastructure for the exchange of language resources, including both data and tools. The application has been designed and is developed as part of the T4ME Network of Excellence. We explain the underlying motivation for such a distributed repository for metadata storage and give a detailed overview on the META-SHARE application and its various components. This includes a discussion of the technical architecture of the system as well as a description of the component-based metadata schema format which has been developed in parallel. Development of the META-SHARE infrastructure adopts state-of-the-art technology and follows an open-source approach, allowing the general community to participate in the development process. The META-SHARE software package including full source code has been released to the public in March 2012. We look forward to present an up-to-date version of the META-SHARE software at the conference.

Keywords: Knowledge Management, Distributed Network, Language Resources

1. Introduction

There exist huge amounts of digital and digitized resources collections (such as, e.g., publications, datasets, multimedia files, processing tools, services and applications). This has drastically transformed the requirements for their publication, archiving, discovery and long-term maintenance. Digital repositories provide an infrastructure for describing and documenting, storing, preserving, and making this information publicly available in an open, user-friendly and trusted way. Such repositories represent an evolution of the digital libraries paradigm towards open access, advanced search capabilities and large-scale distributed architectures.

1.1. Motivation

META-SHARE aims at providing such an open, distributed, secure, and interoperable infrastructure for the Language Technology domain.

Open, since the infrastructure is conceived as an ever-evolving, scalable resource base including free and for-a-fee resources and services;

Distributed, because it will consist of connected network of repository/data center nodes accessible through shared interfaces;

Interoperable, because the resource base will be standards-compliant, trying to overcome any differences regarding format, terminological, and semantic differences;

Secure, since it will guarantee legally sound governance, legal compliance and secure access to licensable resources.

1.2. About META-SHARE

We present META-SHARE version 2, an open network of repositories of Language Resources (LRs), including both language data and language tools, described through a set of metadata, aggregated in central inventories allowing for uniform search and access to resources. Language resources can be both open and with restricted access rights, either for free or for-a-fee.

The META-SHARE network is accessible under both `www.meta-share.eu` and `www.meta-share.org`; the META-SHARE software platform is developed as an open-source project on GitHub. The source code can be found on `github.com/metashare/META-SHARE`.

1.3. Features

More specifically, META-SHARE offers to the user the possibility to:

- search and browse the metadata catalogue;
- view details about a language resource;
- download a language resource;
- view usage general statistics;
- have access as a registered user;
- describe metadata using an editor;
- upload metadata describing a language resource;
- upload a language resource.

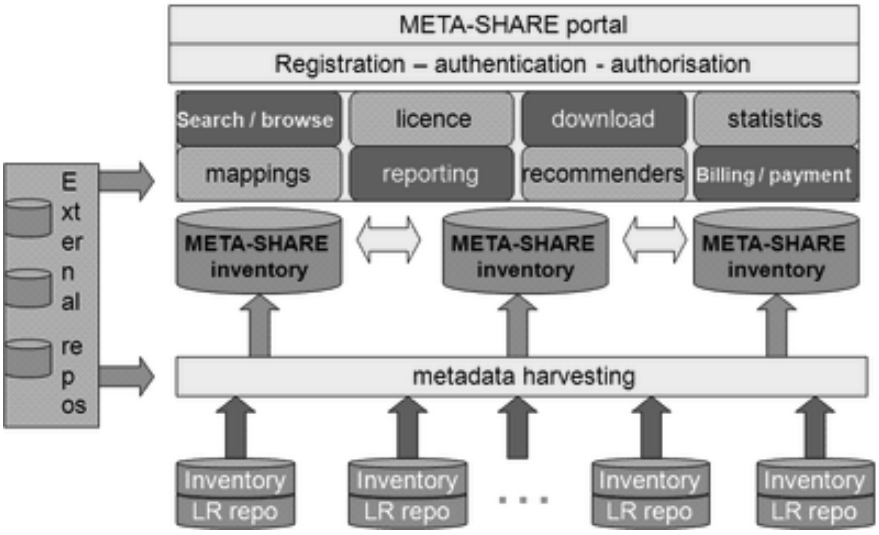

Figure 1: Overview of the META-SHARE architecture.

META-SHARE targets existing but also new and emerging language data, tools and systems required for building and evaluating new technologies, products, and (in an upcoming version) services. The project started by integrating repository nodes which represent the partners of the META-NET consortium. It will gradually be extended to encompass additional repository nodes, other data centres, and more language resources and provide more functionality with the long-term goal of turning into an as largely distributed infrastructure as possible for the benefit of the research community.

In this overview paper, we describe the overall architecture of the META-SHARE network in Section 2. Afterwards, we give a brief introduction to the metadata schema that has been developed for the application in Section 3. Finally, we conclude by giving a summary and an outlook to future work in Section 4.

2. Architecture

The META-SHARE application has been designed to work as a distributed network of so-called "nodes" which allow end users to access the metadata catalogue of available LRs. A schematic overview of the system architecture is depicted in Figure 1.

2.1. META-SHARE Portal

From an end user's perspective, META-SHARE consists of a centralised entry portal and a large metadata catalogue. The portal handles important tasks such as 1) registration, 2) authentication, and 3) authorisation. When accessing the portal, end users are redirected to one of the available META-SHARE repository nodes where they can browse the full metadata catalogue (i.e, the *META-SHARE inventory*). Its metadata contents are synchronised to all of the nodes within the repository network. We will briefly present the other modules that build the overall META-SHARE application in the following subsections.

2.2. Distributed Network

From a bird's eye view, META-SHARE is a network of connected, distributed server nodes, each of which hosts the application software and allows end users to access the shared metadata catalogue.

We have developed the META-SHARE software using the Django framework (`www.djangoproject.com`) which builds on the Python programming language (`http://www.python.org/about/`); both are maintained by a strong and active open-source community and allow for a quick, standards-compliant, and state-of-the-art development cycle. Furthermore, we believe that the use of open-source software allows to extend the development team in a future, especially after the open-source release of the META-SHARE software in March 2012; we are confident that this will help to improve the application and its functionalities over time.

The META-SHARE v2 software has been deployed by all five partners from the corresponding implementation work package in the T4ME project, namely CNR, DFKI, ELDA, FBK, and ILSP. The software has also been distributed to the collaborating PSP projects (CESAR, META-NORD, METANET4U) to setup additional nodes capable of hosting the language resources created in these projects.

2.3. Storage Layer

The metadata descriptions of all resources made available via the META-SHARE network are redundantly replicated throughout the complete network as we believe that the metadata itself should be openly accessible to everyone. Depending on the factual licensing terms of some resource, the actual resource data may only be available at one of the repository nodes. In some cases, such data might not be accessible at all.

The shared metadata information and the optional binary data attached to some of the described language resources are managed by a Django application implementing the so-called "storage layer". This component assigns a network-wide unique identifier to each resource and allows to export metadata from one node inside the META-SHARE network

to other nodes via *synchronisation* which is described in the next section.

2.4. Synchronisation

Each language resource in the local inventory of a META-SHARE node is comprised of two basic layers of information. These are:

1. a *storage object* that contains information about the status of the resource (such as, e.g., its identifier, the creation or modification timestamps, status flags, a checksum for the attached binary data, ...);

2. a *resource object* instance which is the basic object model representation of the language resource.

We have tightly integrated the storage layer and the Django object model implementation of the META-SHARE metadata schema so that it is guaranteed that for each of the resources within the Django database we also have created or updated the corresponding entry inside the storage layer. The storage layer implements a concept of "master copies". These denote language resources that have been created on or imported into the local META-SHARE node. When metadata is exported via XML, only such master copies are considered as all other objects have been copied from some other source and thus should neither be modified nor be redistributed.

2.5. Generated Object Model

The Django object model which implements any language resource compliant to the XML schema described in the next section is automatically generated from the XSD schema files. We decided to work on such an automatic generator approach as previous work with manually written models had been a) very tedious, and b) extremely error-prone. Also, generating the object model from the metadata schema directly eliminates a possible source of errors and ensures that the XSD schema files are the authoritative base for META-SHARE development.

2.6. Statistics Collection

As language resources can be related to other LRs and it hence might be of interest to users of some resource if such related data exists, we have designed and implemented a module for network-wide collection of (anonymised[1]) usage statistics.

2.7. Metadata Catalogue

Figure 2 shows a screenshot of the catalogue search interface which allows filtering of metadata records by a) language, b) type, c) availability, d) license terms, ... while in figure 3 we depict the details for a language resource from the META-SHARE inventory.

[1]We are using the aforementioned network-wide unique identifiers as primary keys; while these identifiers can be resolved by the META-SHARE nodes, for the statistics server that runs on a separate machine these represent non-recoverable "random hexadecimal numbers" which ensures that the statistics collection process does not store sensitive data.

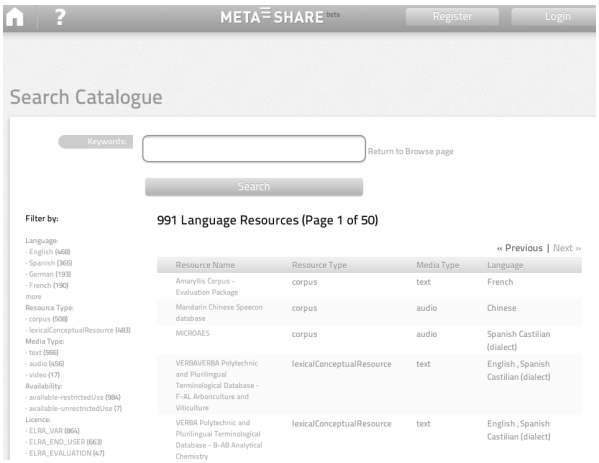

Figure 2: Screenshot of the metadata search interface.

Figure 3: Screenshot of the metadata details view.

2.8. Metadata Editor

Language resource descriptions can be imported from XML files in the META-SHARE XML format. It is also possible to directly create them in the browser with the metadata editor software. The tool for creating new resource descriptions has been designed to follow the component-based approach that has been taken in the design of the XML schema. We provide an advanced user interface that allows validation of entered data and provides feedback to the authoring user in case of non-validating values. A screenshot of the editor user interface can be seen in Figure 4.

3. Metadata Schema

To cater for the complex nature of language resources, our colleagues from work package WP7 in the T4ME project have worked extensively on a new, component-based metadata schema which will be described in more detail in an own publication at LREC, see (Gavrilidou et al., 2012).
In the context of META-SHARE, the term "metadata" refers to descriptions of language resources, encompassing both data sets (textual, multimodal/multimedia and lexical data, grammars, language models, etc.) and tools/technologies/services used for their processing.

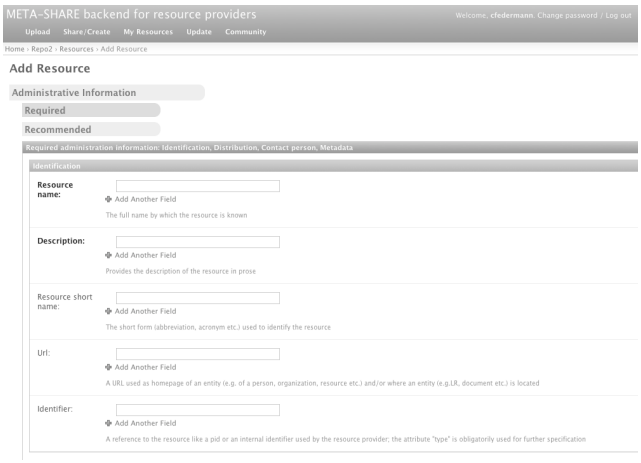

Figure 4: Screenshot of the metadata editor user interface.

The META-SHARE metadata descriptions will constitute the means by which LR users will identify the resources they seek in the META-SHARE context. Thus, the META-SHARE metadata model forms an integral part of the search and retrieval mechanism, with a subset of its elements serving as the access points to the catalogue of LRs. The model is, therefore, as informative and as flexible as possible, allowing for multi-faceted search and viewing of the catalogue, as well as dynamic re-structuring thereof, offering LR consumers the chance to easily and quickly spot the resources they are looking for inside a large bulk of available resources.

In this effort, we build upon previous initiatives so that the model is easily adopted by the target community. The aim is not to create yet another metadata model but rather to adapt existing resource description models to a joint, unified proposal catering for the specific requirements of the overall community.

As a general framework, the mechanism adopted is the component-based mechanism which groups together semantically coherent elements and relations as well as other components. Elements are used to encode specific descriptive features of the resources, while relations are used to link together LRs that are both included in the META-SHARE repository (e.g., original and derived, raw and annotated resources, a language resource and the tool that has been used to create it, etc.), as well as resources with related entities (e.g., documentation manuals, publications, standards used, licences, etc.).

In order to accommodate flexibility, the elements belong to two basic levels of description:

- an initial level providing the basic elements for the description of a resource (minimal schema);
- a second level with a higher degree of granularity (maximal schema), providing more detailed information for each resource.

The maximal META-SHARE metadata model comprises all elements and relations assisting the description of LRs put together in components. Elements will be linked to existing ISOcat DCR data categories and, if they have no counterpart, these will be added with appropriate definitions. Specific profiles will be built for distinct LR types (and subtypes) using the various components, providing also exemplary instantiations (e.g. for wordnet-type resources, for parallel corpora, for treebanks, etc.) as guiding assistance to LRs metadata providers.

Inside work package WP8, we have implemented this metadata schema which forms a fundamental layer of the META-SHARE software. Special care has been given to the serialization of objects from/into XML as this is required for synchronisation between the nodes inside our distributed network.

4. Conclusion

We have described the design and implementation of the META-SHARE software for creating a distributed network of nodes that store and export metadata descriptions of language resources. In particular, we have provided a detailed look into the system architecture, its various layers, and their functionality. Also, we have provided a brief look into the metadata editor which is available for creating and editing language resources, and provided details of the metadata schema that has been developed. META-SHARE has been deployed by five partners from the work package WP8 of the T4ME project and the software has also been distributed to collaborating projects and interested parties. The software itself has been released as open-source package.

Acknowledgments

This work has been funded under the Seventh Framework Programme of the European Commission through the T4ME contract (grant agreement no.: 249119).

We are grateful to the various contributors from the META-NET consortium who have provided helpful comments and criticism during the development of the META-SHARE software package.

We are especially indebted to the following persons who helped with the development of the META-SHARE software package: Clara Bacciu, Davide Gazza, Riccardo del Gratta, Karel Vandas, Yuanyuan Xu, Byron Georgantopoulos, Kostas Perifanos, Jörg Steffen, and Bernd Kiefer.

5. References

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, and Valerie Mapelli. 2012. The meta-share metadata schema for the description of language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May. European Language Resources Association (ELRA). Accepted for publication.

Stelios Piperidis. 2012. The meta-share language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May. European Language Resources Association (ELRA). Accepted for publication.