META-SHARE: One year after

Stelios Piperidis[1], Harris Papageorgiou[1], Christian Spurk[2], Georg Rehm[2], Khalid Choukri[3], Olivier Hamon[3], Nicoletta Calzolari[4], Riccardo del Gratta[4], Bernardo Magnini[5], Christian Girardi[5]

Athena RC/ILSP[1]	DFKI GmbH[2]	ELDA[3]	CNR/ILC[4]	FBK[5]
Athens, Greece	Berlin, Germany	Paris, France	Pisa, Italy	Trento, Italy

E-mail: spip@ilsp.gr, xaris@ilsp.gr, Christian.Spurk@dfki.de, georg.rehm@dfki.de, choukri@elda.org, hamon@elda.org, glottolo@ilc.cnr.it, riccardo.delgratta@ilc.cnr.it, magnini@fbk.eu, cgirardi@fbk.eu

Abstract

This paper presents META-SHARE (www.meta-share.eu), an open language resource infrastructure, and its usage since its Europe-wide deployment in early 2013. META-SHARE is a network of repositories that store language resources (data, tools and processing services) documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access. META-SHARE was developed by META-NET (www.meta-net.eu) and aims to serve as an important component of a language technology marketplace for researchers, developers, professionals and industrial players, catering for the full development cycle of language technology, from research through to innovative products and services. The observed usage in its initial steps, the steadily increasing number of network nodes, resources, users, queries, views and downloads are all encouraging and considered as supportive of the choices made so far. In tandem, take-up activities like direct linking and processing of datasets by language processing services as well as metadata transformation to RDF are expected to open new avenues for data and resources linking and boost the organic growth of the infrastructure while facilitating language technology deployment by much wider research communities and industrial sectors.

Keywords: infrastructures, language resources identification, language resources documentation, metadata, language resources sharing, language resources licensing

1. Introduction

In the greater language technology area, many groups, initiatives and individuals have been advocating for quite some time the need of a language resource and technology infrastructure: an open resource infrastructure, which allows easy sharing of data and tools that are made interoperable and work seamlessly together (Wittenburg et al, 2010; Soria et al 2012). Digital repositories constitute a valuable tool for publishing, archiving, discovery, long-term maintenance and curation of digital data (publications, datasets, multimedia files and processing tools and services), as they provide the infrastructure for describing, documenting, storing, preserving, and making such assets publicly available in an open way.

In this paper we present META-SHARE (www.meta-share.eu, www.meta-share.org), an open language resource infrastructure, and its usage since its Europe-wide deployment in early 2013. META-SHARE is a network of repositories that store language resources (data, tools and processing services) documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access (Piperidis, 2012). META-SHARE was developed by META-NET (www.meta-net.eu) and aims to serve as an important component of a language technology marketplace for researchers, developers, professionals and industrial players, catering for the full development cycle of language technology, from research through to innovative products and services.

Section 2 provides an overview of the META-SHARE design and architecture. Section 3 presents the metadata model and reports initial findings on its usage. Section 4 provides details on the usage of the infrastructure so far, section 5 briefly reports on ongoing and future work, while section 6 draws preliminary conclusions.

2. META-SHARE

META-SHARE is an open, integrated, secure and interoperable infrastructure of language resources for language technologies and other applicative domains (e.g., digital libraries, cognitive systems, robotics, etc). In the realm of META-SHARE, language resources span the whole spectrum from monolingual and multilingual data sets, both structured (e.g. lexica, terminological databases, thesauri) and unstructured (e.g. raw text corpora), as well as language processing tools (e.g. part-of-speech taggers, chunkers, dependency parsers, named entity recognisers, parallel text aligners, etc.).

2.1 Design Principles

META-SHARE is a network of distributed repositories of language resources. The repositories are set up, operated and maintained by META-SHARE members. META-SHARE repositories can have a local or hosting role. Local repositories are set up and maintained by members to store their own resources. Hosting repositories are set up and maintained by members, acting as storage and documentation facilities not only for their own resources, but also for resources developed in organisations not wishing to set up their own META-SHARE repository, including donated and orphan resources.

Every resource in META-SHARE has to be primarily assigned to one of the network's repositories (implementing the notion of a master copy of a resource),

with the member maintaining that repository undertaking its curation. Resources are described according to the META-SHARE metadata schema (Gavriilidou at el, 2012). Actual resources and their metadata are stored in the local repositories. Each repository creates and maintains a local inventory with the metadata records of all its resources, exports such metadata and allows their harvesting. Metadata records are harvested and stored in the META-SHARE central servers, who share metadata and create, host and maintain a central inventory including metadata-based descriptions of all resources available in the distributed network. While resources can be both open or with restricted access rights, free or for-a-fee, metadata are open, available under a Creative Commons Attribution 3.0 licence.

On the user side, distinct profiles have been defined, including related authorisations enabling certain actions, and ensuring security of transactions, such as metadata editing and updating. Users may be registered or non-registered. The former may be divided into end users, providers or administrators of a META-SHARE node. With the exception of non-registered users, all users are given a profile specifying their rights and obligations. Figure 1 presents the META-SHARE architecture.
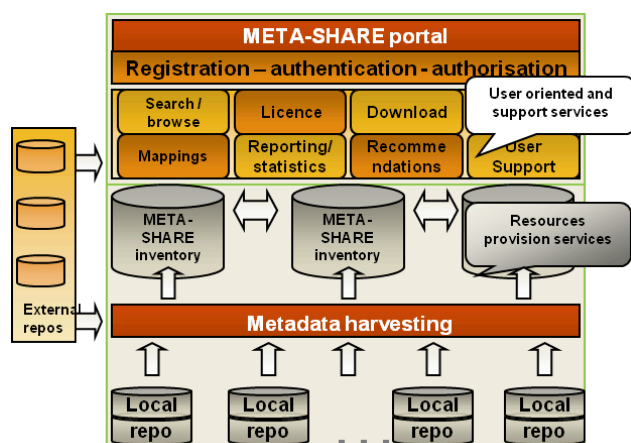

Figure 1: Architecture

Providers of resources can create, store and edit/update resource descriptions by using an editor implementing the META-SHARE metadata model, upload actual resources directly or provide a link to an existing storage. They can choose to tag their resources as *internal* (i.e. visible and accessible only by resource owners and the associated editing group), *ingested* (visible and accessible in the web editor), or *published* (visible by all META-SHARE users registered and non-registered). Providers of resources can also get statistics on the number of views and downloads of their resources, as well as the provenance of the viewers/users. Data centres or other organisations with existing catalogues of resources can get support for mapping their metadata schema onto the infrastructure's model.

Consumers of resources can register, create a user profile and log-in to the network. They can browse and search the central inventory using multifaceted search facilities, access and get information about specific resources and their usage (number of times a resource has been viewed and downloaded), and if a resource fits their needs they can digitally sign its licence and download it by visiting the local or hosting repositories.

2.2 Structure and membership

Members of META-SHARE enter the network aiming to share LRs and promote its principles to the language communities at large. The network is organised on the basis of the level of obligations that the various members commit themselves to. *META-SHARE Network Nodes* are members who set up and maintain META-SHARE repositories, acting as Repository Service Providers. *META-SHARE Managing Nodes* are members that provide core services to the whole network, acting as Core and User Support Service Providers. *META-SHARE Depositors* are members who deposit their LRs to hosting repositories maintained by META-SHARE Network Nodes. *Associate META-SHARE members* are able to use some of the Core Services, such as viewing, and make their metadata available for harvesting, while *Third Parties*, language resource consumers, are able to use the viewing services and acquire LRs from the META-SHARE network.

At the time of writing, META-SHARE consists of 34 member-organisations, 7 of which assume the role of Core and User Support Service Provider, 19 assume the role of Repository Service Provider and 5 participate as Depositors (in Annex 1, we present the list of current members). Three more repositories are being set up. The roles and duties of the members are laid out in a Memorandum of Understanding (http://www.meta-net.eu/meta-share/meta-share-licenses/META-SHARE%20MoU%20v2_0.pdf). The Memorandum of Understanding includes the Agreed Service Level for the Core Services and Repositories of the infrastructure, prescribes the adopted Notice and Take Down policy and includes the META-SHARE membership form.

2.3 Software

For setting up repositories, META-SHARE provides dedicated software (Federmann et al, 2012). We have opted for a platform independent, open-source solution. META-SHARE is implemented using Django (www.djangoproject.com), a Python-based framework, both maintained by strong and active open-source communities. Likewise, META-SHARE software is open source, released under a BSD licence, available on github (https://github.com/metashare/META-SHARE). There have been 5 releases of META-SHARE since early 2012. Version 3.0.1, the latest version, was made available in late January 2013 and has been downloaded 77 times so far.

3. Describing resources

The META-SHARE metadata model has built upon previous initiatives so as to be easily adopted by the

target community. In the model's design, central was the principle of a *minimal core subset of metadata*; the elements that form this subset are considered indispensable in the process of language resource description and are, thus, obligatory. This minimal level is the one at which interoperability with other schemas/typologies is effected. The metadata schema includes *elements*, most linkable to ISOcat Data Categories (ISO 12620), and *relations* used to link together related resources (e.g. original and derived, raw and annotated resources, resources and tools used to create them etc.). The schema comprises all elements and relations required for the description of resources; it refers to any kind of information, including identification parameters, administration (e.g. creation, distribution, licensing) and technical information required for their manipulation, information as to the production and usage (intended and actual), etc. To accommodate flexibility, the elements belong to two basic levels: a) an initial level providing the basic elements for the description of a resource (minimal schema), and b) a second level with a higher degree of granularity (maximal schema), providing more detailed information on each resource. These two levels contain four classes of elements: the first level contains Mandatory and Condition-dependent Mandatory elements (i.e. to be filled in when specific conditions are met), while the second level includes Recommended (i.e. producers are advised to include information on these elements) and Optional elements.

Following the ISOcat DCR model, elements are grouped into semantically coherent components, which, in turn, can include other components. The core of the model is the *resourceInfo* component, which subsumes all relevant components and elements that combine together to provide the resource description. Certain components (e.g. *identificationInfo*, *distributionInfo*, *creationInfo*, *usageInfo*) are common to all resource types. The content component, the annotation component etc., being modality dependent, differ across types. The modality of each type determines the description component, which is used for resources as well as resource parts. Full details on how to use the model are included in its Documentation/User Manual (http://www.meta-net.eu/meta-share/metadata-schema/).

Undeniably the formalization of the appropriate metadata model(s) for describing language resources has been a complex issue that has raised fruitful, almost philosophical discussions, in the last decade or so. In the META-SHARE model we tried to accommodate most of the results of these discussions which involved more than 85 experts. Trying to be as inclusive as possible, the current version of the model includes 711 elements. It is in our interest to run a sort of monitoring service over the use of the different elements and components. Through the META-SHARE statistics service, it appears that from the total of 174 required, 185 recommended and 352 optional elements, 546 have been *used at least once* (Figure 2), while the remaining 165 elements have *not been used* so far (Figure 3).
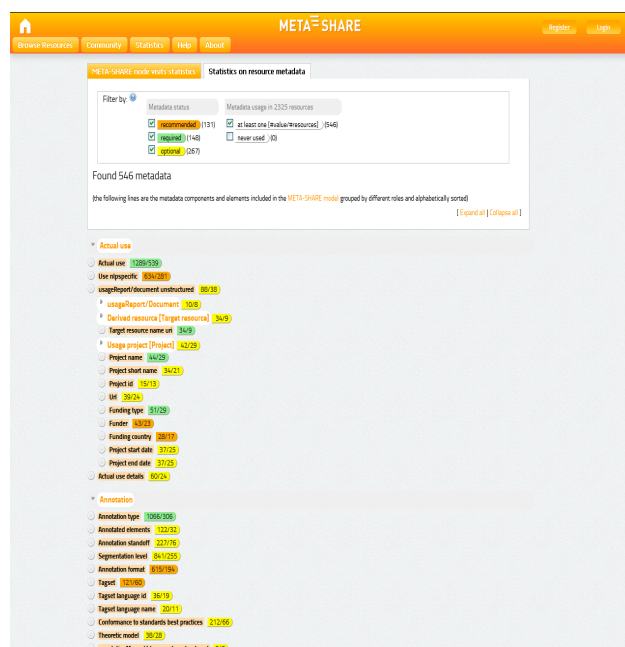

Figure 2: Statistics of metadata model elements used

Examples of elements that have not been used include *audioquality measures*, *sensor technology, image format/capture/creation*, many elements prescribed for n-gram text corpora and language descriptions, as well as inter-annotator agreement measures. Components and elements prescribed for documenting resources related to Sign languages have also not been used so far. These statistics of usage will be used as a guide in refining the model in future versions. From the components and elements heavily used, notable is the case of *DocumentationInfo*, used 760 times in 383 resources paving the way for interconnecting the world of resources with that of the scientific publications.
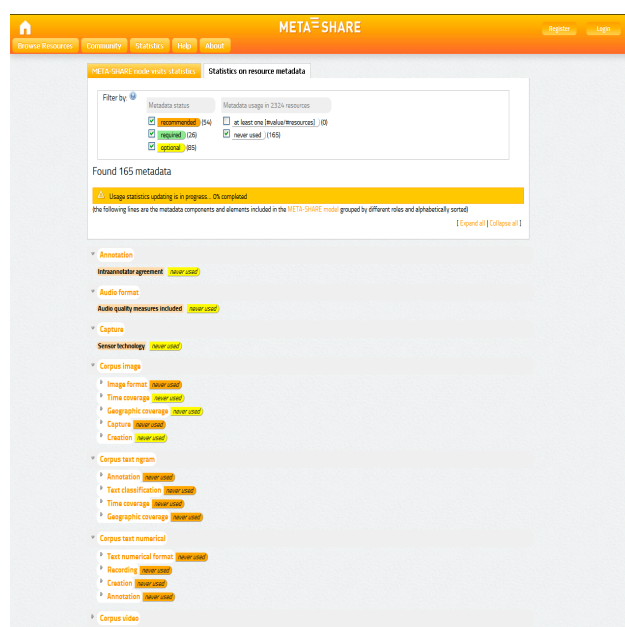

Figure 3: Statistics of metadata model elements not used

Interoperability is effected at the minimal schema level. Initially, we built a mapping service for the ELRA catalogue, and subsequently we developed a mapper for conversion to and from Dublin Core (http://dublincore.org/) and OLAC (http://www.language-archives.org/). We have also mapped the model and incorporated it into the Component Registry of the Component Metadata Infrastructure (CMDI) (http://catalog.clarin.eu/ds/ComponentRegistry/) (Broeder et al, 2010). Last, we have lately implemented an OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting[1]) bridge, as an additional protocol for harvesting and synchronising metadata between repositories, as well as connecting META-SHARE with other similar and/or related infrastructures.

3.1 Legal Framework

An integral part of the resources description is their legal status and availability, including the terms and conditions of use. The proposed, open-access inspired, model licensing scheme (http://www.meta-net.eu/meta-share/licenses) is organised around Creative Commons licences, META-SHARE Commons licences, META-SHARE No redistribution licences, and existing standard open-source software licences. These ready-to-use schemes maximize legal interoperability and are quick and easy to apply, in particular for new resources. The rights of use of the resource, possible restrictions as well as rights and restrictions on the original data are under the control and responsibility of the resource owners and the respective repository. Resources should ideally be available in the public domain. The copyright conditions should be known and resources should ideally be copyright-free or accompanied by one of the permissive licences. Likewise, derivative resources (e.g. annotated texts, lexicons extracted from parallel text) should be open at least for academic/research purposes, allowing re-use, reengineering and repurposing.

Interestingly enough, 440 resources (18.5%) are licensed with a Creative Commons licence, 178 resources (7.5%) with META-SHARE licences and 256 tools/services (11%) with a Free Open Source Software (FOSS) licence. A total of 1241 resources (52%) are restricted for academic-non-commercial use.

4. Population and Usage

At the time of writing META-SHARE makes available 2,392 resources (datasets, tools and services), through 26 repositories, one of them hosted by ELRA including most of the resources of the respective catalogue. Another 555 resources reside in the network in *ingested* (cf. section 2.1 above) status. All resources are described according to the model. In less than one year of use there have been 3,957 metadata update actions, paving the way to as accurate and complete sets of metadata as possible.

---

[1] http://www.openarchives.org/pmh/

Currently, 99 languages are represented in the resources of the network. The top ten languages and the associated number of resources is presented in Figure 4.
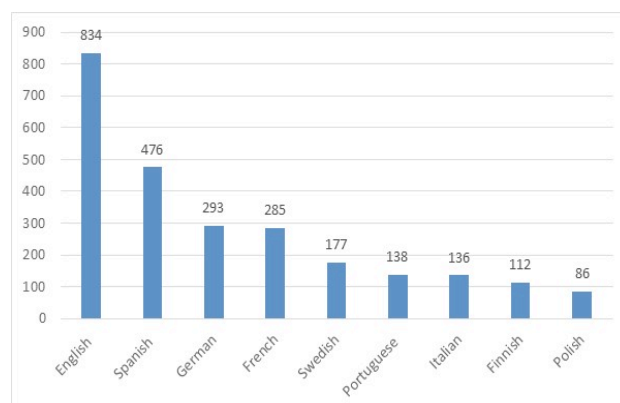

Figure 4: Top ten languages in META-SHARE

The distribution per resource type (i.e. corpora, lexical and conceptual resources, tools/services and language descriptions) is presented in Figure 5.
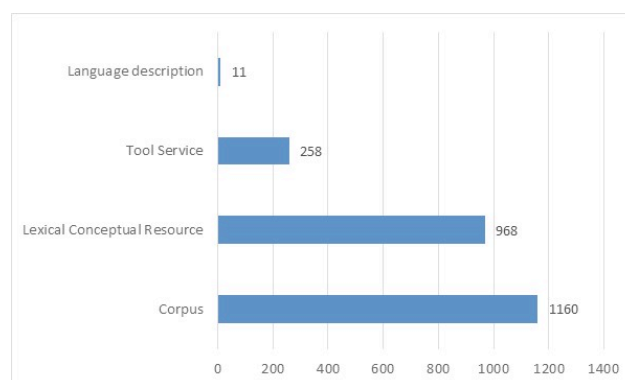

Figure 5: Distribution per resource type

Not surprisingly, the vast majority are classified as text as regards media type, with the other types following far behind (Figure 6).
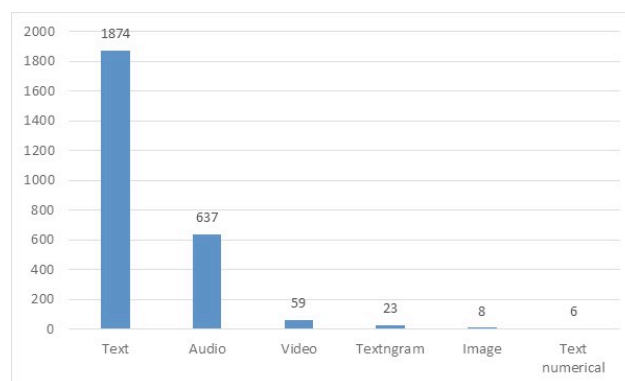

Figure 6: Distribution per media type

Regarding linguality, 1752 are codified as monolingual, 249 as multilingual and 157 as bilingual. Conformance to representation standards is not very encouraging for corpora, with the exception of multilingual ones where

TMX[2] is adhered to, while it is worth noting that from the lexical-conceptual resources the majority follows TBX[3] (145 terminological resources) and LMF[4] (108 lexical resources).

At the time of writing, META-SHARE repositories have been visited by 13,237 users (registered and unregistered) from all over the world. The total of the queries made, included those made by crawling bots, amount to 12,203,565. Metadata records have been viewed 43,164 times, while the number of free downloads amounts to 2,567.

Users can get help through a forum (http://www.meta-share.org/portal/forum/questions/show/all/newest/all) and dedicated helpdesks for technical, metadata-related and legal/licensing issues. A total of >1,500 messages by >130 providers on ~330 different topics have been handled until now. The majority concerned technical issues, reaching a local peak after each software release, while the distribution of metadata and legal issues related messages has been rather uniform over time. In the almost 15 months of operation, the 26 META-SHARE repositories, have achieved 99.99% uptime, excluding scheduled maintenance of the servers.

5. Ongoing and future work

In addition to operating and maintaining the META-SHARE infrastructure network, several members have been working on taking the infrastructure one step further. Activities include inter alia:

- transforming the metadata model into RDF[5] and querying through SPARQL[6],
- implementing and embedding in the metadata model as additional identifier the International Standard Language Resource Number (ISLRN) (Choukri et al, 2012), as well as
- creating specific area (e.g. machine translation, in the framework of the QTLaunchPad project[7]) dedicated repositories and seamlessly linking datasets and relevant language processing services for monolingual and multilingual corpora; data processing, provided as SOAP web services, caters for annotation at the levels of tokenization, sentence splitting, part-of-speech tagging, lemmatization and chunking, for English, German, Portuguese and Greek. These language processing functionalities can be triggered through the option "process" exposed in the metadata records of the relevant monolingual and multilingual parallel corpora.

6. Conclusions

We have described the META-SHARE infrastructure for language resources sharing and exchange. Such an infrastructural attempt is clearly not a one-off process. It is a long-term endeavour by which language resources are recognised as important assets that can boost research, technology and innovation through pooling, openness and sharing. The observed usage in its initial steps, the steadily increasing number of nodes, resources, users, queries, views and downloads are all encouraging and considered as supportive of the choices made so far. In tandem, take-up activities like metadata transformation to RDF are expected to open new avenues for data and resources linking and deployment in the linked data world, while direct linking of resources and processing by associated language processing services are bound to boost the organic growth of the infrastructure and facilitate language technology deployment by much wider research communities and industrial sectors.

7. Acknowledgements

8. References

Broeder, D.; Kemps-Snijders, M.; Van Uytvanck, D.; Windhouwer, M.; Withers, P.; Wittenburg, P. and Zinn, C. (2010). A Data Category Registry- and Component-based Metadata Framework. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (Eds) *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Choukri, K.; Arranz, V.; Hamon, O. and Park, J. (2012), Practical and Technical Aspects for Using the International Standard Language Resource Number. In

---

2

http://www.gala-global.org/oscarStandards/tmx/tmx14b.html
3 http://www.ttt.org/oscarStandards/tbx/
4 http://www.lexicalmarkupframework.org/
5 http://www.w3.org/RDF/
6 http://www.w3.org/TR/rdf-sparql-query/
7 http://www.qt21.eu/launchpad/

Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Eds), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, Istanbul, Turkey. European Language Resources Association (ELRA).

Gavrilidou, M.; Labropoulou, P.; Desypri, E.; Piperidis, S.; Papageorgiou, H.; Monachini, M.; Frontini, F.; Declerck, T.; Francopoulo, G.; Arranz, V. and Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Eds), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, Istanbul, Turkey. European Language Resources Association (ELRA).

Federmann, C.; Georgantopoulos, B.; Girardi, C.; Hamon, O.; Mavroeidis, D.; Minutoli, S. and Schröder, M. (2012). META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Eds), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, Istanbul, Turkey. European Language Resources Association (ELRA).

ISO 12620. (2009). Terminology and other language and content resources -- Specification of data categories and management of a Data Category Registry for language resources. http://www.isocat.org

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Eds), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, Istanbul, Turkey. European Language Resources Association (ELRA).

Soria, C.; Bel, N.; Choukri, K.; Mariani, J.; Monachini, M.; Odijk, JEJM; Piperidis, S.; Quochi, V. and Calzolari, N. (2012) The FLaReNet Strategic Language Resource Agenda. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Eds), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, Istanbul, Turkey. European Language Resources Association (ELRA).

Wittenburg, P., N. Bel, L. Borin, G. Budin, N. Calzolari, E. Hajicova, K. Koskenniemi, L. Lemnitzer, B. Maegaard, M. Piasecki, J.M. Pierrel, S. Piperidis, I. Skadina, D. Tufis, R.v. Veenendaal, T. Váradi, M. Wynne (2010) Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard,

Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias(Eds) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

9.   Annex 1

Member Name	Repository/Node URL
"Athena" RC / Institute for Language and Speech Processing	http://metashare.ilsp.gr:8080
Consiglio Nazionale Ricerche – Istituto di Linguistica Computazionale "Antonio Zampolli"	http://langtech1.ilc.cnr.it:8000
Deutsches Forschungszentrum für Künstliche Intelligenz	http://metashare.dfki.de
Evaluations and Language Resources Distribution Agency	http://metashare.elda.org
Fondazione Bruno Kessler	http://metashare.fbk.eu
Charles University in Prague	http://ufal-point-dev.ms.mff.cuni.cz:9191/
Dublin City University	
Rheinisch-Westfälische Technische Hochschule Aachen	through http://metashare.dfki.de
Research Institute for Linguistics, Hungarian Academy of Sciences	http://metashare.nytud.hu
Department of Telecommunications and Media Informatics, Budapest Technical University	http://metashare.tmit.bme.hu
Institute of Computer Science, Polish Academy of Sciences	http://nlp.ipipan.waw.pl
University of Lodz	http://metashare.ia.uni.lodz.pl
Institute of Linguistics, Faculty of Humanities and Social Science, University of Zagreb	http://meta-share.ffzg.hr/
Faculty of Mathematics, University of Belgrade	http://meta-net.matf.bg.ac.rs:8080/metashare
Ludovit Stur Institute of Linguistics, Slovak Academy of Sciences	http://metashare.korpus.sk
Institute for Bulgarian Language (IBL), Bulgarian Academy of Sciences	http://metashare.ibl.bas.bg
Tilde	http://metashare.tilde.com
Department of General Linguistics, University of Helsinki	http://metashare.csc.fi
Department of Swedish Language, University of Gothenburg	http://spraakbanken.gu.se
Department of Linguistic, University of Bergen	http://metashare.nb.no
Centre for Language Technology, University of Copenhagen	http://metashare.cst.dk
Institute of Computer Science, University of Tartu	http://metashare.ut.ee
School of Humanities, University of Iceland	through http://metashare.tilde.com
Institute of the Lithuanian Language	http://meta-share.lki.lt

Department of Informatics, University of Lisbon	http://metashare.metanet4u.eu/
Spoken Language Systems Lab, Institute for Systems Engineering and Computers	http://metanet4u.l2f.inesc-id.pt/
Department Intelligent Computer Systems, University of Malta	through http://metashare.metanet4u.eu/
Research Institute for Artificial Intelligence, Romanian Academy of Sciences	http://ws.racai.ro:9191/
Faculty of Computer Science, University Alexandru Ioan Cuza	http://metashare.infoiasi.ro/
Center for Language and Speech Technologies and Applications	http://metashare.talp.cat/
Institut Universitari de Lingüistica Aplicada, University Pompeu Fabra	http://metashare.upf.edu/
School of Computer Science, University of Manchester	through http://metashare.metanet4u.eu/
Department of Signal Processing and Communications, University of Vigo	http://firewall.teleco.uvigo.es:56245/
Aholab Signal Processing Laboratory, University of the Basque Country	http://aholab.ehu.es/metashare