CESAR

Central and South-East European Resources

Project no. 271022

Deliverable D5.3b
Sustainability strategy and plans beyond the end of the project

Version No. 1.0

08/02/2013

Document Information

Deliverable number:	5.3b
Deliverable title:	Sustainability strategy and plans beyond the end of the project
Due date of deliverable:	31/01/2013
Actual submission date of deliverable:	08/02/2013
Main Author(s):	Marko Tadić (FFZG)
Participants:	Maciej Ogrodniczuk (IPIPAN) Tibor Pintér (HASRIL) Tamás Váradi (HASRIL) György Szaszák (BME-TMIT) Svetla Koeva (IBL) Radovan Garabík (LSIL) Mladen Stanojević (IPUP) Duško Vitas (UBG)
Internal reviewer:	Tamás Váradi (HASRIL)
Workpackage:	WP5
Workpackage title:	Outreach, awareness and sustainability
Workpackage leader:	FFZG
Dissemination Level:	Public
Version:	1.0
Keywords:	dissemination

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
1.0	2013-02-08		Tamás Váradi (HASRIL)	supervision	
0.9	2013-02-06		Tibor Pintér (HASRIL)	minor adjustments	
0.8	2013-01-30		Marko Tadić (FFZG)	The advanced draft	

EXECUTIVE SUMMARY

This deliverable is the elaborated version of a preliminary deliverable D5.3a and it reports on the plan to ensure the sustainability of META-SHARE, a Language Resource Infrastructure, established within META-NET alliance (T4ME, META-NORD, METANET4U, CESAR). The aim is to describe the infrastructure, the services it offers (and associated functionalities) and draw the list of those that would serve the community for the next 2 to 5 years. The report elaborates a sustainability plan that is supported by the players involved in the projects and others from within the HLT community. This report does not intend to design a Business Model (nor a Business Plan), but provides a list of technical and personnel effort requirements to maintain the nodes in META-SHARE network and accompanying dissemination activities.

It is important for the META-NET Language Resource Exchange Facility, or META-SHARE, to clearly identify its positioning within a value-chain of existing infrastructures serving, at different levels, the Language Technology (LT) community and using LRs to do so. It is important for META-SHARE to identify and fulfill requirements of the LT players (users of LR and LT, LR providers, distributors, producers, evaluators, etc.). It is therefore important to review the services it offers and see what the requirements for their sustainability are, as well as the commitments of the players involved to ensure such sustainability.

To achieve this goal, we first identified the various services offered via META-SHARE and which are important/required by the Language Technology community. We then draw a list of priorities and finally express the intent and achievements so far of various players with respect to the investment they plan to bring in to ensure that the infrastructure is durable.

Sustainability is foreseen here not as a self-sustainability of the platform but seen rather as based on partners investment for a transition period of 2 to 5 years.

Table of Contents

Abbreviations

LR	Language Resources
LRT	Language Resources and Tools
LWP (LWPs)	Language White Paper
LT	Language Technology
PM	person/month
R&D	Research and Development
NLP	National Language Processing
RI	Research Infrastructure
HLT	Human Language Technology
IT	Information Technology
CWG	Communication Work Group
HW/SW	hardware/software

1. Introduction

This deliverable describes the plan and the first steps completed within the CESAR project to ensure the sustainability of META-SHARE, a Language Resource Infrastructure, established within META-NET. The aim is to describe the infrastructure, the services it offers (and its associated functionalities) and draw the list of those that would serve the community for at least next 2 years and that deserves the efforts and involvement of the community to make them lasting as long as necessary. This report is aimed to list not just the foreseen actions to be carried out to ensure the long term sustainability of the objectives of the CESAR project and META-SHARE, but also the actions that were already executed in this direction.

The first section of the deliverable introduces the issue to be described and gives a glimpse of the structure and brief content of the deliverable itself.

The second section elaborates the maintenance and sustainability of the META-SHARE infrastructure, giving an overview of its technical and personal recommendations. This section details the commitments of CESAR (as well as other infrastructures interested in long-time perspectives of META-SHARE) in order to achieve long time reliability and sustainability of META-SHARE in coordination with META-NORD and METANET4U projects.

The third part contains actions enhanced in favour or perspectives of the CESAR project. Here is described the position of each partner in the further life of the project: description of the META-NODES which have been set up by CESAR partners. This section addresses the sustainability of the LRs broken down by categories of corpus resources, lexical/conceptual databases, technology tools/services as well as NooJ (as an open source resource) while points out the position of CESAR as regards the long time perspectives across LRs it supports.

The fourth section delivers the joint effort of CESAR to successfully disseminate the project outcomes. Described are the activities of the completed dissemination campaign, that engaged a wider range of communication channels.

The cooperation with other LT infrastructure initiatives is emphasized in the fifth section and some conclusions set out in the sixth section.

The last (seventh) section contains the scanned copies of the Letter of Intent duly signed by all CESAR partners and a list of already running CESAR META-SHARE centres/nodes.

2 Sustainability of the META-SHARE infrastructure

META-SHARE infrastructure relies on the operation of interlinked META-SHARE nodes that are distributed and autonomously maintained by the participating institutions. With the descriptions below CESAR guarantees the maintenance and the sustainability of META-SHARE infrastructures for the proposed years and for countries involved in CESAR project.

2.1 CESAR META-CENTREs and META-NODEs

To ensure sustainability of the technical resources developed by CESAR, we propose the following organization of support:

- all partners will be responsible for maintenance of resources and tools provided by their organizations and will thus be appointed as META-CENTREs: institutions administering, supporting, updating and ensuring permanence (including backup) of their resources,
- selected partners will maintain META-NODEs, i.e. the META-SHARE applications functioning as CESAR-related points of entry for requests to access descriptions of META-NET resources and tools synchronized with other META-SHARE nodes; each partner establishing their META-NODE will be responsible for maintenance of the server, applying bugfix releases and updates received from META-SHARE, providing backup of the application and data, monitoring service availability and performance etc.
- one dedicated partner will establish a single point of entry for questions related to CESAR resources and tools in the form of an e-mail address similar to META-SHARE helpdesks (e.g. helpdesk-cesar@example.org) and will be responsible for redirecting questions to respective partners.

The META-CENTREs and the META-NODE will be maintained in 24/7 hour mode. The META-NODEs will be provided by IPIPAN, FFZG, HASRIL, IBL, UBG, LSIL and ULODZ.

The CESAR resource helpdesk will be maintained by IPIPAN with the contribution of other parnters in the consortium.

2.2 META-SHARE functionalities and services to offer and sustain

End-user and system management functionality of the infrastructure provided by META-NODEs will be entirely based on functionalities offered by the META-SHARE software platform (downloadable as the latest version 3.0.1 from https://github.com/metashare/META-SHARE/tags). For external users the access will be read-only and will consist of all features that are provided by this software:

- keyword-based LRT search and browse (with standard functionality such as filtering, ordering, paging etc.),
- access to usage statistics (most viewed, top downloaded, most recently updated resources, top and latest queries),
- user registration and login (necessary e.g. to download resources),

- downloading resources (if offered by depositor)

2.3 Technical needs to support of META-SHARE operations

To ensure adequate level of service and support after the end of the CESAR project, equipment dedicated to the maintenance of CESAR resource metadata descriptions and backup version of resources was specified and in some cases also provided. The following configuration consisting of a server and disk matrix has been verified as sufficient for the current needs of the project and for future development of CESAR resources and may be used as a model configuration for managing META-NODE.

Server:

- two x86 processors,
- 32 GB DDR3 1333 UDIMM LV RAM,
- two dual port Gigabit Ethernet NICs with support for TOE and iSCSI,
- two 250 GB SATA 7200 rpm RAID1-configured hotplug drives, compatibility with SATA drives, SAS, SSD, hardware support for RAID 0, 1, 10,
- two redundant hot-plug power packs,
- motherboard-integrated TPM, built-in intrusion sensor, LCD diagnostics panel, OS independent
- RJ-45 connector with remote access to remote monitoring and server
- status information, encrypted connection (SSLv3), user authentication and authorization,
- manufactured in accordance with ISO-9001 and ISO-14001 CE declaration, ISO 9001:2000 for the provision of services,
- 3 years of warranty (with possibility to extension to 5 years) with next working day response time, phone technical support.

Disk matrix:

- 4 GB RAM,
- 12 x 3TB hot-swap HDD, min speed 7200 rpm, drives pre-installed by the manufacturer, MTBF of each of the disks not less than 1,200,000 hours,
- hot-swap redundant power pack and fan,
- 4 x GB Ethernet (ability to work under aggregation / failover / with independent subnets),
- VMware, Citrix, Microsort Server 2003, 2008, 2008 R2, Microsoft Exchange 2010, VMware ® vSphere ESX4 iSCSI and NAS, XenServer ™ 5.5, 5.6 (w / MPIO), iSCSI & NFS certificates, ADS and WebDAV support, NAS rsync replication, secure replication over Internet, replication of data to other devices from the same manufacturer must be a free service,
- telephone technical support carried out by the manufacturer,
- minimum duration of the warranty: 36 months NBD type; the guarantee should cover the entire device, including disks.
- Requirements for human, financial and technical resources to ensure this functionality and services will be elaborated in the following months.

In some cases, e.g. FFZG, the physical configuration was replaced by virtual server, provided by a large high-power computing centre, that can be scaled as needed regarding the number of processors, size of RAM or volume of virtual hard disks.

2.4 Maintenance and support

Administration and maintenance of the META-SHARE node in the distributed network of core nodes consists of:

- keeping the server up to date with respect to security-related software updates,
- installing bugfix releases of META-SHARE,
- monitoring server availability and performance,
- periodically checking log files (e.g., to verify that synchronization between nodes is still operational),
- fixing any problems on maintenance level (e.g., server crashes).

Estimated effort of the administration task by an experienced system administrator is approx. 2 days per month (1 PM per year per META-NODE).
The administration will not cover any META-SHARE application-related implementation, documentation neither bugfixing, which will be provided by META-SHARE.
In the case of configuration of META-NODE on the virtual server(s), some of the services (uptime assurance, backups, server crashes, performance issues, security-related issues, etc.) are taken care by the large computing centre providing virtual server facility, so efforts of administration tasks could even be lower.

CESAR LRT-related support of users will consist of:

- monitoring LRT availability,
- redirecting LRT-related questions to respective partners (resource distributors),
- maintaining communication with partners to track LRT-related issues and build know-how necessary to provide first level of support to external users.

Estimated effort of the support task by an experienced consultant is approx. 0,25 hour per month per resource (with current 251 CESAR resources it makes 0,36 PM per month and 4,27 PM per year).

3. Sustainability activities in the CESAR project

Actions and efforts with aim to enhance the envisaged sustainability of the CESAR META-CENTREs is described in the following section. The proof of reliability of the actions to be carried out are the CESAR partners themselves. The following section describes the previous roles and actions in the LT field of the institutes and universities where a META-NODEs or META-CENTREs were set up.

3.1 Commitments of the CESAR partners

All countries in CESAR have expressed their willingness to maintain a META-CENTRE or META-NODE with the aim to serve as centre of repositories of LRT-s (Bulgaria, Croatia, Hungary, Serbia, Slovakia and Poland). The willingness to set up and the technical requirements of their maintenance for such centres is deeper elaborated in sections 2.1 META-CENTREs and META-NODEs and 3.2 National centres. The commitments of partners to participate in the long time maintenance of the META-SHARE is underpinned with duly signed Letters of Intent in 2012, which can be found in Attachment I (scanned copies), but also in the fact that all partners willing to set up the META-NODE, have already done so by the end of the project.

3.2 National centres

One of the pillars of long time sustainability of the CESAR project benefits are the META-CENTRES or META-NODES established. The META-CENTRES will be responsible for language resources and language technology tools created in their country – they will feature as parts of the European open linguistic infrastructure. This business model is however not the only one featuring in the CESAR countries as partners are involved in other linguistic infrastructure networks, such as CLARIN and ELRA. As META-CENTRES are going to be bases of the LT in each country – for that reason in the near future a rational business model has to be created and adapted with special regards to the CESAR and META-NET infrastructure (meeting all needs of both the providers and end-users). The model of META-CENTRES in CESAR have already been discussed within the partners and with the aim of long term sustainability META-CENTRES have already been established in all CESAR countries.

After the CESAR project ended, both Polish centres serve as META-CENTREs for Polish resources and tools while IPIPAN will provides a META-SHARE Managing Node for all CESAR META-CENTRES. The META-CENTREs and the META-NODE will be maintained in 24/7 hour mode.

Hungary is represented with two partners in CESAR. **HASRIL** as the main pillar of the Hungarian language technology is founder and coordinator of HunCLARIN, which is a part of the European CLARIN network. HASRIL offers and maintains a huge part of Hungarian LRs (as the Hungarian National Corpus or the NooJ processing tools and dictionaries for Hungarian) and will operate as the META-SHARE node. HASRIL is also coordinating the Language and Speech Technology Platform in Hungary, a strategic alliance between centres of excellence in R&D and the leading industrial centres in the sector with aims to provide computational support for natural language communication. HASRIL will serve as the META-CENTRE for Hungarian resources and tools, which will be maintained in 24/7 hours. The address of this node is **http://metashare.nytud.hu/**.

The second Hungarian partner of the CESAR consortium is **BME-TMIT**, the key Hungarian player regarding audio and video spoken language tools and resources. BME-TMIT is a

founding member of the Language and Speech Technology Platform in Hungary. BME-TMIT has participated in the development of the majority of Hungarian audio corpora and will remain responsible for the maintenance and hosting of these resources and tools in the META-NODE for this type of language resources. Its address is **https://cesar.tmit.bme.hu**.

Croatia is represented in CESAR by the University of Zagreb, Faculty of Humanities and Social Sciences (**FFZG**). FFZG is the referral institution in Croatia for computational lingustics, corpus linguistics, NLP and LT. With its two departments (Linguistics and Information Sciences) it covers the area from the sides of both sciences. FFZG takes part in different European projects including RI project CLARIN of which it is the central node for Croatia. FFZG has an expertise in creating and maintaining the largest Croatian corpora and lexical resources such as Croatian National Corpus, Croatian Web Corpus, a number of parallel corpora, Croatian Dependency Treebank, Croatian Morphological Lexicon, Croatian Wordnet, etc. FFZG in collaboration with the University of Zagreb Computing Centre (**SRCE**) established a META-NODE (accessible at **http://meta-share.ffzg.hr/**) using the available virtual server facilities, thus providing high-level computing and top-level maintenance for a META-CENTRE for Croatian resources (or even wider, funding permitting). SRCE, being a central Internet node for Croatia, is of course on-line 24/7/365.

Bulgaria is represented in CESAR by the Institute for Bulgarian (**IBL**) as the main pillar of the Bulgarian language resources. The IBL participates actively in the European infrastructures for the development of language resources and tools. IBL, as an expert on the creation of quality language resources is a member of the European infrastructure CLARIN. Another scientific infrastructure concerning language resources was establish on a national level with the participation (and coordination) of the IBL. The IBL offers and maintains a huge part of Bulgarian LRs (as Bulgarian National Corpus, Bulgarian wordnet, Web based infrastructure for Bulgarian data processing, Proofing tools for Bulgarian, etc.) and will serve as the META-CENTRE for Bulgarian resources and tools and a META-NODE. The META-CENTRE and the META-NODE in IBL will be maintained in 24/7 hours and its address is **http://metashare.ibl.bas.bg/**.

Poland is represented in CESAR by two partners: Institute of Computer Science, Polish Academy of Sciences (**IPIPAN**) and University of Łódź (**ULODZ**). IPIPAN is a leading national center of research in Computer Science, with some focus on the fundamental and applied research in the areas of Artificial Intelligence and Information Systems. CESAR-related research activities are being carried out by the Linguistic Engineering Group at the Department of Artificial Intelligence. As the largest Polish team specializing in Natural Language Processing, Linguistic Tools and Resources, within the last 5 years IPIPAN has been involved in numerous European and national projects, as well as a number of bilateral co-operation projects. Among the tools developed by IPIPAN are taggers, shallow and deep parsers, as well as various machine learning and rule-based information extraction tools. IPIPAN has also developed the first large linguistically annotated corpus of Polish and has recently completed the National Corpus of Polish project, which involves all previous developers of Polish corpora. Within the European CLARIN project, IPIPAN was responsible for a working group dealing with the integration of linguistic tools and resources. Within CESAR, IPIPAN was responsible for creation and is now actively maintaining the "Computational Linguistics in Poland" (CLIP) Web portal (http://clip.ipipan.waw.pl), connecting institutions, people, resources, projects and publications related to language technology. The portal is bringing together linguistic community, industry key players and government representatives offering them clear recommendations on state-of-the-art language resources and technologies for Polish and is currently the largest repository of resources of

that type. After the end of CESAR project IPIPAN will provide a managing META-NODE at the address **http://nlp.ipipan.waw.pl/metashare**.

ULodz is another major research and education centre in Poland. The PELCRA (http://pelcra.ia.uni.lodz.pl) research group based at the university's department of linguistics is a long-standing player on the Polish language resources scene. Over the recent years, its activities have been geared towards the collection of corpus data, including multimodal spoken data and development of language tools and services with applications in research and technology. The group's members have confirmed experience in developing language processing systems for both general and special domain Polish and English texts. A key area of expertise covered by ULodz is the development of scalable HTTP and SOAP language web services, such as the tools providing web access to the National Corpus of Polish. The address of ULODZ META-CENTRE is **http://metashare.ia.uni.lodz.pl/**.

Slovakia is represented by the **Ľ. Štúr** Institute of Linguistics, Slovak Academy of Sciences (**LSIL**). The institute is the leading organisation in the human language technologies and language resources in Slovakia since the establishment of its Slovak National Corpus department. As such, LSIL provides and maintains almost all the existing Slovak language NLP resources as a part of its activity and portfolio. The administration and maintenance of the META-NODE, already running at the address **http://metashare.korpus.sk/**, will be integrated into the departmental tasklist in the scope of the project Construction of Slovak National Corpus and Electronisation of Linguistic Research in Slovakia (2013–2017) to ensure prolongued sustainability.

Serbia is represented by two partners in CESAR project. The University of Belgrade Human Language Technology Group is hosted by the Department for Computer Sciences at the Faculty of Mathematics (**UBG**), a faculty with a tradition of educational and research excellence. The HLT Group was founded more than 30 years ago, in 1978, with the main goal of developing a formal description of Serbian, and producing and exploiting resources and tools for this language. The core of the HLT Group is now composed of researchers from several faculties of the University of Belgrade, but also serves as a hub for research activities involving researches from the University of Novi Sad. The Group has strong relations with the majority of Serbian institutions involved in language technology. In the course of its long existence the HLT Group has developed a considerable amount of language resources and tools. After the end of CESAR project, this partner will be responsible for the administration and maintenance of the META-CENTER (24/7) in Serbia.

Institut Mihajlo Pupin (**IPUPIN**) is a leading Serbian R&D institution in information and communication technologies, the largest and oldest in the whole South Eastern Europe. At Pupin Institute, projects of critical national importance have been conducted, combining systems engineering and information technology to develop innovative solutions in the area of telecommunications and computer networks, knowledge and content technologies and applications, Web services, robotics, management information systems, e-government, ebusiness, e-education, power systems management, water supply management, traffic control, etc. Its service scope covers customized IT solutions, HW/SW outsourcing, technology consulting, engineering, prototyping, and system design and integration. Its task in the CESAR project is to port NooJ to Java platform, thus making it open source and multiple platform application.

3.3 Sustainability of the main types of resources

The main goal of sustainability is to ensure a prevention from: a) a disconnection to the availability of language resources; b) a duplication of work directed to creation of language resources due to the lack of availability, access or information.

The CESAR consortium has concentrated on all features of language resource that can contribute and have an impact on their sustainability (understood as future availability and usage). The consortium set up a number of requirements in order to meet the sustainability of language resources.

1) Language resource are carefully selected - a methodology and criteria that allow partners to assess the quality and importance of language resources are established and carefully followed. The aim is to ensure a balanced coverage of resources for different end users and tasks, groups of products and services.

2) Particular actions are performed to ensure quality and quantity of the selected resources - upgrading, extending and linking the resources, aligning resources across languages.

3) Language resources are made visible and accessible - META-SHARE metadata descriptions are based on established standards, best practice and users needs. Providing exhaustive metadata descriptions enables the users to find out the most suitable resource and to use it in an appropriate way.

3.3.1 Language resources were carefully selected

The consortium selected the best possible resources that will be needed by different groups of end-users. The approach for language resources and tools selection was based on a number of indicators (General evaluation of resources, Total Point Value, Language White Papers) where each language resource was specified according to different groups of criteria. The goal was to ensure as accurate measurement as possible for different quality and quantity parameters.

The General evaluation of resources was carried out in three directions: resource upgrade, extension, and cross-lingual alignment.

For upgraded resources: All selected resources are state-of-the-art representatives of their type for a given language. Equally valuable representatives are all included in the selection. Current status of resources have superior quality at least on regional level without the need of excessive further development. Licensing issues allow free processing and access to resources and resource-related materials to the maximum possible extent or the consortium succeeded in reaching an agreement with respective copyright holders.

For extended/linked resources: The extension of resources provides considerable value to the community, at least on national and/or regional level. The emphasis was on providing building blocks to the existing tools rather than major restructuring. Additional resources were integrated with the existing ones only if they significantly improved the quality of resulting resources. If more than one representative of a certain resource type for a language has been selected, they were very likely interlinked to benefit from strong sides of both solutions. If less developed, but still very popular resources could benefit from the enhancement due to their well-developed equivalent, their enhancement was also considered. Experience of other consortium members/other consortia was used in the process of extension of national resources to provide strong foundation for cross-lingual coverage. Tools that were language-neutral or cross-lingual, were preferred.

For resources aligned across languages: No more than one tool of a certain type for each language was used. Whenever applicable, the largest set of languages was selected. Language

independence was targeted to a great extent while in the same time the quality of a result was of immense concern.

Under the Total Point Value the notions of availability, quality, quantity and standards were further specified and linked with numerical assessment according to previously established qualitative and quantitative criteria and conventions for the measurement. The following Point values have been specified: Availability (scope, price, degree of adaptability); Quality (standard compliance, internal consistency, task-relevance, environment-relevance); Quantity. The established criteria for selecting language resources require Total Point Value lower than or equal to the minimum value of 16. The Total Point Value for resources being selected for the project are calculated before any upgrade work.

The soundness of specification cannot be judged without knowing the broader context of usage, adequacy, and so on, of a certain language resource. To estimate the quality, quantity and importance, every case is thoroughly examined, taking into account regional determinants, popularity of the format outside its home institution, etc. This requires a complex assessment of language resources in the context of the whole set of the established criteria.

Up to M18 114 corpus resources (written and spoken), 45 lexical/conceptual databases and 53 technology tools/services were identified by the CESAR consortium based on the established criteria and were considered to become available according to defined processes in the META-SHARE platform.

Resources per Country	**Total**	**By Resource type**				**By Linguality**			
		Text Corpora	**Audio Corpora**	**Lexical / Conceptual Database**	**technology tool / service**	**Monolingual**	**Bilingual**	**Multilingual**	**Outside the consortium**
Bulgaria	29	7	1	6	15	21	1	7	6
Croatia	23	12	1	4	6	14	1	8	4
Hungary	57	18	19	7	13	49	6	2	22
Poland	40	13	4	12	11	32	6	2	17
Serbia	30	16	3	6	5	21	5	4	6
Slovakia	33	19	1	10	3	22	9	2	16
Total	**212**	**85**	**29**	**45**	**53**	**159**	**28**	**25**	**71**

Table 1. Summary of the reported language resources at M18 of the project

During all three CESAR batches an impressive number of resources were selected and published under the META-SHARE, 251 all together, among them 120 corpus resources; 65 lexical conceptual resources and 66 tools and services.

	RILHAS	TMIT	FFZG	IPIPAN	ULODZ	UGB	PUPIN	IBL	LSIL	∑
Tools / Services	6	3	5	19	5	6	0	16	6	66
Corpora	19	21	12	17	11	10	0	9	21	120
Lexical/Conceptual resources	6	1	9	23	1	3	2	11	9	65
Total	**31**	**25**	**26**	**59**	**17**	**19**	**2**	**36**	**36**	**251**

Table 2. Summary of the language resources for the whole period of the project

3.3.2 Particular actions performed to ensure quality and quantity of the resources

3.3.2.1 Corpus resources

Upgrading corpus resources has been done by means of: upgrading for interoperability (changing annotation format, type, and tagset); metadata-related work (creation, enhancement, conversion, and standardization), production and management of documentation (specification documents, reference documents) and harmonization of documentation (conversion to open formats, reformatting, and linking).

Extending and linking corpus resources has been done by means of: extention or linking across different resources to improve their coverage and increase their suitability for both research and development work; integration of additional resources to improve the quality and quantity of resulting resources.

Aligning corpus resources across languages has been done by means of: cross-lingual alignment of parallel corpora performed at different levels - text parts, sentence, clause, etc.

Quality assessment: measuring the compatibility and adequacy of the resource with a specified one.

Property rights, privacy, consent, and other sensitive issues: all legal aspects were cleared out.

Providing information on language resources: wide dissemination by different channels (scientific and technical publications) - the goal is the resource identification by the potential users.

3.3.2.2 Lexical conceptual resources

Upgrading Lexical conceptual resources has been done by means of: upgrading for interoperability (changing annotation format, type, and tagset); metadata-related work (creation, enhancement, conversion, and standardization), production and management of documentation (specification documents, reference documents) and harmonization of documentation (conversion to open formats, reformatting, and linking).

Extending and linking Lexical conceptual resources has been done by means of: extention or linking across different resources to improve their coverage and increase their suitability for different tasks; integration of additional resources to improve the quality and quantity of resulting lexical conceptual resources.

Aligning Lexical conceptual resources across languages has been done by means of: cross-lingual alignment of lexical conceptual resources, in this case mainly wordnets to the Princeton wordnet, but also of Wikipedia headwords in all CESAR languages and English.

Quality assessment: measuring the compatibility and adequacy of the resource with a specified one.

Property rights, privacy, consent, and other sensitive issues: all legal aspects were cleared out.

Providing information on language resources: wide dissemination by different channels - the goal is the resource identification by the potential users.

3.3.2.3 Technology tools and services

Upgrading Technology tools and services has been done by means of: technology-related upgrade (wrapping, refactoring, etc.), harmonization of documentation, preparation for

maintenance and deployment (debugging, cleaning, building test environments, preparing code repositories), programming tasks (bug-fixing and standardizing API calls), production and management of documentation (specification documents, reference documents) and harmonization of documentation (conversion to open formats, reformatting, and linking)..

Extending and linking Technology tools and services has been done by means of: interlinking of tools performing equal or similar tasks to benefit from strong points of both solutions (unless their usage patterns do not encourage such action), linking of individual tools into processing chains (capable for mono- or multilingual processing), work on languge independency since tools offering language-neutrality or cross-linguality were preferred.

Quality assessment: measuring the compatibility and adequacy of the resource with a specified one.

Property rights, privacy, consent, and other sensitive issues: all legal aspects were cleared out.

Providing information on language resources: wide dissemination by different channels - the goal is the resource identification by the potential users.

3.3.3 Language resources were made visible and accessible

Sustainable availability of identified language resources is directed to overcome the restrictions over the public accessibility caused by different personal, privacy or property rights reasons as well as the practice to report on language resources in research publications not providing detailed description and evaluation data for them.

Providing exhaustive metadata (both technical and descriptive) enables the users to understand the structure, content and main applications of a resource. The CESAR supported the goal of a common and shared resource description between the four projects constituting META-NET (i.e., CESAR, METANET4U and META-NORD, and T4ME). The META-SHARE metadata are 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. The META-SHARE metadata group connected together semantically coherent elements and relations as well as other components. The elements encode descriptive features of language resources, the relations link together resources that are included in the META-SHARE, as well as resources with related entities (e.g. documentation manuals, publications, licences etc.). The elements show two basic levels of description: an initial level providing the basic elements for the description of a resource (minimal schema, required metadata), and a higher degree of granularity (maximal schema, recommended and optional metadata), providing more detailed information. It is important to describe the language resources with respect to the project and framework in which they were produced and (possibly) where and how they have been used. It is also important to document the languages, applications, and usage, for which the resources have been initially designed and have been actually used.

3.3.4 Sustainability of NooJ

NooJ, as one of the most popular linguistic tools (not just in LT community), was one of foci of the CESAR project, and, therefore, one of its tasks was devoted to its translation to open source and its porting to multiple platforms. The task required cross-national collaboration, with Max Silberztein (France) as NooJ's author on one side and the team of Institute Mihajlo

Pupin (Serbia) that was porting NooJ to Java platform on the other. However, after the end of project, the cross-national dimension will be more stressed, because NooJ will become an open source software and many developers from different countries will become involved in applying the existing NooJ solutions or in providing various functional extensions to the existing code. The Institute Mihajlo Pupin will support the open source NooJ community by providing maintenance of the existing code and helping open source NooJ developers.

Under the aegis of the CESAR project it was decided at the NooJ 2012 conference in June 2012 in Paris that the NooJ Community will be organized into a legal entity. The exact format is currently discussed within the community. Such a move is envisaged to provide the necessary permanence and governance structure to ensure that the strategic directions of the now open source system will be in the hands of the newly formed association. This is seen as a vital step towards ensuring sustainability and coherent development of the system in the long run as opposed to the possibility of being handed over to unstructured and uncoordinated redesigning and reprogramming.

Another way to ensure wider visibility, accessibility and sustainability of NooJ was a series of NooJ video tutorials explaining basic functionalities and use of NooJ. The resulting video clips were produced and published on YouTube and CESAR web site, replicated to doubly ensure their easy accessibility, thus contributing to the sustainability of the system.

4. National cooperation and dissemination

4.1 National cooperation

All partners are basic centres of language technology field in their respective country with wide range of contacts in their national scene. The research community at national level was therefore reached by a range of traditional or not so traditional dissemination channels (flyers, posters, public web-page, conferences, events presence, lectures, presentations, demonstrations, press releases, but also video lectures, video presentations and publicity in national media) in order to attract the players to participate in sharing resources and tools through META-CENTRES. Important role at the national level was also played by national language technologies (or similar) societies, that serve as coordinating points or hubs between research institutions or groups and industry. These societies usually organize conferences that have national, regional or wider character, thus being the meeting point of different stakeholders.

Industry at national level is reached through national networks and platforms (such as it is in Hungary) and through the CESAR Road Shows which were organized to mobilize LT community at the national level.

National cooperation is also organized through national infrastructure collaborations (which can be domestic of as part of a wider, international presence) such as the CLARIN or its local forms (such as Hun-CLARIN). In Hungary the national cooperation is organized through common platforms such as the Language and Speech Technology Platform which gathers the main actors in LT field ranged from the Academic part, through universities and business partners. This platform was created to map the Hungarian LT community and enhance activity in the proper field. Research community in Hungary is also gained across the Hun-CLARIN platform which is also a good opportunity for dissemination purposes.

In Croatia, the Croatian Language Technologies Society (CLTS) plays the role of hub where LT activities are tracked and, if possible, co-ordinated at the national level. Members of this professional non-profit society are members of all LT-relevant institutions in Croatia and are participants in the national LT projects, funding of which has just been closed on 2012-10-31. Since no new calls are issued at the moment, it is hard to tell how the LT in general, and CESAR-related activities in particular, will be supported nationally in the future since there are no infrastructure projects running at the moment. Members from EU-funded projects will exercise their dissemination activities throughout the projects' duration and some of these activities will have long-lasting effect relevant for CESAR language resources.

The CLTS is an co-organiser of a regional bi-annual conference Formal Approaches to South-Slavic and Balkan Languages (FASSBL, http://www.fassbl.org) and this conference serves as the meeting point of researchers from computational linguistics or LT and industrial partners for respective languages where all recent research and project activities were presented. The conference is regularly supported by the Ministry of Science, Education and Sports of the Republic of Croatia and from other sources.

The CLTS also maintains the Croatian Language Technologies Portal (http://jthj.ffzg.hr) that delivers information about LT activities nationally and internationally, and is being online since 2000.

The Human Technology Group, supported by the Faculty of Mathematics and the Faculty of Philology at the University of Belgrade coordinates most of the activities related to language technologies in Serbia. These faculties have a lot of experience in organizing international conferences, a number of them related to LT: 9th Intex/Nooj Conference 2006, 29th

International Conference on Lexis & Grammar 2010, QUALICO 2012, Workshop on Computational Linguistics and Natural Language Processing of Balkan Languages at Balkan Conference on Informatics 2012. Research in this field is also supported by the national fundamental project funded by Serbian Ministry of Education and Sciences in the current project period (2010-2014). Research and applications in LT is since 2010 also financed through national multidisciplinary infrastructural projects in which all relevant ICT institutions participate. The Society for Serbian Language gathering various professionals – professors, researchers and lexicographers - is an organization through which ideas about the role and importance of LT are spread. The initiative has been launched to establish the section for Language Technologies under the auspices of this Society.

4.2 Dissemination

Dissemination is a process made by all partners in tight cooperation at national and international level. Dissemination activities at the European and global level were coordinated and harmonised with META-NET dissemination activities in order to maximise the impact by controlled spread of different participants at different events. The joint effort of the project participants could have been seen in the joint presence at conferences and workshops of highest level in LT world.

4.3 Long time dissemination efforts

CESAR's perspective is to be visible at main LT events both at national and international level. This basic idea will remain after project ended. The dissemination (showing efforts for sustainability) will be alive mainly in conferences and events, and spread via traditional ways (posters, flyers).

Some less conventional channels of dissemination will be available long time after the project end. At this moment a series of short NooJ video tutorials describing how to use NooJ was prepared. These video tutorials will remain accessible on-line for the interested audience and will be clearly connected with the CESAR project. They should be considered an integral part of the NooJ as an open source bundle and will be accessible both at NooJ and CESAR web site.

Also, video lectures of CESAR partners' presentations at different conferences and workshops will remain accessible on-line after the project end to continue CESAR presence.

4.4 Language White Papers

Language White Papers (LWPs) were a clear sign of tight interoperability within the META-NET alliance. LWPs were materials for distribution among stakeholders (industry, government, research community) with an aim of raising awareness on language technology – especially in countries where language technology has a weak position in various decision makers' domains.

LWPs were used extensively by the project partners to disseminate information about META-NET and CESAR at the national level to different stakeholders, primarily through the set of one-day high-level events, that were in CESAR named "road shows" (see below). These events represened an ideal opportunity to spread the LWPs as widely as possible. Also, the remaining LWPs will be used at the national and regional events to disseminate information about the META-NET, CESAR and the centres and nodes that will be functioning after the projects ended.

LWPs in the form in which they were finally published, are excellent compendium of state-of-the-art of LT in a given country since a clear comparison with the status in other countries/languages were given, but at the same time these information was presented in a simple and popular manner and they can be immediately used by journalist, policy consultants etc.

4.5 CESAR Road-Shows

Beside the activities targeted to research community, the most important means of enhancing awareness and strengthen sustainability of the project in different communities, i.e., business, society and government was a series of nationally organized high-level awareness events („road shows“) that took place in each country once in the project duration. The schedule of organized road-shows is given below:

- Sofia, Bulgaria (2nd May, 2012)
- Bratislava, Slovakia (7th June, 2012)
- Warsaw, Poland (27th September, 2012)
- Belgrade, Serbia (29th October, 2012)
- Zagreb, Croatia (30th November, 2012)
- Budapest, Hungary (18th January, 2013)

The road shows were a brilliant opportunity to show relevant features of the project and raise awareness for LT of the chosen language community.

If a critical mass of stakeholders at the national level got mobilised, such events could have a potential to become regular events at the national level, e.g. on annual basis. It could be expected that national LT societies would be a natural institutional lieu for organisation of these events, and if they become regular at the levels of different nations, it could be thought about making an international series of these events that could circulate between European countries spreading the knowledge about new solutions and/or achievements in the field of LT. We are not advocating another European-level conference, but a series of similar national events that could be loosely coordinated at the European level and aiming at similar goals – fully LT-supported multilingual EU even after the META-NET projects ended.

5. Cooperation with other infrastructure initiatives

In large number of cases the CESAR partner institutions are (or were) CLARIN nodes and are actively participating or intend to in building the national infrastructure in accordance with CLARIN ERIC expectations, thus seeking the support and/or securing funding from that segment at the national level as well.

6. Conclusion

Although the preliminary version of this deliverable, submitted in 2012-09, the long time sustainability plan of the CESAR-consortium was already elaborated to a certain extent. In this deliverable this plan was pushed forward stating the clear willingness of all partners to commit in long time perspective as can be seen in the fact that META-SHARE centres or nodes have been set up and are already running in each and every CESAR country.

Appendix I. Scanned copies of the 'Letter of intent'

RESEARCH INSTITUTE FOR LINGUISTICS
HUNGARIAN ACADEMY OF SCIENCES

19th June, 2012

To Whom It May Concern:

Letter of Intent

The Research Institute for Linguistics of the Hungarian Academy of Sciences, represented by Director István Kenesei, hereby expresses its non-legally binding intent that the Research Institute for Linguistics of the Hungarian Academy of Sciences is willing and able to participate in the maintenance and evolution of the META-SHARE infrastructure.

We confirm that the Research Institute for Linguistics of the Hungarian Academy of Sciences will

- Set up a META-SHARE repository to host and make available its language resources through the META-SHARE network;
- Continue to host the repository of LRs and serve as a META-SHARE node for a period of 24 months;
- Continue to provide technical and/or user support services, software-based and/or human services provided for a period of 24 months; and
- Start or continue to participate in the META-SHARE software development team (requires Python and Django skills).

Prof. István Kenesei
Director

H-1068 Budapest, Benczúr utca 33.
Phone: + 36 1 321 48 30 / Fax: + 36 1 322 92 97
E-mail: linginst@nytud.hu / Homepage: www.nytud.hu

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
5 Jana Kazimierza Str.,
01-248 Warsaw, Poland

phone: +(48) 22 38-00-500
fax.: +(48) 22 38-00-510
e-mail: ipi@ipipan.waw.pl
www.ipipan.eu

IPI PAN – DN - 32/2012

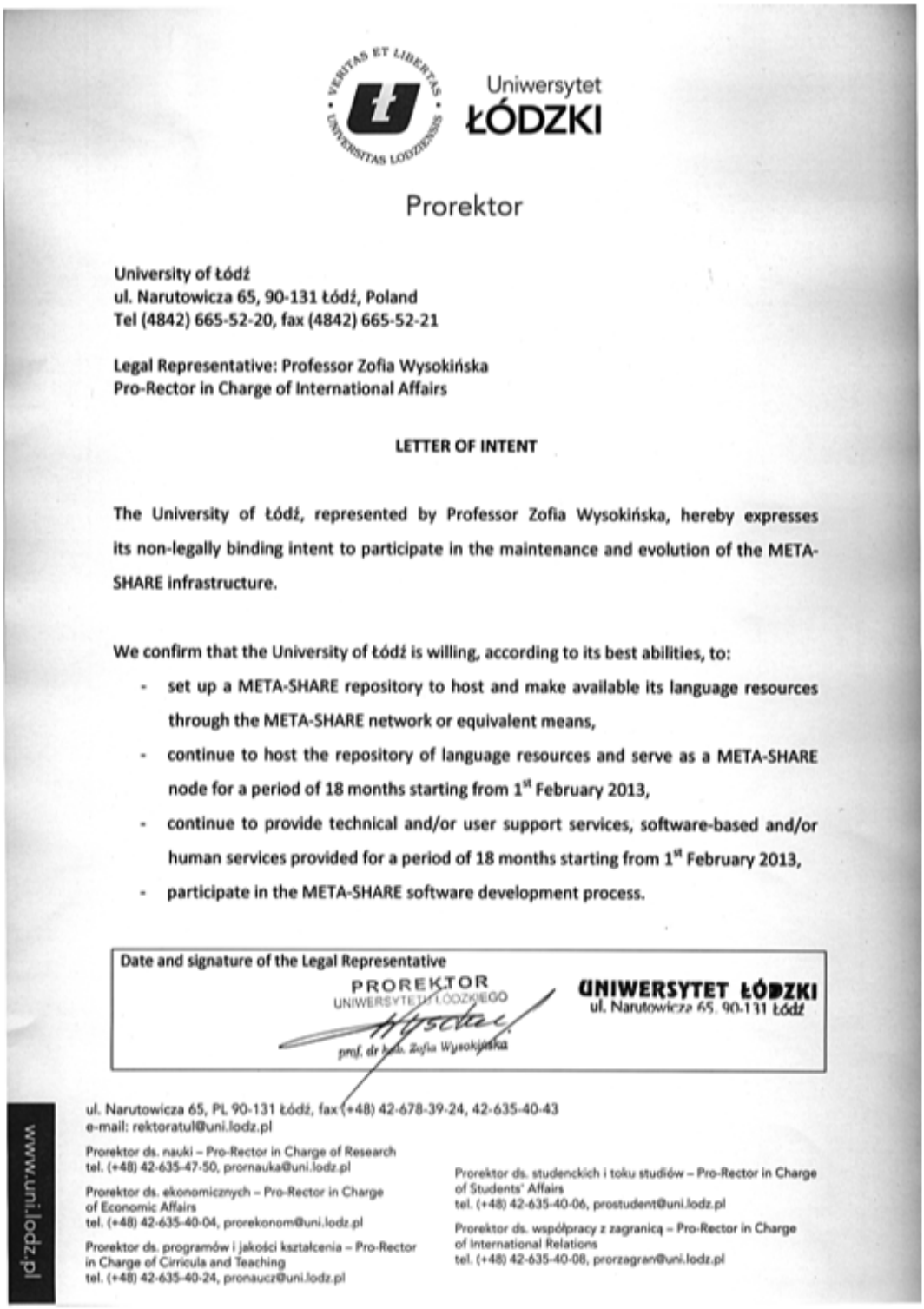
Warsaw, June 19, 2012

Letter of Intent

Institute of Computer Science, Polish Academy of Sciences, represented by prof. Jacek Koronacki, Director, hereby expresses its non-legally binding intent that it is willing and able to participate in the maintenance and evolution of the META-SHARE infrastructure by continuing to host the META-SHARE repository of language resources and serve as a network node for a period of 24 months after the end of the project CESAR, i.e. until 31 January 2015.

DYREKTOR INSTYTUTU

prof. dr hab. inż. Jacek Koronacki



Uniwersytet
ŁÓDZKI

Prorektor

University of Łódź
ul. Narutowicza 65, 90-131 Łódź, Poland
Tel (4842) 665-52-20, fax (4842) 665-52-21

Legal Representative: Professor Zofia Wysokińska
Pro-Rector in Charge of International Affairs

LETTER OF INTENT

The University of Łódź, represented by Professor Zofia Wysokińska, hereby expresses its non-legally binding intent to participate in the maintenance and evolution of the META-SHARE infrastructure.

We confirm that the University of Łódź is willing, according to its best abilities, to:

- set up a META-SHARE repository to host and make available its language resources through the META-SHARE network or equivalent means,
- continue to host the repository of language resources and serve as a META-SHARE node for a period of 18 months starting from 1st February 2013,
- continue to provide technical and/or user support services, software-based and/or human services provided for a period of 18 months starting from 1st February 2013,
- participate in the META-SHARE software development process.

Date and signature of the Legal Representative

PROREKTOR
UNIWERSYTETU ŁÓDZKIEGO
prof. dr hab. Zofia Wysokińska

UNIWERSYTET ŁÓDZKI
ul. Narutowicza 65, 90-131 Łódź

ul. Narutowicza 65, PL 90-131 Łódź, fax (+48) 42-678-39-24, 42-635-40-43
e-mail: rektoratul@uni.lodz.pl

Prorektor ds. nauki – Pro-Rector in Charge of Research
tel. (+48) 42-635-47-50, prornauka@uni.lodz.pl

Prorektor ds. ekonomicznych – Pro-Rector in Charge of Economic Affairs
tel. (+48) 42-635-40-04, prorekonom@uni.lodz.pl

Prorektor ds. programów i jakości kształcenia – Pro-Rector in Charge of Cirricula and Teaching
tel. (+48) 42-635-40-24, pronaucz@uni.lodz.pl

Prorektor ds. studenckich i toku studiów – Pro-Rector in Charge of Students' Affairs
tel. (+48) 42-635-40-06, prostudent@uni.lodz.pl

Prorektor ds. współpracy z zagranicą – Pro-Rector in Charge of International Relations
tel. (+48) 42-635-40-08, prorzagran@uni.lodz.pl

www.uni.lodz.pl

BULGARIAN ACADEMY OF SCIENCES
INSTITUTE FOR BULGARIAN LANGUAGE
"PROF. LYUBOMIR ANDREYCHIN"
52 Shipchenski prohod Blvd. 52, bl. 17
Sofia 1113, tel./fax: 872-23-02
e-mail: ibe@ibl.bas.bg

№ 482 / 18.06.12

Letter of intent

Prof. Svetla Koeva, Ph.D.
Director, Institute for Bulgarian Language
1113 Sofia, 17 Shipchenski prohod Blvd., bl. 52
E-mail: svetla@dcl.bas.bg

The Institute for Bulgarian Language, represented by Svetla Koeva, hereby expresses its non-legally binding intent that the Institute for Bulgarian Language is willing and able to participate in the maintenance and evolution of the META-SHARE infrastructure.

We confirm that the Institute for Bulgarian Language will:

- Set up a META-SHARE repository to host and make available its language resources through the META-SHARE network.
- Continue to host the repository of LRs and serve as a META-SHARE node for a period of 48 months.
- Continue to provide technical and/or user support services, software-based and/or human services provided for a period of 48 months.
- Start or continue to participate in the META-SHARE software development team.

As a non-binding letter of intent, this letter does not create any legal obligations with respect to such matters and has no effect whatsoever until an official agreement on the work envisaged has not been signed and fully executed.

June 18, 2012

Svetla Koeva

Signature

srce
University of Zagreb
University computing centre

Zagreb, 27th June 2012

To whom it may concern:

Letter of Intent

The University computing centre (SRCE), University of Zagreb, represented by dr.sc. Zoran Bekić, Director, hereby express its non-legally binding intent that it is willing and according to its best abilities to participate in the set-up, maintenance and evolution of the META-SHARE infrastructure by hosting the META-SHARE repository of language resources and serve as a network node for a period of 24 months after the end of the project CESAR.

Regards,

Ph.D. Zoran Bekić
Director

Josipa Marohnića 5, 10000 Zagreb, www.srce.unizg.hr, e-mail: ured@srce.hr, tel: (01) 616 55 55, fax: (01) 616 55 59

Letter of Intent

UNIVERSITY OF BELGRADE
11001 BELGRADE,
SERBIA
Studentski trg 16,
P.O.Box 550
Phone (+381 11) 20 27 801
Fax (+381 11) 26 30 151
e-mail: matf@matf.bg.ac.rs
www.matf.bg.ac.rs

Our organisation:
Full name: University of Belgrade, Faculty of Mathematics
Address: Studentski trg 16
Postcode: 11000
Town/city: Belgrade
Country: Serbia
Tel: +381 11 2027801
Fax: +381 11 2630151
Legal Representative: Miodrag Mateljević (dean)

The Faculty of Mathematics, University of Belgrade, represented by Miodrag Mateljević hereby expresses its non legally binding intent that Faculty of Mathematics, University of Belgrade is willing and able to participate in the maintenance of the META-SHARE infrastructure.

We confirm that Faculty of Mathematics, University of Belgrade is willing, according to its best abilities,

- Set up a META-SHARE repository to host and make available its language resources through the META-SHARE network,

- Continue to host the repository of LRs and serve as a META-SHARE node for a period of 2 years,

- Continue to provide technical and/or user support services, software-based and/or human services provided for a period of 2 years

Signature of the Legal Representative of the partner organisation

Name of the Legal Representative: Miodrag Mateljević
Position: dean
Date: 19th June 2012
Place: Belgrade

ĽUDOVÍT ŠTÚR INSTITUTE OF LINGUISTICS
Slovak Academy of Sciences
Panská 26
813 64 Bratislava
Slovakia

Letter of Intent

Our organisation:

Full name:	Ľudovít Štúr Institute of Linguistics, Slovak Academy of Science
Address:	Panská 26
Postcode:	81364
Town/city:	Bratislava
Country:	Slovakia
Tel:	+421 2 5443 1761
Fax:	+421 2 5441 0307
Legal Representative:	Pavol Žigo (director)

The Ľudovít Štúr Institute of Linguistics, Slovak Academy of Science (JÚĽŠ SAV), represented by Pavol Žigo hereby expresses its non legally binding intent that JÚĽŠ SAV is willing and able to participate in the maintenance and evolution of the META-SHARE infrastructure.

We confirm that JÚĽŠ SAV is willing, according to its best abilities,

- set up a META-SHARE repository to host and make available its language resources through the META-SHARE network or equivalent means,
- continue to host the repository of LRs and serve as a META-SHARE node for a period of 18 months,
- continue to provide technical and/or user support services, software-based and/or human services provided for a period of 18 months,
- start to participate in the META-SHARE software development

Signature of the Legal Representative of the partner organisation

Pavol Žigo

Name of the Legal Representative: Pavol Žigo
Position: director
Date: 18th June 2012
Place: Bratislava

Appendix II. List of running CESAR META-SHARE nodes

The following META-SHARE centres are up and running as of 2013-01-31:

- Bulgaria
  - IBL: http://metashare.ibl.bas.bg/
- Croatia
  - FFZG: http://meta-share.ffzg.hr/
- Hungary
  - HASRIL: http://metashare.nytud.hu/
  - BME-TMIT: https://cesar.tmit.bme.hu
- Poland
  - IPIPAN: http://nlp.ipipan.waw.pl/metashare
  - ULODZ: http://metashare.ia.uni.lodz.pl/
- Slovakia
  - LSIL: http://metashare.korpus.sk/