# CESAR

**Central and South-East European Resources**

**Project no. 271022**

## Deliverable D4.5
## Third upload of language resources

**Version No. 1.3**

**31/01/2013**

**Document Information**

| | |
|---|---|
| Deliverable number: | D4.5 |
| Deliverable title: | Third upload of language resources |
| Due date of deliverable: | 31/01/2013 |
| Actual submission date of deliverable: | 31/01/2013 |
| Main Author(s): | György Szaszák (BME-TMIT) |
| Participants: | |
| Internal reviewer: | Tamás Váradi (HASRIL) |
| Workpackage: | WP4 |
| Workpackage title: | Cross-national collaboration and pilot service |
| Workpackage leader: | BME |
| Dissemination Level: | PP |
| Version: | 1.1 |
| Keywords: | upload, batch 3, licence, metadata, META-SHARE nodes |

**History of Versions**

| Version | Date | Status | Name of the Author (Partner) | Contributions | Description/ Approval Level |
|---|---|---|---|---|---|
| 1.3 | 31/01/2013 | | Tamás Váradi (HASRIL) | proofreading | |
| 1.2 | 30/01/2013 | | Tibor Pintér (HASRIL) | proofreading | |
| 1.1 | 25/01/2013 | | György Szaszák (BME) | editing | |
| 1.0 | 23/01/2013 | | György Szaszák (BME) | draft | |

**Executive summary**

This document summarizes the technical background of upload batch 3. Global metadata related considerations, IPR related tasks and work, server node structure and organization is presented. The CESAR community set up one node for the 1st batch, and used the same node as the official node for the 2nd batch. For the 3$^{rd}$ batch and the period after, this official node became CESAR managing node, and several other network nodes were set up.

# Table of Contents

# Abbreviations

| Abbreviation | Term/definition |
|---|---|
| LR | Language Resource |
| LRT | Language Resources and Tools (either language data or tools) |
| IPR | Intellectual Property Rights |
| DoW | Description of Work |
| Editor | META-SHARE server node and editor package V1.0 |
| Partners | Partners of CESAR |

**Table 1. Abbreviations**

# 1. Scope

After metadata description agreement and legal issues clearance, WP4 focused on population and pilot operation of the digital exchange platform codenamed “META-SHARE”. The DoW specifies this work to be carried out in 3 cycles at months 10, 18 and 24, respectively. This document focuses on the third upload batch (3rd cycle) due at M24.

**The tasks covered following the DoW are:**
Resources resulting from WP3 are uploaded to the META-SHARE network, as well as eventually to further appropriate non-commercial platforms (e.g. FLaReNET, CLARIN, LetsMT!, etc.). The physical location and 'hosting' (e.g. central server, owner's own equipment, national/regional data centre managed by a consortium member or a third party) of language resources and tools depends on operational and quality-of-service requirements, the need to provide managed storage services, to monitor accesses and usages, etc., as described in the associated service-level agreements. The consortium complied with the META-NET recommendations and used the META-NET software solutions to implement digital repositories where metadata and/or data are stored or referenced. Understanding that metadata will be harvested by META-NET using the OAI-PMH protocol and used to populate and update the META-SHARE central inventories, the project partners complied with the requirements set out by the harvesters with respect to exporting / making harvestable a set of required metadata elements.

There were no project partners who did not wish to establish and/or maintain an own repository. Resources were 'uploaded' together with their respective descriptions in three stages at M10, M18 and M24. The consortium also participated in early operations of the digital exchange platform, contributed to assessing initial services and provided feedback regarding shortcomings and possible functional and operational improvements.

# 2. Metadata description

## 2.1 Metadata schemes

The CESAR consortium regards the D4.1. deliverable as the basis of metadata description and is dedicated to use metadata schemes agreed between all 3 PSP projects and META-NET.

For the 3rd upload batch, only some minor changes occurred to the metadata schemes. Already for the 2nd batch (July 2012), all metadata description schemes were available for all types (corpus, lexicon, language description, technology/tools) and all media (text, audio, video, image) of resources and tools. The used version of the metadata schemes for the 3rd batch is V3.0. The available V3.0. scheme set consists of a relatively large and complex set of XSD schemes supplied by META-NET and agreed between PSP partners.

### *2.1.1 Updating batch 1 and batch 2 schemes*

Due to some modification of metadata schemes (for example, license notations were changed), batch 1 and 2 schemes had to be updated in M24. The conversion between versions 2.1 and 3.0 was controlled by the contributing Partners, including hand-made checking and validation of all XML metadata description files. Revised files were either reimported into the META-SHARE managing node or imported into the Partners' network node for harvesting by the managing node.

### *2.1.2 Minimal schema*

Partners agreed in providing the metadata description covering at least the minimal schema by the 31th January 2013. However, the description should be as complete as possible and cover possibly non mandatory elements as well in order to provide more detailed information on the resources involved.

## 2.2 Metadata editor

The first release (version 1.0) of the META-SHARE metadata editor (Editor) was made available across all PSP partners by the end of October, 2011. The Editor is part of the META-SHARE server node package, and is intended to be used for metadata annotation, which provides a validated XML description. The current version of the editor is V3.1. which works with schemes V3.0.

CESAR partners have reported several problems to META-SHARE metadata editor software developers during the first upload batch and related to version 1. Version V2.1 and also current version V3.1. of the software were experienced more stable and reliable. Partners had free choice whether to enter metadata via the Editor interface or edit directly the XML metadata descriptions and import them. This latter approach allows for some flexibility, however, validation and re-checking during the import is crucial as offline editing is more prone to errors.

# 3. META-SHARE nodes

## 3.1. Official CESAR nodes for META-SHARE

For batches 1 and 2, Partners agreed to set up one official META-SHARE node in Warsaw, Poland, maintained by IPIPAN. The same node was used for batch 2.

In the mean time, META-SHARE nodes were organized into a hierarchical structure: *managing node*s are synchronized, and provide all META-SHARE metadata and resources, whilst *network node*s are not synchronized, but harvested by a managing node.

As the META-SHARE server software runs in full functionality (including synchronization), CESAR Partners decided to promote the original Warsaw node to become a managing node (CESAR managing node), and set up a network node at each Partner's premises[1] to provide metadata for harvesting by the CESAR managing node, which shares metadata with other META-SHARE managing nodes.

Some of the network nodes will operate from early February as the recent bug-fix version of META-SHARE software was released at the end of January, and installation needs some more time. Metadata to be stored on these is currently imported and hosted by the CESAR managing node, this means that all resources are available by the 31st January 2013.

## 3.2 Data and service security

Backup is done on different levels, allowing for robust and secure data storage.

First, the Warsaw managing node is backed up. The back up process is extended to metadata and all resources and tools. Resources and tools labelled as "unrestricted access" or non available for download (for example, shared on DVD) are backed up by the depositor. Backing up of resources or tools with restricted access is the responsibility of the contributing Partner on its own (the number of this kind of resources is limited).

Secondly, the data and metadata on the managing node is shared via synchronization with other managing nodes.

Beside this, each partner is responsible for backing up its metadata and resources, yielding the third pillar of data security.

---

1

University of Belgrade and Institut Mihaljo Pupin set up one META-SHARE node for Serbia at UBG.

## 3.3 Sustainability

Partners and especially IPIPAN express their wish and commitment to maintain and run the META-SHARE managing node for CESAR after the project ends, at least for a period of two years. This commitment involves all resources referenced in the META-SHARE nodes, but hosted physically elsewhere (according to the letters of intent the Partners submitted for META-FORUM 2012 in Brussels, 2012). However, this document (D4.5) cannot be regarded as a commitment in itself, commitment is guaranteed by the mentioned letters of intention.

## 3.4 Implementation details

The official META-SHARE managing node for CESAR is available at:

http://nlp.ipipan.waw.pl/metashare

All CESAR Partners have received user accounts and passwords to be able to edit their metadata. The server was set up end of October, 2011. Update for version 2.1 was carried out in June, 2012, for version 3.1 in January, 2013.

Communication with META-NET was continuously ongoing via email address helpdesk-technical@meta-share.eu and also via other channels allowing more direct contact to the software developers.

**CESAR network nodes** already in service as of 24 th January are:

*Institute for Bulgarian Language, Bulgarian Academy of Sciences*

http://metashare.ibl.bas.bg/

*Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences*

http://metashare.korpus.sk

*Research Institute for Linguistics, Hungarian Academy of Sciences*

http://metashare.nytud.hu

*University of Lodz*

CESAR http://metashare.ia.uni.lodz.pl

**CESAR network nodes** to enter(ed) in service end of January / early February are:

*Budapest University of Technology and Economics, Dept. of Telecommunications and media Informatics*

CESAR https://cesar.tmit.bme.hu

*University of Zagreb, Faculty of Humanities and Social Sciences*

CESAR http://meta-share.ffzg.hr

*University of Belgrade*

CESAR

# 4. IPR considerations

Licensing is a crucial part of uploading LRTs. License schemes has been continuously developed and codified by META-NET. As for the third upload batch, all necessary basic licence templates and licensing solutions were provided by META-NET. XML schemes and the metadata editor V3.1 also use these templates.

## 4.1 Available licences

The offered licence families are as follows (links provided point to the corresponding META-SHARE document, this list is also available from META-NET at http://www.meta-net.eu/meta-share/licenses):

### *4.1.1 META-SHARE No Redistribution Set*

This set covers all expected combinations of licensing attributes excluding the distribution of the original resource.

- META-SHARE_Commercial_NoRedistribution_For-a-Fee
- META-SHARE_Commercial_NoRedistribution
- META-SHARE Commercial NoRedistribution NoDerivatives_For-a-fee
- META-SHARE Commercial NoRedistribution NoDerivatives
- META-SHARE NonCommercial NoRedistribution NoDerivatives_For-a-fee
- META-SHARE NonCommercial NoRedistribution NoDerivatives
- META-SHARE NonCommercial NoRedistribution_For-a-Fee
- META-SHARE NonCommercial NoRedistribution

- One page overview of the MS-NoRed licences and their attributes

### *4.1.2 META-SHARE Commons Set*

This set covers all expected combinations of licensing attributes, including the distribution of the original resource, but only towards META-SHARE members

- META-SHARE COMMONS_BYNCND
- META-SHARE COMMONS_BYNCSA
- META-SHARE COMMONS_BYNC
- META-SHARE COMMONS_BYND
- META-SHARE COMMONS_BYSA
- META-SHARE COMMONS_BY

- One page overview of the MS-Commons licences and their attributes

#### *4.1.3 Creative Commons Set*

A standard and well documented, widely used legal toolkit for sharing knowledge and data (links below point to the Creative Commons website).

- CC-ZERO
- CC-BY
- CC-BY-SA
- CC-BY-ND
- CC-BY-NC-SA
- CC-BY-NC
- CC-BY-NC-ND

- One page overview of the Creative Commons licences and their attributes

### 4.2. Depositor's agreement

Each Partner should provide a written agreement of the right holder of any of its resources made available via META-SHARE which states that the resource can be licensed in the META-SHARE framework. The agreement should exactly specify the name and the short name of the resource, data of the right holder, the license(s) under which the resource is released. The agreement should be signed and stored by the partner the resource is coming from.

A template for this is provided by META-NET in form of a depositor agreement: META-SHARE Depositor's Agreement

Any special need or problem of any CESAR Partner was discussed via helpdesk-legal@meta-share.eu.

### 4.3 Promoting META-SHARE licences

The wide set of META-SHARE licences allowed for the replacement of the majority of initial CLARIN or END-USER licences with META-SHARE ones, which means a deeper integration and a higher level of standardization, as well as higher compatibility between the licences.

CESAR Partners analysed case-by-case for each LRT planned to be offered so far under a CLARIN or END-USER license to adopt a META-SHARE license instead. Especially, CLIRIN PUB, CLARIN ACA and CLARIN ACA ReD licenses were expected to be at least partly converted into – or co-licensed by – a corresponding META-SHARE licence. This allows dual licensing as an extended option. These new license options were not included in the V1 metadata schemes populated for the 1st batch, but the changes took effect recursively for LRTs involved in the 1st batch (after the update to the schemes V2.1. and now to V3.0.). Most of this work was conducted already by upload batch 2, finalized by upload batch 3.

# 5. Personal data protection

The protection of private and personal data is regarded as a key issue closely linked to IPR clearance of all LRTs involved in upload batches. The basic document related to this issue is recognized as the Directive 95/46/EC of the European Commission. CESAR's guidelines are all based on this document.

## 5.1 General guidelines

A summarized overview of the directive 95/46/EC is available at:
http://europa.eu/legislation_summaries/information_society/data_protection/l14012_en.htm).

Each Partner is responsible for checking whether the released resources comply with the above EC and national level regulation on the protection of private data.

Obfuscation techniques are required to hide personal data in case it is present in the raw material of a language resource. As each resource made available within META-SHARE has its own special characteristics, hereby only general guidelines can be defined which should be carefully adopted for each and every resource. Obfuscation should be carried out such that it ensures that the person cannot be identified by preserving the most data possible.

## 5.2 Personal and private data

All information or data relevant for the explicit identification of a person (name, credit card numbers, address), her/his ethnical belonging, political or religious orientation, personal beliefs, personal privacy and medical condition are regarded as personal data and must be treated and kept according to international and national regulations.

The same stands for all data declared private or secret by the legislation.

## 5.3 Audio and video resources, images

A special concern arises linked to audio and video corpora, that is, speakers may be identifiable after their voice, appearance etc. This identification is regarded to be implicit, i.e. supposes additionally or previously obtained information about the subject (recognize her/him or her/his voice, etc). However, filtering all these data would mean that no speech corpora could be created, for example.

Therefore all such resources should be preferably accompanied by the subject's written or oral and recorded consent to the recording. Further personal data which would allow for the explicit and unambiguous identification of the subject or has relations to the subject's ethnic or politic or religious orientation or his/her medical condition, have to be obfuscated like in other textual resources.

## 5.4 Obfuscation techniques

Personal data privacy usually arises linked to corpora and occasionally linked to lexica. Basically all data regarded as private or sensible has to be deleted from the resource, according to the following principles listed below:

- Preferably data integrity is to be preserved such that tags or other elements refer to the original content.
- An alternate solution is to use false, equivalent data instead of the original one.
- If a part holding private data can be entirely removed from the resource without coverage loss or distortion or replaced by other non-sensible data, this is the straightforward procedure.

In case of audio and video corpora, some additional remarks are necessary:

- If the part containing the private data can be removed (deleted) without corrupting or distorting the remaining data, this is the straightforward procedure.
- False data cannot be inserted.
- Alternatively, in textual transcriptions the obfuscation can be referred to by tags or event markers, referring to the original content if necessary without providing cues, which would allow for its identification.
- Alternatively a beep-like sound or white noise or speech shaped noise can be used to mask the original content; the procedure should be irreversible (preferably delete the original sound and do not mix).
- In case if some speech attributes need to be preserved (prosody for example) a low pass filtering of the private segment can be considered with a cut-off frequency no higher than 300 Hz. However, this usually represent a speech intelligibility of about 20%, therefore the use of this technique is discouraged.

Video and image data can be partly blurred in case they contain private or sensible components. In CESAR contribution, video and image resources represent only a minor part of the resources uploaded, and hence, the licensor is kept responsible to evaluate and choose the corresponding obfuscation technique(s) he/she eventually applies on them.

# 6. Resources uploaded

LRTs to be uploaded in the 3rd batch are listed and presented in details in deliverables D3.3 (documentation of the delivery) and D3.3-B. (actions on resources). Hence, current D4.5. refer to D3.3. and D3.3.-B. regarding the list of the uploaded resources and the actions carried out on them to ensure their extension, standardization, enhancement, linking, etc.