LT_OBSERVATORY – OBSERVATORY FOR LR AND MT IN EUROPE

Acronym: LT_OBSERVATORY

COORDINATION AND SUPPORT ACTION

INFORMATION AND COMMUNICATION TECHNOLOGIES

D1.2 On-line catalogue on resources

GRANT AGREEMENT	644583
DELIVERABLE NUMBER	D1.2
DELIVERABLE TITLE	On-line catalogue on resources
DUE DATE OF DELIVERABLE	31/12/2015
ACTUAL SUBMISSION DATE	31/12/2015 - update 23/12/2016
START DATE OF THE PROJECT	01/01/2015
DURATION	24 months
ORGANIZATION NAME RESPONSIBLE FOR THIS DELIVERABLE	CLARIN ERIC

DISSEMINATION LEVEL		
PU	Public	☒
PP	Restricted to other programme participants (including the Commission Services)	☐
RE	Restricted to a group specified by the consortium (including the Commission Services)	☐
CO	Confidential, only for members of the consortium (including the Commission Services)	☐

TYPE		
R	Document, report	☐
DEM	Demonstrator, pilot, prototype	☐
DEC	Websites, patent fillings, prototype	☐
OTHER		☒

TABLE OF CONTENTS

This project is co-funded by the European Union

DOCUMENT INFO

AUTHORS

Name	Company	E-mail
Bente Maegaard **Lina Henriksen** **Sussi Olsen**	CLARIN	bmaegaard@hum.ku.dk linah@hum.ku.dk saolsen@hum.ku.dk
Margaretha Mazura	EMF	mm@emfs.eu

REVIEWERS/CONTRIBUTORS

Name	Company	E-mail
Vesna Lusicky	UniVie	vesna.lusicky@univie.ac.at
Andrew Joscelyne	LTI	aj@lt-innovate.eu
M^{a} Luz Esparza	ZABALA	mlesparza@zabala.es
All partners	EMF, CLARIN, ZABALA, LTi, UniVie	

DOCUMENT CONTROL

Document version	Date	Change
D1.2 v1.0	23/12/2015	First version submitted to EC
D1.2 v1.1	15/12/2016	First draft for partners' review/completion
D1.2 v2.0	23/12/2016	Final updated version

This project is co-funded by the European Union

1. SUMMARY

From the project summary:

The European Digital Single Market, one of the main goals of Europe 2020, is still fragmented due to language barriers. European society is multilingual; the diversity of its cultural heritage is an opportunity, but hampers transborder eCommerce, social communication and exchange of (cultural) content. Languages without sufficient technological support will become marginalised. These barriers must be overcome by language technology (LT) like Machine Translation (MT) solutions, a need recognized by the future Connecting Europe Facility (CEF).

To support these endeavours to reach an online EU internal market free of language barriers, it is necessary to join, benchmark the quality and facilitate the access to language resources.

Report summary:

This report, attached to the on-line catalogue of language resources, describes the second project deliverable of *WP1 Language resources/catalogues stock-taking*. The first deliverable of WP1 outlined the methodology, the user requirements, the selection criteria, the metadata etc., and it also contained the first collection.

In this report we describe the improvements made since the first deliverable. This work started with identifying and deleting duplicates.

The methodology was slightly modified through metadata optimisation: new metadata were introduced as experience had shown that they would be most useful. Using the modified excel sheet, all language resource (LR) information collected was updated with the new metadata. Additionally all LR information was checked and updated, so that differences in partner understanding of the templates are eliminated.

For the additional resources that came in since the first version, the same procedure was applied.

A “charrette” was organised 4th of December, 2015, with experts who had been looking at a preliminary version of the catalogue as implemented by WP4. This created useful feedback which will be taken into account in the deliverable *D1.3 Best practice guide on LR for automated MT*. However, it is important to note that the feedback was to the WP4 preliminary catalogue version.

Finally, some of the information collected concerns “raw material” for parallel LRs, namely parallel websites, and we have chosen to share this information as well as this can be useful for future work.

This project is co-funded by the European Union

2. AIM

Our aim is to provide a simple, concise and user friendly catalogue of language resources (LRs) that can be used by MT professionals as input for their translation systems. The aim is to sift through existing repositories, select LRs that seem to correspond to pre-agreed usability criteria and have them evaluated by professionals as well. Ultimately, our aim is to provide a methodology and guidelines (which will be part of Deliverable 1.3) that will allow LR suppliers to ensure that more LRs become operationally usable.

3. ON-LINE CATALOGUE

The data collected by WP1 are available at
http://cst.dk/lina/LTO_resources_dec2016.xlsx

The data provided here will become available in the Language Technology Observatory Catalogue provided by WP4.

4. IMPROVEMENTS OF THE COLLECTION

The LR selection and evaluation process has been completed in several stages to ensure an optimal result. Our aim is as mentioned to pinpoint language resources that are "operationally usable" and that will make a difference for MT professionals. Deliverable 1.1 describes the very first step concerning the collection of relevant LR. This report describes the second step where corrections, deletions and additions to the catalogue have been carried out. In this second step the project members have validated and improved the catalogue and besides a committee of user experts have reviewed the catalogue. In order to prepare data for the user review the metadata of some sample resources were uploaded to a newly implemented platform and the user experts[1] reviewed this sample. The experts' feedback was given in a face-to-face meeting on December 4 in Brussels.

In the following we will give an account of all types of improvements – our own improvements as well as those of the expert panel.

[1] Andy Way, Dublin City University
Gert Van Assche, Datamundi
Joachim Van Den Bogaert, CrossLang
Andrzej Zydrón, XTM International
Thierry Etchegoyhen, VICOMTECH

This project is co-funded by the European Union

4.1 LANGUAGE RESOURCE SELECTION CRITERIA

The Language resources initially collected were as described in Deliverable 1.1 selected on the basis of the following criteria:

- *For commercial use.* As our aim is to provide a resource catalogue for MT professionals it is of course essential that the resources can be used for commercial purposes
- *Parallel resources.* In connection with creation of new MT systems monolingual as well as parallel language data are needed. We decided to concentrate our efforts on collection of parallel language data as this is usually considered the most difficult part
- *High quality (preferably best in class).* As regards resource collection we have furthermore aimed for high quality data
- *Low price*. We have preferred data that are either free or can be obtained at a reasonable price.

The user expert review mostly emphasized the importance of the above criteria, but also pointed to other important criteria that we must observe. Some important additional criteria:

- Human validation of resources should have high priority
- Even if it makes good sense that we have focused on parallel data, we should also include some high-quality monolingual data at a later stage
- We should prefer TMX and XLIFF data formats for parallel resources
- An MT developer needs at least 5 million tokens/500,000 segments of language data to build an MT system for a particular domain. This means that so far none of the domains in the collection are sufficiently covered

4.2 ADDITION OF METADATA

In our task of enhancing the value of the collected resources by complementing and optimising the metadata we decided to add some new metadata fields to all the resources. As described in Deliverable 1.1 we initially prepared a list of metadata for the LTO resource based on the outcome of a user study. This original list of metadata is as follows:

- Title
- Resource type
- Creator
- Language
- Availability

- Modality
- URL
- Domain
- Format
- Size
- Production date
- Comment

This list has in this improvement phase been supplemented with the following data: *Description, Tags, Contact person, Format description* (see description of new metadata below). The total list of metadata corresponds to a large extent to the Dublin Core (DC) metadata set for resources, but there are some differences:

- We have chosen *Production date* instead of the DC category called *Date* in order to specify that only the production date of the resource is requested and not the *Change date* or some other date.
- *Size* has been included though it is not part of the DC set as resource size is an important factor in MT system creation
- *Comment* is included as additional information inconsistent with the other data categories is sometimes available and useful, e.g. in relation to the quality.
- *Modality* is included as written as well as spoken resources are expected to be included in the LTO resource.
- *Availablity* has been included as parallel resources often come at a price and sometimes even at a rather expensive price
- *Tags* is included as it will often prove useful in connection with resource search
- DC data categories that are NOT included are: *Contributor, Coverage, Publisher, Relation, Rights, Source* as these are all assessed as dispensable in this context when one of our aims has been to develop a simple, concise and easy-to-use resource collection.

While some of the added metadata fields turned out to be easily filled in, definitely leading to an added metadata value and higher user friendliness, other fields proved hard to find information for.

Contact name: Was found for the majority of the resources. For resources where it is not available, other ways of contacting the resource creator are almost always present such as mail address of the organisation that developed the resource, mail address of the distributor of the resource, or just a contact form to fill in.

Description: This field adds a lot to the user friendliness of a resource. Many resources have good detailed descriptions that give the user a clear understanding of the resource. Unfortunately some resources have very little description apart from language pair and resource type.

Tags: This field is used for keyword-like values, topics that concern the resource in question. Often the values coincide with the values in the domain field but it could also be filled in with information on e.g. genre and text type.

Format description: The specification of the format of a resource is often very poor. The format description was meant to be a way of explaining some special issues about format, character encoding etc. However information for this field has been very sparse, and thus it does not add much value to the present collection.

Comment: Information about the resource, e.g. alignment method, annotations, that does not fit into any of the other metadata fields can be added here. However, the present collection does not have this field filled in very often.

4.3 VALIDATION OF METADATA FOR ALL RESOURCES

All the collected resources have been through a validation process. The metadata fields for all the resources were reviewed, links were checked, and data were corrected when required.

4.4 DELETIONS AND NUMBER OF ADDITIONS

Several partners have been involved in collecting data and though the locations to look for resources were distributed among the partners in order to avoid too many duplicated resources, the collection process still resulted in various duplicates. These have been deleted in the present set of resources.

The total number of resources collected is 149, the distribution of which is shown in the next table:

Parallel corpus	56
Comparable corpus	9
Monolingual corpus	12
Speech corpus	4
Treebank	4

This project is co-funded by the European Union

Tool	15
Terminological resource	25
Wordnet	16
Glossary	3
Thesaurus	3
Translation memory	2

TABLE 1 COLLECTED RESOURCES DISTRIBUTION

4.5 FULL OVERVIEW OF METADATA FIELDS

Below is a list of the metadata fields.

LTO Metadata

Title

Creator

Contact person

Description

Language (mono-, bi- or multilingual)

--select--

Resource type

Availability

--select--

Tags

Modality

Text (do not change)

URL

Domain

Format

Format description

Size

Production date	
Comment	

5. ADDITIONAL SOURCES FOR NEW RESOURCES

The number of relevant resources identified by searching existing catalogues, conference papers etc., is limited if they must comply with the criteria specified in D1.1, and in this report in section 4.1. Existing resources can to some extent be improved, but new resources will also have to be developed in the long run. The project will as mentioned issue guidelines to facilitate this process (D1.3 forthcoming).

We have searched the internet for additional resources that could potentially be developed into parallel corpora, i.e. these are not already usable parallel corpora, but contain the potential to become useful language resources. It should be noted however, that one of the important aspects of creating language resources is to obtain permission from the owners of the materials and the Intellectual Property Rights (IPR) associated with them.

In the following programme/project websites and other sources are listed; some are listed with the available languages, some also with domain.

First, an EC website about financing
http://ec.europa.eu/contracts_grants/microfinance_it.htm (all EU languages)

5.1 ERASMUS+ NATIONAL PROGRAMME WEBSITES:

Erasmus+ is the EC programme for education, training, youth and sport. Erasmus+ is a decentralised programme - the respective websites and documents are de facto bi or multilingual resources: native language(s) and EN.

http://www.bildung.erasmusplus.at/home/EN/ (DE/EN)
http://www.erasmusplus.cy/Default.aspx GR EN
http://www.dzs.cz/de/ CZ EN DE
http://ufm.dk/?set_language=da&cl=da DA/EN
http://www.erasmuspluss.ee/ EE/EN
http://www.cimo.fi/ FI/SE/EN
http://www.iky.gr/erasmusplus/ GR/EN

http://www.mobilnost.hr/ HR/EN
http://www.tpf.hu/ HU/EN
http://www.smpf.lt/lt LT/EN
http://jtba.lt/ LT/EN
http://www.viaa.gov.lv/lat/muzizglitibas_programma/erasmus_plus/erasmus_plus_jaunumi/ LT/EN

5.2 RESULTS FROM LLP PROJECTS IN DIFFERENT LANGUAGES:

LLP was the EC programme for Lifelong Learning until 2014, now continued in Erasmus+.

The corpora below are *de facto* parallel corpora based on the EN original. There are many LLP projects that have multilingual results, in many topics. That may be a valuable first step for rare language pairs or rare subjects to create corpora. One needs to search, but on http://www.adam-europe.eu/adam/ one can search projects according to topics.

Here are some examples:

http://www.e-jobs-observatory.eu/focus_areas/e-culture Cultural skills profiles and guidelines in EN, DE, GR, PT, FR, SI

http://www.e-jobs-observatory.eu/focus_areas/e-learning eLearning skills profiles and guidelines in EN, DE, GR, PL, FR

http://www.e-jobs-observatory.eu/focus_areas/ambient-assisted-living Digital skills for ambient-assisted living, in EN, DE, FR, ES, GR, HU, BG

http://www.e-jobs-observatory.eu/focus_areas/sustainable-ict Skills and competences for jobs in sustainable (green) ICT, in EN, DE, FR, IT, SE

5.3 MULTILINGUAL PROJECT WEBSITES

Alternatively, many of the LLP and EC **project websites** are multilingual, see, just as examples :

http://www.newcap-project.eu/index.php

The CORDIS website is available in EN, DE, FR, ES, IT, PL.

http://cordis.europa.eu/

5.4 PUBLIC SECTOR INFORMATION IN MORE THAN ONE LANGUAGE

Examples :

http://publicdata.belgium.be/ FR, NL
https://www.wien.info Vienna Tourist information in EN, DE, ES, FR, RO, PL, HU, CS, RU, JA, ZH, AR
http://investinaustria.at/en/ Invest in Austria in EN, DE, IT, FR, RU, JA, ZH
http://www.visithelsinki.fi Helsinki Information in FI, SV, ET, EN, DE, FR, IT, ES, RU, JA, ZH
http://www.hel.fi/www/Helsinki/en/administration/enterprises/business/# Helsinki business information in FI, SE, EN, DE, FR, RU
http://www.visitdenmark.dk/ Denmark information in DA, SV, NO, NL, EN, FR, PL, ES, DE, IT, PT (Br), RU, ZH
https://e-justice.europa.eu/content_business_registers_in_member_states-106-at-en.do?member=1 Information on EU business registers, in national language and EN
http://businessculture.org/ Information on business culture in Europe in EN, BG, CS, FI, RO, IT, EL, FR, DE.

Websites and publications of some EU agencies are available in more than one language, and may be considered interesting for specific domains due to the specialization of the individual agencies. The full list of the agencies is available here: http://europa.eu/about-eu/agencies/.

Examples:
http://www.cedefop.europa.eu/en/publications-and-resources/publications Publications in European and some non-European languages, language coverage varies from publication to publication.
echa.europa.eu European Chemicals Agency, available in official EU languages, coverage varies
oami.europa.eu/ohimportal/en/ Office for Harmonisation in the Internal Market, available in official EU languages, coverage varies.

Websites of regional bodies, although some parts (publications, legal documents etc.) may already be included in an available corpus.
Example:
http://www.provinz.bz.it/land/landesregierung/default.asp Autonomous Province Südtirol Bolzano, available in German, Italian, Ladin.

Multilingual websites of higher education institutions, for example:
www.univie.ac.at available in DE, EN

5.5 COMMERCIAL SITES WITH PRODUCT DESCRIPTION IN SEVERAL LANGUAGES OR SEVERAL COUNTRY VERSIONS

The below multilingual websites are publicly available, but they are all protected by IPR – and cannot be used without specific permission. They are all potentially very useful. However, it should be noted, that in the case of localized websites, the individual country versions may reflect localization interventions (e.g. product name adaptations, adaptation of instructions, adaptation of information or products in order to local regulations or different standards) and thus caution may be required when using such websites as sources for parallel corpora.

Useful websites:
www.ikea.com
www.philips.be
www.audi.com
www.volkswagen.com
www.peugeot.com

6. DISCUSSION BASED ON A PRELIMINARY VERSION OF THE CATALOGUE

The user requirements were taken into consideration when creating the methodology, cf. D1.1, Section 2 and in particular *Section 4. User study*. The principles developed there have been guiding the work, but will not be repeated here.

Additionally, as mentioned above, a “charrette”, i.e. a discussion meeting, with experts was held in Brussels on December 4, 2015. The experts had seen a preliminary version of the catalogue as implemented by WP4, and expressed their views on the basis of this. Their comments are of several types. Some of them are related to the fact that they were evaluating a preliminary version of the WP4 implementation, and only a part of the data provided by WP1.

The experts made a comment about the choice of focussing on parallel corpora, as monolingual resources are important as well. The project participants are fully aware that monolingual resources are necessary for SMT, but as this project is limited, and as the most difficult task is to locate useful parallel resources, whereas we believe that monolingual resources are easier available, we have chosen to focus on the parallel resources in this round. As can be seen in table 1 above, the current version contains 9 monolingual corpora; maybe they were not part of the preliminary version seen by the experts.

Some of the experts felt that we should have concentrated only on resources for statistical machine translation, SMT, whereas, as described in D1.1, the project aim is to collect information about LRs relevant for automated translation in a broader sense.

The preliminary version of the implemented catalogue did not make it possible to search using the metadata values. E.g. Title, Domain, Language, Tag, Resource type and other information should be searchable. Otherwise, searching the catalogue becomes too difficult when the catalogue grows.

Overall, the user feedback has been very useful, and the efforts will be continued.

This project is co-funded by the European Union

7. CONCLUSION

With this report the task of collecting information about resources is basically over. But obviously, this does not mean that the catalogue is finalised, it never will be. Work on the catalogue will be ongoing, and the next deliverable of WP1 will tackle the task of producing a *Best practice guide for automated MT.* This should enable various actors to provide more resources for the catalogue and for shared use.

One very important aspect for being able to share LRs is the Intellectual Property Right, IPR. Parallel data available in the public domain are not necessarily open for use by others and for sharing, so it is an important task to obtain the permission from the data owner. In this report we have given some ideas about existing parallel websites, but is has to be stressed that they cannot be used right away. Therefore, the *Best practice guide* will also contain a section on IPR.