LT_OBSERVATORY - OBSERVATORY FOR LR and MT in EUROPE

Acronym: LT_OBSERVATORY

COORDINATION AND SUPPORT ACTION

INFORMATION AND COMMUNICATION TECHNOLOGIES

D1.1 Report on resources

GRANT AGREEMENT	644583
DELIVERABLE NUMBER	D1.1
DELIVERABLE TITLE	Report on resources
DUE DATE OF DELIVERABLE	31/07/2015 - update 31/12/2016
ACTUAL SUBMISSION DATE	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

DOCUMENT INFO

AUTHORS

Name	Company	E-mail
Bente Maegaard Lina Henriksen Sussi Olsen Claus Povlsen	CLARIN	bmaegaard@hum.ku.dk linah@hum.ku.dk saolsen@hum.ku.dk cpovlsen@hum.ku.dk
Andrew Joscelyne	LTI	aj@lt-innovate.eu

REVIEWERS/CONTRIBUTORS

Name	Company	E-mail
All partners	EMF, CLARIN, ZABALA, LTi, UniVie	

DOCUMENT CONTROL

Document version	Date	Change
D1.1 v1.0	30/07/2015	First version submitted to EC
D1.1 v1.1	15/12/2016	First draft for partners' review/completion
D1.1 v1.2	22/12/2016	New version with partners comments
D1.1 v2.0	23/12/2016	Final updated version

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, the first project deliverable of WP1, deals with *Language resources/catalogues stock-taking*. It covers 1) the methodology for collecting language resources for MT development and translation tasks in general, 2) user requirements, 3) information extracted from existing catalogues, 4) information on language resources located through partners' networks etc. It also covers a discussion on the metadata needed for this purpose.

The methodology developed builds partly on experience from earlier EU projects, partly on user needs as expressed by stakeholders in interviews, and at the Dialogue Days in Brussels June 2015 as well as the Charrette in Vienna July 2015 (see separate reports on these). The choice of metadata builds on those two sources (experience from earlier projects and user needs as expressed by stakeholders) as well.

The work has underlined what users repeatedly say: it is very time-consuming to find good and relevant language resources. So, the need for a one-stop shop like the Language Technology Observatory is obvious.

The structure of the report is as follows: After a task description we describe relevant existing catalogues and projects, then the user studies. This leads to a definition of a minimal set of metadata, and selection criteria for the language resources to be part of the Language Technology Observatory catalogue. Chapter 6 discusses quality issues, chapter 7 describes the collection process and chapter 8 provides an overview of the outcome and future work.

2. PROBLEM DEFINITION - TASK DESCRIPTION

Language resources (LR) are indispensable for the development of tools for machine translation (MT) or various kinds of computer-assisted translation (CAT). In particular language corpora, both parallel and monolingual are considered most important for instance for MT. But corpora are expensive and labour-intensive to create or adapt e.g. for MT usability. In addition, the needs for certain LRs for the purpose of developing tools for MT or CAT are quite different depending on the language – or rather on the level of development of MT and CAT systems for different languages. Furthermore the extent of availability of LRs differs considerably from language to language. It is true that LRs are created by EU and national projects and institutions, but they and information about them are scattered across Europe. In order to remedy this lack of overview for the professional user, it is important to apply a user driven approach towards the identification/mapping of best practices in terms of collecting relevant LT resources.

The Language Technology Observatory will provide easy access to information about LRs deemed to be useful for MT and other tools for translation. In order to determine what aspects of an LR are useful for MT practitioners, a user study has been made, providing a guide to the most relevant metadata and the most relevant quality criteria. At the same time, knowledge and best practice has been extracted from previous studies (LetsMT!, META-SHARE etc.) on collecting relevant LT resources.

The purpose of this report is to guide the upcoming efforts in the subsequent WP1 tasks, focusing the efforts on the types of resources and the metadata that are most useful. It is very much a Work in Progress as we gradually extend our understanding of the real concerns of end users of LRs in everyday machine translation contexts.

3. RELEVANT EXISTING CATALOGUES AND PROJECTS

We take a point of departure in work carried out in related projects and co-operations that comprise language resource catalogues: The catalogues we have explored are: CLARIN VLO, META-SHARE, ELRA, OPUS, and TAUS. In addition we also examined major projects such as LetsMT!, LIDER, FALCON, TTC and PANACEA, as well as the European Commission resources of parallel corpora and terminology.

Below we give a description of the various catalogues and subsequently of the projects explored.

CLARIN VLO (the Virtual Language Observatory, http://clarin.eu/content/virtual-language-observatory) is a search facility of metadata for language resources. VLO harvests the metadata descriptions from resources stored in the archives of all CLARIN centres in the various countries, i.e. VLO contains only metadata and links to the actual data. Examples of language resources that can be retrieved through CLARIN VLO are texts (primarily monolingual) of various kinds, sound and video files, images and tools. The resources are in other words of many different kinds, and not all of them will be well suited for MT development. About CLARIN in general, see http://clarin.eu/.

META-SHARE (http://www.meta-net.eu/meta-share) is similar to CLARIN VLO in providing access to language resources throughout Europe, but where CLARIN is a network of centres with repositories, META-SHARE is a distributed network of repositories. META-SHARE resources are within the LT domain and comprise tools and datasets. The number of resources in the catalogue is describes as 2597 Language Resources on the MEAT-SHARE website, including monolingual and multilingual text and speech resources.

The **ELRA Catalogue of Language Resources** (http://www.elra.info/en/catalogues/) is a catalogue for speech, written, multimodal and terminology Language Resources (LRs) for the Human Language Technology (HLT). Many of the resources have been through a validation process. The catalogue comprises both free resources and resources for purchase; members get a discount. ELRA also offers the opportunity to search in other repositories and catalogues. http://www.elra.info/en/.

OPUS - the open parallel corpus (http://opus.lingfil.uu.se/) - is a collection of translated texts from the web. The OPUS project converts and aligns free online data, adds linguistic annotation, and provides the community with a publicly available parallel corpus. The corpus is delivered as an open content package.

TAUS Data is a platform for sharing language data (https://www.taus.net/) . Members can share their translation memories and in return get access to the data of all other members and users. TAUS Data is meant as a super

cloud for the global translation industry, helping to improve translation quality, automation and fuel business innovation.

LetsMT! (https://www.letsmt.eu/Systems.aspx) is a platform for development of statistical machine translation. The platform comprises many corpora of a variety of languages but not all of these are public. LetsMT! resources are described by Dublin Core metadata (with some extensions).

LIDER (http://www.lider-project.eu/) aims at providing an ecosystem for the establishment of a new Linked Open Data (LOD) based ecosystem of free, interlinked, and semantically interoperable language resources (corpora, dictionaries, lexical and syntactic metadata, etc.) and media resources (image, video, etc. metadata) that will allow for free and open exploitation of such resources in multilingual, cross-media content analytics across the EU and beyond, with specific use cases in industries related to social media, financial services, localization, and other multimedia content providers and consumers.

FALCON (http://falcon-project.eu/) is an EU project that aims at building the 'localization web', a decentralised annotated global translation memory and term base. It will establish a Linked Language and Localisation Data standard and provide a platform for the controlled sharing and reuse of language resources, combining open corpora from public bodies with richly annotated output from commercial translation projects.

PANACEA, a STREP Project under EU-FP7 (http://www.panacea-lr.eu/), has developed a factory of Language Resources (LRs) in the form of a production line that automates all steps involved in the acquisition, production, maintenance and updating of the LRs required by Machine Translation and other Language Technologies. The factory is a Web Service-based platform that integrates advanced technological components for:

- Monolingual and Parallel Text Acquisition and Pre-Processing
- Parallel corpora Alignment
- Bilingual Dictionary Production
- Monolingual Rich Information Lexica Production

All LRs are limited to two domains: Environment and labour legislation.

TTC (http://www.ttc-project.eu/) - Terminology extraction, translation tools and comparative corpora, another FP7 project, provides LR in two other domains: wind energy and mobile technology.

CESAR: http://cesar.nytud.hu/about/ - This project (in the framework of the META-SHARE projects) makes available a set of language resources and tools covering the Hungarian, Polish, Croatian, Serbian, Bulgarian and Slovak languages.

EUROTERMBANK:

Following the methodology for evaluation and selection of terminology resources used in the project Eurotermbank (www.eurotermbank.com), the degree of authoritativeness is an important indicator for quality of a terminology resource. When the originators involved in preparing the resource are reputed experts in the terminology field, the quality of the terminology resource is expected to be appropriate. Data originators by highest degree of authoritativeness are legal national and international authorities determined by legislation or jurisdiction, followed by officially authorized harmonization bodies. This authoritativeness principle also ensures that the valid criteria of the methodological approach on which the resource is based are observed.

Finally, we did take the **European Joint Research Centre (JRC)** into account, even if this is neither a catalogue nor a project, as they provide huge amounts of EU parallel data:

JRC – Joint Research Centre (https://ec.europa.eu/jrc/) - is the publisher of the three parallel corpora: JRC-Acquis, DGT-Acquis, and the DCEP-Digital Corpus of the European Parliament, as well as the three translation memories: DGT-TM, EAC-TM, and ECDC-TM, all covering most of the EU languages. This is very useful for the domains and text types covered. (https://ec.europa.eu/jrc/en/language-technologies/).

4. USER STUDY

Apart from identifying existing catalogues and existing resources, the consortium has conducted a limited user study in the language resource (LR) user base of EU stakeholders through interviews with selected LTI members and others. The Dialogue Days in Brussels June 2015 and the Charrette in Vienna July 2015 contributed to this purpose as well.

The general concerns and facts about the availability, quality and usability of LRs for the commercial sector are shared among the user base at large – *free (in the sense of freely accessible, not necessarily without costs), good and usable resources are needed*. But it also became clear that users of LRs (i.e. those deploying an MT engine to translate content for specific purposes) obviously always operate in a context of use that can be characterized most simply for the purposes of our project by the term "industry" or "vertical." (i.e. a commercial or administrative scenario of MT usage). This means that their perspective of identifying and making available LRs for commercial & administrative users will have to start with the *user situation, i.e.* to enable potential end users of LRs to access precisely those LRs that fit their purpose. So, from the existing catalogues only relevant resources should be identified.

The following verticals have been participating in the discussions:

- **Construction** (a major field in Europe, faced with many standards and regulations and much cross-lingual communication due to competition in the sector among major building companies
- **Healthcare** (another complex industry with numerous areas requiring translation)
- **Media monitoring** (for security, marketing, and other business needs)
- **Procurement** (of interest to the DSI constituency)
- **Legal and financial** (wide ranging multi-form needs in a competitive business sector).

These domains will have a specific focus in the collection process, but of course all relevant domains will be taken into account.

Work is ongoing with stakeholders, through follow-up interviews, Dialogue Days and Charrettes, as mentioned, in order to ensure that we have a better understanding of the real needs of users and can identify the various hot points and potential blockages in accessing MT solutions.

4.1 INPUT FROM THE USER STUDY: OBTAINING LANGUAGE RESOURCES

Below we quote some of the observations collected from the stakeholder base.

a) For/from research

There is a general lack of good, cost-free, aligned language data collections for computerized translation processes, especially for language pairs involving smaller language communities (SLCs). Only NGO or government-sourced large *multilingual* corpora are freely available (EU *Acquis*, Canadian Hansard, etc.) but often only for academic research (e.g. Eastern European languages multilingual corpora).

Research institutes who own LRs such as parallel corpora often have no idea about the price/value of their holdings, and sometimes ask for absurdly high licence fees. Industry thinks these should be free as they often result from EC-funded projects (e.g. PAROLE (written), SPEECHDAT (spoken)).

b) In commercial contexts

Enterprises that own LRs do not share their parallel translation data with the world in general as they constitute a strategic value for them. They will make them available to their LSPs (language service providers) when appropriate for a given project, but otherwise not. They own the IP, copyright etc. and they want to keep it.

Companies who are members of TAUS (Translation User Automation Society based in NL) have shared their data via the TAUS repository, which is a paying service for members. The database is searchable for the general public, but data sets cannot be downloaded if you are not a member. Experiments made by some stakeholders suggest that the range of domains covered in the TAUS repository is narrow.

Typical advice given by professionals to their LSP colleagues is as follows: search for potentially useful LRs for your translation tasks on the internet, as there is no single “resource for resources” (which is precisely where the Language Technology Observatory is trying to help).

The problem with this very widespread strategy is that the internet increasingly contains content that has already been translated automatically by Google, Bing etc. which is polluting the linguistic quality of internet-based content in general. Therefore there is no quick method for deciding whether a given LR found on the internet is “original” or “human translated” rather than produced by a machine. Efforts have been made to alert the translation industry, but without any success.

Automated translation jobs in the commercial sector are usually performed on a project by project basis – i.e. for a given language pair or pairs for a particular document or content set. An LSP using MT will therefore typically

ask their client to “give me everything you have” e.g. PDFs, databases, emails, repositories in JSON (the format used for transfer between APIs), data in XML, DDT and XLS files. The risk, apparently, is that the buyer of translation services in the company owning the LRs does not always know what is owned by the organisation (ignorance) or will not send “sensitive” data (privacy concerns).

Multilingual or monolingual LRs?

Most corporate buyers believe *parallel* data LRs (two or more languages side by side) are what is needed to train an SMT system.

But in fact, MT/LSPs suggest that *monolingual* data (in the target language) will become increasingly useful. In statistical MT (e.g. for MOSES, widely used in the EU), a crucial step is to develop a language model for the target language that selects the best translation. Therefore an MT engine needs to be trained on monolingual data as well as parallel data to produce a fluent output.

An additional use of monolingual data is the following: If an LSP can obtain monolingual data from their customer, then they can also have a very good insight into the language used in a (large multinational) company. For example, they may get access to the terminology, and to information about how many people are native writers of the target language, and use words that non-natives don't use, etc.

4.2 PERSPECTIVES ON EVALUATING LRS

There is currently no clear answer from industry about any shared method for evaluating the *quality* of LRs.

There are many aspects of the quality of language resources, and often quality and usefulness for a specific task are interrelated. So we may ask: What *can* be evaluated? What *should* be evaluated?

The answer will partly depend on the evolution of technology in this field, and the capacity of the industry to rise above daily work pressures and find time to collaborate on solutions. We shall ask industry organisations (GALA etc.) about their role in developing methods and scenarios, but it looks as if many decisions in this specific area are taken not by the industry itself but by the *clients* they serve.

Here we provide some of the quality aspects brought up by users (at interviews, Dialogue Days, Charrette cf. earlier):

Time frame of the LR – old lexicons and lists (5, 10, 15 years old) of out-of-date technical terminology will not be relevant to much MT usage today and tomorrow, especially if social media content is to be translated for big data purposes. (Yet it would be appropriate to preserve all these terminology lists for scholars as part of the EU linguistic heritage).

It is widely thought that LR evaluations using some simple list of key parameters for ratings could be crowd-sourced from the user community. But it will inevitably be time-consuming and partial. This is why in commercial contexts LSPs ask for all the data from their customers and then see what works.

It can be noted that ELRA has a full *Validation* procedure for LRs. What is measured is adherence to the standards used, exhaustiveness etc. The validation process is formal (can be checked automatically) as well as manual. See e.g. http://elra.info/en/services-around-lrs/validation/standards-best-practices/: This does not address the value of a language resource for a specific purpose, but it addresses the soundness of the resource as such. This seems to be as far as quality can be measured.

As in our work we are also relying on existing and acknowledged catalogues, we have assumed that their quality criteria are acceptable also to our users. [1]

Basically, when we ask users, it seems that giving **language (pair)**, **domain** and some **basic file format metadata** would be sufficient. However, this relates to usefulness rather than to quality.

So, we believe that if we give users a reasonable list of metadata, they are able to decide for themselves what is useful for them.

User comments on metadata

Users (e.g. LT developers, LSPs, end users using a MT system) usually say they need very little metadata apart from the source, the date, the languages used, plus any unusual encoding information.

Metadata including opinions on the "reliability" or "relevance" of a given LR were not mentioned as useful. This is mainly because users work directly with buyer customer data most of the time, and other LSPs do not necessarily access enterprise LRs. Any further searching is time-consuming and not included in the price.

[1] For further investigation of the evaluation question we will also collaborate with sister projects where applicable. But in the first instance we see that users want to evaluate if a resource is useful for them.

User comments on Clean Data

This is the target for data when LT developers/LSPs talk to their customers about using company LRs. *Clean* means that tags embedded inside documents (especially Word docs) are deleted because they cause problems for the MT engine. The universal preference for MT engines is plain text or XML. This is now standard.

Encrypted data and incrusted OLE data still pose a problem. In the case of character sets, the UTF 8 (and now UTF 16) standard is widely applied. Unicode is the standard for non-alphabetic languages outside the EU. That said, Asian languages may well need to be translated in future within the EU – tourism, global collaboration teams in enterprises, etc. Awareness of this should be noted.

Some experts say that the file format and interoperability problem is now more or less solved. There are numerous filters and converters that allow different formats of text to communicate between different IT systems.

User comments on In-domain data

The "more data is good data" approach much heralded by IBM, Google and others, is now commonplace. But the real question for users is whether or not a given LR is "in-domain" for the translation task in question.

One way to evaluate in-domain LRs is to download small subsets of LRs available (e.g. from the research community) and run a pilot to see how useful the LRs are for the job at hand. This is adequate but time-consuming.

The way the translation operator actually chooses what to use for the translation is not yet clearly described or understood. Maybe it will be beneficial to look at the whole workflow, rather than divide the work into LRs and then the MT system and then the monolingual output. We may learn more from large translation departments, but it should also be noted that there is a lot of development in this field, so new methods for evaluating and choosing may arise.

Enterprise translation memories of technical documentation, although in-domain, become obsolete and decrease in value over time as the company's content changes and as not least technical terminology changes.

Typically, LSPs evaluate in-domain relevance "manually" (i.e. reading it) or by running a pilot MT job. However there are signs in the LT community that the identification of in-domain relevance of a language resource can be achieved automatically, so some suggest:

a) Use a terminology extraction tool and compare the data set of words in the document to be translated with an expected (or benchmark) data set for the customer/job. If the extracted terms show a low match compared with the benchmark, then the LR in question would probably not be very useful for the job at hand.

b) Build an automatic document categoriser/domain analyser. This may now be underway. The idea is to use a resource such as BabelNet (http://babelnet.org/) - a very large multilingual dictionary and semantic network (from an EU project) - to build a rapid *bilingual* dictionary for a document (rather than in (a)). This can then be used for both MT and for mixed man/machine translation solutions, as it gives a rapid semantic fingerprint of the domain and words involved in any document or collection that is processed. A standard such as LEMON RDF (http://lemon-model.net/) can be used to link different sets of online lexical data, making the tool even more powerful.

4.3 FIRST OVERVIEW OF LR USABILITY

Defining LR Usability

Usability is a complex but vital criterion for evaluating and using Language Resources in the context of a one-stop shop catalogue service for the community (one of the objectives of the LT_Observatory project), where relevant information, time and quality are key values for decision-making.

The aspects of usability on which users are focusing, are:
LR usability = ease of download & domain relevance & language pair & availability/cost & time to implementation, where '&' signifies some relation. Many of these aspects have already been covered above.

It is important that usability in this context is understood as being a criterion involved in specifically human decision making. It is entirely possible that language resources will in some future development be selected automatically by digital services operating as part of a broader and deeper language and translation infrastructure, as indicated by some of the EC-funded projects (LIDER and FALCON) listed in section 6. If this innovative approach emerges as a model for LR selection and usage in future years, then the initial human community decision-making carried out under LTO will provide a powerful foundation for any future transformation of LR evaluation into an automated process. Here we mention three important aspects of usability. Language/language pair and domain are of a different nature.

1. **Ease of download**

This refers to the simplicity (number of clicks) of finding the relevant LR on some dedicated LR website. This means creating a "LR access" rating for the various LR repositories, and also promoting those repositories on websites and social media that best serve the translation industry. The quicker the user can find a likely LR, the better the repository/link is as a Language Technology Observatory resource. This sounds very simple, but when users mention it, there is a reason: it seems to be the only way to encourage the use of existing data, and this fits well with the Language Technology Observatory mission.

2. **Availability/cost**

The cost of an LR is obviously an important factor. If the price is very high, it may not make sense to purchase the resource. It is most easy to handle free resources with easy licences, so this is preferred by the user base. The current development where more and more public data become freely available is positive in this respect. However, it has to be taken into account that quality comes with a cost, so if a resource has been validated, e.g. by the private sector, there may be a fee to pay, and this may be a good investment for the user.

3. **Crowdsourcing usability**

It would be useful for everyone to have some device to rate LRs when they have been used, by attaching an evaluation questionnaire to the repository portal. It would give some guide to the benefits of some LRs. Rating info could also be available via social media for the translation industry and others to share these ratings with all-comers and improve the utility of the crowd-sourced opinion.

4.4 USE OF INPUT FROM USER STUDY AND DIALOGUES AT EVENTS

Above we have given a rather open description of the user needs as seen in our discussions with stakeholders. All of the main aspects mentioned have been taken into account in the methodology chosen for the resource collection.

5. METADATA DISCUSSION

It is essential that the LTO catalogue provides users with easy access to the resources they need. This means that the search options of the LTO catalogue must accommodate exactly the information types (metadata) that users are interested in. Therefore the metadata categories selected for the LTO catalogue (i.e. the basic data that must be provided for each resource, summarizing basic information about the resource) are based on the user study, the previous usability check list as well as on experiences from similar projects.

Similarly, it is important that the language resources selected from various existing catalogues for inclusion in the LTO catalogue are exactly the resources that users need for MT-purposes. They must therefore be chosen by means of carefully prepared selection criteria. Subsequently we have "translated" these selection criteria into suitable metadata categories that actually exist in the various catalogues and that have given us the possibility to retrieve exactly the resources that suit our purpose best.

In the following we will first sketch the use of metadata categories in some related projects, then describe metadata categories selected for the LTO catalogue and finally list the selection criteria used to identify language resources to be included in LTO catalogue.

5.1 METADATA CATEGORIES FROM RELATED PROJECTS

As mentioned above (4 User Study) users have expressed some types of information that should describe each language resource to ensure that the resource is interesting in terms of MT. This information describing each resource can be found in the resource metadata. It was therefore important that we identified the metadata types that contain the information types that are crucial in this context and that we decided on the desired values for the minimal as well as the optimal set of metadata categories.

In the below table we compare the use of metadata in three major projects. The metadata comparisons are only approximate as LetsMT! presents a genuine metadata list - and equivalent lists do not exist for CLARIN VLO and META-SHARE as these search facilities are based on resources from a distributed network and as such cover hundreds of metadata.

LetsMT! - DC-metadata categories with description	VLO	METASHARE	Comments
Title (DC)	Collection (names)		These categories are for the name of the resource.

LetsMT! - DC-metadata categories with description	VLO	METASHARE	Comments
		Resource type (corpus, lexical/conceptual, tool service, language description)	This information type may be useful for identification of text resources as VLO and METASHARE contains many types of resources
	Resource type (e.g. text, annotation)	Media Type (e.g. text, audio, image, video, text numerical, text N-gram)	
	Modality (e.g. speech, spoken, written, signs, gestures, etc. and combinations)	Modality type (written, parallel.)	Probably close to the above resource/media type information
Type (DC) (The corpus type can either be monolingual, bilingual or multilingual		- Linguality type (monolingual, bilingual, multilingual) - Multilinguality Type (parallel…)	Type/Linguality type important to identify bi- or multilingual resources
Subject (DC) (closed list)	Subject (e.g. helbred (dk5-61), bankwezen, 02.50 esoterik)	Domain	Here we have the subject or domain of the resource – probably relevant.
Date for text production		Time Coverage	May be relevant for identification of contemporary resources
Text type (closed list)	Genre (open class)		Text type or genre may be relevant
Data provider	Data Provider		May be an indication of quality
Rights (DC) (Description of access policy)		- Availability - License - Restrictions of Use	Important – the resources must be available
Source language Target language	Language	Language	
Translation			States whether the resource is a translation
Original language			States whether the resource is an original
Alignment type			States whether the text is aligned by sentence, section etc.
Corpus size			
	Format (e.g. video/vmw,	-Conformance to	The text format may

LetsMT! - DC-metadata categories with description	VLO	METASHARE	Comments
	text/rtf, audio/wav)	standards/Best Practices - MIME Type	have relevance

5.2 LTO METADATA CATEGORIES

The below list of LTO metadata is the result of the comparison of the user study and the projects investigated. This is to be seen as the minimal set of metadata and all categories should preferably have content. However, since it is not always possible to find the information, we cannot make them mandatory.

Metadata category	Description
Title	Name of the language resource, e.g. *Estonian-Latvian parallel corpus of building product texts*
Type of resource	Resources of different types are interesting for LTO. Initially the list contains five resource types, but more may be added when experience shows. Corpus, classifications scheme, lexical resource, terminology resource, tool
Creator	Creator of the language resource**.** The creator can be an institution as well as a group or a person, e.g. *European Commission* or *Luisa Coheur*
Language(s)	The languages of the resources. The languages must be stated as ISO codes.
Availability	This data category gives basic information about the price of the resource. The 3 options are: *free, reasonable price, expensive*.
Modality	For the time being we only have text in the LTO catalogue and therefore the only option for this data category is *text*. In the future we may have other modalities such as speech and spoken streams in video.
URL	The URL of the language resource.
Domain	Open class containing both text type e.g. *press release, annual report* and domain e.g. *automotive, pharmaceuticals*.
Format	Format information, e.g. *plain text, RTF*.
Size	The size of the language resource and the size unit, e.g. *300,000 words* or *10,000 sentences*.

Metadata category	Description
Production date	The date when the resource was created. Production date can be stated as date or year. Some resources do not have a production data but perhaps only the date of inclusion in the catalogue. A comment about this can be made in the comment field.
Comment	Free text field where the user can provide any information that is considered of relevance.

We have found some extra categories which would be very useful to include. These two categories are important for construction of MT system but are primarily supplementary and make part of an optimal metadata set.

Bi/multilinguality type	In this data category it can be stated whether the bi- or multilingual corpora is *parallel* or *comparable*
Alignment type	In this data category it can be stated the type segment used in connection with alignment: e.g. *sentence, word* or *section*

Finally, cf. 4.3 users have expressed the wish to have a possibility of entering the **evaluation of a resource** on the portal. When several users have used a resource, this will become a valuable part of the LTO portal. META-SHARE shows for each resource how many times it has been downloaded, but it does not give details on user satisfaction, ELRA does not provide details on sales, so this will be an interesting feature.

Although this will tell the users about the value of the resource, this cannot be part of metadata as it is not constant, so it is an input to the portal designers.

5.3 SELECTION CRITERIA FOR COLLECTION OF LANGUAGE RESOURCES

Based on section 4.3 and on experience from previous projects, the below selection criteria have been used as a framework for extraction of useful language resources. The selection criteria have as mentioned been "translated" into the metadata categories that could be identified together with the particular resources. There is e.g. not a metadata category in any of the resources called "longevity", but there may be one called "change_date", "update" or just "date" or something similar. The resources make use of different metadata categories, some standardized and some not, which means that the selection criteria have been "re-translated" many times.

The list below shows the most important selection criteria with a rating from 0-2 where 0 is the least accepted/desired value.

Availability

2 – the resource is available and it is freely, openly available under sensible; Open Source or Creative Commons licenses that allow re-use and re-purposing;

1 – the resource is potentially available but its licenses need negotiation; there may be a cost.

0 – the resource has restricted access.

Languages covered

2 – bilingual, multilingual

1 – monolingual

And we need to cover as many languages as possible

Longevity:

1– resource is actively maintained; the current version is less than 5 years old.

0 – resource is unmaintained.

Validation of resource

2 – extensively tested;

1 – moderately tested resource, with some room of improvement;

0 – untested.

Modality of resource

1 – text – at present we are only selecting written text

0 – other

Ease of download

1 – it is easy to download/obtain the resource, at most 3 clicks

0 – too heavy

6. QUALITY ASSURANCE

Quality is always a key issue in connection with the use of LRs whether they are text corpora, terminology collections or other resources. Many users of LRs say that the quality of LRs varies widely. The perception of high quality however also varies widely from one organization, from one user, or from one task to another (see also 4.2 Perspectives on evaluating LRs). A language corpus considered high quality in one context may be a low quality corpus in another context, and to some extent the same goes for terminology collections. Moreover – and perhaps for the very same reasons - no commonly accepted "LR quality standards" exist which also greatly attributes to the difficulties in deciding what a high quality LR really is.

It is important to keep in mind that corpus and terminology collections are usually created for specific purposes and within particular frameworks, and this is the information that is crucial to pass on to other users of the resources. Here metadata come into the picture; with metadata we can pass on all or most information about the resource in terms of for example creator -, content -, language -, size - and usage information. When an LR is provided with sufficient metadata that thoroughly describe the resource, the user can decide for himself whether it is likely that this resource is a high quality resource in relation to the particular task.
ELRA (European Language Resources Association) has the goal of promoting the creation, validation, standardization and distribution of LRs. ELRA's Validation Committee has developed a methodology for validation of written as well as spoken LRs. The term validation is here to be understood as automatic, semiautomatic or manual quality evaluation of the resource against some criteria. These validation criteria include 1) the documentation of the resource (general or specific explanatory information that describes the resource, contact data, copyright, language etc.), 2) the formal properties of the resource (functional verification checks; can you work with it as described in the documentation) and 3) the content of the resource (a manual check whether the resource contains what is described in the documentation).

We have had an eye to ELRA's recommendations and have prioritized to evaluate and ensure high quality of documentation information. I.e. we have checked the metadata describing the resource and added/corrected metadata where possible. Validation of the formal properties of the resource and especially validation of the content of the resource are outside the scope of this project.

7. THE COLLECTION PROCESS

The methodology for collection of language resources relies on two main features: 1) the identification of existing relevant language resource catalogues and projects as described in *3 Relevant existing catalogues and projects* and 2) the list of selection criteria which is based on the user study as well as experiences from related projects - as described in *5.3 selection criteria for collection of language resources* and 3) input from partners' network.

In this section challenges and special characteristics of the related relevant projects and language resource catalogues are described. Many resources are e.g. only for non-commercial use, many resources lack information about date and maintenance and sometimes resources have no creator but only a license holder. We concentrated on bilingual and multilingual resources.

CLARIN VLO

CLARIN VLO is a user friendly search facility that offers users the possibility of making queries for a wide span of language resources embedded in the repositories of the CLARIN centres. In this context, search patterns such as e.g. *parallel corpora* provided links to many useful data resources that proved relevant for being included in the list of parallel data, constituting the basis for implementing SMT systems.

META-SHARE, including LRS from linked projects such as CESAR, META-NORD: Most LRs on META-SHARE that were downloaded were either totally free or subject to CC with an easy-to-accept CC license upon which the resources could be downloaded. Many LRs though, do not even have an indication how to use them. Some provide a (high) price without the option to view them, some provide a contact person that, alas, in most cases does not respond. Hence, improvement of this practical issue would considerably render LRs useful.

ELRA: http://www.elra.info/en/catalogues/catalogue-language-resources/ The ELRA catalogue has the following content under *Written LR*: 82 monolingual lexica, 62 multilingual lexica and 108 written corpora. Out of the written corpora 39 are multilingual. The ELRA catalogue has a reasonable size, so searching it is reasonably easy. Some of the ELRA LRs are too expensive to be considered here, but many have been chosen as relevant.

OPUS

As mentioned OPUS (http://opus.lingfil.uu.se/) is a collection of translated texts from the web. OPUS provides the research community with open parallel data. It is an easy to use website, licenses are in general clearly marked, and download easy to make. But many of the resources are not available for commercial use.

TAUS

TAUS Data Cloud (https://www.taus.net/data/taus-data-cloud) is a neutral, secure repository platform for sharing, curating, pooling and leveraging language data. It is membership based (cost €2,500 per year to join) and enables members to upload their own data and download other member' data to train their translation engines. There is a free access programme for universities, as well as free access to its term search functionality. The data presumably remains the property of the "owner" who uploads it. It is not available in any other way than becoming a member.

TAUS Data claims to have 2,218 of language pairs in its repository and a total of 62.4 billion words. Many language pairs have been generated by translating a resource via English as a pivot language into a 2nd language. There is no question that for machine translation training alone, TAUS Data is the largest and best organised resource available in terms of quality of the data, maintainability, services offered and effectiveness. It mainly targets organisations that need regular large-volume translation operations in a domain covered by the language resources, and have deep pockets.

TERMINOLOGY

Terminology resources were extracted from CLARIN VLO, META-SHARE and ELRA catalogues, following the methodology for evaluation and selection of terminology resources used in the project Eurotermbank (www.eurotermbank.com), in which the degree of authoritativeness was set as an important indicator for quality of a terminology resource. As terminology resources may fastly become obsolete if not managed and updated appropriately, the most sustainable management of the resources may be assumed by the data originators of the highest degree of authoritativeness. For this reason, the collection of terminology resources extracted from the catalogues has been supplemented by terminology resources of data originators of the highest authoritativeness (international bodies, national terminology centres etc.) that have not been found in the catalogues. Many terminology resources collected are only available through a web interface and the actual availability of the data is restricted or for non-commercial purposes.

EU Projects:

LetsMT! is as mentioned a platform for development of statistical machine translation and not a language resource catalogue as such. The LetsMT! system does include approx. 170 texts/text collections (dependent on how they are counted), but they are not downloadable from the LetsMT! website. However, many resources are listed with the URL of the resource home - where it originates and where it is maintained. Many resources can be downloaded from there. It should be noted that most of the LetsMT! resources are only for non-commercial use and therefore very few are included in the LTO catalogue.

The catalogue includes LRs from other EU projects, e.g. ACCURAT.

PANACEA: http://www.panacea-lr.eu/en/info-for-researchers/
Language resources in English, French and Greek in the environment and labour legislation domain.
Monolingual data contain between 40 and 50 Mio tokens; bilingual data contain between 600.000 and 900.000 tokens. The LRs on labour legislation are stored at ELRA, the LRs on the environment at the repository of the Universitat Pompeu Fabra. All LRs have Creative Common licenses, those of ELRA for non-commercial purpose only, those of UPF are freely downloadable. Their creation date – 2012 – makes them fairly new and therefore, potentially relevant.

TTC: http://www.ttc-project.eu/
The TTC project collected LRs in the area of wind energy and mobile technology in 5 languages: EN, FR, ES, DE, LT plus Chinese and Russian. The LRs are defined as "comparative corpora". They are stored at the website of the University of Nantes and are freely downloadable. Their creation date – 2012 – makes them fairly new and therefore, potentially relevant.

CESAR Project (in the framework of the META-SHARE projects): http://cesar.nytud.hu/about/
The project made available a set of language resources and tools covering the Hungarian, Polish, Croatian, Serbian, Bulgarian and Slovak languages. Resources include interoperable mono and multilingual speech databases, corpora, dictionaries and wordnets and relevant language technology processing tools such as tokenisers, lemmatisers, taggers and parsers. Its practical use differs considerably depending on the creating partner. For example, it was not possible to access any Polish sources, or those of one Bulgarian partner.
It contains monolingual LRs (HU, BG, SK, HR) as well as bilingual and multilingual corpora in the Southeast European languages. Some LRs are free, some are for academic use only. Although monolingual sources were not considered for this document, one example might be interesting: The Hungarian Kindergarten Language Corpus, with the language used by small children, which might be important for Edutainment purposes.

LIDER http://www.lider-project.eu/
LIDER is building a Linked Open Data (LOD) based ecosystem of free, interlinked, and semantically interoperable language resources (corpora, dictionaries, lexical and syntactic metadata, etc.) and media resources (image, video, etc. metadata) that will allow for free and open exploitation of such resources in multilingual, cross-media content analytics across the EU and beyond, with specific use cases in industries related to social media, financial services, localization, and other multimedia content providers and consumers. It is not clear which translation-oriented corpora will be included or which specific language pairs are involved. The main aim is to provide a semantic infrastructure layer for adding value to any type of resource.

FALCON

The FALCON project combines the power of open data on the web with data-driven language technologies to construct the Localization Web. This consists of a network of terms and translations inter-linked to each other and to source and target documents via URLs. FALCON will integrate the resulting web of linked localisation and language data into localisation tool chains using existing data query and access control standards. Metadata from these tools will add value to these data assets, enabling seamless quality monitoring across the value chain and their on-demand leverage in training machine translation and text analytics engines. Like LIDER, it is an "infrastructure" project that attempts to leverage existing semantics resources to enrich "publishing" type resources that appear as HTML code. It has no specific data sources itself, but will if successful enable everyone involved in localising and translating to leverage semantic insights about word meaning from linked data sources.

ELRC

The objective of the European Language Resource Coordination (ELRC) action, launched by the European Commission, is to identify and gather language and translation data relevant to national public services, administrations, and governmental institutions across all 30 European countries participating in the Connecting Europe Facility (CEF) programme. By the end of this project, ELRC had 25 available resources of which 11 have been included in the LTO collection. However for some resources the license conditions were still under negotiation and thus could not yet be taken into account.

JRC

Of the three corpora that JRC publishes, only the DGT-Aquis is available for commercial use. The DGT-Acquis is a family of several multilingual parallel corpora extracted from the Official Journal of the European Union, consisting of documents from the middle of 2004 to the end of 2011 in up to 23 languages.

The JRC-Acquis and the DCEP-Digital Corpus of the European Parliament are only available for research purposes. All three corpora can be downloaded from the website; there are various downloading options and instructions for bilingual extractions.

8. ADDITIONAL COMMENTS

Commercial uses of parallel corpora language resources appear to be focused on individual end user cases or on the re-use of existing translation memories (owned by suppliers or organisations) if they are in-domain. There is obviously no exhaustive list of these resources available.

Terminology resources present a different case. There are a few large banks of terminology that are available for online use. Their main drawback is quality: users say there is out-of-date or erroneous content mixed in with high quality content. Terminology is now often shared by translators in repositories such as Proz, or can be searched online in the TAUS Data Cloud. But here again, there are no metadata that can provide quality controls on accuracy. IATE data can be downloaded via an API for MT use, but most term data cannot.

More generally, there are currently a number of different projects or efforts dedicated to identifying and networking language resources as a necessary asset for the multilingual digital single market. It is imperative that **there is overall agreement on the key goals of these projects and on the manner of evaluating success in reaching their various goals**. By working closely with both the translation/language industry (and its clients) and the academic and research community, LT-Observatory feels well-positioned to aid all parties concerned in reaching agreement on LR usability criteria which we believe will be most useful for starting the next stage of repository analysis, data collection, and inventing the next model of language data dissemination (sharing, market, crowd-sourcing/evaluating, etc.).

In this context, it also appears to be necessary to examine – yet again – the importance of ***legal constraints on data sharing (copyright, IPR, ownership, etc.)*** as there is still a lack of clarity in the minds of many about the rights of Europeans with respect to using LRs. This will also involve critiquing certain current practices that appear to add further barriers to building a healthy automated translation culture at a time of maximum need.

9. CONCLUSIONS AND FUTURE WORK

9.1 AMOUNT OF RESOURCES

During this phase and using the methodology described above, information about 146 resources have been collected. This comprises 25 terminological resources, 75 corpus resources, 16 wordnets, 15 tools, and some thesauruses and translations memories. According to the selection criteria (building on users' expressed needs) only information about resources that are available for commercial use (or general research use) and that are either free or with a moderate licence fee will go into the LTO portal.

The most important conclusion is that many resources exist which are useful for MT and similar work, but the majority are for (academic) research or educational use only, and as such **not available for commercial use**.
If companies have collected useful language resources for their own purposes, this is an asset that they do not easily share with their competitors. During the user study companies expressed that if they need LR they search the web and use what they find, this is often the fastest and cheapest way. It should be noted that it may happen that some of the resources collected this way are actually not available for commercial or any other use because of IPR problems, as mentioned in the previous chapter.

The consortium is continuing the work, and more resources are being retrieved until the end of the year and made available on the LT Observatory.

9.2 LANGUAGES COVERED

An investigation of the languages covered by the language resources identified until present has been made:

Language list	No of resources	Language list	No of resources
Bg	12	**it**	18
Bs	1	**lt**	17
Ca	1	**lv**	18
Cs	12	**mk**	1
Da	16	**mt**	7
De	32	**nl**	14
El	29	**no**	4
En	114	**pl**	11
Es	39	**pt**	22
Et	16	**ro**	15
Fi	17	**ru**	2
Fr	64	**sk**	10

Language list	No of resources	Language list	No of resources
Ga	6	**sl**	15
Gl	4	**sq**	1
Hr	15	**sr**	1
Hu	13	**sv**	17
		tr	1

TABLE 1 COVERAGE OF LANGUAGES IN THE RESOURCES IDENTIFIED.

Table 1 shows the list of languages covered by the identified resources and the number of resources each language appears in. For some large resources smaller languages such as Greenlandic, Icelandic and Latin have been reported as 'other'.

As illustrated, the languages covered are mostly official EU languages, some EU minority languages such as Catalan or Galician and a few resources comprising languages from neighbouring countries such as Russia, Norway, Serbia, Albania, etc. The list clearly illustrates that the large languages are included in most resources. Out of 149 resources and tools identified, English forms part of the 114, French of 64, Spanish of 39 and German of 32. For the other official EU languages the number of resources varies from 12 to 29.

9.3 DOMAINS COVERED

Domain is normally crucial when looking for relevant resources. Therefore 'domain' is one of the LTO minimal metadata. However, no standard agreed classification of domains exist that can be used. JRC provides Eurovoc, a comprehensive classification of domains, but this is too detailed and therefore complicated to use. The consortium decided to start with an open class of domains, in order to see which domains are most relevant. We believe that a small number of domains will be the most useful, always with the possibility to add more domains when needed. Here e.g. the LetsMT! domains may be useful: *Law, Finance, Business, Information technology and data processing, Electronics, Industrial manufacturing, Biotechnology and health, Environment, Energy, Transport, Communications systems, Tourism, National and international organizations and affairs, Education*. Or alternatively the domain classification of TAUS Data Association: *Computer Hardware, Computer Software, Consumer Electronics, Financials, Industrial Manufacturing, Legal Services, Pharmaceuticals and Biotechnology, Professional and Business Services, Telecommunications, Undefined*.

The collection of references to domains in the current database basically can be split up into two categories. One that represents general language usage and one that covers various subject areas. An example of the former category is various newspaper corpora including *Newswire* data. Domain information adhering to the subject area category consists of *traveling tourism, computer science, medicine, economy, law, and environment.* Compared to

the domains mentioned by the users (section 4), we see a pretty small overlap: *health* and *law* (if *healthcare* can be taken to overlap with *medicine*).

As the current description of the LTO domain metadata refers to subject as well as text type/genre, text type has also been used for this field; this means that both *Europarl* and *subtitles* appear as domains. Europarl is a transcription of spoken data and therefore not really the same type of language as written texts; similarly subtitles are produced under strict conditions where vocabulary choices are made because of space limitations. This use of the domain field is in line with current use of the term *domain* (the Twitter domain, the Europarl domain) and may or may not be useful. This is a discussion point for future meetings with the users.

9.4 FUTURE WORK

Methodology update: This work package started out with the discussion and the development of the methodology to be used. This methodology was made building on user studies and previous experience. However, the methodology has been modified underway, e.g. the minimal set of metadata we started out with was too minimal, and more has been added. The consortium is aware that more modifications may arise when more discussions with users are made and more data is collected.
E.g. the consortium believes a small list of domains should be suggested, as we believe this will improve the usability. This will be discussed with users.

More resources: The consortium will be continuing the identification of resources that will be made available on the LT Observatory in December.

Valorisation: One of the next steps in the process is to select LRs that can be of better value if modified. One of the actions will be to add or modify domain information along one of the lines described above.

Cleaning up: As resources can be identified through different search paths, the current list contains a few duplicates; these will be removed. Additionally we will check the information about all LRs, and see if it lives up to our selection criteria, e.g. the list currently contains information about some resources that are available for academic research only. It may be checked with the users if they want this information anyway.

Discussion with users: In parallel discussions with users will take place; in particular we now have a good starting point for a better discussion with users about the value and evaluation of language resources.

Coverage and gaps: Deliverable D1.4 will discuss the coverage of the language resources identified and will identify gaps.

To sum up: The collection task has been very useful as it has put the methodology to a test. The amount of resources is not as large as it might have been, but we need to remember that the selection criteria are strict – no resources which are not for (potential) commercial use – and more resources may still be added. The talks with users have also given useful input for the portal to be built, e.g. the idea of a tool for user evaluation.