ELRC DATA

SMART 2015/1091 Tools and Resources for CEF Automated Translation Lot 3

Andrejs Vasiļjevs, Tilde

LRB Meeting

Berlin, March 28, 2017

European Language
Resource Coordination

ELRC DATA Objective

ELRC DATA is a continuation and expansion of ELRC activities in the acquisition of additional Language Resources and related refinement/processing services (with e.g. anonymisation) and their provision to the Language Resource Repository of CEF Automated Translation platform.

Short-term objective

- to acquire high-quality IPR-cleared language resources
- in all CEF languages
- in topical areas relevant to CEF DSIs.

Long-term objective

To **contribute** and **complement** language resource collection in the Member States (ELRC Networking), in view to reach an acceptable level of automated translation quality in key areas of CEF DSIs, where multilingual functionalities are needed.

ELRC + L3 Tasks

- **Language Resource identification**
- **Language Resource processing**
- **Language Resource compiling and production**
- **On-site assistance to data providers**

Language Resource Identification

- Identification of language resources
- Includes manual as well as automatic discovery of domain-specific data from the web
- Identify the licensing conditions and the right-holder(s) of the resources
- Collection of language resources
- Dissemination activities

Language Resource Identification

- Your support is needed and appreciated!
- Prioritization in respect to the needs of MT@EC
- Identify potential institutions/data holders
- Metadata preparation – description of data
- Identify the need for onsite assistance

Language Resource identification

Group 1	Group 2	Group 3
Languages where EC don't have enough data and want as much as possible	Languages where EC could use more data	Languages where EC have a lot of data and need more data only for the purposes of domain specific MT engines
Croatian	Bulgarian	English
Icelandic	Czech	French
Irish (Gaelic)	Danish	German
Norwegian	Dutch	Portuguese
	Estonian	
	Finish	
	Greek	
	Hungarian	
	Italian	
	Latvian	
	Lithuanian	
	Maltese	
	Polish	
	Romanian	
	Slovak	
	Slovene	
	Spanish	
	Swedish	

Priorities and requirements for Language Resource identification

- Linguistic data of primary interest for MT@EC:
  - parallel bilingual or multilingual corpora with any of the CEF languages combined with French or English
- Language pairs not involving English:
  - not directly useful at the moment
  - potentially useful for eTranslation/MT@EC
  - every case should be discussed on an individual basis
- Variety of language or different variant or orthography should be distinguished
  - E.g., whether the data is Norwegian Bokmål or Norwegian Nynorsk
  - Although tools will be used to detect spelling variations (e.g., Portuguese before the spelling reform or after), those should also be marked, if known
- Monolingual data, glossaries etc. should also be collected and processed. These types of resources are just not a top priority for the DGT right now.

LR Stakeholder Network

- LR Stakeholder Network members will serve as a supporting contact point in their respective sector or field of operations, providing a crucial nexus for identifying resources:

  * Public Services Data
  * EU Parliamentary Data
  * Media Data
  * Culture Data
  * Terminology Data
  * Research Data
  * Corporate Data
  * Localization Services Data
  * Lexicons
  * Publishing Data
  * Justice
  * Other sectors relevant to CEF.AT

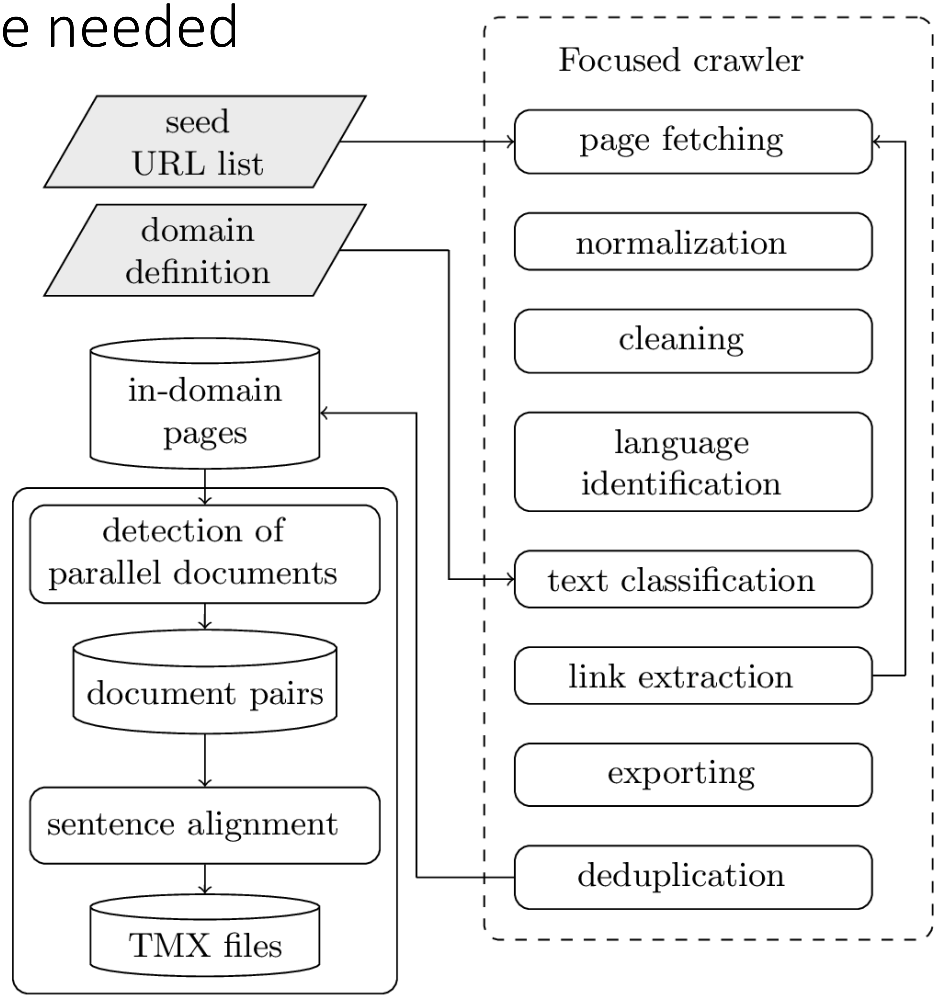
LR Stakeholder Network

Sector	Invited Stakeholder
Media Data	EurActiv, www.euractiv.com
Culture Data	Europeana, www.europeana.eu
Terminology Data	International Network for Terminology (TermNet, www.termnet.org), International Information Centre for Terminology (Infoterm, www.infoterm.org)
Research Data	META-NET, www.meta-net.eu; CLARIN
Corporate Data	Big Data Value Association (BDVA, www.bdva.eu)
Localization Services Data	Globalization and Localization Association (GALA, www.gala-global.org)
Lexicons	European Association of Lexicography (EURALEX, www.euralex.org), European Network of e-Lexicography (ENeL, www.elexicography.eu)
Publishing Data	FEP (Federation of European Publishers)
Justice	e-Sens, http://www.esens.eu/

Crawling of parallel data from the Web

European Language Resource Coordination

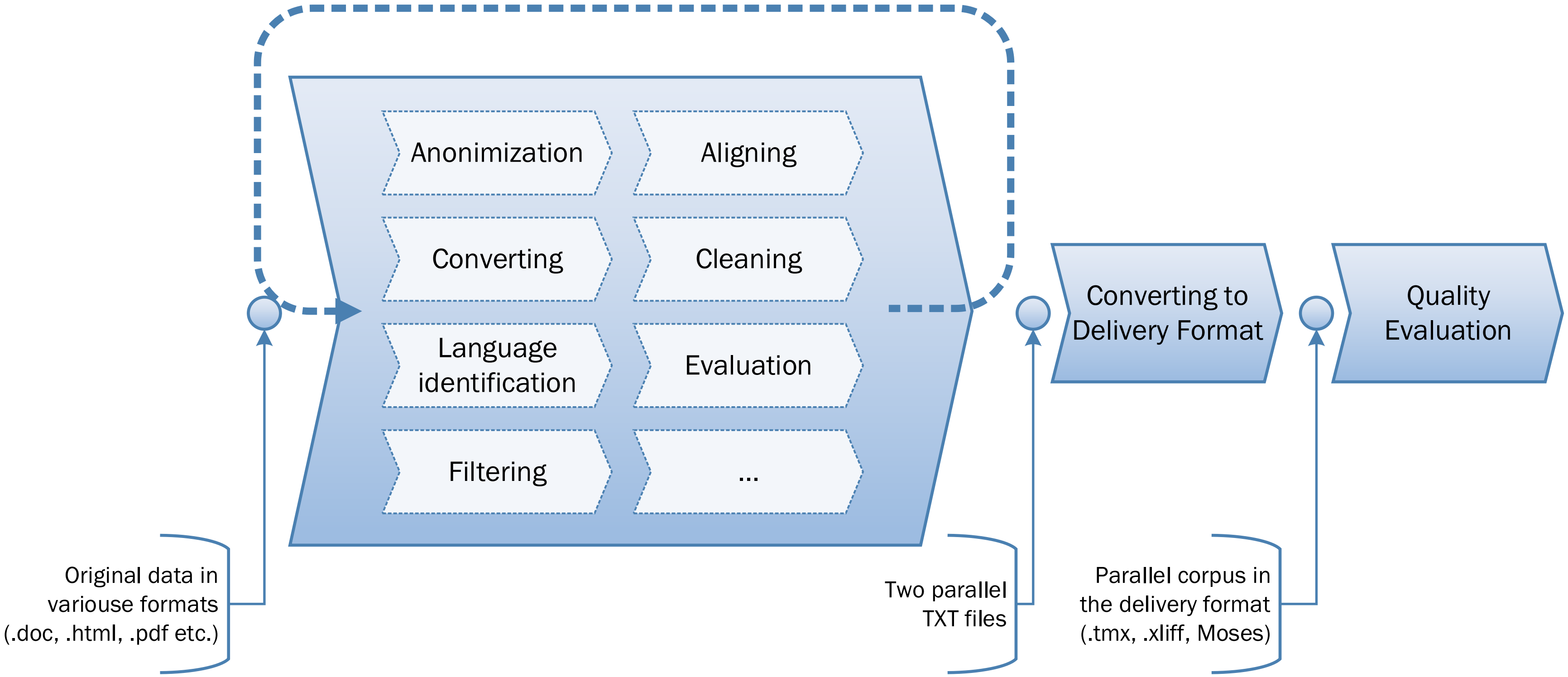
- Seed URLs are needed



Language Resource Processing

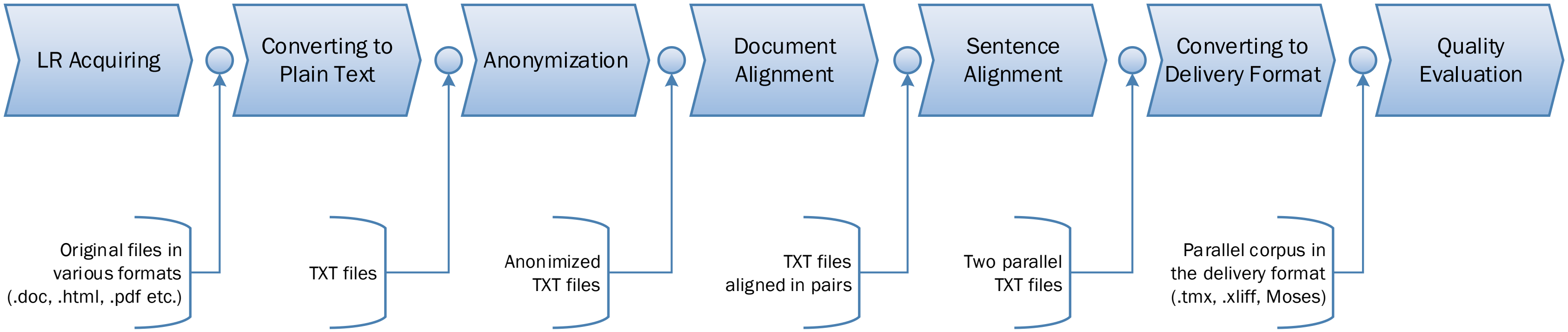
- Processing of language resources to get the data directly applicable for MT training
- Language Resources to be processed:
  - Collected so far at ELRC-SHARE
  - New LRs identified and collected in the next 3 years
- Includes anonymization of language resource datasets
- Language resource evaluation and meta-data validation

General language resource processing workflow


Anonimization
Aligning
Converting
Cleaning
Language identification
Evaluation
Filtering
...
Converting to Delivery Format
Quality Evaluation
Original data in variouse formats (.doc, .html, .pdf etc.)
Two parallel TXT files
Parallel corpus in the delivery format (.tmx, .xliff, Moses)

Language Resource Processing

- Example processing workflows for anonymization and reformatting of source data


LR Acquiring
Converting to Plain Text
Anonymization
Document Alignment
Sentence Alignment
Converting to Delivery Format
Quality Evaluation
Original files in various formats (.doc, .html, .pdf etc.)
TXT files
Anonimized TXT files
TXT files aligned in pairs
Two parallel TXT files
Parallel corpus in the delivery format (.tmx, .xliff, Moses)

General parallel corpus processing workflow

Anonymization of datasets

- Anonymization steps:
  - Automatic text anonymization (removing names)
  - De-identification - removing information which would enable a reader to identify whom a text is about
  - Includes entity recognition – e.g., detecting names of patients and relatives, hospitals, phone numbers, but also the jobs performed by the patient or their relatives, etc.
- Human validation of anonymized data
- For the very sensitive information, experts may carry anonymization within the resource provider's location so as to prevent any information leaks before anonymization

Language Resource Compiling and Production

- Identification of gaps in the priority areas/languages with a critical need for additional data
- Combination of automatic, semi-automatic and manual processes to create new parallel corpora to fill these gaps
- Quality evaluation of bilingual parallel corpora

European Language
Resource Coordination

On-site assistance

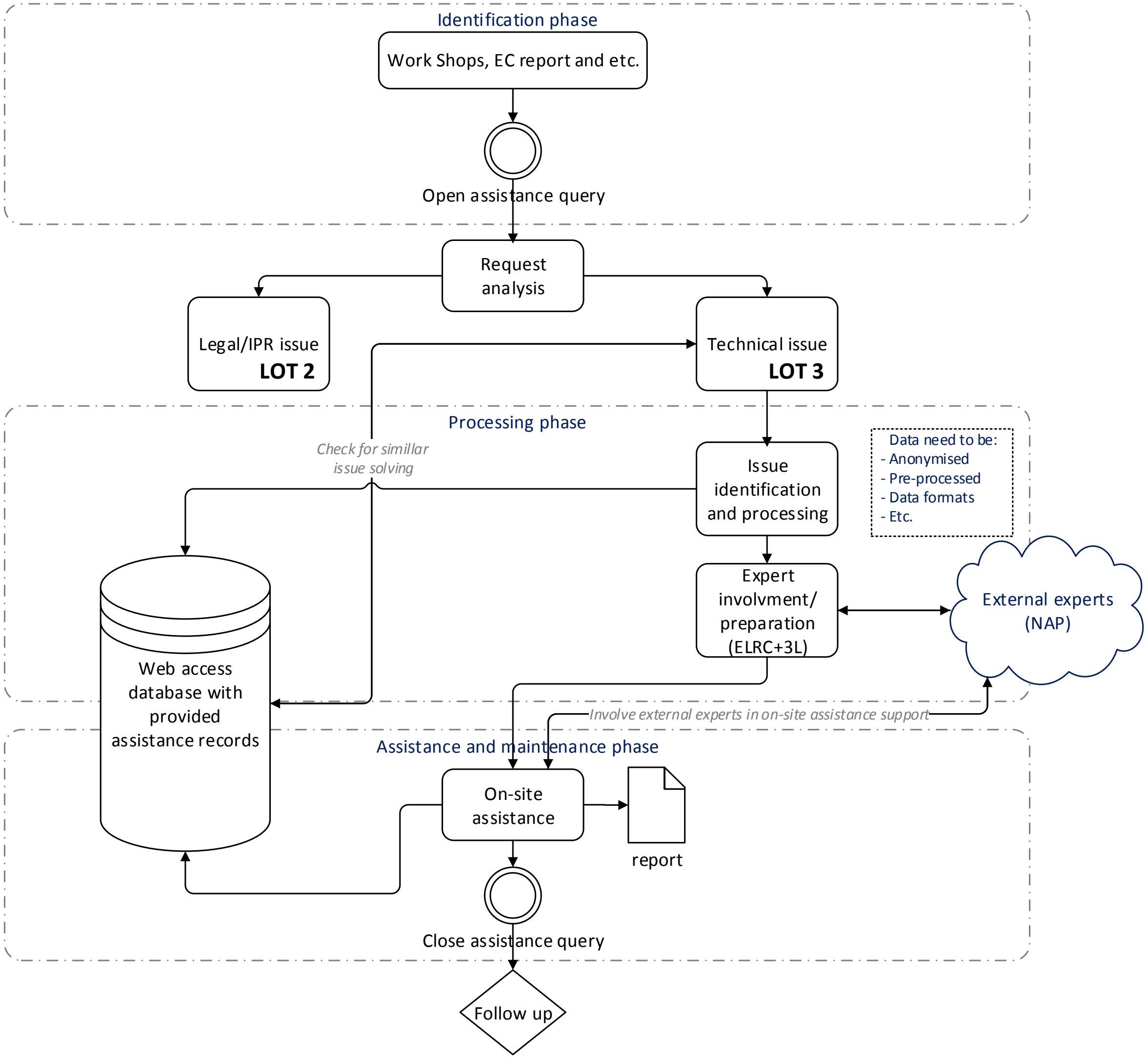
- Possibility to arrange on-site support at the data-holder institution

On-site assistance

- On-site assistance tasks:
  - Assessment of data
  - Filtering and cleaning of data
  - Anonymization
  - Reformatting and processing
  - Consultations on data management

- Support with language/data specific issues
  Support from External Experts – ELRC NAPs
- Follow up activities
- Each case of on-site assistance typically results in new LR(s) collected

On-site assistance workflow


Identification phase
Work Shops, EC report and etc.
Open assistance query
Request analysis
Legal/IPR issue
LOT 2
Technical issue
LOT 3
Processing phase
Check for simillar issue solving
Issue identification and processing
Data need to be:
- Anonymised
- Pre-processed
- Data formats
- Etc.
Web access database with provided assistance records
Expert involvment/ preparation (ELRC+3L)
External experts (NAP)
Involve external experts in on-site assistance support
Assistance and maintenance phase
On-site assistance
report
Close assistance query
Follow up

Engagement of National Anchor Points

- Support in identifying resources from the various domains required for MT@EC
- Support in LR processing with language/domain/technology specific issues
- Support in creation of new LRs with language/domain expertise
- Support for providing on-site assistance to language resources owners in their relevant countries

European Language
Resource Coordination

Thank you!

andrejs@tilde.com