European Language Resource Coordination
Connecting Europe Facility

European Commission

# LANGUAGE DATA VALIDATION AND CURATION IN ELRC

**Victoria Arranz & Mickaël Rigault (ELDA)**
**http://www.elra.info/**

European Language Resource Coordination
Connecting Europe Facility

European Commission

# WHY LANGUAGE DATA …?

# WHY ARE LANGUAGE RESOURCES SO IMPORTANT?

- Because they are worth having!
  - Language Technology (LT) is a key market for Europe - and language data are crucial for the development of LT
  - CEF eTranslation Platform supports public services across Europe

European Language Resource Coordination
Connecting Europe Facility

European Commission

# DATA COLLECTION WITHIN ELRC

# LANGUAGE RESOURCES (LR) COLLECTED WITHIN ELRC

- ELRC collects in particular the following types of language data:
  - Corpora, i.e. a set of documents or a text in one or more languages
  - Language or translation models
  - Lexical and Conceptual resources
- From whom or how is data being collected in ELRC?
  - External donors (e.g., public organisations, other EC-funded projects)
  - Webcrawling (whenever legally feasible):

    ➔ Webcrawling report: http://www.elra.info/en/dissemination/legal-issues-webcrawling-report/

European Language Resource Coordination
Connecting Europe Facility

European Commission

# DATA VALIDATION & CURATION WITHIN ELRC

# DATA PROCESSING WITHIN ELRC

- Each LR is analysed and processed by ELRC experts to ensure compliance with the **Language Resources Data Formats Specification** agreed with the EC
- According to this specification, data should be provided in the following form:
  - **Parallel data:** TMX format, UTF-8 encoding, without optional data fields (e.g. translator id, adjacent segments) without non-printable control characters
  - **Monolingual corpora:** plain text format without any additional annotation, in UTF-8 encoding, single file by language and resource, segmented into paragraphs
  - **Terminology resources:** TBX format

**Explanatory notes:**

- **TMX** stands for „Translation Memory eXchange". TMX is an XML specification for the exchange of translation memories
- **TBX** stands for "TermBase eXchange". TBX is an XML-based format for the representation and exchange of terminology data

# DATA VALIDATION WITHIN ELRC

- **Validation** = quality control of a LR against a list of relevant criteria

- Because of the different processes of gathering the data and their varying quality level, validation may be conducted in **two different ways**:
  - **Quick Content Check (QCC)**: for high-quality data (e.g., human translations)
  - **Extended Content Validation**: e.g., for data derived from automatic processing or for high-quality data needing further processing

  ➔ Important note: Both ways include first the **technical validation** of the resource and then the **legal validation** (IPR Clearance)

- **ELRC data reports**: Validation and processing reports are prepared following each validation

# DATA VALIDATION WITHIN ELRC: QCC

- **Quick Content Check (QCC)**: for high quality data (e.g. human translations):
  - check compliance of data with the ELRC objectives and **scope**
  - check that **files** are not corrupt
  - check the **format** of provided data
  - check that the **metadata** fields have been correctly filled in and are compliant with the data **content**, and
  - check whether the legal information provided is compliant with the ELRC scope (➔ **legal validation**)

# Data Validation within ELRC: Extended Content Validation

- **Extended validation:** for data derived from automatic processing or for high-quality data needing further processing

  - **Automatic validation**:
    - Check processing steps applied for data cleaning (e.g., PDF2text conversions)
    - Identify specific parts of the data that must be removed or checked more in depth
    - Applying a series of processing steps which are aimed at cleaning the data, removing TUs whose quality can be deemed as poor

  - **Manual validation** (by humans/editors):
    - On a random sample (3-5% of data size)
    - Course-grained: assign “acceptable” or “non-acceptable” labels over the sample
    - Fine-grained:
      - For particular cases where error type distinctions can help or are crucial
      - Sample is proofread and error types are listed

## Step 1: Automatic validation/filtering

**TU filtering**

Alignment score/ outliers/ length ratio/ different numbers in TUVs/ spell checker

List of sources (URLs)

**Choose a source (URL) and language pair to extract a sample**

List of TUs selected

**URL prioritisation**

Check alignment score means by crawled website per language pair > if low: potential indicator of bad quality

Source and language pair selected

Check variance/mean ratio of alignment scores

> If high: heterogeneous data → extract larger sample

Head over to Step 2

## Step 2: Human evaluation

Extract a random sample of 3 to 5% of a set of TU from the same source (URL) and having the same language pair

Sample to be validated

Have the sample proofread and tagged by error type:

1 - Language
2 - Alignment
3 - Tokenisation
4 - MT-translated
5 - Translation error
6 - Free translation

Sample

Filter out manually-checked TUs identified as containing
1 - language identification, 2 - alignment, 3 - tokenisation, 4 - MT translation, 5 - translation errors

Flag TUs from the same source of problematic TUs

Final data set

If high percentage of
1 - language identification,
2 - alignment and
3 - tokenisation
errors

# ELRC Workflow for Extended Content Validation

European Language Resource Coordination
Connecting Europe Facility

European Commission

# LEGAL VALIDATION

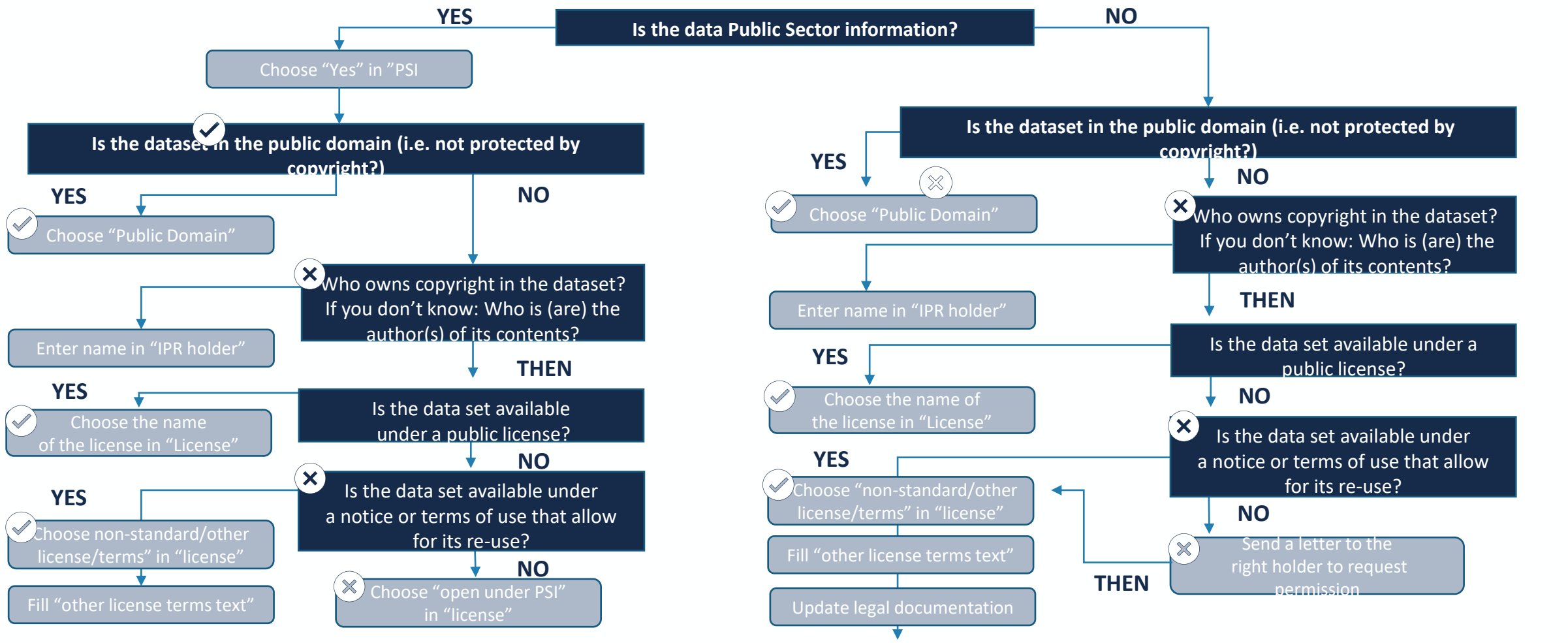
# DATA VALIDATION WITHIN ELRC: LEGAL VALIDATION

- **Objective:** To allow reuse and redistribution of Language Resources by ensuring their legal status

- **Key aspects** to be addressed and assessed
  - Public Sector Information Directive (PSI)
  - Copyright
  - Public license
  - Check the terms of use

# DATA VALIDATION WITHIN ELRC: LEGAL VALIDATION

- **Key questions to be answered**:
  - Does the data fall within the scope of the **Public Sector Information Directive (PSI)**? ➔ Public Sector Information Directive 2003/98/EC (modified in 2013 by the Directive 2013/37/UE and recast in 2019 by the Directive 2019/1024)
  - Is the data protected by **copyright**? ➔ National laws may contain rules excluding certain works from copyright protection (e.g.: Court decisions, legal texts,…)
  - If the data is protected by copyright, can I identify the owner of the copyright or the author of the work? ➔ to **obtain a license**
  - Is the data available under a **public license**? ➔ For example, certain datasets are made available by the owner of copyright under a license that allows reuse or redistribution free of charge (e.g., CC licenses, NCGL 1.0, OGL 3.0 etc.)
  - If no public license is clearly marked on the document ➔ **check the terms of use** or if any documentation may help you determine the conditions of reuse of the material




## ELRC Workflow for Legal Validation

# HANDLING PERSONAL DATA: ANONYMISATION

- Challenge for LR sharing: **personal data**
- Detecting and removing LRs with **personal data:** anonymisation tests
- Several solutions are being considered
- **MAPA (Multilingual Anonymisation Toolkit for Public Administrations)**
  - De-identification of personal data (person names, addresses and contact details, numbers – bank accounts, ID cards, NHS numbers, among others)
  - 24 EU languages (mono- and multilingual models)
  - Medical and legal domains
  - Available through a secured docker that can be installed at Public Administrations (use of Domibus security)
  - Connected to eTranslation
- For further information: **ELRC Helpdesk** (www.lr-coordination.eu/helpdesk)

European Language Resource Coordination
Connecting Europe Facility

European Commission

# ANY QUESTIONS …?

European Language Resource Coordination
Connecting Europe Facility

European Commission

# THANK YOU FOR YOUR ATTENTION!

Website: www.lr-coordination.eu

Twitter: @LR_Coordination

Email: info@lr-coordination.eu