Task 7 (ELRC)
Task 2 (ELRC+3)

Validation Guidelines

Author(s): Priscille Schneller (ELDA)
Meritxell Fernández-Barrera (ELDA)
Valérie Mapelli (ELDA)
Vladimir Popescu (ELDA)
Khalid Choukri (ELDA)
Victoria Arranz (ELDA)
Maria Giagkou (ILSP)
Prokopis Prokopidis (ILSP)
Vassilis Papavassiliou (ILSP)
Roberts Rozis (Tilde)

Dissemination Level: Internal

Version No.: <V6.2>
Validation guidelines updated

Date: Validation report updated
Processing report added

2018-07-20
Contents

1 Introduction and Goal 4
2 Validator profile 4
3 Validation of donated data 6
  3.1 Compliance with ELRC scope 6
  3.2 Quick content check 7
4 Validation of both automatically-processed donated data and crawled data 7
  4.1 Procedure 7
  4.2 Content validation 8
    4.2.1 Content validation workflow 8
    4.2.2 Content validation procedure 10
      4.2.2.1 Automatic validation / filtering 10
        4.2.2.1.1 At the TU level (TU filtering) 10
        4.2.2.1.2 At the document level (URL prioritization) 11
        4.2.2.1.3 Determining the size of the sample 11
      4.2.2.2 Validation by human experts 11
        4.2.2.2.1 What do validators need to do? 11
        4.2.2.2.2 Actions following human validation 13
    4.2.3 Final filtering 14
5 Validation of LR Metadata 16
6 Legal validation 18
  6.1 Assessment of the PSI scope and conditions for re-use 18
    6.1.1 PSI scope validation workflow 18
    6.1.2 PSI scope validation procedure 22
  6.2 Resource licence and conditions of use 22
7 ELRC data reports 24
  7.1 Validation report 24
  7.2 Processing report 25
ANNEX 1: Template for ELRC data reports

1. Report Validation

2. Processing Report
1 Introduction and Goal

“Validation” is understood as the quality control of a Language Resource (LR) against a list of relevant criteria. Within the ELRC, ELRC+2 and ELRC+3 projects, a number of Language Resources are being gathered, thanks to the contribution of external players from the public sector (donated LRs). Due to the high number of LRs required within these projects (several hundreds), the ELRC consortium decided to complement the donated LRs with LRs produced from scratch through a website crawling process. All gathered LRs need to be submitted to a validation process to check their conformity with the original requirements of the project.

The goal of these guidelines is to provide a methodology for validating donated and crawled data collected within the project, towards the constitution of parallel corpora. This version of the guidelines is an updated version of the document produced within the initial ELRC action. This new version takes into account the validation needs for processed data and, thus, provides an updated version of both validation guidelines and validation report, as well as a new document called processing report. Both validation report and processing report are submitted jointly and constitute a single document named ELRC Data Reports.

The validation of donated and crawled data should be conducted in different ways. Indeed, we assume that some of the donated data consist of qualitative data in terms of contents (in particular asserted translations for multilingual data, data produced by human experts) but require a technical- and legal-oriented evaluation.

Their validation should thus consist in:
- checking the compliance of data with ELRC scope
- checking the format of provided data
- checking if legal information provided is compliant with ELRC scope

As other kinds of donated data (e.g. sets of PDF or DOC(X) documents), as well as the crawled data come from an automatic processing, their validation necessitates a deeper content validation whereas the technical part already corresponds to the Language Resource production requirements. The legal validation of crawled data is also required. The validation of automatically-processed donated data and of crawled data should be carried out along the following steps:
- For crawled data only: check if the crawled websites are under the scope of the Public Sector Information (PSI) Directive so as to make sure that the content can be re-used.
- Perform a series of automatic validation steps, based on spell checking, TU length ratio and TU alignment score-based outlier detection, etc.
- Estimate the quality of the translation with a team of language expert validators.

For all types of data, the ELRC-Share repository should be filled in and corresponding metadata should be validated as well.

Finally, for each data set, a validation report should be provided asserting the quality of all data. Furthermore, for all those data sets undergoing processing, a processing report should also be provided, detailing the processing steps followed.

2 Validator profile
Representatives in ELRC-Share repository responsible for the validation of data are as follows:

- DFKI: Austria, Czech Republic, Germany, Hungary, Luxemburg, Netherlands
- ELDA: Belgium, France, Ireland, Italy, Malta, Portugal, Spain
- ILSP: Bulgaria, Croatia, Cyprus, Greece, Poland, Romania, Slovakia, Slovenia
- Tilde: Denmark, Estonia, Finland, Iceland, Latvia, Lithuania, Norway, Sweden

Thus, each representative should appoint internal validators to proceed with each step of the validation process. General issues brought up by internal validators and not covered within the present Validation Guidelines may require further consortium advice and should be brought up through the elrc-data@dfki.de mailing list. Legal issues not covered by the present Validation Guidelines should be brought up to the Legal Helpdesk managed by ELDA (help@cef-at-helpdesk.org, or telephone: +33 970 440 522, or skype ID: CEF-AT-Helpdesk).

The validator should be familiar with the ELRC objectives and the PSI context. He/She should also have the right language skills when reviewing languages that are not in his/her mother tongue.
3 Validation of donated data

We assume that the donated data is of two sorts:

- Some of it consists of qualitative data in terms of content (in particular asserted translations for multilingual data, data produced by human experts) but require a technical- and legal-oriented validations. A full validation protocol would require a lot of time while not being necessary, given the sources of the data. Therefore, the ELRC team would like to have some first indications and confirmation of the quality of a yet non-validated LR, hence this “Quick Content Check” (QCC). The QCC mainly checks the metadata elements against a number of minimal requirements that depend on the type of the resources as listed in section 3.2.

- Other donated data may still consist of high-quality content, but come as a series of PDF or Word DOC(X) documents, which need to be automatically processed so that translation memories are extracted from them. The details of these automatic processing steps are outside the scope of these guidelines and strongly depend on the nature of the data at hand (e.g. do the document also contain scanned images?). Nonetheless, the processing steps undertaken for such LRs in ELRC+3 are documented in the LR Processing Report, detailed in Annex 1. These donated data, as they are automatically processed, need to undergo a content validation process, sketched below and described more thoroughly in Section 4.2.

In ELRC+3, donated data transformed into translation memories through automated means need to undergo a content validation process, akin to the process undertaken for crawled data and described in Section 4.2. This content validation has two parts:

- An automatic part, which consists in applying a series of processing steps which are aimed at cleaning the data, removing TUs whose quality can be deemed as poor by automated means. This process can be a variation of the procedure presented in Section 4.2.1.

- A manual part, which should be applied only if the Editor(s) deem it necessary, e.g. for high-priority under-resourced languages where data quality should compensate for the lack of quantity. This manual validation process can be performed either:
  - At a coarse-grained level, whereby a human annotator labels a sample of TUs as “Acceptable” or “Non-acceptable” and, if e.g. more than 10 % of the TUs in the sample have been labeled as “Non-acceptable”, then the whole LR is discarded. The sample creation for this type of validation follows the same procedure as that of the sample used for the fine-grained validation below (cf. Section 4.2.2.2).
  - At a fine-grained level, along the lines described in Section 4.2.2.2, only in particular cases, e.g. for high-volume high-priority LRs where the error type distinctions specified in Section 4.2.2.2 can help maximize the TU recall (e.g. by taking TUs marked as “Machine-translated text” or “Free translation” into account).

3.1 Compliance with ELRC scope

Donated data should be compliant with original ELRC scope defined as follows:

- Data should come from public institutions or be relevant to the general administrative/regulatory domain
- Data from the European Commission should be checked before being used: eTranslation should be asked whether such data are of interest to the EC (as they may already have it).
European Language Resource Coordination

Data Quality Indicators

- Language of the data should be in at least one of the 24 languages from the 30 EU member states and the official languages of countries associated to the CEF programme listed as follows: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Slovakian, Slovenian, Spanish, Swedish

3.2 Quick content check

The aim is to validate 3 types of data:
- Multilingual corpora,
- Monolingual corpora,
- Terminologies.

To proceed with this task, validator should download and check data provided in archive.zip file. The following validation procedure should be carried out:

- Validator checks if the format is readable. If files are provided under proprietary formats (like Trados format), submitter of the resource should be contacted to obtain a converted version into a non-proprietary format. Another example is data consisting of scanned PDF files, which are accepted only if automatic Optical Character Recognition (OCR) can allow one to obtain reasonable-quality text. If the LR is likely to contain significant amounts of text and the OCR process does not yield reasonable-quality results, then the Validator should contact the Submitter of the LR to obtain a converted version into a reusable format.
- Validator checks if the data content is acceptable “a priori” with no deep analysis. For instance:
  o Validator should indicate if files are empty,
  o For parallel corpora, alignment should be quickly manually checked on a small number of entries (at the level of document, sentence, etc.) to see if segments seem appropriate between source and target language.

4 Validation of both automatically-processed donated data and crawled data

4.1 Procedure

Once automatically obtained from donated documents or from crawled web sites, bilingual texts are automatically aligned and TMX files containing translation units (TUs) are produced for each language pair.

The list of crawled URLs is manually checked to assess if the websites are under the PSI scope. Content from websites that do not fall under the PSI directive is excluded.
Then TMX files are validated to assess quality of the content of the TUs. Errors on TUs are reported and the percentage of each error type is calculated per website. On this basis, the following actions are carried out for each website:

- Automatically remove TUs assigned with errors
- Annotate the other TUs from the same website with an indication on the probability of finding the same errors.

It must be noted that indications of source / target language in the TMX files are not considered as source / target language as this can be understood in a translation work. TMX files were produced automatically without knowing which language is the source or target language. Checking the provenance of the crawled data, in particular the extension of the URL, could help in deducing the source and target languages.

### 4.2 Content validation

#### 4.2.1 Content validation workflow

Validation of the content consists of both an automatic and a manual procedure:

- the automatic validation / filtering is conducted to identify specific parts of the data that must be removed or checked more in depth,
- the human validation is carried out on a random subset of 3 to 5% of a set of TMX files from the same source (website crawled) and with the same language pair.

Figure 2 depicts the workflow we propose for content validation.
Figure 2 Workflow for fine-grained content validation of crawled data. It also applies to automatically-processed donated data provided that one substitutes “source” and “URL” with “donated dataset”.
4.2.2 Content validation procedure

4.2.2.1 Automatic validation / filtering

Translation units (TUs) are automatically aligned by the ILSP-FC crawler (http://nlp.ilsp.gr/redmine/projects/ilsp-fc)\(^1\) for crawled data, or by other sentence-level aligners such as Maligna (https://github.com/loomchild/maligna) for donated data, and several scores are provided at the TU level. An automatic validation and / or filtering can be completed by checking the metadata provided:\(^2\):

- The alignment score (indicated by the tag `<prop type="score">`), an estimate of the text similarity between the target segment and the source segment. The higher this score is, the higher the chance for the target segment to be close to the source segment.
- The length ratio (indicated by the tag `<prop type="lengthRatio">`), computing ratio between length (number of characters) of the source segment and the target segment. Segments with a length ratio close to 1 have similar length whereas segments having a ratio far from 1 have a big difference in terms of length, which could reveal segments are not well aligned.
- Different numbers in translation unit variants (TUVs) (indicated by the tag `<prop type="info">different numbers in TUVs</prop>`), indicating when numbers in the target segment are different compared to the source segment (for segments containing numbers in the text).

These scores can be taken into consideration to detect potential alignment or translation issues. Furthermore aggregation of these scores at the document level (TUs from the same source) can give an indication on the estimated quality of the content.

Besides checking the metadata of a TU, examination of its content also contributes to providing “cleaner” resources. For instance, duplicate TUs, TUs with identical TUVs, very short TUVs (less than 2 tokens) are of no or limited use for training MT systems.

4.2.2.1.1 At the TU level (TU filtering)

One option is to establish a threshold of alignment score under which TUs are considered as ‘bad quality TUs’ and then, remove them. Human validation of several samples of data can help to assess the threshold to be defined.

We also suggest to filter outliers (TUs having an alignment score distant to the median of alignment scores of TUs from the same source). The outlier detection is based on a median-relative deviation analysis. It allows removing TUs which are potentially badly aligned.

Length ratios can be used to give an indication on the correct alignment of a TU. However each language has specificities in terms of word’s length and this may not be a good indicator.

---

\(^1\) The methodology here explained is based on the use of several specific tools that have been used within the ELRC actions by a number of partners. Nevertheless, it can be applied to, and carried out by, any other compatible tools achieving the same type of results/output.

\(^2\) These scores are obtained by enriching the TMX output provided by aligners such as Maligna. For crawled data, these scores are readily added by the ILSP-FC tool. For donated data, TMX files can be enriched by basic TMX XML document manipulations based on the TU contents.
They are not considered at the moment, but experiments could be conducted to check this indicator’s relevance.

The indication of TUs containing “different numbers in TUVs” is another way of detecting badly aligned TUs, although some TUs containing different numbers may be correct (for instance: “8 pm” in English is correctly aligned with “20 H” in French). We suggest to decide whether such TUs must be kept or removed depending on the results of the human validation: all the TUs having different numbers in TUVs could be removed from a TMX with TUs from the same source if all the identified TUs having ‘different numbers in TUVs’ in a sample were rejected through the human validation.

4.2.2.1.2 At the document level (URL prioritization)

Another option is to identify which websites need to be prioritized during the manual validation. To give an estimated quality indicator of the content of each crawled website, means and variances are calculated for alignment scores and for length ratios of TUs from the same source. Then human validation can be conducted on specific parts of the crawled content, for instance TMX files generated from websites with a low or a good quality level indicator.

As for donated datasets, they all need to be validated.

4.2.2.1.3 Determining the size of the sample

The ratio between alignment score variance and mean gives an indication on how homogeneous the data is: the lower this ratio is, the more homogeneous the data is. For the human validation, it is recommended to extract several samples from heterogeneous data so as to get a more representative subset of the data.

4.2.2.2 Validation by human experts

The manual validation takes place over a selected sample of data. The indications detailed below are used for the selection of the material to be validated as well as for its validation.

These guidelines are derived from those used within the TC-STAR project (http://www.tc-star.org) and adapted to the project requirements, so that all the content reaches the same quality level.

4.2.2.2.1 What do validators need to do?

For each website, a random subset of the corpus is selected, until the number of TUs adds up to about 3% of the source dataset (even a higher percentage for small data sets). Then, the validation corpus containing both source and target segments is submitted to the validators.

The validation task consists in proofreading the texts and, whenever a problematic point arises, labeling the problematic TU (with a label from the list of problems detailed in the table further down).

The aim of the validator is to evaluate:

- if the content of the segment is in the language indicated in the metadata,
- if the content of the source language segment is well aligned with the target language segment,
- if the translation does not seem to be the result of automated translation and is of good quality.

The following technical issues should be taken into account:
1. Files to be validated are provided to validators in .txt format. Validators are expected to submit their files in the same format as the original.

2. The sentences to be validated look as follows:

```
[id ; alignment score ; (different number in TUVs)]
Source segment
Target segment
Blank line
```

3. Notification of errors is provided per TU. If no remark is provided by the validator, this format will remain the same. However, if a segment contains an error, then a new line is inserted starting with “#” right after the segment. After the “#” follows the type of error. The resulting format is as follows:

```
[id ; alignment score ; (different number in TUVs)]
Source segment
Target segment
# Error type
Blank line
```

4. To ensure consistency from one validator to another, the following system has been adopted for grading translations. Validators should use the following types/labels to tag problematic cases:

<table>
<thead>
<tr>
<th>Label</th>
<th>Shortcut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrong language identification</td>
<td>L</td>
</tr>
<tr>
<td>Incorrect alignment</td>
<td>A</td>
</tr>
<tr>
<td>Wrong tokenization</td>
<td>T</td>
</tr>
<tr>
<td>MT translation</td>
<td>MT</td>
</tr>
<tr>
<td>Translation error</td>
<td>E</td>
</tr>
<tr>
<td>Free translation</td>
<td>F</td>
</tr>
</tbody>
</table>

**Wrong language identification** means the crawler tools failed in identifying the right language.

**Incorrect alignment** refers to segments having a different content due to wrong alignment.

**Wrong tokenization** means the text has not been tokenized properly by the crawler tools (no separator between words).

---

3 “different number in TUVs” is an optional field indicated for TUs tagged as such in the TMX file.
**MT translation** refers to content identified as having been translated through a Machine Translation system. A few hints\(^4\) to detect if this is the case:

- grammar errors such as gender and number agreement;
- words that are not to be translated (trademarks for instance Nike Air => if ‘Air’ is translated in the target language instead of being kept unmodified);
- inconsistencies (use of different words for referring to the same object/person);
- translation errors showing there is no human behind.

**Translation error** refers to:

- Lexical errors (omitted/added words or wrong choice of lexical item, due to misinterpretation or mistranslation),
- Syntactic error (grammatical errors such as problems with verb tense, coreference and inflection, misinterpretation of the grammatical relationships among the words in the text).
- Poor usage of language (awkward, unidiomatic usage of the target language and failure to use commonly recognized titles and terms). It could be due to MT translation.

**Free translation** means a non-literal translation in the sense of having the content completely reformulated in one language (for editorial purposes for instance). This is a correct translation but in a different style or form. This includes figures of speech such as metaphors, anaphors, etc. We consider it as important to tag such cases for data that will be used for training MT systems.

5. Only **one type of error** should be attributed to each TU. The hierarchy presented below should be followed in the identification of errors:

   a. The **wrong language identification** is the most important error to tag, followed by **incorrect alignment** and then, **wrong tokenization**. These features are related to the automatic processing by the crawling / alignment tools. If any of these errors are detected, the translation itself will not be checked more in depth.

   b. When no errors in the format are found, the content of the TU should be checked, focusing first on potential **MT-translated content** (in this case all the TUs from such a source should be discarded). Then major **translation errors** must be noted. If none of the errors described above are detected and the translation contains minor differences in formulation it must be noted as **free translation**.

6. It is essential that the given translation receives the benefit of the doubt. Only clear errors should be indicated.

4.2.2.2.2 **Actions following human validation**

The percentage of TUs identified as containing errors is calculated for each error type.

---

For data from manually-checked samples presenting a high rate of language identification, tokenization and alignment errors, the following actions can be taken to improve the quality of the crawled data:

➢ For **language identification** or **tokenization** errors, we may use a spell-checker on the TUs from the sample. TUs containing more than 50% of tokens\(^5\) considered as incorrect will be filtered out and the validator should check if fewer errors are found on the remaining TUs.

➢ For **alignment** errors, we may filter out TUs which alignment score is high compared to the average of alignment scores of TUs from the same source / filter out outliers (see section 2.3.2.1.1).

### 4.2.3 Final filtering

To produce the final resource, we automatize the process and filter TUs that must be discarded or flagged. The following actions should be carried out:

- automatically remove TUs assigned with errors,
- annotate the remaining TUs from the same website, based on the probability of finding the same errors.

This should be done through the definition of error thresholds for the 5 main error types (wrong language identification, alignment, tokenization, MT translation and translation errors). The “free translation” indication is considered as an informative feature and TUs assigned to this error should be systematically flagged.

Two thresholds need to be specified: an inferior threshold (th\_inf) and a superior threshold (th\_sup). These thresholds are expressed as percentages of TUs containing errors out of the total number of manually-checked TUs associated to a website and a language pair.

For each error:

- **the inferior threshold (< th\_inf)** defines the limit under which the sample is considered to be of good quality. We assume the remaining TUs from the same source will have the same quality and we flag them as ‘unlikely’ to contain such error.

- **the superior threshold (Strictly > th\_sup)** defines the limit above which the sample is considered of bad quality, which means the remaining TUs from the same source may also be of bad quality, so all the TUs from the same source are removed.

- for TUs from samples with a percentage of error between the inferior and the superior threshold (Strictly > th\_inf and < or = th\_sup), we assume the remaining TUs from the same source may contain such an error and they are flagged as ‘likely’ to contain such error.

The thresholds need to be defined for each language pair. Comparing several samples of data will help to assess these thresholds.

The automation is implemented in the ‘tmxbuilder’ tool of the ELDA Crawled data Management Toolkit (ELDA-CMTK)\(^6\). The rules implemented are presented below:

<table>
<thead>
<tr>
<th>Percentage of TUs</th>
<th>TUs discarded</th>
<th>TU flagged</th>
<th>Label</th>
</tr>
</thead>
</table>

---

\(^5\) This percentage depends on the language pair and the website’s content.

\(^6\) ELDA-CMTK is available under GPL-3.0 at [https://github.com/ELDAELRA/elda_cmtk](https://github.com/ELDAELRA/elda_cmtk)
## Data Quality Indicators

<table>
<thead>
<tr>
<th>containing errors for each error type compared to thresholds</th>
<th>TUs assigned with an error</th>
<th>Other TUs from the same source</th>
<th>Likely/Unlikely/Undetermined</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; th_inf</td>
<td>TUs assigned with an error</td>
<td>Other TUs from the same source</td>
<td>Unlikely</td>
</tr>
<tr>
<td>Strictly &gt; th_inf and &lt; or = th_sup</td>
<td>TUs assigned with an error</td>
<td>Other TUs from the same source</td>
<td>Likely</td>
</tr>
<tr>
<td>Strictly &gt; th_sup</td>
<td>TUs from the same source as the ones assigned with an error</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>No percentage indicated</td>
<td>No TUs</td>
<td>All the TUs</td>
<td>Undetermined</td>
</tr>
</tbody>
</table>

In the final TMX file each TU will be assigned to a `<prop>` element per error type:

- `<prop type="languageIdentificationErrors">[Likely|Unlikely|Undertermined]</prop>`
- `<prop type="alignmentErrors">[Likely|Unlikely|Undertermined]</prop>`
- `<prop type="tokenizationErrors">[Likely|Unlikely|Undertermined]</prop>`
- `<prop type="machineTranslatedTexts">[Likely|Unlikely|Undertermined]</prop>`
- `<prop type="translationErrors">[Likely|Unlikely|Undertermined]</prop>`
- `<prop type="freeTranslation">[Yes|No|Unknown]</prop>`
5 Validation of LR Metadata

Information on all packaged LRs should be reported into the ELRC-SHARE repository (https://elrc-share.ilsp.gr).

The ELRC-SHARE schema includes metadata fields for the description of LRs, such as textual corpora, computational lexica, terminological resources, grammars etc. It includes common elements for all resources as well as distinct elements depending on the type of the resource. A subset of the elements is mandatory to comply with the project requirements.

Free text fields may be completed in any European language. However, "resource name" and "description" fields should also be translated into English.

Validator should check if all metadata fields are filled in correctly and are compliant with provided data contents (e.g. size of data, languages, etc.). Whenever needed, validator will require for corrections or completion of missing mandatory fields. Interaction between validator and provider should be carried out in order to obtain correct information.

If applicable, information should be provided on what pre-processing was done. If pre-processing was performed, the raw material must be provided, too. Mandatory fields are the following:

General information:
- Resource name
- Resource type:
  - corpus (e.g. monolingual corpus, bilingual corpora, translation memories etc.)
  - lexical/conceptual resource (e.g. dictionaries / lexica / ontologies / NE gazetteers)
  - language description (e.g. grammars)
- PSI - Public Sector Information (yes/no)

Distribution:
- Licence

Contact person(s):
- Surname
- Email

Information required depending on resource type:
- Written Corpus:
  - Linguality type
  - Language name
  - Size + size unit
  - Mime type (Resource format(s) (e.g. doc/tmx/xml etc.))
  - Character encoding
- Lexical/conceptual resource (e.g. dictionaries / lexica / ontologies / NE gazetteers) :
  - Lexical conceptual resource type
  - Encoding level
  - Linguality type
  - Language name
  - Size + size unit
  - Mime type (Resource format(s) (e.g. doc/tmx/xml etc.))
- Language description (e.g. grammars) :
  - Language description type
  - Encoding level
  - Linguality type
- Language name
- Size + size unit
- Mime type (Resource format(s) (e.g. doc/tmx/xml etc.))
6 Legal validation

6.1 Assessment of the PSI scope and conditions for re-use

6.1.1 PSI scope validation workflow

The crawled websites are assessed according to the Directive 2003/98/EC, known as the 'PSI Directive', which entered into force on 31 December 2003 and was revised by Directive 2013/37/EU, which entered into force on 17 July 2013\(^7\). According to the Directive, Member States should ensure that documents held by public sector bodies and which are accessible according to national access regimes should be re-usable for commercial or non-commercial purposes. The Directive does not apply to documents subject to third-party copyright and some public bodies are excluded from the scope of the Directive (e.g. broadcasters, cultural establishments, educational or research institutions). The Directive allows public sector bodies to impose conditions on re-use (e.g. acknowledgement of the source, data integrity). Specific conditions for re-use can be imposed through a standard licence (e.g. CC, open data licence) or through ad hoc terms of use.

Below we suggest a workflow for checking if a website is under the scope of the PSI Directive (stage 1) and if any specific conditions for re-use are to be taken into consideration (stage 2). For all websites which do not fall within PSI scope, we recommend to filter them out from the crawled data, unless the content they publish is clearly licensed under an open licence allowing re-use, such as CC.

---

So far PSI validation involves only stage 1. The cases falling outside PSI scope depicted in the workflow are probably the most common but they are not exhaustive (the Directive excludes some other domains, e.g. the parts of documents containing only logos, crests and insignia).

The possibility of partially automatizing the identification of specific conditions imposed on re-use (stage 2) was envisaged. Out of 1104 crawled URLs, a random sample of 36 URLs was manually analysed to extract two types of expressions. Firstly, expressions which are typically used on the main page to redirect to the page containing the legal notice. Secondly, expressions which are found in the legal notice and which identify specific legal conditions imposed on the re-use of data.

Expressions of the first type were identified in seven languages (Italian, Spanish, French, Dutch, English, Portuguese and German).

Such a list of expressions would allow to automatically extract the texts containing the legal notice and thus to automatically generate a corpus of conditions of data re-use.
Table 1. Expressions extracted from a sample of 36 URLs which link to the page containing the legal notice

Expressions of the second type were identified in four languages (Italian, French, Portuguese and Spanish) and classified into four categories which correspond to four different conditions imposed on the re-use of data:

(i) attribution: obligation to give attribution to the author;
(ii) non-commercial use: only non-commercial use allowed;
(iii) data integrity: original data cannot be altered;
(iv) previous authorisation: use only allowed if previous authorisation obtained from the data owner.

Certain lexical families seem to be representative of some of the aforementioned legal conditions (e.g. for the attribution condition, the verb “citer” and the noun “citation” in French, “à condition de citer clairement la source”, “citation explicite du site [...] comme source des données”). However, previous lemmatisation will be required to be able to match representative expressions with legal notice texts. This requires that lemmatisers are available in all the CEF languages.

Furthermore, in general, the expression of legal conditions is not standardised, not even within the same language. Instead, there is rather a considerable variation (e.g. for the non-commercial use condition, in French, we have examples such as “gratuité de la diffusion” and “Toute utilisation à des fins commerciales ou publicitaire est interdite”). This heterogeneity requires thus building a representative corpus of all the crawled URLs and a further analysis of the actual linguistic variation within each condition and language.

Last but not least, the highest legal risk lies in the existence of other legal conditions on re-use beyond the four conditions already mentioned. Therefore, an exhaustive analysis of all
the potential conditions on re-use should also be envisaged. This entails a considerable complexity, given the number of languages and legal systems to be taken into account.

<table>
<thead>
<tr>
<th>BY</th>
<th>à condition de citer clairement la source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>de la citation expresse du site <a href="http://www.senat.fr">www.senat.fr</a> comme source, accompagnée d’un lien hypertexte le cas échéant, ainsi que de toutes les informations utiles s’y rapportant (nom des sénateurs, titre du rapport…)</td>
</tr>
<tr>
<td></td>
<td>citation explicite du site […] comme source des données reproduites et mentions permettant de déterminer la nature et l’origine des documents reproduits (date, intitulé, auteur, …)</td>
</tr>
<tr>
<td></td>
<td>Citation explicite du ministère de la Justice comme source et précision de la date</td>
</tr>
<tr>
<td></td>
<td>la citation de la source</td>
</tr>
<tr>
<td></td>
<td>mentionner clairement la source et la date du document (ou, à défaut, sa date de mise à jour) ; faire apparaître de manière claire et lisible le titre et l’adresse url de la page reproduite</td>
</tr>
</tbody>
</table>

| NC                     | sous réserve : |
|                        | - de la gratuité de la diffusion |
|                        | Gratuité de la diffusion |
|                        | Toute utilisation à des fins commerciales ou publicitaires est interdite. |
|                        | à l’exclusion de toute utilisation à des fins commerciales ou publicitaires. |

| Data integrity         | sous réserve du respect de l’intégrité des documents (aucune modification ni altération) |
|                        | respect de l’intégrité des données, qui suppose que la teneur et la portée des textes et décisions reproduits ne soient pas altérées, notamment par des retraitements (modifications des contenus, insertion de commentaires sans que ceux-ci puissent être clairement distingués du texte authentique, coupes altérant le sens du document) de nature à induire le lecteur en erreur |
|                        | Respect de l’intégrité des documents reproduits : aucune modification ni altération d’aucune sorte |
|                        | le respect de l’intégrité de l’information reproduite |
|                        | respecter l’intégrité des documents et données reproduits : ils ne doivent être ni modifiés ni altérés |

| Previous authorisation  | La reproduction de tout ou partie de ce site sur support électronique est interdite sauf autorisation expresse du directeur de la publication. |

Table 2. Example of legal expressions in French expressing four different legal conditions on the re-use of data (extracted from a sample of 36 URLs)

---

8 Despite the relative harmonisation aimed at by the PSI Directive, the national laws implementing the Directive can introduce particularities. Furthermore, there are extant differences between national intellectual property regimes.
6.1.2 PSI scope validation procedure

To proceed with PSI validation, the list of URLs crawled with the ILSP-FC crawler (http://nlp.ilsp.gr/redmine/projects/ilsp-fc/) should be manually checked for PSI scope on the basis of the above-mentioned workflow (limited to stage 1 only).

Three types of results may be reported by validator:
- URLs fall under PSI scope: URLs that comply to validation workflow, stage 1.
- URLs do not fall under PSI scope: URLs that do not comply to validation workflow, stage 1.
- URL status is unclear: for some URLs it may be unclear whether they fall within PSI scope, for instance for those which belong to state-owned companies and which therefore belong to a public body but are purely commercial in nature. In this case a more detailed analysis of national PSI regulation is required in order to identify the kind of legal entities to which this regulation applies. This fine-grained analysis is not performed at this stage.

Validators should use the following labels to tag this information:

<table>
<thead>
<tr>
<th>Type</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLs fall under PSI scope</td>
<td>Yes</td>
</tr>
<tr>
<td>URLs do not fall under PSI scope</td>
<td>No</td>
</tr>
<tr>
<td>URL status is unclear</td>
<td>Unclear</td>
</tr>
</tbody>
</table>

This information will then be kept for future exploitation, in particular to enable the selection of PSI-compliant data during the following validation and production phases\(^9\). To facilitate this future exploitation, information may be inserted in the corresponding field (entitled “Within PSI Scope”) from the source database (http://sources.lr-coordination.eu/).

6.2 Resource licence and conditions of use

Resource licence and conditions of use are indicated in the completed Metadata of the ELRC-SHARE repository. Validator should check the following information:

1) If metadata “PSI - Public Sector Information” has been validated as correct (see section 5):
   a. If ticked, no further information is required (except if value “under review” is selected) we assume that the resource can be used within the ELRC project.
   b. If not ticked, Licence information should be provided (see following steps).

2) If the resource is identified as Open Data:
   a. appropriate value should be selected in the “Licence field”,
   b. if attribution is required, IPR Holder(s) should be identified in the “IPR holder” field.

3) If the resource is identified as restricted and a licence has been signed or terms of use identified:
   a. “Non Standard/Other Licence/Terms” value should be selected in “Licence” field and related text should be provided either in “Other Licence Text” or to ELRC consortium,

\(^9\) For data non-validated (TUs with none of the 3 labels Yes/No/Unclear attributed), the label Unknown is automatically attributed during the processing via the ELDA Crawled data Management Toolkit (CMTK) <https://github.com/ELDAELRA/elda_cmtk>
b. IPR Holder(s) should be identified in the “IPR holder” field.

4) If the resource is identified as private or confidential, either “Personal Data Included” or “Sensitive Data Included” box should be selected.

If further assistance is required, the legal helpdesk will be contacted.
7 ELRC data reports

All produced language resources will be documented with a validation report, as well as with a processing report, the latter to be filled in if such processing has been carried out on the resource. The file containing both reports will be uploaded into the ELRC-SHARE repository together with the submitted LR.

7.1 Validation report

A validation report should be created and updated at each new validation, allowing the follow-up of the validation procedure and the interaction between ELDA and the validation experts. One validation report will be attached to each validated LR. Template of the validation report is given in Annex 1, Section 1, as part of the ELRC Data Reports document.

If any modification is carried out during the validation process, such modification should be easily traced as follows:
- information about transformation of data should be kept in the validation report,
- both raw and processed data will be kept in the ELRC-Share repository.

Status of the validation report will be taken in charge by a Validation Manager. Although Validators and Validation Manager should be ideally distinct, we may accept the same person who should do the process in separate and consecutive steps. Three validation status may be selected:

1) If modifications are needed to have the data accepted, Validation Manager will select status “Changes required” and submit it to the data team for applying corresponding changes. A new validation will be carried out to check if modifications have been well applied.

2) If the validation results comply with all project requirements, Validation Manager will select status “Validated”.

3) If the validation report does not comply with the project requirements, even after corrections, Validation Manager will select status “Rejected”.

Final validation report should be provided with the final set of validated data for future exploitation of the data. It does not need to be shared with data donators.

Once the full validation process has been finalized and validation report stored in the ELRC-SHARE repository, the resources will be classified according to the 3 different edition status:

- resources that have been submitted with the minimum metadata requirements are kept as “ingested”,
- if the resource has passed the validation process positively, the corresponding metadata and attached data should be moved into the “published” status in the ELRC-SHARE repository,
- remaining resources that do not comply to the above conditions should be kept as “internal”.
7.2 Processing report

A processing report should be produced for each processed LR, so that data users can check what processing steps have been followed. A template of the processing report is also provided in Annex 1, Section 2, as part of the ELRC Data Reports document.
ANNEX 1: Template for ELRC data reports
1. Validation Report

Summary sheet

The validation results for this resource are as follows (please refer to the Validation Guidelines for the meaning of the various items):

<table>
<thead>
<tr>
<th>Validation steps</th>
<th>Validated (check box if yes)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) ELRC scope (see section 1 for details)</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>2) Quick content check (see section 2 for details)</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>3) LR Metadata (see section 3 for details)</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>4) Legal issues (see section 4 for details)</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>5) Content validation (see section 5 for details)</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>6) Declaration on the list of pre-existing rights (see section 6 for details)</td>
<td>☐</td>
<td></td>
</tr>
</tbody>
</table>

If relevant, for details about the processing of the LR, see section 2 (Processing Report) at the end of this document.
1. Compliance with ELRC scope

<table>
<thead>
<tr>
<th>Validated (check box if yes)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data origin (comes from public institutions or relevant to the general administrative/regulatory domain and does not come from the European Commission)</td>
<td>☐</td>
</tr>
<tr>
<td>Language(s) of the data content (not the documentation)</td>
<td>☐</td>
</tr>
</tbody>
</table>

2. Quick content check

<table>
<thead>
<tr>
<th>Validated (check box if yes)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability of files</td>
<td>☐</td>
</tr>
<tr>
<td>Data content acceptability (no empty files, correct alignment for parallel corpora, …)</td>
<td>☐</td>
</tr>
</tbody>
</table>

3. Validation of LR Metadata

a. General information

<table>
<thead>
<tr>
<th>Validated (check box if yes)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language used in free text fields are CEF languages</td>
<td>☐</td>
</tr>
<tr>
<td>Does the &quot;resource name&quot; field contain an English version?</td>
<td>☐</td>
</tr>
<tr>
<td>Does Language(s) in &quot;description&quot; field contain an English version?</td>
<td>☐</td>
</tr>
<tr>
<td>Is there any information mentioning Pre-processing done by the provider?</td>
<td>☐</td>
</tr>
<tr>
<td>Is there any information mentioning Pre-processing done through ELRC services?</td>
<td>☐</td>
</tr>
</tbody>
</table>

---

10 Parallel / multilingual corpora LRs should contain English and, at least, one of the following languages: Bulgarian, Croatian, Czech, Danish, Dutch, Estonian, Finnish, French, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Slovakian, Slovenian, Spanish or Swedish. Monolingual corpora and terminology LRs should contain, at least, one of the following languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Slovakian, Slovenian, Spanish, Swedish.
Has any conversion been performed on this resource so as to make it directly useful for training MT engines of the Automated Translation platform?

<table>
<thead>
<tr>
<th>Mandatory metadata field names</th>
<th>Current value</th>
<th>Correct</th>
<th>Wrong</th>
<th>Missing</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource name</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resource type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSI - Public Sector Information</td>
<td></td>
<td></td>
<td></td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>License</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contact person – surname</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contact person - email</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linguality type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical conceptual resource or Language description type (n/a for corpora)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language(s) name</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encoding level (n/a for corpora)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Character encoding (applicable for corpora only)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mime type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other metadata field names (to be listed if completed by submitter)</td>
<td>Current value</td>
<td>Correct</td>
<td>Wrong</td>
<td>Comments</td>
<td></td>
</tr>
<tr>
<td>Domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conformance to classification scheme</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multilinguality type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribution text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allows Uses Besides DGT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPR Holder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relation type and ID of related resource</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. Accuracy of completed metadata with respect to provided LR
4. Legal validation
   a. If “PSI - Public Sector Information” metadata checkbox is ticked

<table>
<thead>
<tr>
<th>Validated (check box if yes)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;License field&quot; value is identified (any value except “Under Review”)</td>
<td>☐</td>
</tr>
<tr>
<td>If attribution is required, IPR Holder(s) is identified in the “IPR holder” field</td>
<td>☐</td>
</tr>
<tr>
<td>Privacy/Confidentiality (if the resource is identified as private or confidential, is “Personal Data Included” or “Sensitive Data Included” box ticked?)</td>
<td>☐</td>
</tr>
</tbody>
</table>

   b. If “PSI - Public Sector Information” metadata checkbox is not ticked

<table>
<thead>
<tr>
<th>Validated (check box if yes)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;License field&quot; value is identified (any value except “Under Review”)</td>
<td>☐</td>
</tr>
<tr>
<td>If attribution is required, IPR Holder(s) is identified in the “IPR holder” field</td>
<td>☐</td>
</tr>
<tr>
<td>Privacy/Confidentiality (if the resource is identified as private or confidential, is “Personal Data Included” or “Sensitive Data Included” box ticked?)</td>
<td>☐</td>
</tr>
</tbody>
</table>

5. Content Validation

<table>
<thead>
<tr>
<th>AUTOMATIC VALIDATION</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Has spell checking-based TU filtering been done?</td>
<td>Yes ☐ No ☐</td>
</tr>
<tr>
<td>Has alignment score outlier detection-based TU filtering been done?</td>
<td>Yes ☐ No ☐</td>
</tr>
<tr>
<td>Has TU length ratio-based filtering been done?</td>
<td>Yes ☐ No ☐</td>
</tr>
<tr>
<td>Have any other content validation steps been applied? If yes, list them in the columns to the right, one content validation step per row (add further rows if needed)</td>
<td>Yes ☐ No ☐</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MANUAL VALIDATION</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Has manual TU validation been done?</td>
<td>Yes ☐ No ☐</td>
</tr>
<tr>
<td>If yes, indicate manually-annotated sample percentage (in terms of the number of TUs)</td>
<td>&lt; 1 % ☐</td>
</tr>
<tr>
<td>1-3 % ☐</td>
<td></td>
</tr>
</tbody>
</table>
Has fine-grained error annotation been done?  
Yes ☐ No ☐

If yes, indicate error type likelihoods (if available)  
Language identification error  ☐  ☐  ☐  ☐
Tokenisation error  ☐  ☐  ☐  ☐
Translation error  ☐  ☐  ☐  ☐
Machine-translated text  ☐  ☐  ☐  ☐
Free translation  ☐  ☐  ☐  ☐
Character formatting error  ☐  ☐  ☐  ☐
Alignment error  ☐  ☐  ☐  ☐

Have any other content validation steps been applied?  
Yes ☐ No ☐

If yes, list them in the columns to the right, one content validation step per row (add further rows if needed)

6. Declaration on the list of pre-existing rights

<table>
<thead>
<tr>
<th>No.</th>
<th>Options</th>
<th>Selected option</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The results of this LR are free of rights or claims from creators or from any third parties for any use. The contracting authority may envisage and declare that the results do not contain any pre-existing rights to the results or parts of the results or to pre-existing materials as defined in the above-mentioned contract.</td>
<td>☐</td>
</tr>
<tr>
<td>2</td>
<td>The results of this LR and the pre-existing material incorporated in the results are free of rights or claims from creators or from any third parties for any use. The contracting authority may envisage and declare that the results contain the following pre-existing rights:</td>
<td>☐</td>
</tr>
</tbody>
</table>

For Option 2 complete the table below – one line per pre-existing right

<table>
<thead>
<tr>
<th>Result concerned</th>
<th>Pre-existing material concerned</th>
<th>Rights to pre-existing material</th>
<th>Identification of rights’ holder</th>
</tr>
</thead>
</table>
2. Processing Report

This report provides details on the processing steps carried out on the resource referred to above. This information is filled in by the same LR validator.

<table>
<thead>
<tr>
<th>Processing action</th>
<th>Check if true</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the processed resource originate from ELRC sources?</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>Has automatic text extraction from scanned documents (via Optical Character Recognition – OCR) been performed?</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>Has automatic text extraction from PDF or DOC(X) documents been performed?</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>Has automatic document pair detection been performed?</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>Has automatic sentence-level alignment been performed?</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>Has TMX cleaning been performed?</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>Have any other processing steps been carried out? If yes, list them in the columns to the right, one processing step per row (add further rows if needed)</td>
<td>☐</td>
<td></td>
</tr>
</tbody>
</table>
