CURLICAT

Curated Multilingual Language Resources for CEF.AT

Agreement number: INEA/CEF/ICT/A2019/1926831

Action No: 2019-EU-IA-0034

CEF
programme

# Deliverable 5.1
# Metadata Harmonisation in CURLICAT

**Version 1.0**

**2022-05-31**

## Document Information

| | |
|---|---|
| Activity: | Activity 5: METADATA Harmonisation |

| | |
|---|---|
| Deliverable number: | DELIVERABLES:<br>D5.1 COMMON METADATA SCHEMA<br>D5.2 MAPPING THE ATTRIBUTES IN THE ORIGINAL METADATA TO THE ATTRIBUTES OF THE COMMONLY AGREED METADATA SCHEMA<br>D5.3 VALIDATED COMMON METADATA FOR ALL THE SEVEN LARGE SCALE MONOLINGUAL CORPORA |
| Deliverable title: | Metadata Harmonisation |
| Indicative submission date: | 2022-05-31 |
| Actual submission date of deliverable: | 2022-05-31 |
| Main Author(s): | Elena Irimia, Bence Nyéki, Bartłomiej Nitoń, Andraž Repar, Radovan Garabik, Radu Ion, Svetla Koeva, Martin Yalamov, Dan Tufiș |
| Participants: | Tsvetana Dimitrova |
| Version: | V1.0 |

## History of versions

| Version | Date | Status | Name of the Author (Partner) | Contributions | Description/ Approval Level |
|---|---|---|---|---|---|
| V1.0 | 03.12.2021 | Draft | ICIA | Elena Irimia,<br>Bartłomiej Nitoń,<br>Andraž Repar,<br>Radovan Garabik,<br>Svetla Koeva,<br>Vanja Štefanec,<br>Dan Tufiș | |

## EXECUTIVE SUMMARY

This deliverable provides a detailed documentation of the work of harmonising the metadata for the CURLICAT corpus – a collection of multilingual corpora for seven consortium languages: Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovene, at least 20 million words per language, selected from the national corpora and supplemented with additional, IPR-cleared data. Providing metadata for such big corpora is particularly essential for allowing users to design sub-corpora based on different criteria like language, domain, publication date, literary style, etc., according to their specific interest. Each national corpora in CURLICAT comes with their own metadata framework and particular attention has to be paid to defining a common set of metadata fields and value types. This document contains descriptions of each original metadata schema for each corpus in the project (see section 1.1, 2.1, 3.1, 4.1, 5.1, 6.1 and 7.1) and the mapping of these schemata (see sections 1.2, 2.2, 3.2, 4.2, 5.2, 6.2 and 7.2) to the commonly agreed CURLICAT schema (see section 8). Validation activities for the metadata are described (see sections 1.3, 2.3, 3.3, 4.3, 5.3, 6.3 and 7.3).

## Table of contents

## Introduction

Metadata fields contain information about documents in the corpora and they represent the means for extracting texts from larger collections according to specific criteria (domain, style, language, etc.) and for exploiting the richness of corpora. Given their importance, the aim of this activity was to create harmonised metadata for all the resources provided through this project: the entire collection of texts in 7 languages, containing more than 150 million words has to be structurally searchable by means of the same criteria.

Each language subcorpus in CURLICAT comes with a specific metadata schema and our objective is to provide a harmonised common CURLICAT schema: deciding on the format for metadata encoding (e.g. XML separate file or in the heart of the CONLL-U format text documents), settling for the included fields, defining value types for these fields, mapping specific metadata fields to the fields in the common schema, converting the language corpora metadata to the CURLICAT agreed format and technically and semantically validating the metadata.

The activity was carried out along three tasks in the project:

T5.1: Setting the common (CURLICAT) metadata schema,

T5.2: Mapping components from individual metadata schema to the components from the

commonly agreed metadata schema and

T5.3 Turning all the individual metadata attached to the delivered text files into the common

metadata format.

# 1. The Bulgarian metadata description (T5.2, T5.3)

## 1.1. Original metadata description (T5.2)

The **Bulgarian CURLICAT corpus** consists of texts from different sources, provided with appropriate licences for distribution. We used three general types of sources with regard to the metadata extraction:

- [1] Sources with very rich metadata structure, such as the Bulgarian National Corpus;
- [2] Sources with a shallow metadata structure, such as some public repositories with open and copyright free data;
- [3] Sources with no metadata structure but with extractable metadata values, such as copyright free blogs, public domain websites, etc.

The metadata schemata of these sources are described below.

### 1.1.1. Bulgarian National Corpus metadata schema

In the **Bulgarian National Corpus**, the classification suggested by Burnard (2005) is adopted as a baseline description of the text metadata: a) **editorial** – information about texts in relation to their original source (source, author, date of publishing, etc.; here we included information about language, direction of translation, name of the translator, etc.); b) **descriptive–classificatory** information such as style, domain, and genre; c) **administrative** – documentary information about the texts and the corpus, such as its availability, revision status, etc.; d) **analytical** – various levels of annotation; e) **statistical** – number of tokens, words, general words, domain-specific words, lemmas, noun phrases, phrases, clauses, sentences, etc. In addition to Burnard's classification we include various statistical information (Koeva et al. 2012: 81).

The metadata description of the texts in the Bulgarian National Corpus is stored into 25 categories that are compliant with the established standards (Burnard, 2005), although defined for the particular needs of the Bulgarian National Corpus. Metadata are mostly derived automatically, using two main techniques: a) extracting information from the html or xml markup of the original files collected from the Internet, and b) keyword-based heuristics. Html pages usually contain specifically tagged editorial information such as author, title, and date of publishing that are easily extractable from the html source. On the average, in each metadata record in the BulNC, 17.79 categories are non-empty (71.16%),

Some of the metadata categories are optional and here we present: a) obligatory metadata categories within the Bulgarian National Corpus, and b) optional metadata categories within the

Bulgarian National Corpus corresponding with the categories from the CURLICAT metadata schema.

| **Title** | **Value: type string** |
|---|---|

A string that represents the document title.

| **Author** | **Value: type string** |
|---|---|

A string that represents the author's name.

| **TranslatorName** | **Value: type string** |
|---|---|

If the text in the document is a translation, the value of this field is the name of the translator.

| **Style** | **Value: type string** |
|---|---|

This style is chosen from the predefined values: *Administrative, Science, Fiction, Journalism, Popular science, Informal, Informal/Fiction, Science/Administrative* (*Healthcare*)*, Undefined.*

| **Domain** | **Value: type string** |
|---|---|

The field describes the thematic domain to which the document belongs: Politics, Law, Education, Economy, Health, Military, Culture and arts, Sports, Ecology, Social policy, Relations, Undefined, etc. The number of domains distributed by style is presented in Table 1.

| **Genre** | **Value: type string** |
|---|---|

This genre is chosen from the predefined values: *novel, story, subtitles, agreement, article, manual, debates, contract, report, interview, minutes, undefined, etc.* The number of genres distributed by style is presented in Table 1.

| **PublicationDate** | **Value: type string** |
|---|---|

This field encodes the date (year) when the associated document was published.

| **CollectionDate** | **Value: type string** |
|---|---|

This field encodes the date (year) when the associated document was acquired.

| | |
|---|---|
| **SourceType** | **Value: type string** |

The values of this field are strings representing the source: internet, publishing house, etc.

| | |
|---|---|
| **Source** | **Value: type string** |

The values of this field are strings representing the source address: internet link, etc.

| | |
|---|---|
| **TextForm** | **Value: type string** |

The modality is chosen from the predefined values: *written, spoken.*

| | |
|---|---|
| **Language** | **Value: type string** |

This attribute could take values in the range of *language names.*

| | |
|---|---|
| **Quality** | **Value: type string** |

The values of this field are strings representing the document format from which the text is extracted: xml, html, doc, txt, etc.

| | |
|---|---|
| **LicensingTerm** | **Value: type string** |

This field encodes the ownership of the text.

| | |
|---|---|
| **WordNumber** | **Value: type digit string** |

The values of this field are digit strings representing the number of words within the document.

| | |
|---|---|
| **ParagraphNumber** | **Value: number** |

The values of this field are digit strings representing the number of paragraphs within the document.

| | |
|---|---|
| **SentenceNumber** | **Value: number** |

The values of this field are digit strings representing the number of sentences within the document.

| Style | Number of domains | Number of Genres |
|---|---|---|
| **Administrative** | 11 | 16 |
| **Science** | 21 | 15 |
| **Popular science** | 25 | 7 |
| **Journalism** | 19 | 12 |
| **Fiction** | 13 | 25 |
| **Informal** | NA | NA |
| **Informal/Fiction** | 17 | 1 |
| **Science/Administrative** | 21 | 16 |

**Table 1:** Bulgarian National Corpus: number of domains and genres distributed by style

Other categories of metadata that we do not describe in detail here are: edited version (if the text is edited or not), normalization version (if the text is normalized or not), number of original texts in the sample, overlapping of the text with original sample (i.e., exact match, paragraph, random excerption, etc.), author's information (age, sex, nationality, native language), direction of translation, name of publisher, place of publishing, text edition (first edition, second edition), parallel text (yes or no), text origin (original, translation), translator's information (age, sex, nationality, native language), title of the original text, notes (any additional information), keywords, administrative information about the file and access to it, etc. (Koeva et al. 2016) Detailed metadata allows comprehensive classification and easy selection of texts when creating subcorpora according to certain criteria (e.g. thematic domain, year of publication, authorship, translation, etc.).

### 1.1.2. Sources with a shallow metadata structure

The sources in this category contain the **National portal for open science (NPOS)** and the **repositories provided by some Bulgarian universities**. The documents are organised according to: the collections they belong to, issue date, authors, titles, keywords, and scientific domains. In addition, the National portal for open science provides information about the providers, licensing terms and the language. We collected documents with the following criteria: they are in Bulgarian, with clear and free copyright and (preferably) they belong to the CURLICAT selected domains. The available metadata come from the compilation of the documents.

| | | |
|---|---|---|
| **Title** | **All** | **Value: type string** |

A string that represents the document title.

| | | |
|---|---|---|
| **Author** | **All** | **Value: type string** |

A string that represents the author's name.

| | | |
|---|---|---|
| **Domain** | **All** | **Value: type string** |

A string that represents the document domain. In different repositories the name of the metadata category is different.

| | | |
|---|---|---|
| **PublicationDate** | **All** | **Value: type string** |

This field encodes the date (year) when the document was published at the repository. In different repositories the name of the metadata category is different.

| | | |
|---|---|---|
| **Source** | **All** | **Value: type string** |

The values of this field are strings representing the source address: internet link, etc. In different repositories the name of the metadata category is different.

| | | |
|---|---|---|
| **Language** | **BPOS** | **Value: type string** |

This attribute could take values in the range of *language names.*

| | | |
|---|---|---|
| **Provider** | **BPOS** | **Value: type string** |

This field encodes the names of the academic institutions to which the author is affiliated.

| **LicensingTerm** | **BPOS** | **Value: type string** |
|---|---|---|

This field encodes the licence under which the document is published: Attribution 4.0 International, Attribution-NoDerivatives 4.0 International, Attribution-NonCommercial 4.0 International, Attribution-NonCommercial-NoDerivatives 4.0 International, Attribution-NonCommercial-ShareAlike 4.0 International, Attribution-ShareAlike 4.0 International, etc.

### 1.1.3. Sources with no metadata structure

The documents from such sources were collected, if they have clear and free copyright, in addition to the documents from the Bulgarian National Corpus. Documents were selected if the CURLICAT obligatory metadata values could be extracted from them automatically, and if they belong to the thematic domains represented in the project.

## 1.2. Mapping with the common metadata schema (described in Section 8) (T5.2)

The following set of core obligatory metadata (common for CURLICAT) is used for all Bulgarian documents:

- *Identifier* – unique identifier of the document within all the collections, created using bg language code as a prefix;
- *Language* – language code of the Bulgarian sub-corpora (‘bg’);
- *Licence* – the conditions for use: CC licence (e.g. *CC BY-SA 4.0*) or source-specific licences;
- *PublicationDate* – the date of the original publication of the document, in ISO 8601 format;
- *DocumentTitle* – an informative, human readable title (name) of the document;
- *Type* – specifies the type of the source document (e.g. book, chapter, paper, newspaper article, blogpost, etc.);
- *Source* – the name of the organisation that published the source document, be it a Journal, Publishing House, Blog, Website, etc., in the original language;
- *Domain* – classification of particular thematic domain selected from the predefined list of CURLICAT domains and based on the domain metadata fields in the source corpora;

- *NumberWords* – the total number of words in the document;
- *NumberSentences* – the total number of sentences in the document;
- *NumberTokens* – the total number of tokens in the document.

The following set of metadata is optional for the Bulgarian documents:

- *Author* – the name/s of the person/s that created the text in the source document;
- *Url* – the original individual address the document was accessed at, if applicable;
- *Style* – the literary style of the text in the document, selected from a predefined list: imaginative, administrative, law, journalistic, etc;
- *Subdomain* – a further classification of the documents into narrower categories, e.g. scientific fields for the Science domain, or cultural fields for the Culture domain;

There are also some metadata specific for the Bulgarian dataset that we decided to keep:

- EuroVoc – automatic classification to Eurovoc classes;
- CollectionDate – the date of collection of the document in the ISO 8601 format;
- LicenseLink – the link to the licence at the source webpage, if available;
- ParagraphNumber – the total number of paragraphs in the document.

| **EuroVoc** | **Value: type string** |
|---|---|

A string that represents the automatic classification to the Eurovoc classes. The values have the following format: 40/0.1875 24/0.1719 12/0.125, where initial digits point to the EuroVoc class and the fractions indicate the confidence.

| **CollectionDate** | **Value: ISO 8601 format date** |
|---|---|

The date of the collection of the document, in ISO 8601 format.

| **LicenseLink** | **Value: type string** |
|---|---|

The link to the licence, specified at the source webpage is provided,  if available.

| **ParagraphNumber** | **Value: type digit string** |
|---|---|

The total number of paragraphs in the document.

The following activities were performed to harmonise the metadata of Bulgarian documents with the accepted conventions.

1) Some limited number of metadata categories and values remain unchanged as they coincide with the accepted format (U);

2) Some original metadata categories and values are directly mapped to the CURLICAT metadata and values (M);

3) Some metadata values are automatically extracted (E). The main techniques for automatic extraction of metadata are: a) metatextual procedures, which consist of information extraction from the html/xml markup of the original files; and b) textual procedures, which consist of text analysis and heuristics using a set of language resources.

The following metadata is extracted automatically from html sources: author, document title, publishing date. The publication (creation) date is also extracted from pdf files. Classification information includes the domain of the documents, their genre and type and results from text analysis. In some cases the source may contain classificatory labels according to an adopted domain and/or genre classification on the source, e.g. texts on a news website can be classified into editorials and articles of various domains – Economy, Sport, etc.

Documents not classified to thematic domains are classified with the Bulgarian MARCELL classifier, which combines statistical and predictive modelling.  Bulgarian Statistical classifier groups documents containing EuroVoc terms related to one Top Level Domain. In addition, IATE pointers to EuroVoc Micro Thesauruses or Top Level Domains are taken into account if a particular term is not presented in EuroVoc. The Statistical classifier is designed to work as a multi-label classifier providing confidence measures for the correctness of assigned classes. It relies on data pre-processing, which is part of the Bulgarian Language Processing Chain (Koeva et al. 2020). A second classification method (Classification predictive modelling based on document titles) complements the Statistical classifier and both are integrated in the Bulgarian pipeline.

4. Some metadata values are automatically generated (G). Here is the statistical information – it is derived from processing the text, and includes the number of words, tokens, sentences, etc. Administrative metadata such as the document identifier, language code and source is also generated.

| **Metadata** | **Source type** [1] | **Source type** [2] | **Source type** [3] |
|---|---|---|---|
| *Identifier* | G | G | G |
| *Language* | M | M/G | G |
| *Licence* | U/M | U/M | E |
| *PublicationDate* | U/M | U/M | E |
| *Source* | U/M | U/M | G |
| *Domain* | U/M | U/M/E | E |
| *NumberWords* | U | G | G |
| *NumberSentences* | U | G | G |
| *NumberTokens* | G | G | G |
| *Author* | U/M | U/M | E |
| *Url* | U/M | U/M | E |

| | | | |
|---|---|---|---|
| *Style* | U/M | NA | NA |
| *Subdomain* | U/M | U/M/E | E |

T**able 2:** Main activities in the harmonisation of metadata in Bulgarian documents, U = unchanged, M = mapped; E = extracted; G = generated, NA = not available.

### 1.3. Metadata validation activities (T 5.3)

#### 1.3.1. Technical validation

Technical validation ensures that metadata is in the correct format. This includes the format of the file the metadata is included in, and the format of the metadata categories and values. Some of the values are inherited from the source metadata, some are extracted from the sources, and some are calculated or generated. In all three cases the accepted format for the metadata values is applied. Some other technical conventions are also included: the order of the metadata records and the empty fields.

#### 1.3.2. Semantic validation

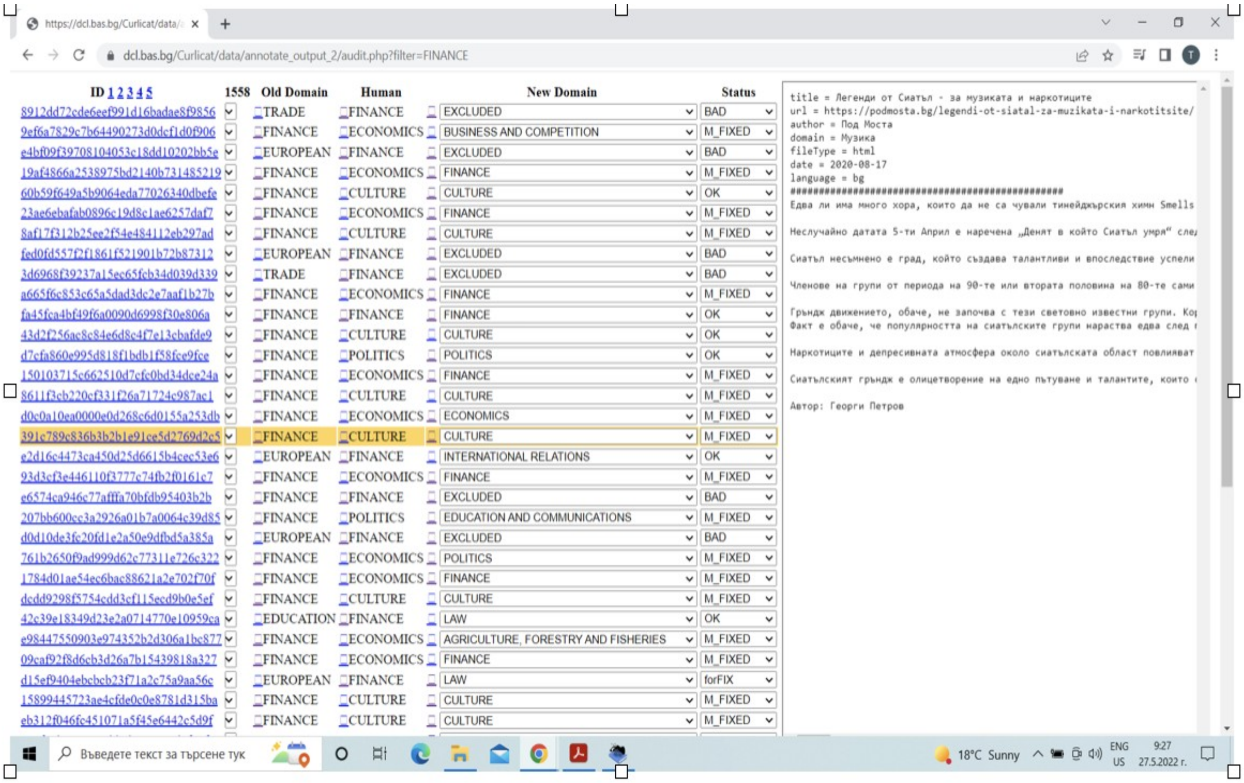
Semantic validation is mainly directed to the classification of documents to different domains. There are three basic cases: documents obtained with a classification to a thematic domain; documents that were automatically classified with the Bulgarian MARCELL classifier, and documents that are manually assigned with a thematic domain.

A web application (specially designed for this task) is used for viewing text files in the CURLICAT corpus, selecting those that can be used (and discarding those that can not) and editing the domain attribution metadata value of each file. The first column on the left gives the file name in the corpus; the next two columns show information about the domain attribution (first one is automatically assigned by the MARCELL classifier, and the second one is assigned by a human, however not to a particular file, but to all source documents or parts of them if they can be clearly associated with a domain), while the third column is used for new domain attribution (if the previous attributions need to be corrected).
The last column on the right is used for signalling the status of the file, as follows:

- BAD (and EXCLUDED) files are to be discarded (these were mostly files in languages other than Bulgarian – in English and in Russian, and files with text that has not been properly identified – missing paragraphs, partial sentences, too many special symbols due to tables, formulae, etc.).
- OK files are good to be used and further processed .

- M_FIXED is for files that have been manually edited/fixed: the panel on the right shows the text in the file. If the text in the file is in need of further editing, it can be done outside the application, with any text editor.
- forFIX is for files that were deemed to be able to be automatically fixed further.

The application was used for editing information in about 2319 text files (metadata for 904 text files were manually edited).


https://dcl.bas.bg/Curlicat/data/
dcl.bas.bg/Curlicat/data/annotate_output_2/audit.php?filter=FINANCE
Въведете текст за търсене тук
18°C Sunny
ENG
US
9:27
27.5.2022 г.

| ID 1 2 3 4 5 | 1558 | Old Domain | Human | New Domain | Status |
|---|---|---|---|---|---|
| 8912dd72cde6eef991d16badae8f9856 | ☑ | TRADE | FINANCE | EXCLUDED | BAD |
| 9ef6a7829c7b64490273d0def1d0f906 | ☑ | FINANCE | ECONOMICS | BUSINESS AND COMPETITION | M_FIXED |
| e4bf09f39708104053c18dd10202bb5e | ☑ | EUROPEAN | FINANCE | EXCLUDED | BAD |
| 19af4866a2538975bd2140b731485219 | ☑ | FINANCE | ECONOMICS | FINANCE | M_FIXED |
| 60b59f649a5b9064eda77026340dbefe | ☑ | FINANCE | CULTURE | CULTURE | OK |
| 23ae6ebafab0896c19d8c1ae6257daf7 | ☑ | FINANCE | ECONOMICS | FINANCE | M_FIXED |
| 8af17f312b25ee2f54e484112eb297ad | ☑ | FINANCE | CULTURE | CULTURE | M_FIXED |
| fed0fd557f2f1861f521901b72b87312 | ☑ | EUROPEAN | FINANCE | EXCLUDED | BAD |
| 3d6968f39237a15ec65fcb34d039d339 | ☑ | TRADE | FINANCE | EXCLUDED | BAD |
| a665f6c853c65a5dad3dc2e7aaf1b27b | ☑ | FINANCE | ECONOMICS | FINANCE | M_FIXED |
| fa45fca4bf49f6a0090d6998f30e806a | ☑ | FINANCE | FINANCE | FINANCE | OK |
| 43d2f256ac8c84e6d8c4f7e13cbafde9 | ☑ | FINANCE | CULTURE | CULTURE | OK |
| d7cfa860e995d818f1bdb1f58fce9fce | ☑ | FINANCE | POLITICS | POLITICS | OK |
| 150103715c662510d7cfc0bd34dce24a | ☑ | FINANCE | ECONOMICS | FINANCE | M_FIXED |
| 8611f3cb220cf331f26a71724c987ac1 | ☑ | FINANCE | CULTURE | CULTURE | M_FIXED |
| d0c0a10ea0000e0d268c6d0155a253db | ☑ | FINANCE | ECONOMICS | ECONOMICS | M_FIXED |
| 391c789c836b3b2b1e91ce5d2769d2c5 | ☑ | FINANCE | CULTURE | CULTURE | M_FIXED |
| e2d16c4473ca450d25d6615b4cec53e6 | ☑ | EUROPEAN | FINANCE | INTERNATIONAL RELATIONS | OK |
| 93d3cf3e446110f3777c74fb2f0161c7 | ☑ | FINANCE | ECONOMICS | FINANCE | M_FIXED |
| e6574ca946c77afffa70bfdb95403b2b | ☑ | FINANCE | FINANCE | EXCLUDED | BAD |
| 207bb600cc3a2926a01b7a0064c39d85 | ☑ | FINANCE | POLITICS | EDUCATION AND COMMUNICATIONS | M_FIXED |
| d0d10de3fc20fd1e2a50e9dfbd5a385a | ☑ | EUROPEAN | FINANCE | EXCLUDED | BAD |
| 761b2650f9ad999d62c77311e726c322 | ☑ | FINANCE | ECONOMICS | POLITICS | M_FIXED |
| 1784d01ae54ec6bac88621a2e702f70f | ☑ | FINANCE | ECONOMICS | FINANCE | M_FIXED |
| dcdd9298f5754cdd3cf115ecd9b0e5ef | ☑ | FINANCE | CULTURE | CULTURE | M_FIXED |
| 42c39e18349d23e2a0714770e10959ca | ☑ | EDUCATION | FINANCE | LAW | OK |
| e98447550903e974352b2d306a1bc877 | ☑ | FINANCE | ECONOMICS | AGRICULTURE, FORESTRY AND FISHERIES | M_FIXED |
| 09caf92f8d6cb3d26a7b15439818a327 | ☑ | FINANCE | ECONOMICS | FINANCE | M_FIXED |
| d15ef9404ebcbcb23f71a2c75a9aa56c | ☑ | EUROPEAN | FINANCE | LAW | forFIX |
| 15899445723ae4cfde0c0e8781d315ba | ☑ | FINANCE | CULTURE | CULTURE | M_FIXED |
| eb312f046fc451071a5f45e6442c5d9f | ☑ | FINANCE | CULTURE | CULTURE | M_FIXED |

```
title = Легенди от Сиатъл - за музиката и наркотиците
url = https://podmosta.bg/legendi-ot-siatal-za-muzikata-i-narkotitsite/
author = Под Моста
domain = Музика
fileType = html
date = 2020-08-17
language = bg
################################################
Едва ли има много хора, които да не са чували тинейджърския химн Smells

Неслучайно датата 5-ти Април е наречена „Денят в който Сиатъл умря“ сле

Сиатъл несъмнено е град, който създава талантливи и впоследствие успели

Членове на групи от периода на 90-те или втората половина на 80-те сами

Гръндж движението, обаче, не започва с тези световно известни групи. Ко
Факт е обаче, че популярността на сиатълските групи нараства едва след

Наркотиците и депресивната атмосфера около сиатълската област повлияват

Сиатълският гръндж е олицетворение на едно пътуване и талантите, които

Автор: Георги Петров
```

Figure 1. Web application for the evaluation and correction of the domain metadata value

# 2. The Croatian metadata description (T5.2, T5.3)

## 2.1. Original metadata description (T5.2)

The Croatian CURLICAT corpus is composed of three distinct sources:

1. selection of documents from the Croatian National Corpus (HNK), covering the domains of culture and economy;
2. selection of documents from the MedCorA Corpus;
3. selection of scientific papers from HRČAK, the Open Access central portal of Croatian scientific journals.[1]

### 2.1.1 Croatian National Corpus (HNK) metadata schema

**Croatian National Corpus** has been collected for a longer period of time and the metadata description slightly varies depending on the particular document collecting round. However, two of the attributes are common to all documents and these are type and file.

| **Type** | **Value: type string** |
|---|---|

A string that represents the type of document, whether it is an article, document, etc.

| **File** | **Value: type string** |
|---|---|

A string that represents the filename of the source document. In essence, it can be understood as an identifier of the document since it is unique throughout the corpus. Also, it has a structured format from which other metadata attributes, i.e. source name, domain or publication date, can be reconstructed, although they are not explicitly assigned to the document. E.g. from the value “vj20100630kul08”, we can reconstruct the source name (daily newspaper “Vjesnik”), published date (“2010-06-30”) and domain (“culture”) while “08” stands for the eight article in this domain in that daily issue.

### 2.1.2 MedCorA Corpus metadata schema

**MedCorA Corpus** was collected from HALMED, Croatian Agency for Medicinal Products and Medical Devices and is composed of pharmaceutical instructions for medicinal products. Its metadata description contains the following attributes:

| **Identifier** | **Value: type string** |
|---|---|

A string that uniquely identifies the document in the corpus.

[1] https://hrcak.srce.hr

| | |
|---|---|
| **Name** | **Value: type string** |

A string that represents the name of the medicinal product.

| | |
|---|---|
| **Authorisation_number** | **Value: type string** |

A string that represents the Marketing Authorisation Number for the medicinal product.

| | |
|---|---|
| **Date** | **Value: ISO 8601 format date** |

A ISO 8601 formatted string representing the date of approval for the medicinal product.

| | |
|---|---|
| **URL** | **Value: type string** |

A string that represents the URL to the document.

### 2.1.3 HRČAK metadata schema

**HRČAK**, as a digital repository, is fully compliant with the OpenAIRE Guidelines for Literature Repository Managers 3.0. It features a very elaborate metadata schema available in multiple formats, including MODS. This repository was harvested primarily for the purpose of the CURLICAT project, so the chosen set of metadata attributes matches the CURLICAT common metadata schema.

## 2.2. Mapping with the common metadata schema (described in Section 8) (T5.2)

With respect to the metadata schema proposed in the CURLICAT project we use the following mapping for different sources of Croatian data.

### 2.2.1 Mapping the Croatian National Corpus metadata to the CURLICAT metadata schema

#### 2.2.1.1 Obligatory attributes

| HNK | CURLICAT | Details |
|---|---|---|
| - | *Identifier* | Value of the file attribute is added to the common prefix “hr-hnk-”. |
| - | *Language* | Always ‘hr’. |
| - | *Licence* | Always “other freely redistributable”. |

| | | |
|---|---|---|
| *File* | *PublicationDate* | If not reconstructible from the file attribute, corpus publishing date is used. |
| - | *DocumentTitle* | “N/A” if not reconstructible from the source XML element attribute <head type=”na”>. |
| - | *Type* | Always “newspaper article”. |
| *File* | *Source* | Mapping to a closed set of sources, based on the value reconstructed from the file attribute. |
| *File* | *Domain* | Mapping to “culture” or “economy”, based on the value reconstructed from the file attribute. |
| - | *Number of sentences, words, punctuation marks and tokens* | Calculated automatically during processing. |

### 2.2.1.2 Optional attributes

| | | |
|---|---|---|
| - | *Author* | “N/A” if not reconstructible from the source XML element <byline>. |
| - | *SourceType* | Always “Newspaper”. |
| - | *Keywords* | Omitted. |
| - | *Url* | “N/A” if not reconstructible from the XML element attribute <doc url=”...”>. |
| - | *Style* | Always “journalistic”. |

| | | |
|---|---|---|
| - | *Subdomain* | Omitted. |
| - | *ISSN_ISBN_EISBN* | Omitted. |

### 2.2.2 Mapping the MedCorA Corpus metadata to the CURLICAT metadata schema

#### 2.2.2.1 Obligatory attributes

| MedCorA | CURLICAT | Details |
|---|---|---|
| *Identifier* | *Identifier* | Value of the identifier attribute is added to the common prefix “hr-med-”. |
| - | *Language* | Always ‘hr’. |
| - | *License* | Always “other freely redistributable”. |
| *Date* | *PublicationDate* | Value of the date attribute. |
| *Name* | *DocumentTitle* | Value of the name attribute. |
| - | *Type* | Always “document”. |
| - | *Source* | Always “HALMED”. |
| - | *Domain* | Always “health”. |
| - | *Number of sentences, words, punctuation marks and tokens* | Calculated automatically during processing. |

#### 2.2.2.2 Optional attributes

| | | |
|---|---|---|
| - | *Author* | Omitted. |
| - | *SourceType* | Always “Other”. |

| | | |
|---|---|---|
| - | *Keywords* | Omitted. |
| *Url* | *Url* | Value of the Url attribute. |
| - | *Style* | Always “scientific”. |
| - | *Subdomain* | Omitted. |
| - | *ISSN_ISBN_EISBN* | Omitted. |

### 2.2.3 Mapping HRČAK metadata to the CURLICAT metadata schema

#### 2.2.3.1 Obligatory attributes

| **Hrčak** | **CURLICAT** | **Details** |
|---|---|---|
| *Identifier* | *Identifier* | Document's repository OAI identifier is added to the common prefix “hr-hrcak-”. |
| - | *Language* | Always ‘hr’. |
| *Licence* | *Licence* | “Other freely redistributable” if attribute not present in repository metadata. |
| *Date* | *PublicationDate* | Attribute present in repository metadata. |
| *Name* | *DocumentTitle* | Attribute present in repository metadata. |
| - | *Type* | Always “paper”. |
| *Source* | *Source* | Attribute present in repository metadata. |

| | | |
|---|---|---|
| - | *Domain* | Manually mapped to CURLICAT domains. |
| - | *Number of sentences, words, punctuation marks and tokens* | Calculated automatically during processing. |

### 2.2.3.2 Optional attributes

| | | |
|---|---|---|
| *Author* | *Author* | Attribute present in repository metadata. |
| - | *SourceType* | Always “Other”. |
| Keywords | *Keywords* | “N/A” if attribute not present in repository metadata. |
| *Url* | *Url* | Attribute present in repository metadata. |
| - | *Style* | Always “scientific”. |
| - | *Subdomain* | Omitted. |
| *ISSN* | *ISSN_ISBN_EISBN* | Attribute present in repository metadata. |

## 2.3. Metadata Validation Activities (T5.3)

During automatic corpus generation, various methods are performed to ensure that all the obligatory metadata attributes are present and in correct format. Also, during document selection, most of the metadata attributes were manually checked and verified.

# 3. The Hungarian metadata description (T5.2, T5.3)

## 3.1. Original metadata description (T5.2)

The Hungarian CURLICAT corpus consists of texts from the following sources:

- Wikipedia articles from the scientific subcorpus of the Hungarian National Corpus (MNSZ2, HNC)[2]
- Digitized books from the publishers Akadémiai Kiadó and Osiris Kiadó provided by Arcanum[3]
- Journals from the REAL-J repository of the Library and Information Centre, Hungarian Academy of Sciences[4]

The differences between the list of sources above and the list in Deliverable 1.1 *Collection of multilingual corpora* are due to IPR clearance. Some of the originally selected documents needed to be discarded from the Hungarian CURLICAT corpus owing to legal issues, but other texts available under Creative Commons licences were collected. Further materials may be added to the corpus in the forthcoming stages of work.

The sources listed above provided the metadata schemata described below.

### 3.1.1. Hungarian National Corpus metadata schema

The Wikipedia articles collected from the Hungarian National Corpus were available in a single XML file that contained metadata information. The DTD schema of the metadata header is given below.

```
<!ELEMENT cesHeader (fileDesc, encodingDesc, profileDesc, revisionDesc)>
<!ATTLIST cesHeader lang CDATA #REQUIRED
        type CDATA #REQUIRED
        status CDATA #REQUIRED
        version CDATA #REQUIRED
        TEIform CDATA #REQUIRED>
<!ELEMENT fileDesc (titleStmt, editionStmt, extent, publicationStmt, sourceDesc)>
```

[2] http://corpus.nytud.hu/mnsz/index_eng.html
[3] https://www.arcanum.com/en
[4] http://real-j.mtak.hu/

```
<!ELEMENT titleStmt (h.title, respStmt?)>
<!ELEMENT h.title (#PCDATA)>
<!ELEMENT respStmt (respName+, respType+)>
<!ELEMENT respName (#PCDATA)>
<!ELEMENT respType (#PCDATA)>
<!ELEMENT editionStmt (#PCDATA)>
<!ELEMENT extent (wordCount, byteCount)>
<!ELEMENT wordCount (#PCDATA)>
<!ELEMENT byteCount (#PCDATA)>
<!ELEMENT publicationStmt (distributor, pubAddress, eAddress?, availability, pubDate)>
<!ELEMENT distributor (#PCDATA)>
<!ELEMENT pubAddress (#PCDATA)>
<!ELEMENT eAddress (#PCDATA)>
<!ATTLIST eAddress type CDATA #REQUIRED>
<!ELEMENT availability (#PCDATA)>
<!ATTLIST availability region CDATA #IMPLIED status CDATA #IMPLIED>
<!ELEMENT pubDate (#PCDATA)>
<!ELEMENT sourceDesc (biblFull | biblStruct)>
<!ATTLIST sourceDesc Default CDATA #IMPLIED>
<!ELEMENT biblFull (titleStmt, publicationStmt, sourceDesc)>
<!ELEMENT biblStruct (monogr)>
<!ELEMENT monogr (h.title, edition, imprint)>
<!ELEMENT edition (#PCDATA)>
<!ELEMENT imprint (publisher, pubDate, pubPlace)>
<!ELEMENT publisher (#PCDATA)>
<!ELEMENT pubPlace (#PCDATA)>
<!ELEMENT encodingDesc (projectDesc)>
<!ELEMENT projectDesc (#PCDATA)>
<!ATTLIST projectDesc Default CDATA #IMPLIED>
<!ELEMENT profileDesc (langUsage)>
<!ELEMENT langUsage (language+)>
<!ELEMENT language (#PCDATA)>
<!ATTLIST language id #REQUIRED iso639 #REQUIRED>
```

```
<!ELEMENT> revisionDesc (change)>
<!ELEMENT change (changeDate, respName, h.item)>
<!ELEMENT changeDate (#PCDATA)>
<!ELEMENT h.item (#PCDATA)>
```

The semantics of these elements are specified by the TEI standard[5]. Some of the element names slightly differ from the corresponding standard TEI names. They are mapped to TEI elements or attributes in Table 3.

| HNC | TEI |
|---|---|
| element *cesHeader* | element *teiHeader* |
| element *h.title* | element *title* |
| element *pubDate* | element *date* |
| element *pubAddress* | element *address* |
| element *eAddress* | element *address* |
| element *respName* | element *name* |
| element *respType* | element *resp* |
| element *changeDate* | attribute *when* of element *change* |

**Table 3:** XML element names from the HNC and the corresponding TEI elements.

The elements *wordCount*, *byteCount* and *h.item* could not be directly mapped to TEI elements by their semantics. Their description is provided in Table 4.

| Element | Description |
|---|---|
| *wordCount* | The number of words in the document |
| *byteCount* | The document size in bytes. |
| *h.item* | Arbitrary additional information |

**Table 4:** Description of XML elements from the HNC without a direct mapping to TEI elements.

[5] https://tei-c.org/

For more clarity in section 3.2.1, the semantics of the elements relevant to the CURLICAT corpus are given in Table 5 (these specifications correspond to the TEI standard).

| Element | Description |
|---|---|
| *h.title* | The title of any kind of document |
| *pubDate* | A date in any format |
| *respName* | A proper noun or noun phrase |
| *respType* | Describes a person's or organisation's role in the production or the distribution of a work |

**Table 5:** Description of the XML elements from the HNC that are relevant to the CURLICAT corpus.

### 3.1.2. Arcanum metadata schema

The documents from the Arcanum database were described in a single TSV document that includes the following fields:

| | |
|---|---|
| **publication_title** | **Value: type string** |

A string that represents the document title.

| | |
|---|---|
| **print_identifier** | **Value: type string** |

The ISBN of the document. This data was not always provided in the metadata table.

| | |
|---|---|
| **online_identifier** | **NA** |

The value of this field was always empty in the metadata table.

| | |
|---|---|
| **date_first_issue_online** | **Value: type 4-digit number** |

Publication year of the document.

| | |
|---|---|
| **num_first_vol_online** | **NA** |

The value of this field was always empty in the metadata table.

| **num_first_issue_online** | **NA** |
|---|---|

The value of this field was always empty in the metadata table.

| **date_last_issue_online** | **NA** |
|---|---|

The value of this field was always empty in the metadata table.

| **num_last_vol_online** | **NA** |
|---|---|

The value of this field was always empty in the metadata table.

| **num_last_issue_online** | **NA** |
|---|---|

The value of this field was always empty in the metadata table.

| **title_url** | **Value: type string** |
|---|---|

URL to document view in PDF format.

| **first_author** | **NA** |
|---|---|

The value of this field was always empty in the metadata table.

| **title_id** | **Value: type string** |
|---|---|

A string used as a document identifier by Arcanum.

| **embargo_info** | **NA** |
|---|---|

The value of this field was always empty in the metadata table.

| **coverage_depth** | **Value: type string, predefined value:** *fulltext* |
|---|---|

The value of this field is always *fulltext*.

| | |
|---|---|
| **coverage_notes** | **NA** |

The value of this field was always empty in the metadata table.

| | |
|---|---|
| **publisher_name** | **Value: type string** |

A string that represents the publisher name.

| | |
|---|---|
| **metadata_url** | **Value: type string** |

URL to the metadata record. This data was not always provided in the metadata table.

### 3.1.3 The REAL-J metadata schema

No metadata description was available for the documents downloaded from the REAL-J repository. As described in section 3.2.3, the metadata was manually extracted from the source documents and the website.

## 3.2. Mapping with the common metadata schema (described in section 8) (T 5.2)

The Hungarian CURLICAT corpus provides all the obligatory fields of the common metadata schema (values are *always* provided):

- *Identifier*
- *Language (*constant value: ‘hu’*)*
- *PublicationDate*
- *DocumentTitle*
- *ArticleTitle*
- *Type*
- *Source*
- *Domain* (‘Culture’, ‘Economy’, ‘Science’ or ‘Social issues’)
- *Licence*
- *Number of sentences, words, punctuation marks and tokens*

Furthermore, the corpus includes the following optional metadata field of the common schema:

- *Author* (It can be N/A.)

Finally, the following local fields (specific for the Hungarian CURLICAT corpus) were added to the metadata schema:

- *Editor*: A string representing one or multiple proper names, the editor(s) of a collection of works. It can be N/A.
- *RespName*: A string representing one or multiple proper names. It refers to the persons or organizations that had any role in the distribution of the *source* data (not the CURLICAT corpus). It can be N/A.

The fact that the value of the field *Language* could always be set to 'hu' is due to the Hungarian text processing pipeline which included a language identification step. Sentences not identified as Hungarian sentences were removed from the corpus.

The value of *Number of sentences, words, punctuation marks and tokens* was always calculated automatically based on the output of the text processing pipeline.

The metadata had to be obtained in different ways depending on the source. Details are provided in sections 3.2.1-3.2.3.

### 3.2.1 Mapping the HNC metadata to the CURLICAT metadata schema

The collection of Wikipedia texts was not split into articles. It was added to the Hungarian CURLICAT corpus as a single document.

Table 6 specifies how the HNC metadata elements were mapped to CURLICAT metadata fields. As there can be multiple elements with the same name in the HNC XML header, the full path is provided to the relevant element from the top of the hierarchy to descendants. For example, *cesHeader/fileDesc* points to the *fileDesc* element that is a child of the root.

Each CURLICAT metadata field was filled in manually unless specified otherwise in the table. The HNC metadata elements missing from the table were ignored.

| HNC | CURLICAT | Details |
|---|---|---|
| - | *Identifier* | Generated automatically. |
| - | *Language* | Set to 'hu'. |
| *cesHeader/fileDesc/sourceDesc/biblFull/ publicationStmt/pubDate* | *PublicationDate* | Only years were specified. |
| *cesHeader/fileDesc/sourceDesc/biblFull/* | *DocumentTitle* | Copied directly from the source |

| | | |
|---|---|---|
| *titleStmt/h.title* | | metadata. |
| - | *ArticleTitle* | N/A as the Wikipedia collection was not split into articles. |
| - | *Type* | Set to ‘Wikipedia articles’. |
| - | *Source* | Set to ‘Wikipedia’. |
| - | *Domain* | Set to ‘Science’. |
| *cesHeader/fileDesc/publicationStmt/ availability* | *Licence* | The description from the HNC metadata was replaced with the proper licence specification. |
| - | *Number of sentences, words, punctuation marks and tokens* | Calculated automatically after the text was processed. |
| - | *Author* | N/A |
| - | *Editor* | N/A |
| *cesHeader/fileDesc/titleStmt/respStmt/ respName* | *RespName* | The names of the project leaders were selected. The roles were specified by the *cesHeader/fileDesc/titleStmt/ respStmt/respType* elements. |

**Table 6:** The mapping between the HNC XML header elements and the CURLICAT metadata fields.

### 3.2.2 Mapping the Arcanum metadata to the CURLICAT metadata schema

Table 7 specifies how the Arcanum metadata fields were mapped to CURLICAT metadata fields. The Arcanum metadata fields missing from the table were ignored.

| Arcanum | Curlicat | Details |
|---|---|---|

| | | |
|---|---|---|
| - | *Identifier* | Generated automatically. |
| - | *Language* | Set to 'hu'. |
| *date_first_issue_online* | *PublicationDate* | Copied directly from the source metadata. |
| *publication_title* | *DocumentTitle* | Copied directly from the source metadata. |
| *publication_title* | *ArticleTitle* | The same as *DocumentTitle* if the document is a monograph. Set to N/A if it is a collection of works. |
| - | *Type* | Set to 'book'. |
| *publisher_name* | *Source* | Copied directly from the source metadata. |
| - | *Domain* | Set manually. |
| - | *Licence* | Set manually based on the agreement with Arcanum and the CURLICAT project requirements. |
| - | *Number of sentences, words, punctuation marks and tokens* | Calculated automatically after the text was processed. |
| *title_url* | *Author* | The title_url prefix was always the lowercased author or editor name without diacritics. The correct proper names (with diacritics) were restored by a rule-based system. The results were checked manually. |

| | | |
|---|---|---|
| *title_url* | *Editor* | The title_url prefix was always the lowercased author or editor name without diacritics. The substring ‘szerk.’ (ed.) distinguished editors from authors. The correct proper names (with diacritics) were restored by a rule-based system. The results were checked manually. |
| - | *RespName* | N/A |

**Table 7:** The mapping between the Arcanum and the CURLICAT metadata fields.

### 3.2.3 Mapping the REAL-J metadata to the CURLICAT metadata schema

In the case of the REAL-J repository, no metadata table was acquired from the source. Consequently, the metadata was collected manually from the source documents and from the website of the repository.

However, some of the fields could be filled in with constant values; these are given in Table 8.

| **CURLICAT metadata field** | **Value** |
|---|---|
| *Language* | hu |
| Type | journal |
| Source | REAL-J |

**Table 8:** CURLICAT metadata fields filled in with constant values for documents from REAL-J.

Similarly to the documents from HNC and Arcanum, the *Identifier* was generated automatically for each document. The quantitative features (*number of sentences, words, punctuation marks and tokens*) were always calculated automatically after processing the source texts.

### 3.3. Metadata Validation Activities (T 5.3)

The metadata mapping described in section 3.2. resulted in 3 TSV metadata tables (one for each source) with the CURLICAT metadata fields. For each table, a set of regular expressions was defined to check the values in the fields. Creating separate sets of regular expressions for different tables was necessary as the sets of possible metadata field values could vary depending on the source. For example, the value of the field *Source* could be either ‘Akadémiai Kiadó’ or `Osiris Kiadó` for the documents acquired from Arcanum but the same field was allowed to take arbitrary string values for the documents collected from REAL-J. When the regular expressions did not match the value of a field, corrections were made manually.

Another metadata validation step involved checking the *Domain* field. When the metadata tables were compiled, this field was filled in with the values assigned to the documents by a single annotator. Afterwards, the documents were classified into the four possible domains by another annotator as well. The annotations were compared and a third annotator made the final decision in case of disagreement. The metadata tables were updated correspondingly.

Finally, the three metadata tables were merged, resulting in a table with 443 records. (The low number of records is due to the fact that the Hungarian CURLICAT corpus consists of large documents: books, a collection of Wikipedia articles, journal issues.) It was used to automatically generate the headers of the CONLL files which contained the processed and analysed texts of the CURLICAT corpus.

## 4. The Polish metadata description (T5.2, T5.3)

### 4.1. Original metadata description (T5.2)

The Polish CURLICAT corpus consists of texts (mainly abstracts, titles and fragments of full texts extracted from articles in PDF format) from the Library of Science (https://bibliotekanauki.pl/), a platform providing open access to full texts of articles published in Polish scientific journals and full texts of selected scientific books together with rich bibliographic metadata.

The original fields, acquired over the programmatic interface endpoints provided by the Library of Science platform in JSON format, were imported (almost one-to-one) into a relational database using a modified version of the collector tool (http://git.nlp.ipipan.waw.pl/Marcell/collector/tree/curlicat). Available fields and their possible values are described below.

| | |
|---|---|
| **language** | **Value: 3-letters language code** |

A string that represents the full text language.

| | |
|---|---|
| **id** | **Value: positive integer** |

Original document id from the Library of Science platform.

| | |
|---|---|
| **mainTitle** | **Subfields:**<br>**language:** 3-letters language code<br>**text:** title text |

Main title of the document with language code.

| | |
|---|---|
| **mainTitleTranslation(s)** | **Subfields:**<br>**language:** 3-letters language code |

| | |
|---|---|
| | **text:** title text |

Possible translations of the main title. Can occur multiple times for a single document.

| | |
|---|---|
| **mainAbstract** | **Subfields:**<br>**language:** 3-letters language code<br>**text:** abstract text |

Main abstract of the document with language code.

| | |
|---|---|
| **abstractTranslation(s)** | **Subfields:**<br>**language:** 3-letters language code<br>**text:** abstract text |

Possible translations of the main abstract. Can occur multiple times for a single document.

| | |
|---|---|
| **date(s)** | **Subfields:**<br>**type: PUBLISHED**, **ACCEPTED, RECEIVED** or **RELEASED_ONLINE**<br>**date:**<br>**year:** integer<br>**month:** integer<br>**day:** integer |

Possible dates associated with the document. Can occur multiple times for a single document. **Type** represents if the date refers to publication, acceptance, receipt or online release. **Date** consists at least of **year** value. If **month** or **day** is unknown they are filled with 0 value.

| | |
|---|---|
| **keyword(s)** | **Subfields:**<br>**language:** 3-letters language code<br>**text:** keyword text |

Keyword associated with the document. Can occur multiple times for a single document.

| **contributor(s)** | **Subfields:**<br>**firsName:** first name string<br>**lastName:** last name string<br>**role: AUTHOR**, **REVIEWER** or **TRANSLATOR**<br>**biography:** biography string<br>**email:** email address string<br>**orcid:** orcid number<br>**publicationInstitutions:** list of institutions associated with the contributor |
|---|---|

Contributor associated with the document. Can occur multiple times for a single document. **Role** represents if the contributor is one of the authors, reviewers or translators of the article.

| **bibEntry(ies)** | **Value: string** |
|---|---|

Bibliographic entry associated with the document. Can occur multiple times for a single document.

| **scientificDiscipline(s)** | **Subfields:**<br>**id:** original id of the discipline in the Library of Science platform<br>**namePl:** discipline name in Polish language<br>**nameEn:** discipline name in English language |
|---|---|

Scientific discipline associated with the document. Can occur multiple times for a single document. Scientific disciplines are grouped by ScientificFields in a way presented in the table below.

| **scientificField:**<br>**id:** original id of the scientific field in the Library of Science platform<br>**namePl:** field name in Polish language<br>**nameEn:** field name in English language | **scientificDisciplines [nameEn]** |
|---|---|
| **id:** 1<br>**namePl:** Nauki humanistyczne<br>**nameEn:** Humanities | archeology; art sciences; history; linguistics; literature; philosophy; science about culture and religion |
| **id:** 2<br>**namePl:** Nauki inżynieryjno-techniczne<br>**nameEn:** Engineering and technical sciences | architecture and urban planning; automation, electronics and electrical engineering; biomedical engineering; chemical engineering; civil engineering and transport; environmental engineering, mining and energy; material engineering; mechanical engineering; technical IT and telecommunications |
| **id:** 3<br>**namePl:** Nauki medyczne i o zdrowiu<br>**nameEn:** Medical and health sciences | health sciences; medical sciences; pharmaceutical sciences; physical culture sciences |
| **id:** 4<br>**namePl:** Nauki rolnicze<br>**nameEn:** Agricultural sciences | agriculture and gardening; food and nutrition technology; forestry sciences; veterinary medicine; zootechnics and fishing |
| **id:** 5<br>**namePl:** Nauki społeczne<br>**nameEn:** Social sciences | economics and finance; education; legal sciences; management and quality; politics and administration; psychology; security sciences; social communication and media; socio-economic geography and spatial economy; sociological sciences; the canonic law |
| **id:** 6 | astronomy; biological sciences; chemical |

| | |
|---|---|
| **namePl:** Nauki ścisłe i przyrodnicze<br>**nameEn:** Exact and natural sciences | sciences; Earth and the environment sciences; informatics; mathematics; physical sciences |
| **id:** 7<br>**namePl:** Nauki teologiczne<br>**nameEn:** Theological sciences | theological sciences |
| **id:** 8<br>**namePl:** Sztuka<br>**nameEn:** Art | fine arts and conservation of works of art |

| **remarks** | **Value: string** |
|---|---|

A string that represents additional remarks to the document. For example about source of financing.

| **fullTextFile(s)** | **Subfields:**<br>**key:** relative link to the document source file<br>**format:** most likely 'PDF' string |
|---|---|

Most likely a single field within document definition with information about relative path to source PDF with full text.

| **license** | **Value: string, null for unknown** |
|---|---|

A string that defines the license of full text. For example: *CC BY - Creative Commons Uznanie Autorstwa 4.0*

| | |
|---|---|
| **DOI** | **Value: string** |

A string representing the *digital object identifier* of the document.

| | |
|---|---|
| **pageRange** | **Value: string** |

A string representing the page range of the document in the journal where it was published.

| | |
|---|---|
| **type** | **Value: ORIGINAL_SCIENTIFIC_TEXT** or **REVIEW** |

A string further specifying the type of the document. **ORIGINAL_SCIENTIFIC_TEXT** for scientific texts and **REVIEW** for scientific texts reviews.

| | |
|---|---|
| **issueInfo** | **Subfields:**<br>**id:** original id of the issue in the Library of Science platform<br>**year:** issue year<br>**volume:** issue volume<br>**number:** issue number<br>**coverId:** issue cover id<br>**journalInfo:** see **journalInfo** field description below |

Groups information about publishing company, journal and issue where document was published.

| | |
|---|---|
| **journalInfo** | **Subfields:**<br>**id:** original id of the journal in the Library of Science platform<br>**title:** journal title<br>**issn:** International Standard Serial |

| | |
|---|---|
| | Number of the journal<br>**eissn:** electronic ISSN of the journal<br>**publishingCompanyInfo:** see **publishingCompanyInfo** field description below |

| | |
|---|---|
| **publishingCompanyInfo** | **Subfields:**<br>**id:** original id of the publishing company in the Library of Science platform<br>**name:** name of the publishing company |

### 4.2. Mapping with the common metadata schema (described in Section 8) (T5.2)

#### Metadata linking and enhancement

The headers of the acquired article files already contained rich metadata which were mapped to a relational model used to manage the Polish CURLICAT dataset.

The following set of core obligatory metadata is used for all documents:

- id – unique identifier of the document within all the corpora, following CoNLL-U conventions; created using **pl** language code, internal source marker (**bn** - the Library of Science) and original document **id** (from the Library of Science platform); ex. pl-bn-476093
- Language – the ISO 639-1 language code of the polish sub-corpora ('pl')
- PublicationDate – the primary date of the document, the publication date in the ISO 8601 format, with accuracy given by source metadata (at least the year)
- DocumentTitle – and informative, human readable title (name) of the document, in the original language (**mainTitle** marked with POL language or **mainTitleTranslation** marked with POL language)
- Type – type of the document, in English, based on document content "paper" (fragments extracted from full article PDF) or "abstract" (only abstract and title of the article)
- Source – title of the journal, **title** subfield from the **journalInfo**

- Domain – CURLICAT domain mapped from sets of **scientificDiscipline(s)** available in the source corpora (detailed mapping described in Table 9 below to local metadata fields description)
- No_of_sentences, No_of_words, No_of_punctuation, No_of_tokens – the total number of sentences, words, punctuation marks and tokens (words + punctuation marks) in the document.

The optional metadata:

- Author – list of authors separated by the pipe character (“|”); **contributors** with **role AUTHOR**

- *SourceType:* the type of organization that published the source document, always a “Publishing House”
- Url – URL of the source document; concatenation of link to the full texts API (https://bibliotekanauki.pl/api/full-texts/) and **fullTextFile key**
- Keywords – separated by the pipe character (“|”) and/or commas, in the polish language (**keywords** marked with POL language)
- Style – always “scientific”
- Subdomain – list of **scientificDiscipline(s)** separated by the pipe character (“|”) in the **enName** translation

The local metadata fields included in the Polish dataset:

- ScientificField – list of **scientificField(s)**, separated by the pipe character (“|”) in the **enName** translation
- EnTitle – title in English language (**mainTitle** marked with ENG language or **mainTitleTranslation** marked with ENG language)
- EnKeywords – keywords separated by the pipe character (“|”) and/or commas, in English language (**keywords** marked with ENG language)
- EnAbstract – article abstract in English language (**mainAbstract** marked with ENG language or **abstractTranslation** marked with ENG language)
- PublishingCompany – publishing company **name** from the **publishingCompanyInfo** field
- IssueYear – **year** from the **issueInfo** field
- IssueVolume – **volume** from the **issueInfo** field
- IssueNumber – **number** from the **issueInfo** field
- PageRange – **pageRange** string
- License – full text **license**
- Reviewer – separated by the pipe character (“|”); **contributors** with **role REVIEWER**
- Translator – separated by the pipe character (“|”); **contributors** with **role TRANSLATOR**
- OriginalType – the **type** of the document from source corpora; here: **ORIGINAL_SCIENTIFIC_TEXT** or **REVIEW**.

Sets of **scientificDiscipline(s)** were mapped to CURLICAT domain fields as bellow:

| CURLICAT Domain | scientificDiscipline(s) |
|---|---|
| Culture | archeology |
| Culture | archeology \| art sciences |
| Culture | archeology \| art sciences \| history |
| Culture | archeology \| art sciences \| history \| literature \| science about culture and religion |
| Culture | archeology \| art sciences \| history \| science about culture and religion |
| Culture | archeology \| history |
| Culture | architecture and urban planning \| art sciences |
| Culture | art sciences |
| Culture | art sciences \| history |
| Culture | art sciences \| history \| linguistics \| literature \| science about culture and religion |
| Culture | art sciences \| history \| literature \| philosophy \| science about culture and religion |
| Culture | art sciences \| history \| literature \| science about culture and religion |
| Culture | art sciences \| history \| science about culture and religion |
| Culture | art sciences \| history \| social communication and media |
| Culture | art sciences \| legal sciences |
| Culture | art sciences \| linguistics \| literature \| science about culture and religion |
| Culture | art sciences \| literature |
| Culture | art sciences \| literature \| science about culture and religion |
| Culture | art sciences \| science about culture and religion |
| Culture | history |
| Culture | history \| legal sciences |
| Culture | history \| legal sciences \| management and quality \| social communication and media \| sociological sciences |
| Culture | history \| legal sciences \| philosophy \| science about culture and religion \| sociological sciences \| the canonic law |
| Culture | history \| legal sciences \| sociological sciences |
| Culture | history \| linguistics |
| Culture | history \| linguistics \| literature |
| Culture | history \| linguistics \| literature \| science about culture and religion \| sociological sciences |
| Culture | history \| linguistics \| literature \| social communication and media |
| Culture | history \| linguistics \| science about culture and religion \| sociological sciences |
| Culture | history \| literature |
| Culture | history \| literature \| philosophy |
| Culture | history \| literature \| science about culture and religion |
| Culture | history \| literature \| science about culture and religion \| sociological sciences |
| Culture | history \| philosophy |
| Culture | history \| philosophy \| science about culture and religion |
| Culture | history \| philosophy \| social communication and media |
| Culture | history \| philosophy \| the canonic law |

| | |
|---|---|
| Culture | history \| science about culture and religion |
| Culture | history \| science about culture and religion \| social communication and media |
| Culture | history \| science about culture and religion \| sociological sciences |
| Culture | history \| social communication and media |
| Culture | legal sciences \| philosophy |
| Culture | linguistics |
| Culture | linguistics \| literature |
| Culture | linguistics \| literature \| science about culture and religion |
| Culture | linguistics \| literature \| social communication and media |
| Culture | linguistics \| philosophy |
| Culture | linguistics \| philosophy \| psychology |
| Culture | linguistics \| philosophy \| science about culture and religion \| social communication and media \| sociological sciences |
| Culture | linguistics \| science about culture and religion \| sociological sciences |
| Culture | linguistics \| social communication and media |
| Culture | linguistics \| social communication and media \| sociological sciences |
| Culture | literature |
| Culture | literature \| philosophy \| psychology \| sociological sciences |
| Culture | literature \| science about culture and religion |
| Culture | literature \| science about culture and religion \| sociological sciences |
| Culture | management and quality \| philosophy \| sociological sciences |
| Culture | management and quality \| science about culture and religion |
| Culture | philosophy |
| Culture | philosophy \| psychology |
| Culture | philosophy \| science about culture and religion |
| Culture | philosophy \| social communication and media \| sociological sciences |
| Culture | philosophy \| sociological sciences |
| Culture | science about culture and religion |
| Culture | science about culture and religion \| social communication and media |
| Culture | science about culture and religion \| social communication and media \| sociological sciences |
| Culture | science about culture and religion \| sociological sciences |
| Economy | Earth and the environment sciences \| civil engineering and transport \| environmental engineering, mining and energy \| socio-economic geography and spatial economy |
| Economy | art sciences \| economics and finance \| history \| linguistics \| literature \| philosophy \| science about culture and religion |
| Economy | economics and finance |
| Economy | economics and finance \| history \| sociological sciences |
| Economy | economics and finance \| legal sciences |
| Economy | economics and finance \| legal sciences \| management and quality |
| Economy | economics and finance \| legal sciences \| sociological sciences |
| Economy | economics and finance \| management and quality |
| Economy | economics and finance \| management and quality \| sociological sciences |

| | |
|---|---|
| Economy | economics and finance \| science about culture and religion |
| Economy | economics and finance \| social communication and media |
| Economy | economics and finance \| sociological sciences |
| Economy | socio-economic geography and spatial economy |
| Education | art sciences \| education |
| Education | art sciences \| education \| history \| linguistics \| literature \| psychology \| science about culture and religion \| sociological sciences |
| Education | art sciences \| education \| history \| psychology \| social communication and media \| sociological sciences |
| Education | art sciences \| education \| linguistics \| literature \| psychology |
| Education | art sciences \| education \| literature \| science about culture and religion |
| Education | economics and finance \| education \| history \| politics and administration \| sociological sciences |
| Education | economics and finance \| education \| literature \| management and quality \| philosophy \| politics and administration \| social communication and media \| sociological sciences |
| Education | education |
| Education | education \| history \| literature |
| Education | education \| history \| philosophy \| sociological sciences |
| Education | education \| history \| philosophy \| the canonic law |
| Education | education \| history \| psychology |
| Education | education \| history \| sociological sciences |
| Education | education \| legal sciences \| politics and administration \| security sciences \| sociological sciences |
| Education | education \| legal sciences \| sociological sciences |
| Education | education \| linguistics |
| Education | education \| linguistics \| literature \| politics and administration \| science about culture and religion |
| Education | education \| linguistics \| literature \| social communication and media |
| Education | education \| linguistics \| sociological sciences |
| Education | education \| management and quality |
| Education | education \| management and quality \| science about culture and religion \| security sciences \| sociological sciences |
| Education | education \| politics and administration \| psychology |
| Education | education \| psychology |
| Education | education \| psychology \| sociological sciences |
| Education | education \| science about culture and religion \| social communication and media |
| Education | education \| science about culture and religion \| social communication and media \| sociological sciences |
| Education | education \| social communication and media |
| Education | education \| sociological sciences |
| Health | Earth and the environment sciences \| health sciences \| medical sciences |
| Health | biomedical engineering |
| Health | biomedical engineering \| health sciences \| medical sciences \| pharmaceutical sciences |
| Health | biomedical engineering \| medical sciences |
| Health | biomedical engineering \| medical sciences \| pharmaceutical sciences |
| Health | economics and finance \| literature \| management and quality \| physical culture sciences \| politics and administration \| science about culture and religion \| sociological sciences |

| | |
|---|---|
| Health | health sciences |
| Health | health sciences \| medical sciences |
| Health | health sciences \| physical culture sciences |
| Health | history \| medical sciences |
| Health | legal sciences \| medical sciences |
| Health | linguistics \| medical sciences \| psychology |
| Health | management and quality \| physical culture sciences |
| Health | management and quality \| physical culture sciences \| sociological sciences |
| Health | medical sciences |
| Health | medical sciences \| pharmaceutical sciences |
| Health | medical sciences \| pharmaceutical sciences \| psychology |
| Health | medical sciences \| physical culture sciences |
| Health | pharmaceutical sciences |
| Health | physical culture sciences |
| Nature | Earth and the environment sciences \| biological sciences |
| Nature | agriculture and gardening |
| Nature | biological sciences |
| Nature | forestry sciences |
| Nature | veterinary medicine |
| Nature | zootechnics and fishing |
| Politics | archeology \| history \| philosophy \| politics and administration \| sociological sciences |
| Politics | art sciences \| history \| politics and administration |
| Politics | art sciences \| literature \| politics and administration \| sociological sciences |
| Politics | economics and finance \| legal sciences \| management and quality \| politics and administration |
| Politics | economics and finance \| legal sciences \| politics and administration |
| Politics | economics and finance \| management and quality \| politics and administration |
| Politics | economics and finance \| management and quality \| politics and administration \| security sciences |
| Politics | economics and finance \| management and quality \| politics and administration \| sociological sciences |
| Politics | economics and finance \| philosophy \| politics and administration \| sociological sciences |
| Politics | economics and finance \| politics and administration |
| Politics | economics and finance \| politics and administration \| sociological sciences |
| Politics | history \| philosophy \| politics and administration \| psychology \| sociological sciences |
| Politics | history \| politics and administration \| science about culture and religion \| security sciences \| social communication and media |
| Politics | history \| politics and administration \| science about culture and religion \| security sciences \| sociological sciences |
| Politics | history \| politics and administration \| science about culture and religion \| sociological sciences |
| Politics | history \| politics and administration \| security sciences |
| Politics | history \| politics and administration \| sociological sciences |
| Politics | legal sciences \| management and quality \| politics and administration |
| Politics | legal sciences \| management and quality \| politics and administration \| security sciences |

| | |
|---|---|
| Politics | legal sciences \| management and quality \| politics and administration \| sociological sciences |
| Politics | legal sciences \| politics and administration |
| Politics | legal sciences \| politics and administration \| security sciences |
| Politics | legal sciences \| politics and administration \| security sciences \| sociological sciences |
| Politics | legal sciences \| politics and administration \| sociological sciences |
| Politics | management and quality \| politics and administration \| security sciences |
| Politics | politics and administration |
| Politics | politics and administration \| psychology \| sociological sciences |
| Politics | politics and administration \| science about culture and religion |
| Politics | politics and administration \| science about culture and religion \| sociological sciences |
| Politics | politics and administration \| security sciences |
| Politics | politics and administration \| security sciences \| social communication and media |
| Politics | politics and administration \| security sciences \| social communication and media \| sociological sciences |
| Politics | politics and administration \| social communication and media |
| Politics | politics and administration \| sociological sciences |
| Science | Earth and the environment sciences |
| Science | architecture and urban planning |
| Science | automation, electronics and electrical engineering |
| Science | chemical engineering |
| Science | chemical sciences |
| Science | civil engineering and transport |
| Science | environmental engineering, mining and energy |
| Science | food and nutrition technology |
| Science | informatics |
| Science | material engineering |
| Science | mathematics |
| Science | mathematics \| technical IT and telecommunications |
| Science | mechanical engineering |
| Science | physical sciences |
| Science | technical IT and telecommunications |
| Social issues | legal sciences |
| Social issues | legal sciences \| management and quality |
| Social issues | legal sciences \| security sciences |
| Social issues | legal sciences \| the canonic law |
| Social issues | management and quality |
| Social issues | management and quality \| social communication and media |
| Social issues | psychology |
| Social issues | security sciences |
| Social issues | security sciences \| sociological sciences |
| Social issues | social communication and media |

| | |
|---|---|
| Social issues | social communication and media \| sociological sciences |
| Social issues | sociological sciences |
| Social issues | the canonic law |

**Table 9**: Mapping of scientific disciplines to CURLICAT domains

### 4.3. Metadata validation activities (T 5.3)

Metadata validation in the Polish pipeline was done at three points:

- when importing metadata from JSONlines file (acquired over the programmatic interface provided by the Library of Science platform) into a relational database using a collector tool (http://git.nlp.ipipan.waw.pl/Marcell/collector/tree/curlicat)
- using collector mechanisms and during exporting metadata and documents content from collector tool to the CONLL-U+ format
- and finally cleaning and mapping metadata to the final version of corpora by additional script.

During importing metadata to the collector tool all of the source metadata fields were:

- verified in terms of their value types (are they strings, integers, lists or more complex objects) and mapped to relational database structure in the most simplistic way (ex. key, value pairs)
- checked if they cover initially defined obligatory CURLICAT metadata fields such as “id”, “title”, “date”, “language”, etc.
- checked in terms of their occurrence (are they available for all documents, most of the documents, rather appear occasionally or are always empty)
- dates were converted from **year**, **month**, **day** fields to ISO 8601 format.

Using collector mechanisms and during exporting metadata to CONLL-U+ format additional checks were made, including:

- licence checking, only texts with CC-BY and CC-BY-SA licences can be distributed as full text fragments, otherwise title and abstract are taken as a document content
- titles, title translations, abstract and abstract translations marked as “POL” in the Library of Science were verified (using Google’s Compact Language Detector v3) if they are actually in Polish not only marked as Polish
- metadata fields with values forming lists were exported using “|” delimiter

- missing fields or ones with empty values were omitted.

Additional script was written to check, clean and prepare final corpora version, its role is to:

- remove documents without known title
- remove metadata lines with UNKNOWN value
- map initially defined metadata field names to ones harmonised across all language corpora
- ensure that metadata values are valid UTF-8 strings
- collapse multiple spaces in the metadata values
- remove leading and trailing whitespaces from the metadata values
- check for completeness of all the obligatory metadata
- map **scientificDiscipline(s)** to CURLICAT domains.

# 5. The Romanian metadata description (T5.2, T5.3)

## 5.1. Original metadata description (T5.2)

The Romanian National Corpus (CoRoLa) metadata are stored as CMDI-based XML files, separated from the text files, but in files sharing the same name with different (.xml vs. .txt) extensions.

The metadata fields available in CoRoLa are the following:

| **DocumentTitle** | **Value: type string** |
|---|---|

The values of this field can be strings representing the title of the book, journal, volume, etc. contained by the associated original text document. The value can be "-" when the information is not available.

| **ArticleTitle** | **Value: type string** |
|---|---|

The values of this field can be strings representing the title of the chapter or article when the original text document was a collection of texts. DocumentTitle and ArticleTitle can have the same value in case of single text original documents (books, online newspaper articles and blogspots, wikipedia articles, etc.). The value can be "-" when the information is not available.

| **AuthorName** | **Value: type string** |
|---|---|

The values of this field can be strings representing the name of the author/authors of the text in the CoRoLa document (the authors of the chapters or articles coming from collections, or of the whole original document if it was a single text document). The value can be "-" when the information is not available.

| **PublicationDate** | **Value: type 4-digit number** |
|---|---|

The values of this field can be 4-digit numbers representing the year of publication for the corresponding text in the document (book, journal, volume, newspaper article, etc.).

| | |
|---|---|
| **Source** | **Value: type string, possible predefined values:**<br>*Journal, Publishing House, Blog, Website, Other* |

This field encodes the type of the source that provided the document: like a journal, a publishing house, a blog, etc. The values are restricted to the list: *Journal, Publishing House, Blog, Website, Other.*

| | |
|---|---|
| **SourceName** | **Value: type string** |

The values of this field can be strings representing the name of the source: the name of the journal, publishing house, etc.; the web address of the news site or blog. The name of the source always has to be written in the same form, avoiding abbreviation, changing capitalization or any other variation in the value. E.g, the name of the publishing house will not be accompanied by any description: "Editura Polirom" ("Polirom Publishing House") is not correct, "Polirom" is correct.

| | |
|---|---|
| **TranslatorName** | **Value: type string** |

If the text in the document is a translation, and not an original Romanian text, the value of this field will be the name of the translator. Otherwise, the value will be "-".

| | |
|---|---|
| **Medium** | **Value: type string, predefined value:** *written* |

This attribute's value is always *„written"*. In the CURLICAT's context, it can be superfluous.

| | |
|---|---|
| **DocumentType** | **Value: type string, predefined values:** *BlogPost, Booklet, Book, inBook, inCollection, inProceedings, Manual, Newspaper article, Other, Proceedings, Techreport, Unpublished* |

This field describes the type of the document, whether being a blogpost or newspaper article, a whole proceedings volume or only an article in a proceedings volume, a book or a chapter in a book, etc. The values are predefined string values.

| | |
|---|---|
| **DocumentTextStyle** | **Value: type string, predefined values:** *Journalistic, Science, Administrative, Imaginative, Memoirs, Law* |

This field encodes a literary style classification of the documents in most corpora. The values are of type string and there is a predefined set of possible values: *Journalistic, Science, Administrative, Imaginative, Memoirs, and Law.*

| | |
|---|---|
| **DocumentTextDomain** | **Value: type string, predefined values:** *Arts and Culture, Nature, Society, Science, Other, "-".* |

The field describes a large domain for the document. This attribute has to be correlated with the DocumentTextSubDomain attribute (see below), with clear restrictions on domain and subdomain associations. The values of DocumentTextDomain are strings in the set: *Arts and Culture, Nature, Society, Science, Other, "-".* The association restriction between DocumentTextDomain values and DocumentTextSubDomain values are presented in Table 10 below: e.g, if DocumentTextDomain = Arts and Culture, DocumentTextSubDomain value can only be a string in the set of predefined values: *Music, Literature, Art History, Folklore, Film, Architecture, Sculpture, Painting and Drawing, Design, Fashion, Theatre, Dance, Other.*

Values of DocumentTextDomain and DocumentTextSubDomain are always "-" when DocumentTextStyle is *Imaginative*: belletristic, poetic and other types of imaginative texts cannot be classified into a specific domain or subdomain.

| **DocumentTextDomain** | **DocumentTextSubDomain** |
|---|---|

| | **Value: type string** |
|---|---|
| **Arts and Culture** | **predefined values:** *Music, Literature, Art History, Folklore, Film, Architecture, Sculpture, Painting and Drawing, Design, Fashion, Theatre, Dance, Other* |
| **Nature** | **predefined values:** *Environment, Natural Disasters, Universe, Natural Resources, Other* |
| **Society** | **predefined values:** *Politics, Law, Administration, Economy, Army, Health, Sport, Family, Gossip, Social Events, Education, Social Movements, Tourism, Religion, Entertainment, Other* |
| **Science** | **predefined values:** *Mathematics, Informatics, Logics, Standards, Medicine, Archeology, Engineering, Technics/technology, Aeronautics, Agronomy, Metrology, Criminalistics, Constructions, Military Science, Pharmacology, Enology, Pedagogy, Geography, Economy, History, Psychology, Sociology, Ethnology, Anthropology, Religious Studies and Theology, Juridical Sciences, Linguistics, Political Sciences, Philosophy, Philology., Biology, Physics, Astronomy, Chemistry, Other* |

**Table 10:** Classification of domains according to subdomains

| | |
|---|---|
| **SubjectLanguage** | **Value: type string**<br>Predefined value for CoRoLA documents: *Romanian* |

The field encodes the language of the text in the document.

| **ISSN-ISBN** | **Value: type string** |
|---|---|

This field encodes the International Standard Serial Number of the publication of the document. If the information is unavailable, the value of the field is "-". In the case of a digital edition of a book, its ISBN may be eISBN.

| **CollectionDate** | **Value:type  4-digit number** |
|---|---|

This field encodes the year when the associated document was acquired.

### 5.2. Mapping with the common metadata schema (described in section 8) (T 5.2)

To comply with the CURLICAT metadata schema,

1. Some fields were renounced of: Medium, CollectionDate, TranslatorName, AuthorName (the last two for IPR issues);
2. some of this fields were adapted
   - the date type format became ISO 8601 format;
   - The "SubjectLanguage" attribute name became "Language" in the CURLICAT schema;  the string predefined value of this field was replaced with the ISO 639-1 code for Romanian (ro);
3. Information to be described in the new count fields was obtained from the documents, by automatically counting the number of sentences, words, punctuation and tokens;
4. A License field complying with the CURLICAT project requirements concerning corpus distribution was generated;
5. An Identifier field was generated in the format: ro-c-numeric_identifier, where numeric identifier is based on the original file name of the document in CoRoLa;
6. CoRoLa DocumentTextDomain and DocumentTextSubDomain fields were mapped to CURLICAT domain fields as in Table 11 bellow:

| **CURLICAT Domain** | **DocumentTextDomain** | **DocumentTextSubDomain** |
|---|---|---|
| Culture | Arts and Culture | Literature |
| Culture | Arts and Culture | Film |
| Culture | Arts and Culture | Music |

| | | |
|---|---|---|
| Culture | Arts and Culture | Painting and Drawing |
| Culture | Arts and Culture | Other |
| Culture | Arts and Culture | Theatre |
| Culture | Arts and Culture | Sculpture |
| Culture | Arts and Culture | Architecture |
| Culture | Arts and Culture | Dance |
| Culture | Arts and Culture | Design |
| Culture | Arts and Culture | Fashion |
| Culture | Arts and Culture | Art History |
| Culture | Arts and Culture | Folklore |
| Economy | Science | Economy |
| Economy | Society | Economy |
| Education | Science | Pedagogy |
| Education | Society | Education |
| Health | Science | Medicine |
| Health | Science | Pharmacology |
| Health | Society | Health |
| Nature | Nature | Environment |
| Nature | Nature | Natural Disasters |
| Nature | Nature | Other |
| Nature | Nature | Natural Resources |
| Nature | Nature | Universe |
| Nature | Science | Astronomy |
| Nature | Science | Chemistry |
| Nature | Science | Biology |
| Nature | Science | Agronomy |
| Nature | Science | Geography |
| Nature | Society | Tourism |

| | | |
|---|---|---|
| Politics | Science | Political Sciences |
| Politics | Society | Politics |
| Science | Science | History |
| Science | Science | Philosophy |
| Science | Science | Religious Studies and Theology |
| Science | Science | Physics |
| Science | Science | Informatics |
| Science | Science | Archeology |
| Science | Science | Mathematics |
| Science | Science | Other |
| Science | Science | Philology |
| Science | Science | Sociology |
| Science | Science | Anthropology |
| Science | Science | Linguistics |
| Science | Science | Psychology |
| Science | Science | Technics/technology |
| Science | Science | Ethnology |
| Science | Science | Logics |
| Science | Science | Constructions |
| Science | Science | Standards |
| Science | Science | Engineering |
| Science | Science | Aeronautics |
| Science | Science | Metrology |
| Science | Science | Criminalistics |
| Science | Science | Military Science |
| Science | Science | Enology |
| Science | Science | Juridical Sciences |
| Science | Science | Philology |

**Table 11:** Mapping of CoRoLa domains and subdomains to CURLICAT domains

The mapping of the fields between the CoRoLa schema (column 2) and the common schema (column 1), together with the automatically generated or computed fields are presented in Table 12 below:

| **CURLICAT fields** | **CoRoLa fields** | **Other** |
|---|---|---|
| **Obligatory fields** | | |
| Identifier | | Automatically generated |
| Language | SubjectLanguage | |
| Publication | PublicationDate | |
| DocumentTitle | DocumentTitle | |
| ArticleTitle | ArticleTitle | |
| Type | DocumentType | |
| Source | SourceName | |
| Domain | mapping to DocumentTextDomain + DocumentTextSubdomain | |
| No_of_sentences | | Automatically computed |
| No_of_words | | Automatically computed |
| No_of_punctuation | | Automatically computed |
| No_of_tokens | | Automatically computed |
| Licence | | Commonly agreed |
| **Optional Fields** | | |
| Author | | Not available (N/A), due to IPR |
| Url | if available, when SourceName is a url | |
| SourceType | Source | |
| Style | DocumentTextStyle | |
| Keywords | | N/A |
| Subdomain | DocumentTextSubdomain | |
| Issn_isbn_eisbn | ISSN-ISBN | |

**Table 12:** Mapping of metadata fields from CoRoLa to CURLICAT common schema

### 5.3. Metadata validation activities (T5.3)

#### 5.3.1. Technical validation

Technical validation ensures that the metadata are in the correct format. This includes the format of the file the metadata is included in, and the format of the metadata values. When converting the original standoff XML metadata from CoRoLa  to the CONLLU header format adopted in CURLICAT, we follow the presence of the fields within the XML file. New fields like Identifier, No_of_sentences, No_of_words, No_of_punctuation, No_of_tokens or  License are automatically generated and inserted into the positions commonly agreed in the schema, therefore keeping an uniform order of the fields.
If fields are missing or duplicated, the conversion process throws an exception, requiring manual intervention for correcting the supplied metadata file. Value types are checked for each field and values to be selected from predefined lists are also automatically verified.

#### 5.3.2 Semantic validation

Computing value frequency lists for the corpus on all metadata fields allowed for manually verifying and correcting the outliers (values appearing only once or a few times, with the exception of the author and title fields).

To assure the correct document classification according to the CURLICAT domains, an extensive process of manual and automatic validation and correction was applied on the CoRoLa metadata fields Domain and Subdomain (which were then mapped to CURLICAT domains as presented in Table 11). For all the data acquired from private providers (3,042 documents), the validation of the metadata was done completely manually.  For 27,640 documents coming from the Romanian Wikipedia corpus, the classification according to CoRoLa domains and subdomains was done automatically, by mapping Wikipedia document categories which are given at the end of each Wikipedia document. If we use them as directions in the CoRoLa domain/subdomain space, we can build vectors that can point to the right CoRoLa domain and subdomain for a Wikipedia document. An example of a list of categories for the document entitled “Basset Hound” is the following: Hounds, Dog breeds originating in France, Dog breeds originating in England, FCI breeds, Hunting with hounds, Scent hounds.

From each Wikipedia document we automatically extracted the top 50 keywords, using the YAKE keyword extractor (Campos et al., 2020). Its Python 3 implementation can be found at https://pypi.org/project/yake/.  We then constructed a map from the Wikipedia categories of the document to the extracted keywords. Each Wikipedia category is a key in this map and the value is a frequency map of keywords that are associated with the category. As this was a process of correcting/validating already existent metadata values, the 27,640 documents were already classified in the five CoRoLa domains mentioned earlier: *Arts and Culture, Nature,* *iety, Science, Other*. Thus, we built the category to keyword map for each set of documents in a CoRoLa domain. Here’s an example:

**Category:Arme (Weapons)**

| | |
|---|---|
| amiralul | 1 |
| arbalete | 1 |
| arc | 2 |
| arcul | 2 |
| arcului | 1 |
| arcuri | 1 |
| arderea | 1 |
| arderii | 1 |
| arma | 3 |
| armament | 1 |
| armata | 2 |
| armatei | 2 |
| arme | 4 |
| armele | 1 |
| armelor | 3 |
| armura | 1 |
| armă | 6 |
| artileria | 1 |
| artilerie | 1 |
| aruncătoarele | 1 |
| aruncătorul | 3 |
| asalt | 1 |
| ascuțit | 1 |
| autopropulsat | 1 |
| autopropulsată | 2 |

```
avion 1

baionete    1

baliste     1

balistic    1

balistă     1

bastoane    1

blindat     2

calibru     1

calibrul    1

capabil     1

cheiroballistra   1

cuțite      1

explozivă   1

fier  1

foc   3

inamic      1

inamice     2

lamei 1

lamă  2

lansare     1

lansatoare  1

lansatoarele      1

lansator    2

lansează    2

laser 1

luptei      1

luptă 5
```

| | |
|---|---|
| marinei | 1 |
| militare | 2 |
| mitralieră | 1 |
| mortier | 1 |
| **muniție** | **2** |
| **munițiilor** | **2** |
| navă | 1 |
| nazistă | 4 |
| pumnal | 1 |
| pumnale | 1 |
| **pușca** | **2** |
| pușcă | 1 |
| **puștii** | **1** |
| rachete | 2 |
| raza | 1 |
| rus | 1 |
| război | 9 |
| războiul | 2 |
| războiului | 2 |
| sabia | 1 |
| sabie | 1 |
| scară | 1 |
| scurtă | 1 |
| **semișenilat** | **2** |
| **semișenilatul** | **1** |
| senzorilor | 1 |
| senzorul | 1 |

| | |
|---|---|
| sistem | 1 |
| sistemele | 1 |
| spada | 1 |
| spade | 1 |
| spadei | 1 |
| spadă | 1 |
| stilet | 1 |
| stilete | 1 |
| stiletului | 1 |
| subansamblurile | 1 |
| submarină | 1 |
| sulițele | 1 |
| săbii | 1 |
| săgeți | 2 |
| tactica | 1 |
| tanc | 3 |
| tancul | 2 |
| tancului | 2 |
| tancuri | 2 |
| tancurile | 2 |
| teaca | 1 |
| tehnica | 1 |
| telescopice | 1 |
| tir | 1 |
| tirul | 2 |
| torpila | 2 |
| torpile | 1 |

```
torpilele   1
```

We also build the inverse map, from the keyword to the possible Wikipedia categories. We name the two hashmaps the `wiki2word` and `word2wiki` maps respectively.

The main idea of finding the CoRoLa domain and subdomain of a Wikipedia document is to find the largest keyword subset of the document that identifies a given Wikipedia category that, in turn, maps to a CoRoLa category (we will call a CoRoLa domain/subdomain a CoRoLa category). Because we have many more Wikipedia categories than CoRoLa categories (Table 13 shows the number of Wikipedia categories per CoRola domain), we have a many to one mapping. For each CoRoLa domain, we manually mapped a subset of frequent Wikipedia categories to CoRoLa categories.

| **CURLICAT domain** | **Wikipedia category count** |
|---|---|
| Arts and Culture | 9,509 |
| Nature | 62 |
| Science | 24,101 |
| Society | 2,498 |

**Table 13**: Wikipedia category counts per CoRoLa domain

The mapping algorithm executes the following steps:

1. The input document is tokenized so that keywords can be extracted.
2. A 0-initialised vector with the number of Wikipedia categories is created for each keyword. For the Wikipedia category present in the document, the corresponding cell value in this vector is the normalised frequency of the Wikipedia category for the given keyword from the `word2wiki` hashmap, described previously.
3. We automatically cluster all keywords vectors using the KMeans algorithm from the scikit-learn Python package. The number of clusters is automatically selected using the silhouette analysis score (see this article[6] for a description of the method).

[6] https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html?highlight=silhouette

4. For each found cluster:
    a. For each Wikipedia category that was manually mapped to a CURLICAT category,
    b. For the intersection of the set of keywords in the cluster and the set of keywords corresponding to the category,
    c. Each keyword in the intersection contributes a score for the CURLICAT category by adding a value equal to $word2wiki[kw][c] * TF(kw) / DF(kw)$, where $word2wiki[kw][c]$ is the frequency of the keyword in the given Wikipedia category, $TF(kw)$ is the frequency of the keyword in all Wikipedia documents and $DF(kw)$ is the number of documents in which the keyword appears.
    d. The CoRoLa category score is multiplied by the length of the intersection set computed above, at step b.

The mapping algorithm can detect new CoRoLA categories for a document. Even if we directly mapped some Wikipedia categories to CoRoLa categories (if the mapping has been done manually), with the mapping algorithm we can discover new CoRoLa categories that correspond to the given document. Each document has thus two types of mappings: “manual” if a Wikipedia category of the document was manually mapped to a CoRoLa category or “automatic” if the mapping algorithm finds new CoRoLa categories corresponding to the document. A manually mapped CoRoLa category may also be found by the mapping algorithm (having the biggest score). Thus, we can compute the percent of the documents for which the best mapping is both manual and automatic, as a performance score of the mapping algorithm (see Table 14, below).

Table 14 shows the Wikipedia to CoRoLa category mapping for three types of keywords: lemmatized keywords, stemmed keywords and unprocessed keywords. One would think that, by doing lemmatization and/or stemming, performance would improve but Table 14 shows that this is not the case.

| **Keyword transformation** | **Wikipedia to CoRoLa category mapping performance** |
|---|---|
| Lemmatization | 93.1% |
| Stemming | 91.34% |
| No transformation | **96.66%** |

**Table 14**: Mapping algorithm performance

One explanation of this result is that lemmatization and stemming were done using string operations (i.e., longest common subsequences of similar word forms) as the Wikipedia documents were not POS tagged first. This operation is probably too destructive and the meaning of the remaining root/lemma is lost.

We give an example of a Wikipedia document, whose Wikipedia categories were not among those which were manually mapped to CoRoLa categories and for which the correct CoRoLa categories have been automatically inferred by the mapping algorithm. The Wikipedia document is “Istoria scrisului” (https://ro.wikipedia.org/wiki/Istoria_scrisului) and the CoRoLa categories are as follows:

```
Automatic: Science/History 139.18152

Automatic: Science/Linguistics   77.66515
```

We see that the mapping algorithm correctly inferred both the linguistics category and the history category, given the fact that the document is called (and it’s about) the history of writing.

The next step was mapping all the Domain and Subdomain pairs, which were automatically assigned to the documents, to a CURLICAT domain. For this purpose, all the scores assigned to a CoRoLa category that can be mapped to the same CURLICAT domain according to the mapping in Table 11 above are summed. In the example above, both the Linguistics and History subdomain of the CoRoLa Science domain have been mapped to the CURLICAT domain Science and the assigned CURLICAT category would be Science, with a score of 139.18152 + 77.66515 =216.84667.

More than one CURLICAT domain can be associated to a specific file, but we selected only one domain by consequently:

- Selecting only manual assigned categories as candidates, when they are available;
- Keeping as associated CURLICAT domain only the one corresponding to the category with the highest score.

Following the correction process, some documents included in the first version of the corpus (see D1 deliverable) proved to be not suitable for distribution in CURLICAT - since their newly attributed domain was outside CURLICAT’s target - and were eliminated. The current version of Romanian CURLICAT corpus has 26,477 documents.

## 6. The Slovak metadata (T5.2, T5.3)

There are several subcorpora of the Slovak National Corpus selected as sources for the CURLICAT project. These subcorpora use different annotations - the main corpus *prim-9.0* has detailed metadata with thorough style/genre annotation, while other, monothematic corpora use simpler annotation with data relevant for the respective corpora.

### 6.1. Original metadata description (T5.2)

#### 6.1.1. Corpora *wiki-2018-03* & *wiki-2019-08*

These corpora use a very simple metadata schema, consisting only of following keys:

| **id** | **Value: type string** |
|---|---|

Short unique identifier of the document. The first two characters encode the source: wp stands for Wikipedia, np for Necyklopédia, followed by a string of several digits. Example: wp1410

| **timestamp** | **Value: type string** |
|---|---|

Timestamp of the last edit of the page, in ISO 8601 format (date and time in UTC). Example: 2018-07-15T20:10:51Z

| **title** | **Value: type string** |
|---|---|

Title of the page (human readable).

#### 6.1.2. Corpus *od-justice-1.0*

| **url** | **Value: type string** |
|---|---|

URL of the document, at the time of data acquisition

|  |  |
|---|---|
| **court** | **Value: type string** |

Name of the court issuing the statement, in human readable form

|  |  |
|---|---|
| **zn** | **Value: type string** |

Official identifier of the court statement

|  |  |
|---|---|
| **date** | **Value: type string** |

Date of the statement publication, in ISO 8601 format (date)

|  |  |
|---|---|
| **tokcount** | **Value: type unsigned integer** |

Size of the document in tokens.

### 6.1.3. Corpus *prim-9.0*

All the metadata keys have two forms: the real and the visual one. The real is stored in the metadata annotation, the visual is displayed in the corpus manager. The visual form is limited to four characters and is motivated by the need for a nice, aligned display.

In the following tables, the visual form of the keys is shown in parentheses.

|  |  |
|---|---|
| **Name (Name)** | **Value: type string** |

Name (title) of the document. As given in the document.

|  |  |
|---|---|
| **Origname (OrgN)** | **Value: type string** |

If the document is a translation, the original name (title) of the document in the original language, else empty.

| **Author (Auth)** | **Value: type string** |
|---|---|

Author of the document. As given in the document (including errors, if any).

| **Origauthor (OrgA)** | **Value: type string** |
|---|---|

Author of the document. Original name, e.g in the original language, or with errors fixed.

| **Translator (Trnr)** | **Value: type string** |
|---|---|

Name of the translator. YYY if the document is not a translation.

| **Translation (Trnn)** | **Value: type enum(trn,org,ftr)** |
|---|---|

Translation or original: trn : translation; org : original Slovak; ftr: Free translation; YYY : Mix of translated texts and original Slovak

| **ISBN (ISBN)** | **Value: type string** |
|---|---|

ISBN, empty if not assigned or not annotated

| **ISSN (ISSN)** | **Value: type string** |
|---|---|

ISSN, empty if not assigned or not annotated

| **SourceId (ScId)** | **Value: type string** |
|---|---|

Unique identification string of the source (in the archive of source documents). By convention, it starts with a date of document acquisition.

| | |
|---|---|
| **Id (Id)** | **Value: type string** |

Unique identification string of this document

| | |
|---|---|
| **Rhyme (Rhym)** | **Value: type enum(nrh,rhy)** |

Text is in rhymes: nrh : unrhymed; rhy:    rhymed; MIX:    Partially rhymed

| | |
|---|---|
| **Type (Type)** | **Value: type enum(img,inf,prf,liv)** |

Text type: Img: Belles-lettres, artistic literature, fiction; inf : Journalistic, informative type;  Prf: Professional texts; Liv: Live communication

| | |
|---|---|
| **Subtype (SubT)** | **Value: type enum(poe,pro,dra,pub,adv,adm,sci,pop,txb,enc,man,spk,wri)** |

Text subtype

- For **Type=img**: poe:  poetry; pro:  prose; dra:  drama
- For **Type=inf**: pub: Journalistic text; adv: Commercials, advertisement; Adm: Administrative text
- For **Type=prf**:  sci: Scientific literature, articles, professional journals, university textbooks;  pop: Popular science, hobbyist journals; txb: Elementary and secondary education textbooks;  enc: Encyclopaediae and other (alphabetically) sorted documents, man: Manuals, instructions of use
- For **Type=liv**: spk: Spoken (transcribed);  wri: Live written communication (internet, teletype, communication of hearing impaired people via writing etc.)

| | |
|---|---|
| **Genre (Genr)** | **Value: type enum(ver,son,scd,scf,scr,nov,col,ess,mem,let,sen,mon,hnd,dis,std,abs,tcl,rfl,ref,lct,dsc,crs,crt,opn,ins,doc,ann,rst,lpt,anl,pbb,spc)** |

- Genre of the text

ver:  poem; son: song, libretto; scd:  theater play;  scf:  movie script, movie subtitles; scr : radio broadcast (transcript); nov: novel; col: short story, anthology; ess: essay; mem: memoirs; let: letter; chr: chronicle; sen: very short genres (citations, quotations, aphorisms etc.); mon: monograph; hnd: handbook ; dis: dissertation; std: study; abs: abstract; tcl: article; rfl: reflection; ref: report; lct: lecture; dsc: discussion, debate; crs: characteristics; crt: review (consumer); opn: review (scientific, professional); ins: instruction of use; doc: minutes, protocol, treaty; ann: edict, announcement, questionnaire; lst: list, schedule, programme,rules, tables of contents; rpt: report, interview; anl: (analytics) foreword, comment, gloss, review, critique, discussion, caricature; pbb: (belles lettres) feuilleton, feature, column; spc: talks (political, occasional), sermons.

| | |
|---|---|
| **Subgenre (SubG)** | **Value: type enum(crm,scf,bel,jun,trv,fac)** |

Subgenre. Specified only for Genre **nov**, **col**, **ess**.

crm: crime, thriller, espionage; scf: sci-fi, fantasy; bel: belles lettres; jun: juvenile, young adult; trv: travelogue; fac: factual literature.

| | |
|---|---|
| **Domain (Domn)** | **Value: type enum(ars,hum,law,nat,tec,ecn,blf,lif,ins,plt)** |

Professional domain

ars: at science; hum: humanities; law: law; nat: natural science; tec: technical; ecn: economy, management; blf: belief, supernatural; lif: life style; ins: interdisciplinary; plt: politics.

| | |
|---|---|
| **Subdomain (SubD)** | **Value: type enum(mus,cin,arc,art,the,lit,his,psy,edu,soc,phi,inf,pol,eth,cul,bil,jud,jur,agr,med,pha,zoo,bot,bio,che,mat,ggr,phy,met,geo,env,traene,ind,com,bui,sta,eco,mng,mer,rel,teo,exc,hou,fsh,spo,sct,amu,min,reg,cnl,clt)** |

Subdomain - further elaboration on the professional domain.

for **Domain=ars**: mus: music, opera, operetta, ballet; cin: movies, cinema; arc: architecture; art: painting, photography, sculpture; the: theatre; lit: literature.

for **Domain=hum**: his: history, archaeology; psy: psychology; edu: pedagogical; soc:

sociology, communication, mass media; phi: philosophy; inf: library science; pol: politology; lin: linguistics; eth: ethnology, ethnography; cul: culturology.

for **Domain=law**: bil: law, edicts; jud: judicature, court decision; jur: other

for **Domain=nat**: agr: agriculture; med: medicine; pha: pharmacy; zoo: zoology; bot: botanics; bio: biology; che: chemistry; mat: mathematics; ggr: geography; phy: physics (incl. astronomy); met: meteorology; geo: geology; env: environmentalism, ecology.

for **Domain=tec**: tra: traffic, telecommunication; ene: energetics; ind: industry; com: computers, computer science; bui: civil engineering; sta: normalisation, standardisation

for **Domain=ecn**: eco: economy, banking, commerce; mng: management; mer: merchandise

for **Domain=blf**: rel: religion, belief, religion sects; teo: theology; exc: supernatural, occult, magic, astrology.

for Domain=lif; hou: home (garden, home improvement, kitchen, handwork); fsh: fashion; spo: sport; sct: social life; amu; games, hobbies, free time, travel: min: minorities; reg: regional; cnl:advice, counselling; clt: culture.

| **Medium (Medi)** | **Value: type string** |
|---|---|

Medium

lib: book (primarily published as a physical book); ebk: E-book (primarily published as an e-book); nws: newspaper; jou: journal; ste: lecture notes; net: internet and other (pre-internet) networks. Internet newspaper, webpages, e-mail, usenet posts, discussion forums, interactive communication; for: forms; occ: occasional publications, proceeding; npu: unpublished, manuscripts; tvf: television, cinema; rad: radio.

| **Authsex (AutS)** | **Value: type enum(msc,fem,MSS,MIX,YYY)** |
|---|---|

Author gender (sex)

msc: male; fem: female; MSS: other; MIX: mixture; YYY: none; XXX: unknown.

| **Lang (Lang)** | **Value: type const string = “slk”** |
|---|---|

language of the document, ISO 639-2 code; should be slk

| | |
|---|---|
| **Varieta (Vari)** | **Value: type enum(std,nst)** |

Language variant, usually literary Slovak

std: standard (literary) Slovak; nst: non-standard Slovak

| | |
|---|---|
| **Paragraphs (Para)** | **Value: type enum(tru,fls)** |

Is the text segmented into paragraphs?

tru: segmented; fls: paragraph segmenting lost.

| | |
|---|---|
| **Emphasis (Emph)** | **Value: type enum(tru,fls)** |

Does the text keep information about emphasis?

tru: yes; fls: no.

| | |
|---|---|
| **Diacritics (Dcrt)** | **Value: type enum(tru,fls)** |

Is the text written using (correct) diacritics?

tru: yes; fls: no.

| | |
|---|---|
| **Transsex (TrnS)** | **Value: type enum(msc,fem,MSS,MIX,YYY)** |

Gender (sex) of the translator, see Authsex.

| | |
|---|---|
| **Origlang (OrgL)** | **Value: type string** |

language of the original document (if this is a translation), ISO 639-3 code

Translations through a different language are denoted using the „>“ U+003C LESS-THAN SIGN character. Example: eng>ger

| | |
|---|---|
| **Date (Date)** | **Value: type string** |

Publishing date

| **Dateorig (OrgD)** | **Value: type string** |
|---|---|

Date of the first issue (creation of the document); date of publication of the original, if this is a translation

| **Conglomerate (Cong)** | **Value: type string** |
|---|---|

Unique identifier of the set of documents this document belongs to

| **Bogocong (Bogo)** | **Value: type string** |
|---|---|

Shortened (few characters long) Conglomerate

| **Comment (Comn)** | **Value: type string** |
|---|---|

Arbitrary comment

| **Corrected (Corr)** | **Value: type enum(tru,fls)** |
|---|---|

Was the document proofread? tru: yes; fls: no.

| **Bibliography (Bibl)** | **Value: type string** |
|---|---|

Bibliography

### 6.2. Mapping with the common metadata schema (described in section 8) (T5.2)

The metadata in the Slovak National Corpus annotation follows a simple key-value structure, with a fixed set of annotation keys and a singular string value (can be an empty string) assigned to a particular key. A key has two names: a long one, used in the metadata, and a short (at most 4 letter long) one, used purely for readout in corpus managers, to keep the width of displayed metadata reasonable. The long and short names are otherwise isomorphic.
Several keys contain freeform values (arbitrary string), but there is a set of keys containing only values out of a given set specific for that key ("enums"). These enums are generally three

characters long, by convention lowercase is used for “proper” annotation tied to the specific key, while uppercase for generalised annotation applicable to various keys.

The annotation is logically two-level for many of the keys, however the annotation format flattens these levels into a standard *key:value* file. One practical consequence of this arrangement is that not all combinations of values are valid - for a given Type, Genre or Domain, only selected Subtype, Subgenre and Subdomain are legitimate.

The conversion of these metadata to the CURLICAT format has been straightforward, by replacing the relevant keys with the CURLICAT ones, as described in Table 15 for the *Domain* and Table 16 for Type (which corresponds to the SNK *medium* key). Note that not all values are covered in the final CURLICAT corpus.

The following set of core obligatory metadata is used for all documents:

- Identifier – unique identifier of the document within all the corpora, following CoNLL-U conventions; consisting of the language code, internal source identifier and a document *id* (an alphanumeric string), separated by the U+002D HYPHEN-MINUS character.
- Language – the ISO 639-1 language code of the sub-corpus (always *sk*)
- PublicationDate – the primary date of the document, the publication date in the ISO 8601 format, with accuracy given by source metadata (at least the year, at most the day)
- DocumentTitle – informative, human readable title (name) of the document or collection of documents
- ArticleTitle – informative, human readable title (name) of the individual document, if applicable (i.e. if part of a collection)
- Type – type of the document, e.g. *book*, *article*
- Source – the name of the organisation that published the source document, title of the journal etc.
- Domain – CURLICAT domain mapped from the original corpus metadata (detailed mapping described in the table below local metadata fields description)
- No_of_sentences, No_of_words, No_of_punctuation, No_of_tokens – the total number of sentences, words, punctuation marks and tokens (words + punctuation marks) in the document.
- Licence – licence of the text, using abbreviations for well-known licences, including the version, if applicable (e.g. *CC BY-SA 4.0*); or the string “*other freely redistributable*” for source-specific licences.

The optional metadata:

- Author – author of the document, as listed in bibliography or similar data
- Url – URL of the source document at the time of acquisition; might point to a collection or a higher level webpage

| SNK domain | CURLICAT domain |
|---|---|
| hum | Science |
| ars | Culture |
| ins | General |
| tec | Science |
| nat | Nature |
| plt | Politics |
| law | Law |
| ecn | Economy |
| lif | Culture |

**Table 15:** mapping of the Slovak National Corpus (SNK) domains to the CURLICAT domain

| SNK media | CURLICAT type |
|---|---|
| ebk | book |
| jou | journal |
| lib | book |
| net | internet |
| npu | other |
| ste | other |
| for | other |
| tvf | other |
| rad | other |
| nws | news |
| occ | other |

**Table 16:** mapping of the Slovak National Corpus (SNK) media values to the CURLICAT type

## 6.3. Metadata validation activities (T5.3)

### 6.3.1. Technical validation

Technical validation ensures that the metadata are in the correct format. This includes the format of the file the metadata is included in, and the format of the metadata values.

In the pipeline of the Slovak corpus, there are two validation points at various steps of the pipeline:

- When converting from raw text output into a common XML format with paragraph delimiters and metadata in the header, the metadata is validated to be a valid UTF-8 string. Whitespace and non-printable characters are converted to spaces, multiple spaces are collapsed, html entities (characters &, < and >) are escaped, leading and trailing whitespace is removed.
- When converting vertical files (output of lemmatization and MSD annotation) into CONLLU+ format, the metadata are checked for completeness (i.e. if all the metadata fields are present, even if empty valued) and duplicates (two metadata fields with the same key).

### 6.3.2. Semantic validation

Semantic validation consists of heuristical and statistical validation of the metadata values. Heuristical tests verify that the value is not empty, and for dates (publication date and collection date) is tested for being in the interval [1993, current_day]; URLs are tested to start with the strings either http:// or https:// (other URL protocols were not used in collecting the sources). Statistical validation comprises making a frequency list of metadata values for each key, and manually verifying the outliers (values appearing only once or a few times, with the exception of the author and title fields); and in making a distribution of characters and manually verifying metadata with alphabetical characters not present in the standard Slovak orthography and with uncommon symbols.

## 7. The Slovenian metadata (T5.2, T 5.3)

### 7.1. Original metadata description (T 5.2)

Gigafida 2.0 is a reference corpus of written Slovene. It comprises daily news, magazines, a selection of web texts (a certain portion of which covers news texts as well), and different types of publications (fiction, school books, and non-fiction). The texts have been selected and automatically processed with the aim of creating a corpus that represents a sample of modern standard Slovene and can be used for research in linguistics and other branches of humanities, for compiling modern dictionaries, grammars, and learning materials, as well as for developing language technologies for Slovene. The metadata is stored alongside the texts in the central corpus database. For the CURLICAT project, most of the data will be selected from the Gigafida 2.0 corpus and the remainder will be processed with the same pipeline to ensure compatibility. Hence, this chapter effectively deals with mapping the Gigafida 2.0 metadata to the metadata set defined in the CURLICAT project.

| **Text ID** | **Value: type string** |
|---|---|

This field encodes a unique ID of a document contained in the corpus.

| **Author** | **Value: type string** |
|---|---|

This field encodes the name of the document author. It can be empty if not available.

| **Title** | **Value: type string** |
|---|---|

This field encodes the title of the document or article.

| | |
|---|---|
| **Year of publication** | **Value: type number** |

This field encodes the year when the document was published.

| | |
|---|---|
| **Source** | **Value: type string** |

This field encodes the na ere the document was published. It can be empty if not available.

| | |
|---|---|
| **Publisher** | **Value: type string** |

This field encodes the name of the publisher of the publication where the document was published. It can be empty if not available.

| | |
|---|---|
| **Text type** | **Value: type enum (newspapers, magazines, internet, non-fiction, fiction, other)** |

This field encodes the name of the type of the original document, with predefined values from the list in the table above. It can be empty if not available.

| | |
|---|---|
| **Number of paragraphs** | **Value: type number** |

This field encodes the number of paragraphs in the document.

| | |
|---|---|
| **Number of sentences** | **Value: type number** |

This field encodes the number of sentences in the document.

| | |
|---|---|
| **Number of words** | **Value: type number** |

This field encodes the number of words in the document.

| | |
|---|---|
| **Number of punctuation characters** | **Value: type number** |

This field encodes the number of punctuation characters in the document.

| | |
|---|---|
| **Number of tokens** | **Value: type number** |

This field encodes the number of tokens in the document.

| | |
|---|---|
| **Technical text standardness** | **Value: type number** |

This field encodes the numeric score assigned by an automatic technical standardness evaluation algorithm (LJUBEŠIĆ et al, 2015).

| | |
|---|---|
| **Linguistic text standardness** | **Value: type number** |

This field encodes the numeric score assigned by an automatic linguistic standardness evaluation algorithm.

| **Domain** | **Value: type enum (culture, economics, finance, education, health, politics)** |
|---|---|

This field encodes the domain of the document contained on the corpus.

### 7.2. Mapping with the common metadata schema (described in section 8) (T5.2)

With respect to the metadata schema proposed in the CURLICAT project we propose the following mapping (see Table 17). We were able to map all obligatory metadata, as well as the optional "Author" metadata field. Some obligatory metadata were readily available in the Gigafida 2.0 database, such as PublicationDate, Type, Source, No_of_sentences, No_of_words, No_of_punctuation, No_of_tokens, while others had to be generated based on other similar data types. For example, Gigafida does not have separate document title and article title fields but in vast majority of cases, articles are self-contained documents in the database, so it is possible to use the "Title" metadata for both DocumentTitle and ArticleTitle since there is (almost) no overlap. Finally, some fields will always have the same default value (e.g., Language and License).

| **CURLICAT schema** | **Gigafida 2.0** |
|---|---|
| Identifier | Text ID |
| Language | "sl" |
| License | "CC-BY-SA 4.0" |
| PublicationDate | Year of publication |

| | |
|---|---|
| DocumentTitle | Title |
| ArticleTitle | Title |
| Type | Text type |
| Source | Source |
| Domain | Domain |
| No_of_sentences, No_of_words, No_of_punctuation, No_of_tokens | Number of sentences, Number of words, Number of punctuation characters, Number of tokens |
| Author | Author |
| SourceType | N/A |
| Keywords | N/A |
| URL | N/A |
| Style | N/A |
| Subdomain | N/A |

| | |
|---|---|
| Issn_isbn_eisbn | N/A |

**Table 17:** Mapping of Gigafida 2.0 metadata fields to the CURLICAT common metadata schema

### 7.3. Metadata Validation Activities (T5.3)

When new documents are added to the Gigafida 2.0 database, their metadata is validated from a technical and semantic point of view. The technical aspect of the validation is covered by the database schema containing the corpus and custom import processes: imported strings are converted into UTF-8 encoding, checked for completeness and deduplicated. In terms of semantic validation, various statistical and heuristic tests are run on a regular basis, generating lists of suspect entries to be checked manually.

# 8. Common metadata schema (T5.1)

Principles of metadata encoding from CEF-project MARCELL (Váradi et al., 2020) are to be followed also in the current endeavour of creating a common metadata annotation schema. Metadata is, therefore, seen as a collection of information classified as *obligatory* (all partners have to provide it), *optional* (the field can be missing or containing an empty value in some language corpora), or *local* (annotation specific for a given language corpus, included for fidelity to the original source annotation) and associated to each document in CURLICAT corpus.

*Obligatory* metadata fields in the CURLICAT schema, that were easily provided for all language corpora, are the following :

- *Identifier:* is a short string uniquely identifying the document in its language corpora, in the format *lc-string-dddddd*, where lc is the language code, string is an internal code marker specifying the document provenience (e.g. bn - for the Polish Library of Science or crl - for CoRoLa, the Romanian National Corpus ) original identifier or file name;

- *Language:* the ISO 639-1 codes of the specific represented languages;
- *Licence* – licence of the text, using abbreviations for well-known licences, including the version, if applicable (e.g. *CC BY-SA 4.0*); or the string "*other freely redistributable*" for source-specific licences.
- *PublicationDate:* the date of the original publication of the document, in ISO 8601 format;
- *DocumentTitle and ArticleTitle:* the human readable title of the source document, in the original language, e.g the title of the book, chapter, paper, newspaper article etc. based on which the document was created;
- *Type:* further specifies the type of the source document, in English e.g. book, chapter, paper, newspaper article, blogpost, etc.
- *Source:* the name of the organisation that published the source document, be it a Journal, Publishing House, Blog, Website, etc., in the original language;
- *Domain:* the domain covered in the document, in English, selected from the predefined list of CURLICAT domains and based on the domain metadata fields in the source corpora;

- *No_of_sentences, No_of_words, No_of_punctuation, No_of_tokens:* the total number of sentences, words, punctuation marks and tokens (words + punctuation marks) in the document.

The *optional* metadata fields we added to the metadata schema are:

- *Author:* the name/s of the person/s that created the text in the source document;
- *SourceType:* the type of organisation that published the source document, selected from a predefined list *(*Newspaper/Publishing House/Blog/Website/Other)
- *Keywords:* contains several keywords related to the content of the document;
- *Url:* is the original individual address the document was accessed at, if applicable;
- *Style:* the literary style of the text in the document, selected from a predefined list: imaginative, memoirs, administrative, law, journalistic, etc;
- *Subdomain:* a further classification of the documents into narrower categories, e.g. scientific fields for the Science domain, or cultural fields for the Culture domain;
- Issn_isbn_eisbn: the International Standard Serial Number or International Standard Book Number of the source document.

Some local fields that different partners included come from:

- the scientific publications descriptions that comprise the Polish corpus: *title in English, abstract in English, issue volume, issue number, page range, full text licence, reviewers*, etc.
- Fields specific to the Hungarian corpus: *Editor*: A string representing one or multiple proper names, the editor(s) of a collection of works; *RespName*: A string representing one or multiple proper names that refers to the persons or organisations that had any role in the distribution of the *source* data (not the CURLICAT corpus);

Such elaborated metadata schema will allow easy selection of relevant subcorpora, using metadata value as a criterion, thus facilitating the training of different in-domain language models.

All types of metadata fields, including the obligatory ones, may have unavailable information for specific language corpus or specific documents in the collections. In such cases, the value is N/A (not available) for whatever field.

The format for providing the harmonised metadata is in the header of each CONLL-U Plus processed text document, as previously done in MARCELL Project. Scripts were implemented by each partner for extracting the metadata from original sources, mapping them to the common CURLICAT schema and printing them as headers in the CONLL-U+ documents.

# Bibliographical references

Burnard, L. (2005), Metadata for corpus work. In *Developing Linguistic Corpora: a Guide to Good Practice*, chapter Metadata for corpus work, Oxford: Oxbow Books.

Campos, R., Mangaravite, V., Pasquali, A., Jatowt, A., Jorge, A., Nunes, C. and Jatowt, A. (2020). YAKE! Keyword Extraction from Single Documents using Multiple Local Features. In Information Sciences Journal. Elsevier, Vol 509, pp. 257-289.

Koeva, S., Stoyanova, I., Leseva, S., Dimitrova, Ts., Dekova, R. and Tarpomanova, E. (2012) The Bulgarian National Corpus: Theory and Practice in Corpus Design. In *Journal of Language Modeling*, 2012, Vol. 0, No. 1, pp. 65-110.

Koeva S., Stoyanova, I., Todorova, M., Leseva, S., Dimitrova, T.. (2016) Metadata Extraction, Representation and Management within the Bulgarian National Corpus. In *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora* 2016, Portorož (LREC-2016 workshop), ELDA, pp. 33-39.

Koeva, S., Obreshkov, N., Yalamov, M. Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, 2020, pp. 6988-6994.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S., & Škrjanec, I. (2015, September). Predicting the level of text standardness in user-generated content. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 371-378).

Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., & Brank, J. (2020). The MARCELL Legislative Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 3761–3768.