Metadata quality and common mistakes in CLARIN-PL

Marcin Oleksy

Check-lists

1. CMDI best practice guide:
    - 1.1. is the CMDI file schema-valid?
    - 1.2. are the header fields complete?
        - 1.2.1. is there a uniqe MdSelfLink?
        - 1.2.2. is there an MdCollectionDisplayName?
    - 1.3. does it contain ResourceProxy elements?
        - 1.3.1. is there a link to a LandingPage when available?
        - 1.3.2. is there an indication of the mime type?
    - 1.4. is the file too big (e.g. several megabytes) to be useful? (suggest higher granularity)
    - 1.5. sparseness: what about files that hardly contain any information?
    - 1.6. what is the information entropy? (lots of very similar files might be an indication of a suboptimal modelling)
    - 1.7. do some elements contain multiple values within a single string? (better to repeat the element instead of e.g. having comma-separated enumerations)
    - 1.8. if there are multilingual elements, is the xml:lang attribute used to indicate the language?

2. Inspected/Checked aspects Metadata Quality Assessement Service:
    - 2.1. Schema level
        - 2.1.1. presence of "required" data categories
        - 2.1.2. ratio of elements with data categories
        - 2.1.3. size?

2.2. Instance level
2.2.1. availability of the schema
2.2.2. validity of the record wrt to the schema
2.2.3. links are resolvable
2.2.4. filled-in ratio?
2.2.4.1. how many of the elements defined by schema are actually populated with information
2.2.5. values conform to a controlled vocabulary
2.2.6. size e.g. overall size (measured in characters)

Required data categories - proposal

1. VLO
1.1. location country
http://hdl.handle.net/11459/CCR_C-2532_d004b0a6-fd1d-3ca3-abf1-1e6aeb3e37b2
1.2. mime type
http://hdl.handle.net/11459/CCR_C-2571_2be2e583-e5af-34c2-3673-93359ec1f7df
1.3. genre
http://hdl.handle.net/11459/CCR_C-2470_d191f2b2-6339-f031-b534-70d526b28357
1.4. sub genre
http://hdl.handle.net/11459/CCR_C-3899_c6c608e7-cb2e-1832-09ff-aee36e1f2ed4
1.5. metadata tag
http://hdl.handle.net/11459/CCR_C-5436_6ab57c2c-5f8d-3561-6db6-d75da23d2637
1.6. language ID
http://hdl.handle.net/11459/CCR_C-2482_08eded24-4086-7e3f-88e5-e0807fb01e17
1.7. language name
http://hdl.handle.net/11459/CCR_C-2484_669684e7-cb9e-ea96-59cb-a25fe89b9b9d
1.8. language usage
http://hdl.handle.net/11459/CCR_C-5361_ba085ec1-9746-52bf-8cc1-3c300ce16eb8
1.9. language
http://hdl.handle.net/11459/CCR_C-5358_3cd089fe-ad03-6181-b20c-635ea41ed818
1.10. availability
http://hdl.handle.net/11459/CCR_C-2453_1f0c3ea5-7966-ae11-d3c6-448424d4e6e8
1.11. life cycle status
http://hdl.handle.net/11459/CCR_C-3818_8c4aec73-1654-7565-9575-c4a17425ee29
1.12. modalities
http://hdl.handle.net/11459/CCR_C-2490_44bc38a3-1799-4149-c791-40ac0176f0ff
1.13. organization
http://hdl.handle.net/11459/CCR_C-2459_fc4e74d6-84de-c8cd-1ae8-2c2be5ee90b1
1.14. project name
http://hdl.handle.net/11459/CCR_C-2536_13fc5f10-c14a-1f64-a669-32736f6d3ef5
1.15. project title
http://hdl.handle.net/11459/CCR_C-2537_fa206273-223a-f4fa-dde3-ba59b965701f
1.16. resource class
http://hdl.handle.net/11459/CCR_C-3806_e55e9ed6-b099-c21d-a634-3c7f4d22a215
1.17. TEI Header type
http://hdl.handle.net/11459/CCR_C-5424_3200a38b-344e-41de-e539-f71f80c38df8
1.18. domain of use
http://hdl.handle.net/11459/CCR_C-6147_ebed915e-f911-f128-cddc-466aa41c9c73

1.19. classification code
http://hdl.handle.net/11459/CCR_C-5316_2c6244b4-4f10-5e8e-49b6-26fbf7004791

1.20. Time coverage
http://hdl.handle.net/11459/CCR_C-3664_eb600f47-5123-efbe-251b-d952c65fc847

1.21. End range
http://hdl.handle.net/11459/CCR_C-3655_bc4c2656-2946-0be9-49f0-021a811e531b

1.22. Start range
http://hdl.handle.net/11459/CCR_C-3654_f1608e88-95e6-4233-5d21-5312e76de32d

1.23. IPR holder
http://hdl.handle.net/11459/CCR_C-6709_cb3572ed-ffd3-04f1-c145-b9c1f26bfc82

1.24. Legal Owner
http://hdl.handle.net/11459/CCR_C-2956_519a4aab-2f76-0fd3-090e-f0d6b81a7dbb

1.25. availability
http://hdl.handle.net/11459/CCR_C-2453_1f0c3ea5-7966-ae11-d3c6-448424d4e6e8

1.26. rights

1.27. source country

2. Metadata Quality Assessement Service

2.1. resource title or name
2.2. modality
2.3. resource class
2.4. genre?
2.5. keywords or tags
2.6. country?
2.7. contact person,
2.8. publication year
2.9. availability / licence

Value normalization

1. CLAVAS
2. VLO preprocessor
3. Suggested date format

Common mistakes

1. **Missing language tags**

   In some cases it is necessary to indicate the language of the metadata element content (value) by a language tag for multilingual elements. It is a xml:lang attribute, for example:

   ```
   xml:lang="eng"
   ```

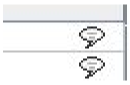
   Thus, for multilingual elements the metadata element should be coded:

   ```
   <Organisation xml:lang="eng">Wrocław University of Technology</Organisation>
   ```

   instead of (incomplete):

   ```
   *<Organisation>Wrocław University of Technology</Organisation>
   ```

   The user should make sure that all multilingual elements contain complete language tag. There is a special icon (bubble) in Arbil on the right side of the cell. It indicates that user has to specify the language by clicking on the bubble and selecting a proper language from the list:

2. **Confusing language of the metadata element and metadata element: “language”**

   A xml:lang attribute isn’t equal to metadata element describing the language:

   ```
   <SubjectLanguage>
    <Language>
      <LanguageName xml:lang="eng">Polish</LanguageName>
      <ISO639>
        <iso-639-3-code>pol</iso-639-3-code>
      </ISO639>
    </Language>
   </SubjectLanguage>
   ```

   xml:lang attribute refers to the language of the metadata entry, while the CMDI element containing “language” in the name usually refers to the language of the indicated object (resource, tool e.g. language supported by the tool etc.). If a resource contains polish texts or a tool supports polish language, the value of metadata element describing the language should be: “Polish”. Even if a metadata description is edited in English (as above). From another perspective, if a description is edited in English, xml:lang attribute should be: xml:lang="eng", even if a resource contains polish texts or a tool supports polish language.

3. **A several values in one metadata element**

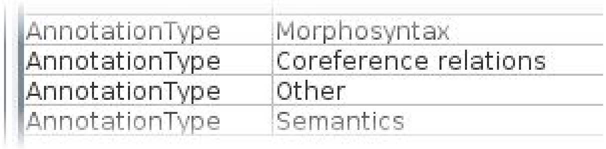
In some cases complete metadata description involves indication of several values within the category (e.g. annotation types in the corpus, which has been annotated on several layers), but user should not mention all values (with a comma etc.) in one xml element (one occurrence of CMDI element):

*<AnnotationType>Morphosyntax, Coreference relations, Semantics</AnnotationType>

Each value should appear in a separate xml element:

<AnnotationType>Morphosyntax</AnnotationType>
<AnnotationType>Coreference relations</AnnotationType>
<AnnotationType>Semantics</AnnotationType>

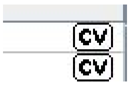
= separate cell in Arbil:



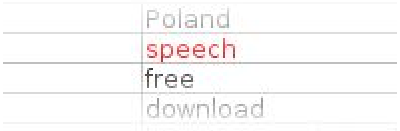
AnnotationType	Morphosyntax
AnnotationType	Coreference relations
AnnotationType	Other
AnnotationType	Semantics

4. **Ignoring the closed/controlled vocabulary**

Some of the categories are connected with a closed vocabulary. Their value may be indicated only by selecting the proper one from the list. Such categories are marked by "CV" symbol in Arbil (on the right side of the cell).

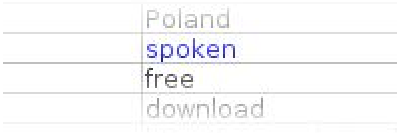


Arbil also alerts the values that are not on the closed vocabulary list. They are marked by a red colour of the font:



	Poland
	speech
	free
	download

Proper entry (the value derived from the closed/controlled vocabulary) turns in blue:



	Poland
	spoken
	free
	download