# Jakość metadanych i częste błędy

oprac. Marcin Oleksy



# 1. Check-listy

1.1. CMDI best practice guide:
- 1.1.1. Czy schemat CMDI jest ważny/obowiązujący?
- 1.1.2. Czy są kompletne pola nagłówka, w tym:
  - 1.1.2.1. unikalny MdSelfLink?
  - 1.1.2.2. MdCollectionDisplayName?
- 1.1.3. Czy schemat/instancja zawiera elementy ResourceProxy?
  - 1.1.3.1. Czy jest link do strony docelowej (LandingPage), jeśli jest dostępna?
  - 1.1.3.2. Czy jest wskazany typ MIME (mime type)?
- 1.1.4. Czy plik nie jest za duży, by być użytecznym?
- 1.1.5. Jaka jest entropia informacji? (→ wiele podobnych plików może wskazywać na nieoptymalne modelowanie)

1.2. Zalecenia Metadata Quality Assessement Service:
- 1.2.1. Poziom schematu
  - 1.2.1.1. obecność "wymaganych" kategorii danych
  - 1.2.1.2. stosunek elementów do kategorii danych
  - 1.2.1.3. rozmiar
- 1.2.2. Poziom instancji
  - 1.2.2.1. dostępność schematu
  - 1.2.2.2. ważność zapisu w odniesieniu do schematu
  - 1.2.2.3. rozpoznawalność linków
  - 1.2.2.4. współczynnik wypełnienia - jak wiele elementów zdefiniowanych w schemacie zostało wypełnionych informacjami
  - 1.2.2.5. dostosowanie wartości do słownika kontrolowanego
  - 1.2.2.6. odpowiedni rozmiar

# 2. Wymagane kategorie danych

2.1. przeglądarka VLO

2.1.1. location country
http://hdl.handle.net/11459/CCR_C-2532_d004b0a6-fd1d-3ca3-abf1-1e6aeb3e37b2

2.1.2. mime type
http://hdl.handle.net/11459/CCR_C-2571_2be2e583-e5af-34c2-3673-93359ec1f7df

2.1.3. genre
http://hdl.handle.net/11459/CCR_C-2470_d191f2b2-6339-f031-b534-70d526b28357

2.1.4. sub genre
http://hdl.handle.net/11459/CCR_C-3899_c6c608e7-cb2e-1832-09ff-aee36e1f2ed4

2.1.5. metadata tag
http://hdl.handle.net/11459/CCR_C-5436_6ab57c2c-5f8d-3561-6db6-d75da23d2637

2.1.6. language ID
http://hdl.handle.net/11459/CCR_C-2482_08eded24-4086-7e3f-88e5-e0807fb01e17

2.1.7. language name
http://hdl.handle.net/11459/CCR_C-2484_669684e7-cb9e-ea96-59cb-a25fe89b9b9d

2.1.8. language usage
http://hdl.handle.net/11459/CCR_C-5361_ba085ec1-9746-52bf-8cc1-3c300ce16eb8

2.1.9. language
http://hdl.handle.net/11459/CCR_C-5358_3cd089fe-ad03-6181-b20c-635ea41ed818

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

2.1.11. life cycle status
http://hdl.handle.net/11459/CCR_C-3818_8c4aec73-1654-7565-9575-c4a17425ee29

2.1.12. modalities
http://hdl.handle.net/11459/CCR_C-2490_44bc38a3-1799-4149-c791-40ac0176f0ff

2.1.13. organization
http://hdl.handle.net/11459/CCR_C-2459_fc4e74d6-84de-c8cd-1ae8-2c2be5ee90b1

2.1.14. project name
http://hdl.handle.net/11459/CCR_C-2536_13fc5f10-c14a-1f64-a669-32736f6d3ef5

2.1.15. project title
http://hdl.handle.net/11459/CCR_C-2537_fa206273-223a-f4fa-dde3-ba59b965701f

2.1.16. resource class
http://hdl.handle.net/11459/CCR_C-3806_e55e9ed6-b099-c21d-a634-3c7f4d22a215
2.1.17. TEI Header type
http://hdl.handle.net/11459/CCR_C-5424_3200a38b-344e-41de-e539-f71f80c38df8
2.1.18. domain of use
http://hdl.handle.net/11459/CCR_C-6147_ebed915e-f911-f128-cddc-466aa41c9c73
2.1.19. classification code
http://hdl.handle.net/11459/CCR_C-5316_2c6244b4-4f10-5e8e-49b6-26fbf7004791
2.1.20. Time coverage
http://hdl.handle.net/11459/CCR_C-3664_eb600f47-5123-efbe-251b-d952c65fc847
2.1.21. End range
http://hdl.handle.net/11459/CCR_C-3655_bc4c2656-2946-0be9-49f0-021a811e531b
2.1.22. Start range
http://hdl.handle.net/11459/CCR_C-3654_f1608e88-95e6-4233-5d21-5312e76de32d
2.1.23. IPR holder
http://hdl.handle.net/11459/CCR_C-6709_cb3572ed-ffd3-04f1-c145-b9c1f26bfc82
2.1.24. Legal Owner
http://hdl.handle.net/11459/CCR_C-2956_519a4aab-2f76-0fd3-090e-f0d6b81a7dbb
2.1.25. availability
http://hdl.handle.net/11459/CCR_C-2453_1f0c3ea5-7966-ae11-d3c6-448424d4e6e8
2.1.26. rights
2.1.27. source country

2.2. Metadata Quality Assessement Service
2.2.1. resource title or name
2.2.2. modality
2.2.3. resource class
2.2.4. genre?
2.2.5. keywords or tags
2.2.6. country?
2.2.7. contact person,
2.2.8. publication year
2.2.9. availability / licence

# 3. Normalizacja wartości - słowniki

3.1. CLAVAS
3.2. VLO preprocessor
3.3. Proponowany format zapisu dat

# 4. Częste błędy

## 4.1. Brak określenia języka, w którym został dokonany wpis

W odniesieniu do znacznej części elementów (wartości metadanych) wymagane jest określenie języka, w którym został dokonany wpis. Jest to atrybut, który ma postać:

xml:lang="kod_języka"

np. w odniesieniu do języka angielskiego:

xml:lang="eng"

a więc pełny wpis dotyczący tego typu kategorii powinien wyglądać, np:

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

nie zaś:

<Organisation>Wrocław University of Technology</Organisation>

Edytując metadane należy więc zwrócić uwagę na pola, w których deklaracja dotycząca języka jest wymagana. W edytorze Arbil po prawej stronie takiego pola znajduje się dymek, którego kliknięcie spowoduje rozwinięcie listy dostępnych języków:

## 4.2. Kilka wartości w ramach jednego elementu

W pewnych sytuacjach możliwe jest podanie kilku wartości w ramach danej kategorii (np. w przypadku typów anotacji w danym korpusie/dokumencie). Nie należy jednak wymieniać tych wartości po przecinkach, np.

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

Dla każdej wartości należy utworzyć osobne wystąpienie:

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

W Arbilu będą to osobne pola:

| | |
|---|---|
| AnnotationType | Morphosyntax |
| AnnotationType | Coreference relations |
| AnnotationType | Other |
| AnnotationType | Semantics |

### 4.3. **Pominięcie słownika zamkniętego**

Część kategorii może zostać określona jedynie poprzez wybranie odpowiedniej wartości z zamkniętej listy. W edytorze Arbil po prawej stronie takiego pola znajduje się prostokąt z symbolem “CV”, którego kliknięcie spowoduje rozwinięcie odpowiedniej listy.


Wprowadzenie innej wartości niż określona w słowniku jest komunikowane przez Arbil czerwonym kolorem czcionki, np.


Poprawny wpis (w tym przypadku wartość wybrana z listy) to kolor niebieski, np.


Jeśli pełny opis wymaga wskazania kilku wartości, stosujemy zalecenia opisane w punkcie 4.2.