Open Datasets for Evaluating the Interpretation of Bibliographic Records

Joffrey Decourselle[1], Fabien Duchateau[1], Trond Aalberg[2], Naimdjon Takhirov[3] and Nicolas Lumineau[1]

[1] LIRIS, UMR5205, Université Lyon 1
Lyon, France
firstname.lastname@liris.cnrs.fr

[2] NTNU
Trondheim, Norway
trondaal@idi.ntnu.no

[3] Westerdals - Oslo School of Arts, Communication and Technology - Faculty of Technology - Oslo, Norway
taknai@westerdals.no

1 - Background

FRBRization is a metadata migration process which aims at extracting FRBR entities from MARC records.

- Crucial for the adoption of Semantic Web technologies in libraries
- Many tools proposed to perform the migration during the last decades
- No benchmark to compare and evaluate these tools

We provide two open datasets dedicated to the evaluation of FRBRization tools considering different specificities of MARC catalog like cataloguing practices, inconsistencies and bibliographic patterns.

2 – Specificities of MARC records

Cataloguing practices and inconsistencies:

- **Missing information** (missing of publication info or authoritative data leading to misunderstandings).
- **Linkage errors** (All errors in title or responsibility identifiers leading to dead links between records).
- **Cataloguing practices and norms** (Specific form of data in the record, e.g., ISBD punctuation)

Bibliographic patterns:

- **Core pattern** (basic bibliographic cases)
- **Augmentation pattern** (any addition of a Work)
- **Derivation pattern** (Intellectual modification)
- **Aggregation pattern** (whole-part relationships)
- **Complementary pattern** (other related works)

3 – Open Datasets

Including both MARC files and FRBR gold standard

- **T42** allows the evaluation of a migration tool in terms of bibliographic patterns and cataloging issues.
- **BIB-RCAT** offers a larger collection for evaluating the interpretation of MARC records in a real-world context.

http://bib-r.github.io/

Features	T42	BIB-RCAT
Number of unit tests	42	-
Number of collections	126	3
Number of languages	3	1
Number of media types	8	4
Average MARC records	10 / test	560
Average fields / records	18	17
Average FRBR entities	73 / test	1922
Average FRBR properties	241 / test	9517

4 – Extract of a unit test from T42

Example of derivation patterns in FRBR (adaptation and translations)