Open Datasets for Evaluating the Interpretation of Bibliographic Records
Joffrey Decourselle, Fabien Duchateau, Trond Aalberg, Naimdjon Takhirov, Nicolas Lumineau

To cite this version:

HAL Id: hal-01302830
https://hal.archives-ouvertes.fr/hal-01302830v2
Submitted on 18 Oct 2016
Open Datasets for Evaluating the Interpretation of Bibliographic Records

Joffrey Decourselle¹, Fabien Duchateau¹, Trond Aalberg², Naimdjon Takhirov³ and Nicolas Lumineau¹

1 LIRIS, UMR5205, Université Lyon 1
2 NTNU Trondheim, Norway
3 Westerdals - Oslo School of Arts, Communication and Technology - Faculty of Technology - Oslo, Norway
firstname.lastname@liris.cnrs.fr
trondaal@idi.ntnu.no
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 cataloging practices, inconsistencies and bibliographic patterns.

2 – Specificities of MARC records
Cataloging 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).
• Cataloging 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

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

4 – Extract of a unit test from T42
Example of derivation patterns in FRBR (adaptation and translations)