BIB-R: a Benchmark for the Interpretation of Bibliographic Records

Joffrey Decourselle, Fabien Duchateau, Trond Aalberg,
Naimdjon Takhirov, Nicolas Lumineau

07/09/2016 - TPDL, Hannover

From MARC to… FRBR

MARC Record

```
020 $c 13,5€
041 $a eng
100 $a Robert Louis Stevenson
245 $a Strange Case of Dr. Jekyll and Mr. Hyde
300 $b Colorful illustrations
```

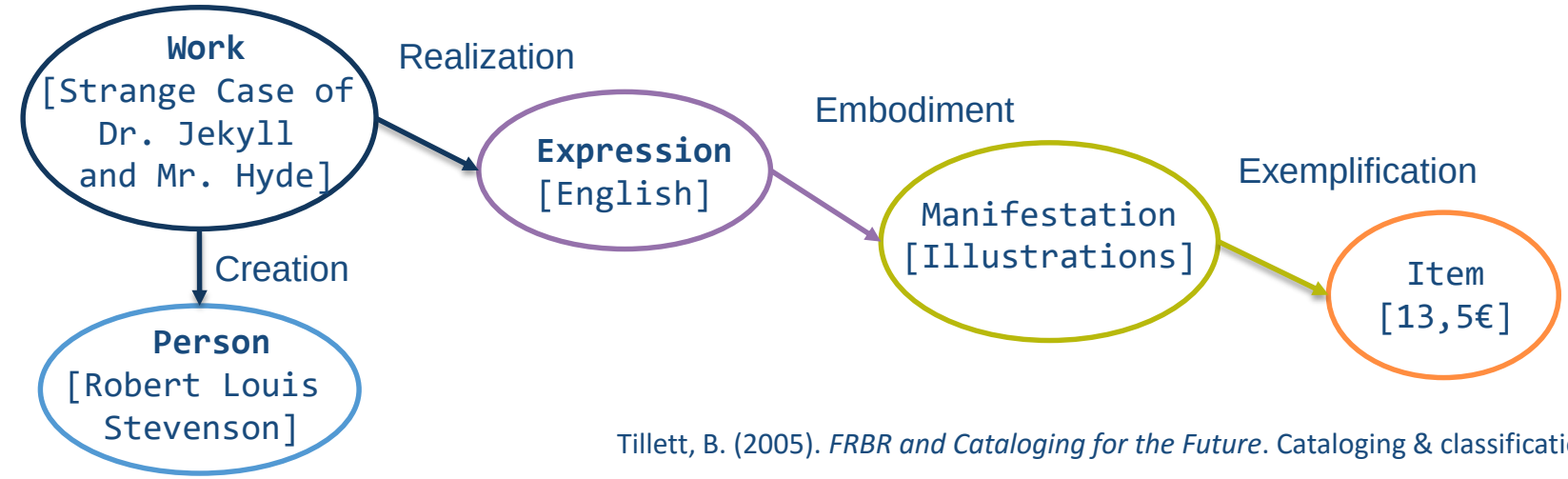
Tennant, R. (2002). MARC must die. Library Journal - New York

From MARC to… FRBR

MARC Record

```
020 $c 13,5€
041 $a eng
100 $a Robert Louis Stevenson
245 $a Strange Case of Dr. Jekyll and Mr. Hyde
300 $b Colorful illustrations
```

FRBR



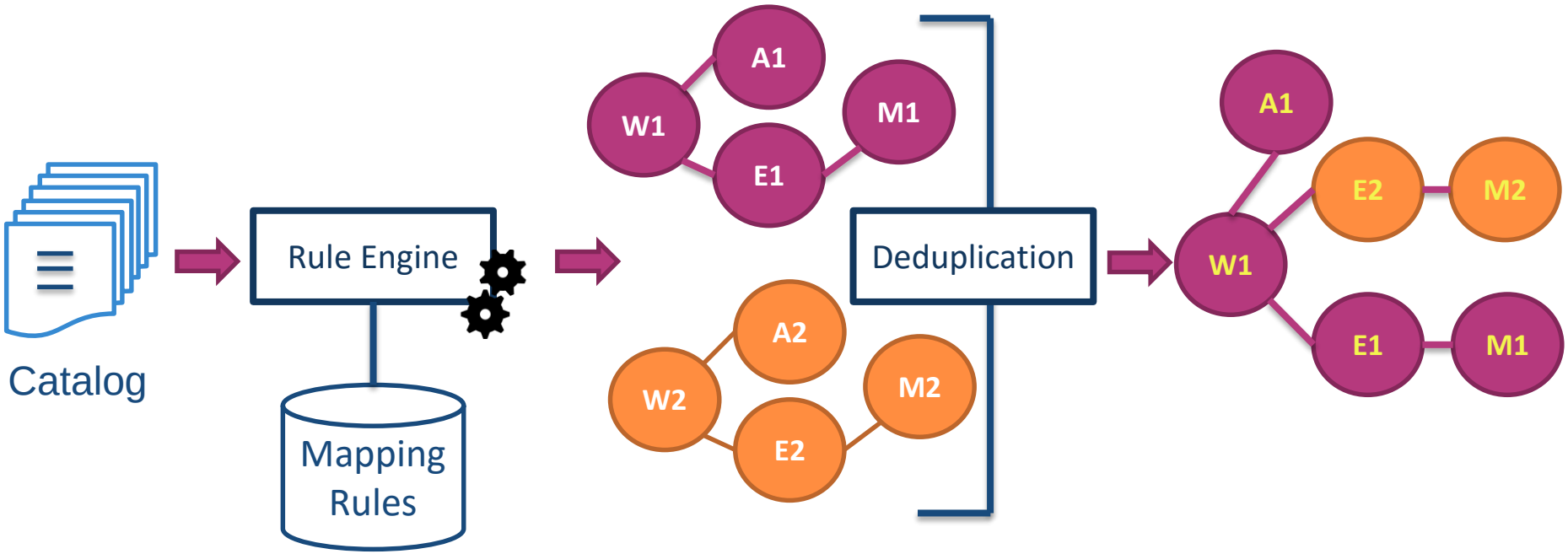
Tillett, B. (2005). *FRBR and Cataloging for the Future*. Cataloging & classification quarterly

FRBRisation process

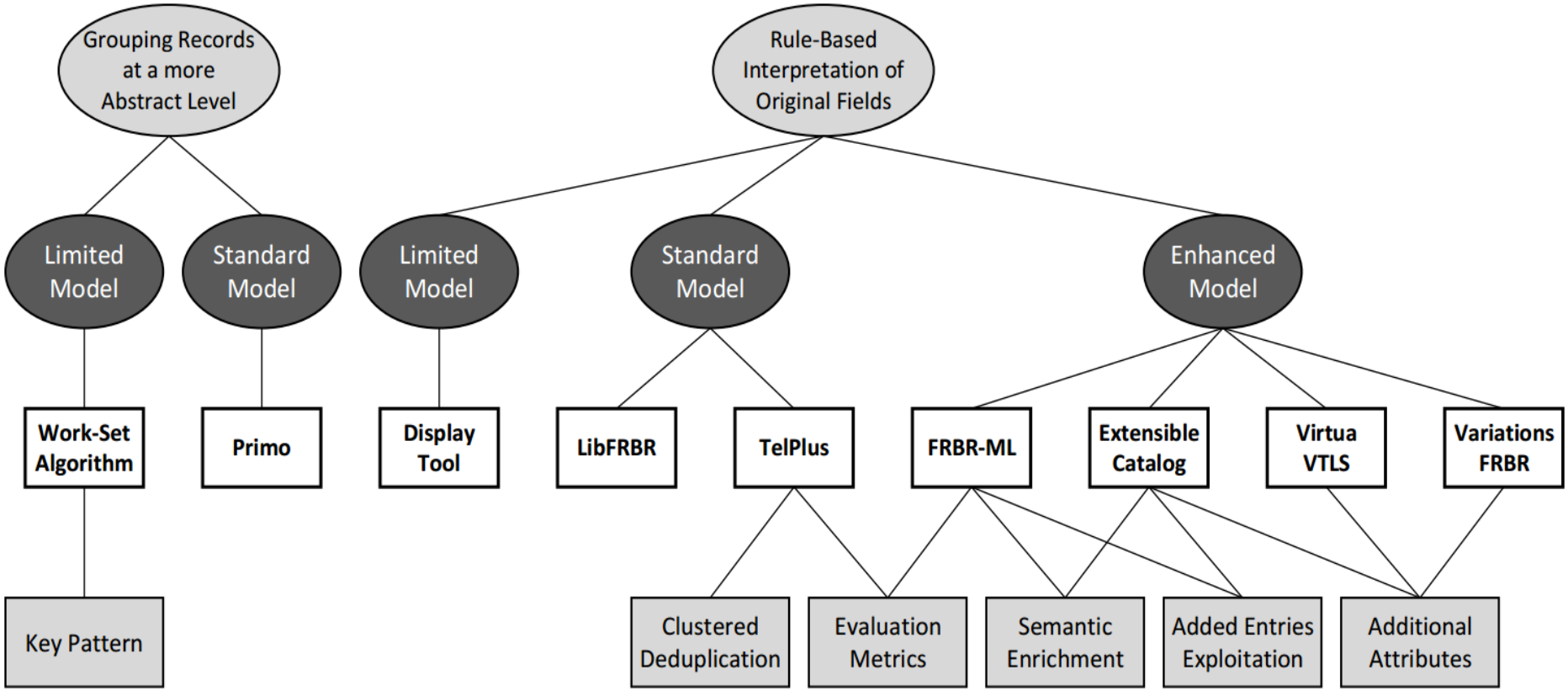
Pre-FRBRization
- Tuning
- Preparation

FRBRization
- Entity/property extraction
- Deduplication

Post-FRBRization
- Validation
- Enrichment



State of the art of FRBRization techniques



Decourselle, J., Duchateau, F., Lumineau, N. (2015). *A Survey of FRBRization Techniques*. TPDL

Related Work for evaluating FRBRisation

Process and evaluation metrics for FRBRisation

- Takhirov, N., Aalberg, T., Duchateau, F., Žumer, M. (2012). *FRBR-ML: A FRBR-based framework for semantic interoperability*. Semantic Web.

Requirements for Bibliographic records

- Manguinhas, H. M. Á., Freire, N. M. A., Borbinha, J. L. B. (2010). *FRBRization of MARC records in multiple catalogs*. In JCDL

Challenges of FRBRisation through use-cases

- Aalberg, T., & Žumer, M. (2013). *The value of MARC data, or, challenges of frbrisation*. Journal of documentation

Motivation

- **Comparison of existing solutions**
  - Need for metrics according to the bibliographic patterns
  - No qualitative comparison between tools
- **Datasets for FRBRisation**
  - Too small or simple
  - Not representative of specific FRBRisation cases

Contributions

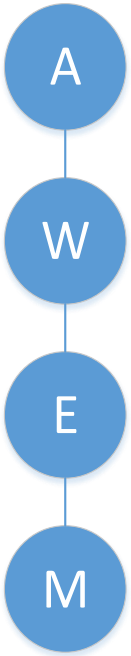
- **Definition of dedicated metrics**
  - Pre-FRBRisation (issues, cataloguing practices, …)
  - FRBRisation (rules usage, performance, …)
  - Post-FRBRisation (completeness, consistency, …)
- **Open datasets with FRBR ground truth**
  - T42 (multiple records collections focused on migration cases)
  - BIBR-CAT (larger collection representative of real work catalog)
- **Experiments on three recent FRBRisation tools**

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

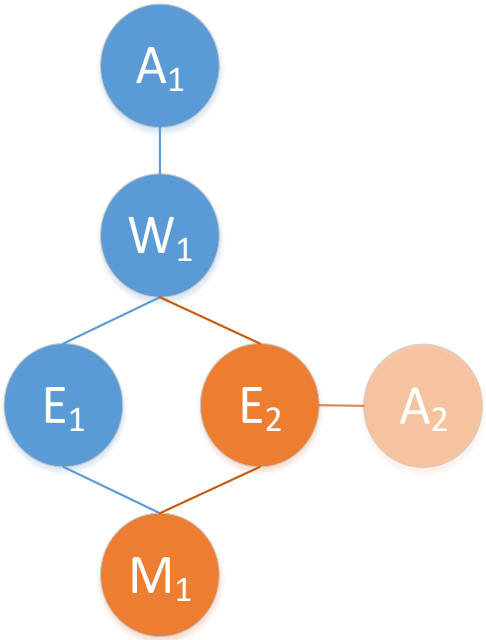
Metrics – Datasets – Experiments

Hidden bibliographic patterns in MARC

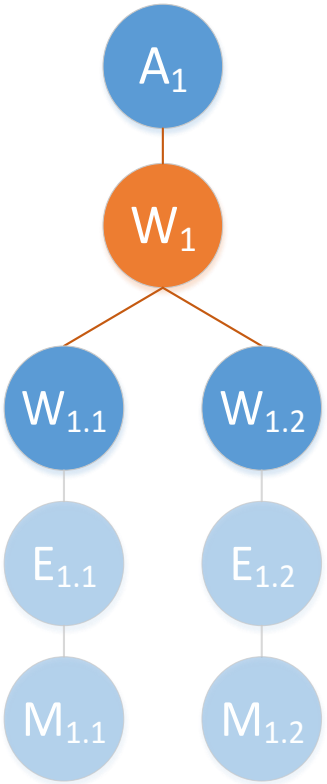
Core



Derivation



Aggregation



Riva, P. (2004). *Mapping MARC 21 linking entry fields to FRBR and Tillett's taxonomy of bibliographic relationships*. Library resources & technical services

Inconsistencies and cataloguing practices

```
101 $a no $c en
200 $a Ringenes herre = The Lord Of The Ring
$f J.R.R. Tolkien
$g trans. by Eilev Groven
210 $a Oslo ; Paris $c Tiden Norsk Forlag $d 2006
500 $a
997 $k 1543218621
```

Manguinhas, H. M. Á., Freire, N. M. A., Borbinha, J. L. B. (2010). *FRBRization of MARC records in multiple catalogs*. JCDL

Pre-FRBRisation Metrics

Metrics to compare the specificities of a catalog with the rules of a FRBRisation tool.

- **Pattern analysis**
  - COR, AUG, AGG, …
- **Inconsistencies & Cataloguing practices**
  - MID, MPD, MUT, MOT, …
- **Rules (usage, conflicts)**
  - MR, CR, …

Pre-FRBRisation Metrics (examples)

```
041 $a no $c en
100 $a J.R.R. Tolkien
240 $a The Lord Of The Ring
245 $a Ringenes herre $f
700 $a Roche, Daniel $4 trl
```

- **DER:** Percentage of records that describe a **Derivation** pattern
- **MUT:** Percentage of records where the **Uniform Title** is missing

FRBRisation Metrics

Metrics to evaluate the efficiency of a FRBRisation tool.

- **Rules application**
  - **NRT:** Number of rules applied
- **Performance**
  - **ETC:** Execution time of the entity/relationship creation
  - **ETD:** Execution time for deduplication

Post-FRBRisation Metrics

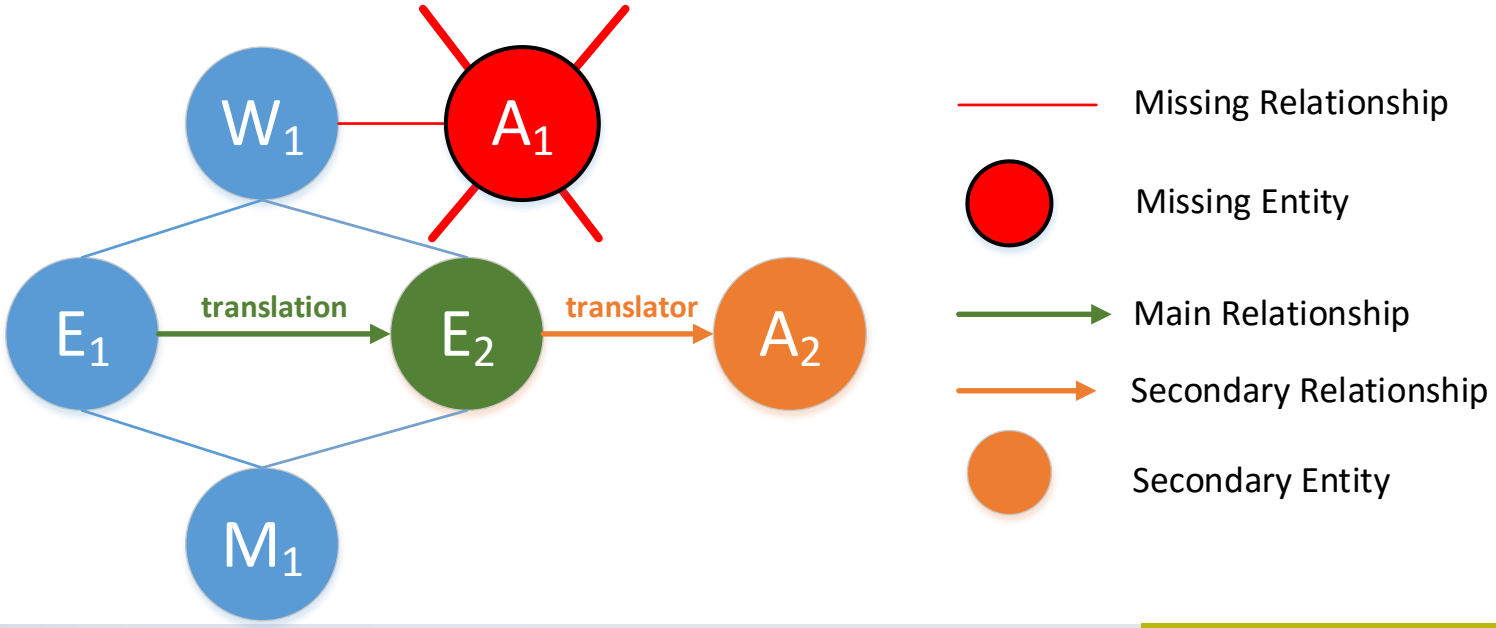
Metrics to compare the FRBRisation result with the FRBR expert.

- **Completeness**
  - MD, IAD, WSD
- **Pattern detection**
  - MEND, MRND, ESE, …

Post-FRBRisation Metrics (examples)

Example of related metrics

- **MEND**: Main entity of a specific pattern is not detected
- **MRND**: Main relationship of a specific pattern is not detected
- **ESE**: Secondary element (entity or relationship) is not detected
- **MD-E/MD-R:** Missing entity / relationship



Metrics – **Datasets** – Experiments

Datasets

- **T42**
  - 42 tests, 5 categories of bibliographic patterns
    - 1.x for Core pattern, 2.x for Augmentation, …
  - Each test combines one bibliographic pattern and one inconsistency/cataloguing practice
    - e.g., 3.5 for Derivation with Missing Uniform Title

- **BIBR-CAT**
  - One collection closer to real-world catalogs
  - Mix of bibliographic patterns and issues

Datasets

- Files provided in XML formats
- MARC21, UNIMARC & FRBR/RDA
- Hosted on GitHub: **http://bib-r.github.io/**

Feature	T42	BIB-RCAT
Number of 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 / record	18	17
Average (FRBR) entities	73/test	1922
Average (FRBR) properties	241/test	9517

Metrics – Datasets – **Experiments**

FRBRisation Tools

- **Variations VFRBR** (Indiana University)
  - Hardcoded rules

Washington, M., Notess, M., & Dunn, J. W. (2011). *Taking Music Metadata from MARC to FRBR to RDF.* International Conference on Dublin Core and Metadata Applications

- **Extensible Catalog** (Organization / Consortium)
  - Hardcoded rules (harvesting limited to OAI-PMH)

Bowen, J. B. (2010). *Moving library metadata toward linked data: Opportunities provided by the eXtensible catalog*. International Conference on Dublin Core and Metadata Applications

- **FRBR-ML** (NTNU)
  - Declarative rules

Takhirov, N., Aalberg, T., Duchateau, F., & Žumer, M. (2012). *FRBR-ML: A FRBR-based framework for semantic interoperability*. Semantic Web

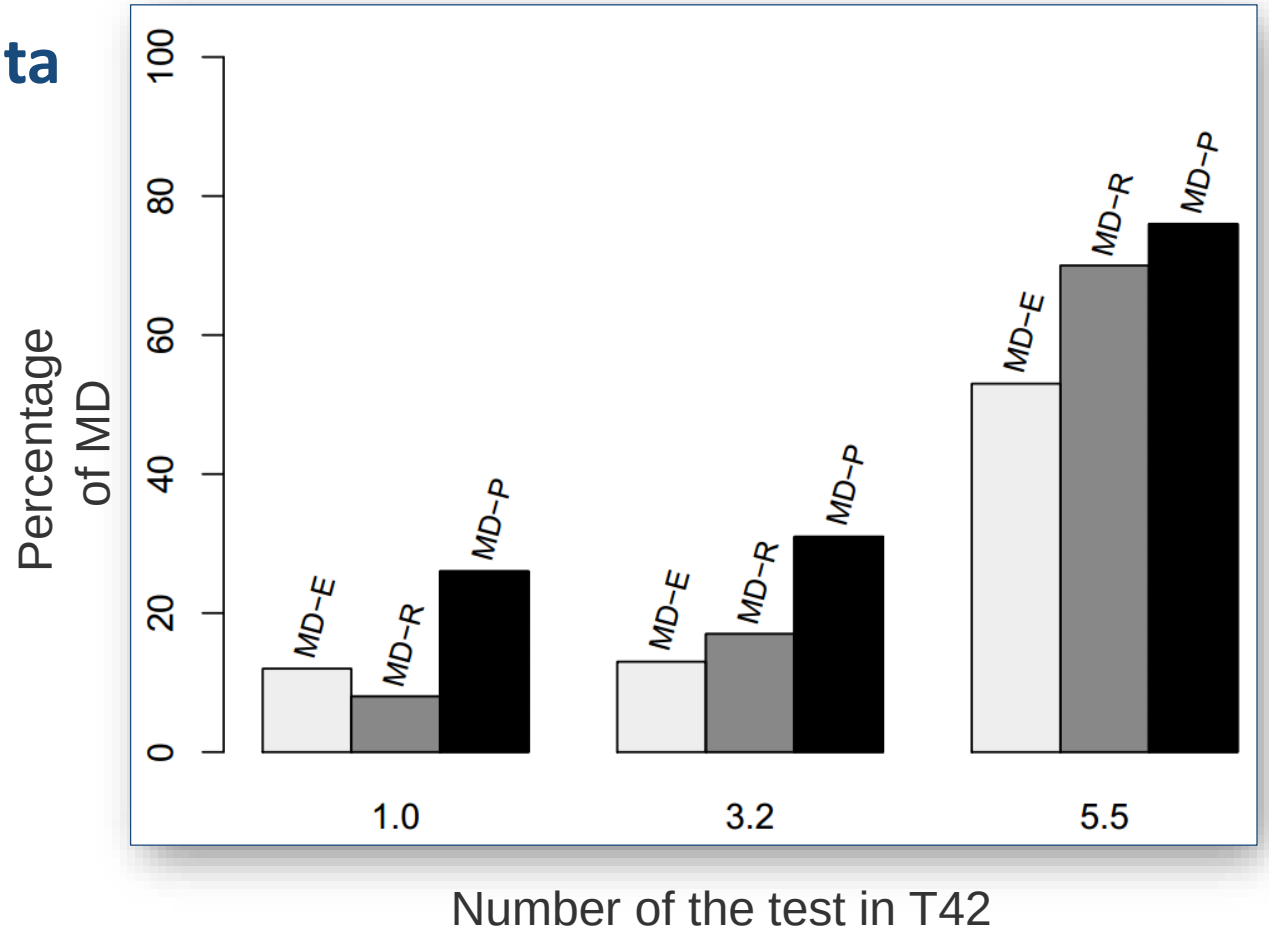
Experiments

- **Assessing strengths and weaknesses**
  - Three tools applied to the 42 tests of T42
  - Metrics from Post-FRBRization
- **Comparing tools in real-world context**
  - Three tools applied to BIBR-CAT
  - Metrics from FRBRization & Post-FRBRization
- **Facilitating the tuning**
  - Only for FRBR-ML (declarative rules) applied to BIBR-CAT
  - Tuning based on Pre-FRBRization metrics

Experiment 1 (T42)

Evaluating completeness with FRBR-ML

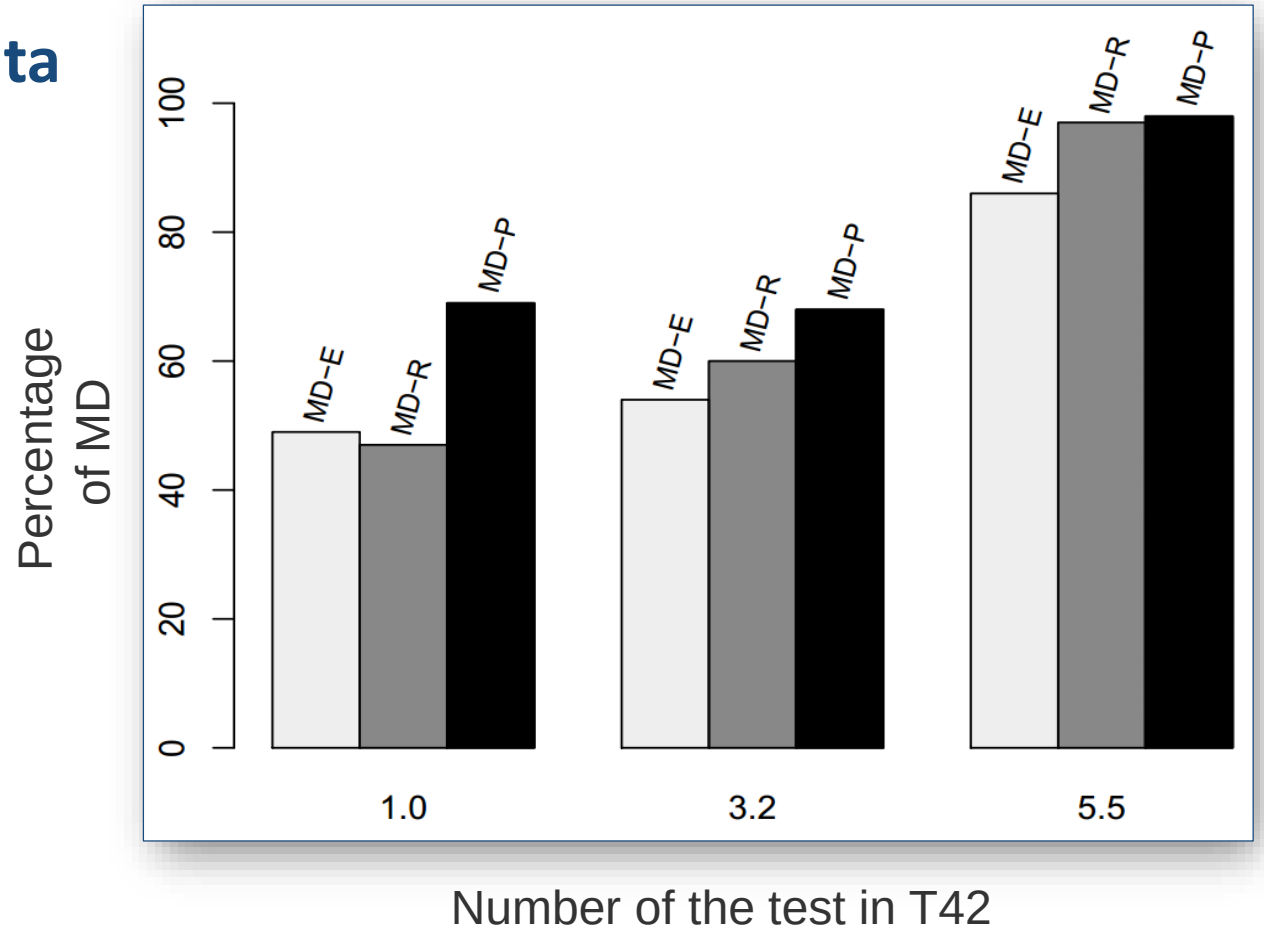
- **MD**: Missing Data
  - **E**: entity
  - **R**: relationship
  - **P**: property



Experiment 1 (T42)

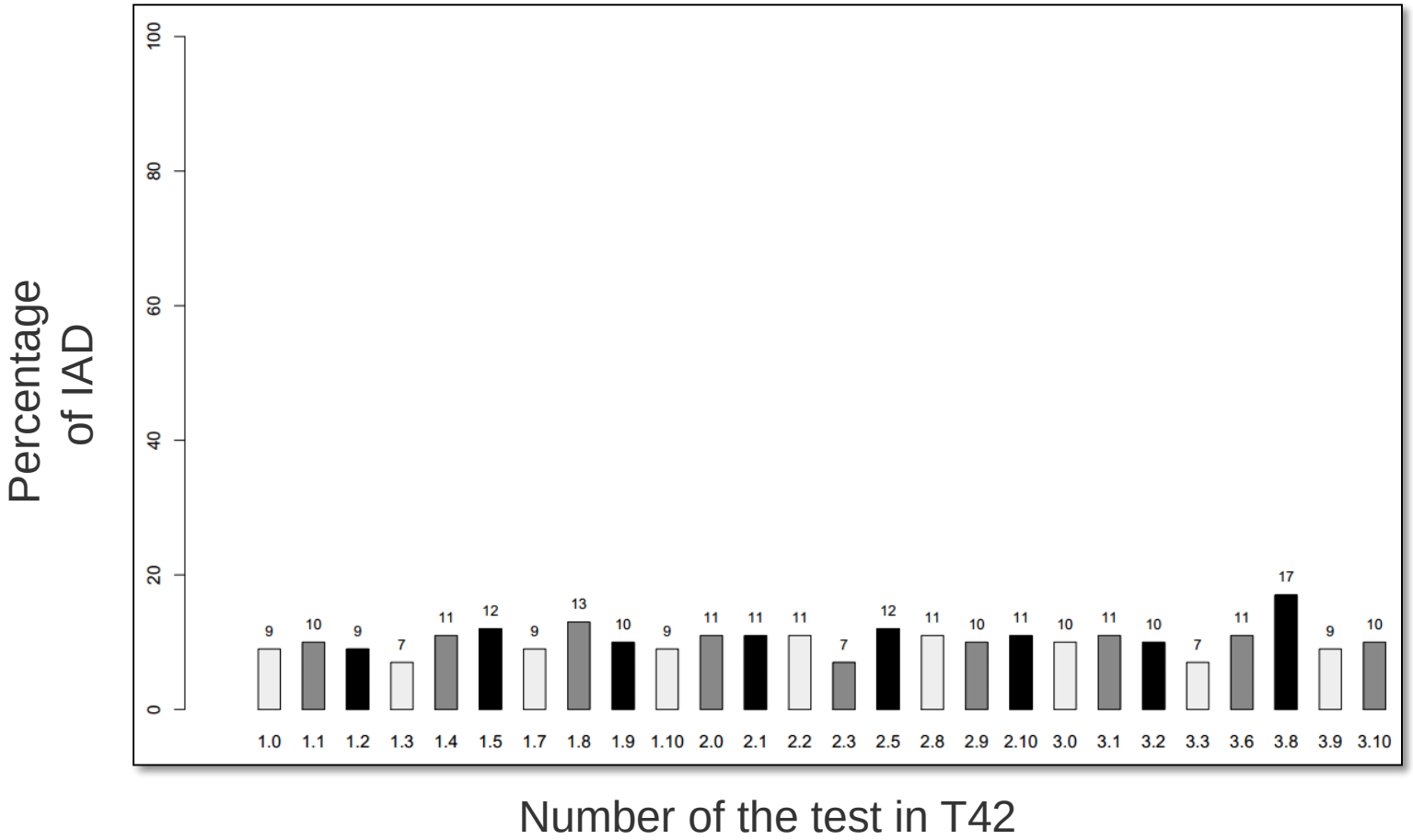
Evaluating completeness with VFRBR

- **MD**: Missing Data
  - **E**: entity
  - **R**: relationship
  - **P**: property



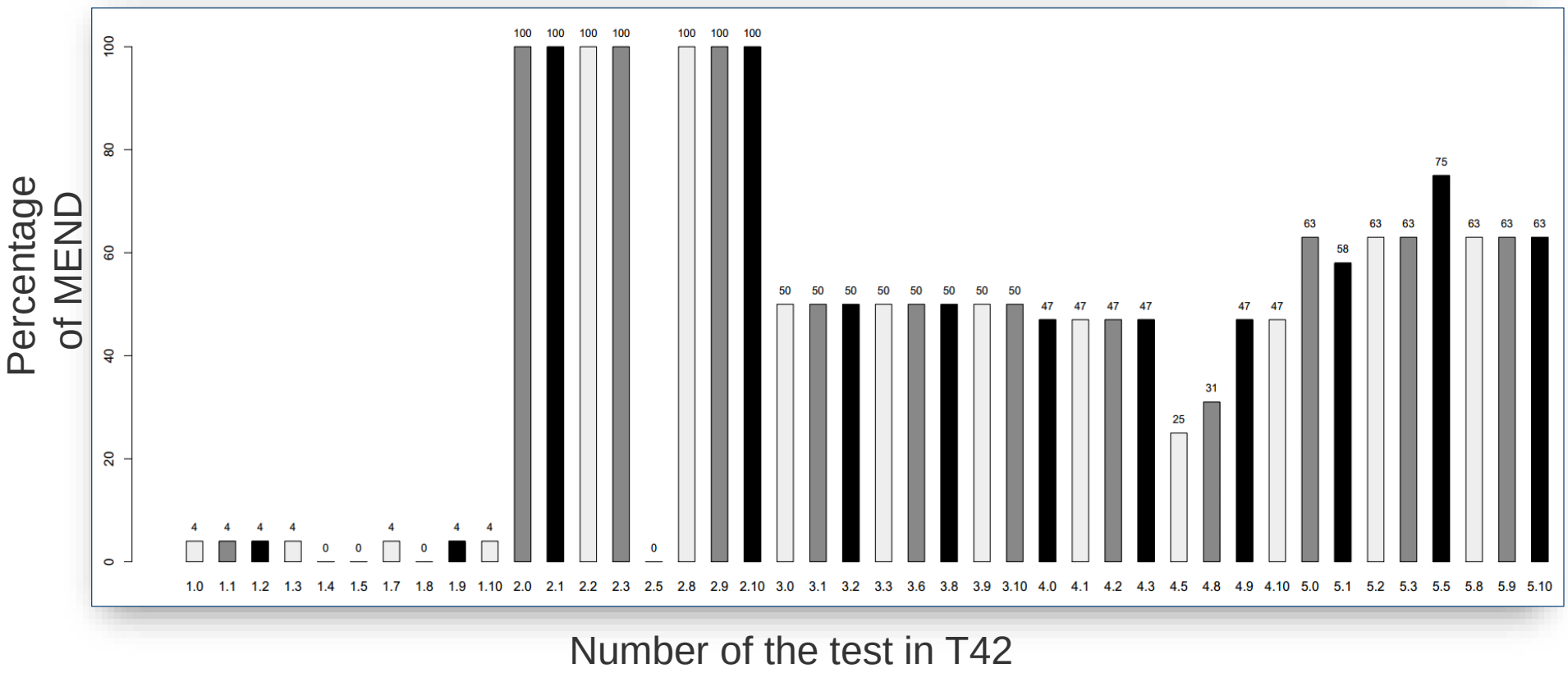
Experiment 1 (T42)

Incorrectly Added Data with Extensible Catalog



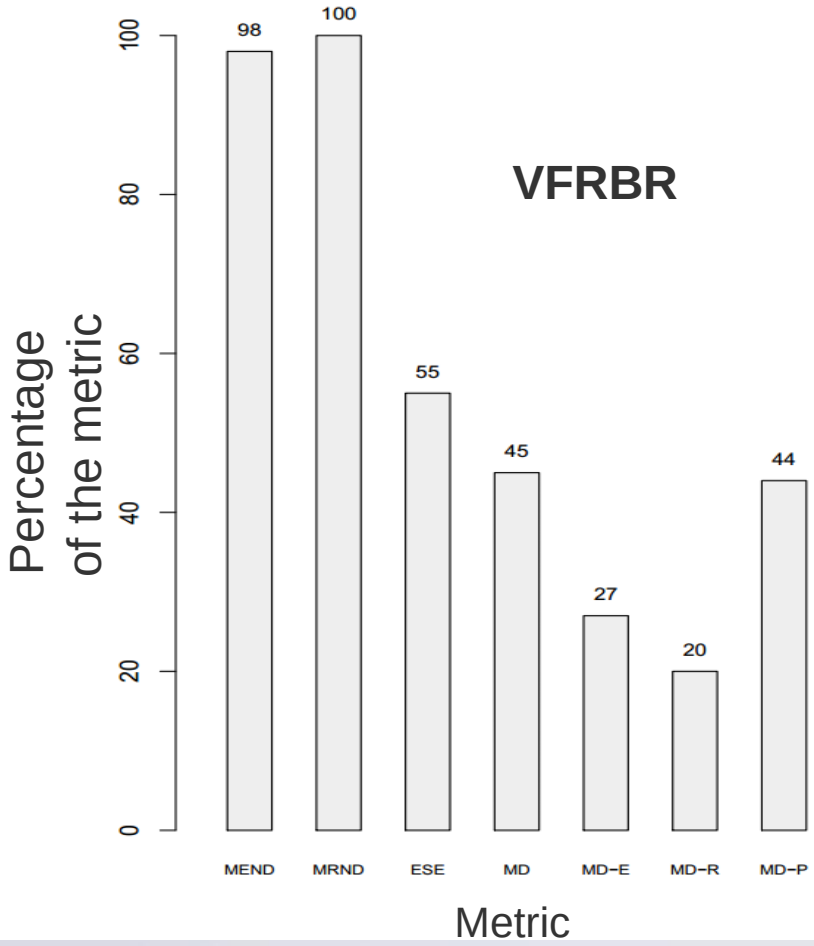
Experiment 1 (T42)

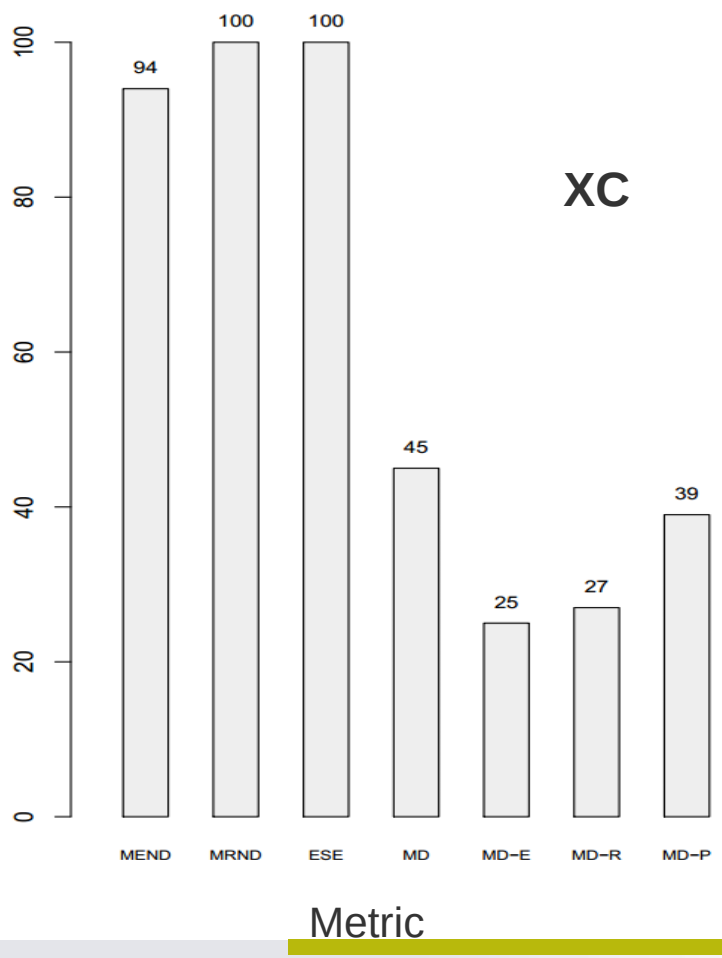
(Pattern) Main Entity Not Detected with FRBR-ML



Experiment 2 (BIBR-CAT)

Evaluation of the quality (multiple metrics)





Experiment 2 (BIBR-CAT)

- **Summary of evaluation results for the three tools**

	FRBR-ML	VFRBR	XC
MEND	94%	98%	94%
MRND	100%	100%	100%
ESE	99%	55%	100%
MD	44%	45%	45%
IAD	0%	0%	0%
WSD	0%	0%	0%
Time	2.8s	44.9s	2.8s

Experiment 3 (BIBR-CAT with tuned FRBR-ML)

- **Based on analysis feedback from pre-FRBRisation metrics**
- **Tuning performed by one expert for 4 hours**

	FRBR-ML	FRBR-ML tuned
MEND	94%	1%
MRND	100%	29%
ESE	99%	21%
MD	44%	13%
IAD	0%	0%
WSD	0%	0%
Time	2.8s	3.4s

Discussion

- **Experiments results:** http://bib-r.github.io/experiments.pdf
- **Analysis of evaluation results**
  - Limited bibliographic pattern detection
  - Difficulty to implement some metrics (e.g., IAD, WSD)
- **Keys for further improvements**
  - Enhanced tuning with pre-FRBRisation metrics
  - Detection of bibliographic patterns
  - Visualization and interactions on migration rules

Conclusion

- **BIB-R benchmark**
  - Definition of new metrics (Pre-FRBRization, FRBRization & Post-FRBRization)
  - Two open Datasets (T42 & BIBR-CAT)
  - Experimental results with VFRBR, XC & FRBR-ML
- **Ongoing works**
  - Creation of new datasets with ground truth
  - Design of a novel FRBRisation solution

Thank you !

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

To get more details about our projects:

http://liris.cnrs.fr/diricks/

syrtis by progilone

http://www.progilone.fr/en/syrtis