Library Data models under the lens of interoperability and quality

Christos Papatheodorou

Department of Archives, Library Science and Museology,

Ionian University,

Corfu, Greece

papatheodor@ionio.gr

DBIS

database & information systems group

ionian university

12th MTSR

23 - 26 October 2018

Limassol, Cyprus

Data Quality

- We are moving towards a data driven world
  - from healthcare to retail and finance, data is collected, analyzed and used to make decisions.
- Data Science involves collecting, cleansing and integrating data prior of analysis.
- The quality of data is critical and affects the outcome of all data science related tasks.

Library Data

- Any type of digital information that describes resources or supports their discovery and is produced or curated by libraries.
- Catalogues, bibliographies, vocabularies and metadata elements.
- Diversity of metadata schemas, vocabularies.
- Huge volumes of data due to aggregation services
  - Europeana, Research Data, Research Infrastructures.

Library Linked Data

- Library data published as Linked Data.
- LLD exploit the benefits of the Web standards:
  - Uniform Resource Identifiers (URIs), which identify and address resources;
  - HTTP protocol for interlinking URIs so as users can discover more resources;
  - RDF for describing and organizing the resources;
  - SPARQL to retrieve RDF data.

Current situation

- Significant efforts
  - W3C Library Linked Data Incubator Group
  - National Library of Spain
  - British Library
  - Library of Congress, Bibliographic Framework Initiative
- Library data models
  - FRBR, FRBRoo, EDM, BIBFRAME
- Low level semantic interoperability

Requirements

- Moving from Records to Entities, Properties and Relationships.
- Interlinking: improved capabilities for discovering library and non-library information resources from other trusted sources.
- Interoperability: common vocabularies and harmonization with commonly accepted conceptual models.

Questions

- How effectively could MARC records to be converted to instances of LLD models?
- Assessment of conversion process.
- Assessment of interoperability between the LLD:
  - Facilitate interlinking.
  - Facilitate explore and navigation user tasks.

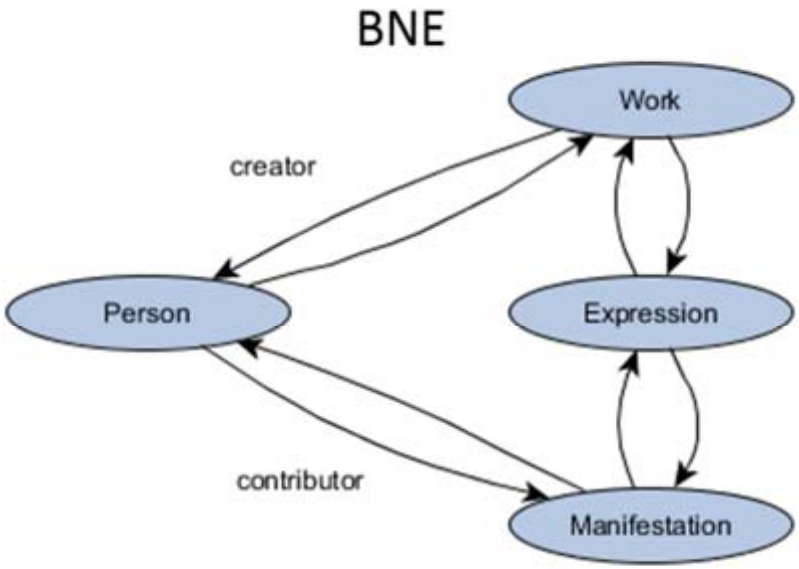
An Experiment in 2016

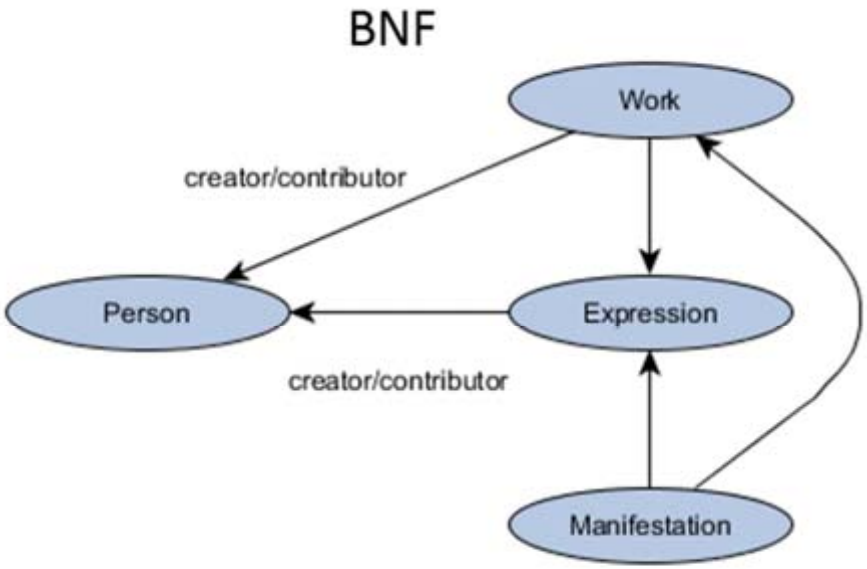
- Linked Datasets published by
  - Bibliothèque Nationale de France (BNF),
  - British Library (BNB),
  - Biblioteca National De Espana (BNE), and
  - Deutsche Nationalbibliothek (DNB).
- The data follow in general the LOD standards.

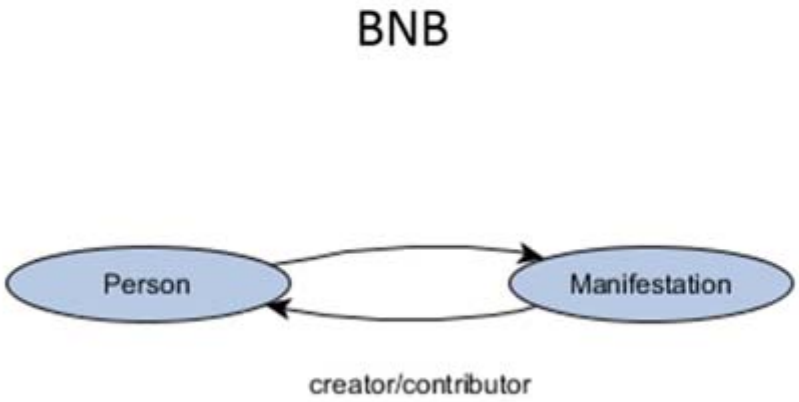
Transforming MARC records to LLD

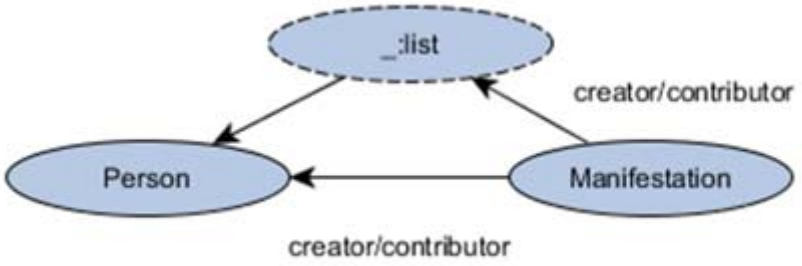
- Missing information,
- Accuracy issues,
- Inconsistencies that generate
  - matching issues in URIs and
  - different RDF graph structures.

Transforming MARC records to LLD









Transforming MARC records to LLD

- Low degree of interlinking (actual no LOD)
  - among the sets with the most links
  - fewer external links than the “top linkers” worldwide
  - isolated sets, was quite high on a general level.
- Only viaf.org is shared by all sets.
- 8 targets are shared by more than two sets, from the 28 identified across the sets.
- 3of 1,141 unique property and class terms are used by the 4 libraries (owl:sameAs, rdf:type, and dct:language), 13 terms by 3 sets, and 34 by 2.

Metadata Quality issues

- Cataloguing practices.
- Modeling and Cataloguing policies.
- Generate Interoperability issues and Interlinking issues.
- Diversity of semantics between Library Data Models: need for interoperability between them.

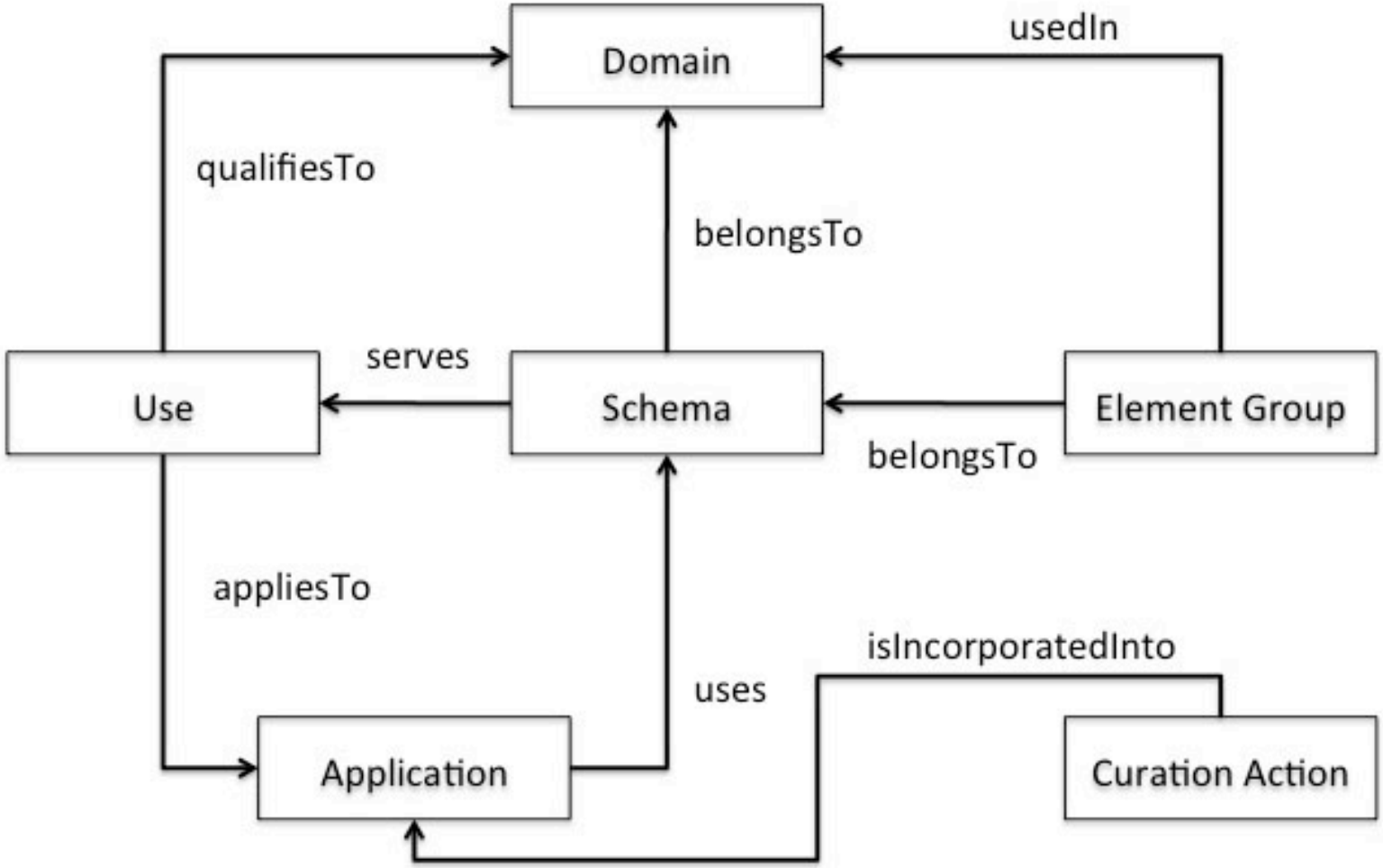
Metadata Quality

- Quality: compliance to standards.
- Metadata quality affects data discovery and retrieval and other operations and workflows that are metadata driven.
  - such as data integration
- Data quality is determined in terms of a set of specific criteria: completeness, validity, consistency, timeliness, accuracy, etc.

Metadata Quality Assessment

- Metadata quality is context dependent:
  - a metadata record that is 30% complete (according to its corresponding schema) might be of better quality for a specific application and domain than another record that is 70% complete.
- Multi-dimentionality:
  - Domain: each schema fits to more than one domain.
  - Application: the usage of a metadata schema is based on different user needs and applications.
  - The structure: the schema, the mandatory and optional elements.

Metadata Quality Assessment: Framework



Metadata Quality Assessment

Criteria: Completeness

- The percentage of completion of the elements of a schema.
  - Element groups (i=1,…,N): mandatory elements, recommended elements, optional elements.
  - Quanta (j=1,…,Q): (j_1) completeness of the mandatory set of elements, (j_2) completeness of the 'recommended' element set and (j_3) completeness of optional elements.
  - γ(d, u, a) weighting function, depends on the context classes domain (d), use (u) and application (a).

$$COMP = \frac{\sum_{i=1}^{N}\sum_{j=1}^{Q}\gamma_j(d,u,a)comp(eg_{ij})}{\sum_{i=1}^{N}\sum_{j=1}^{Q}\gamma_j(d,u,a)max(comp(eg_{ij}))}$$

Metadata Quality Assessment

Criteria: Accuracy

- How accurate is the information provided to describe a certain element.
  - An address is more accurate than a place label and less accurate than a point (encoded in latitude/longitude).
- Different accuracy quanta may correspond to each element group.
- A function of accuracy assigns values 0 or 1 to each element group that belongs to a quantum

$$ACCU = \frac{\sum_{i=1}^{N} \sum_{j=1}^{Q_i} \alpha_j(d,u,a) accu(eg_{ij})}{\sum_{i=1}^{N} \sum_{j=1}^{Q_i} \alpha_j(d,u,a) max(accu(eg_{ij}))}$$

Metadata Quality Assessment

Criteria: Consinstency

- The metadata values are consistent with the acceptable types of the metadata elements described by the metadata schema.
- The elements of a schema are used in a consistent manner across a metadata record.
  - For example, in an academic repository if the contributor element (dc:contributor) is only used to define the committee of the reviewers of a thesis.
  - Manual assessment.

$$CONS = k\frac{\sum_{i=1}^{M} cons(e_i)}{\sum_{i=1}^{M} \max(cons(e_i))} + l\frac{\sum_{i=1}^{M} \delta_i(d,u,a)}{\sum_{i=1}^{M} \max(\delta_i(d,u,a))}$$

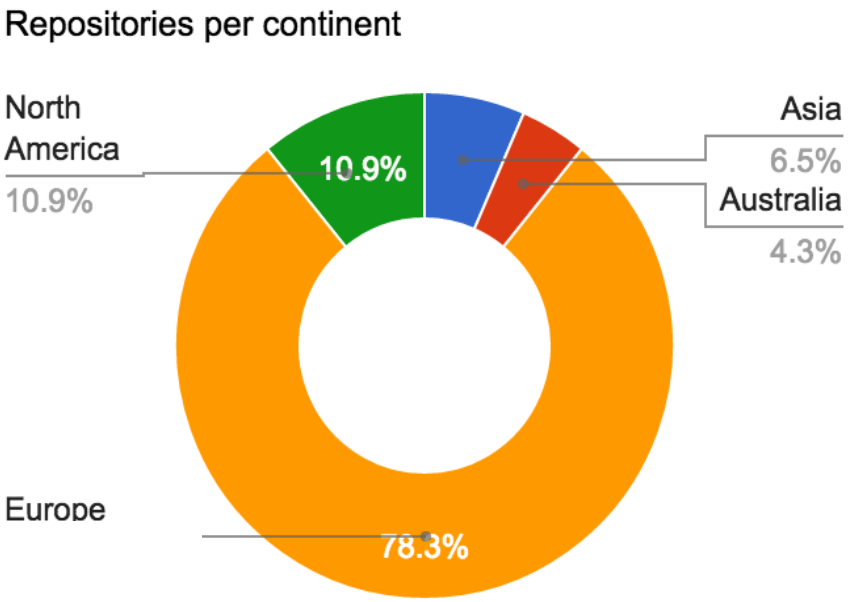
… and other contextual measures

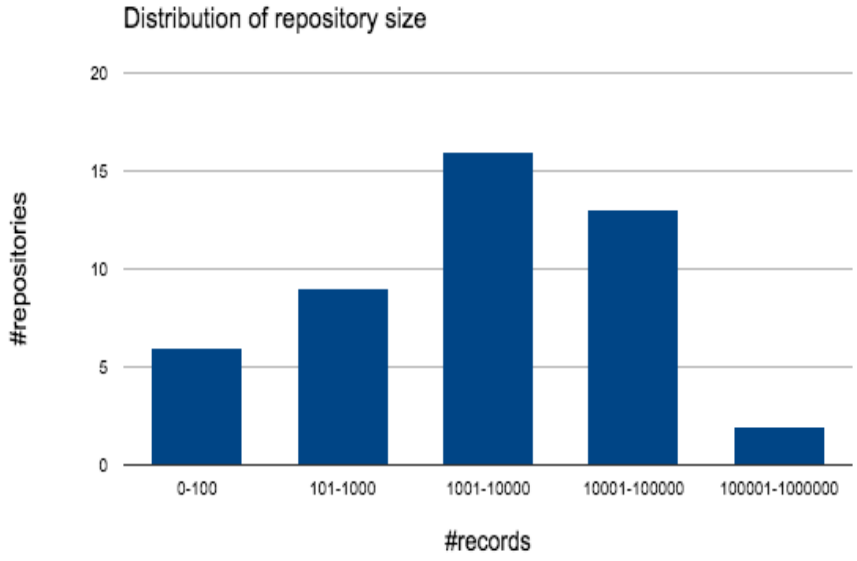
- **Appropriateness**: whether the values provided are appropriate for the targeted use.
- **Auditability**: Indicates whether the record can be tracked back to its original form.

$$QUALITY = w_1 COMP + w_2 ACCU + w_3 CONS + w_4 APPR + w_5 AUDI$$

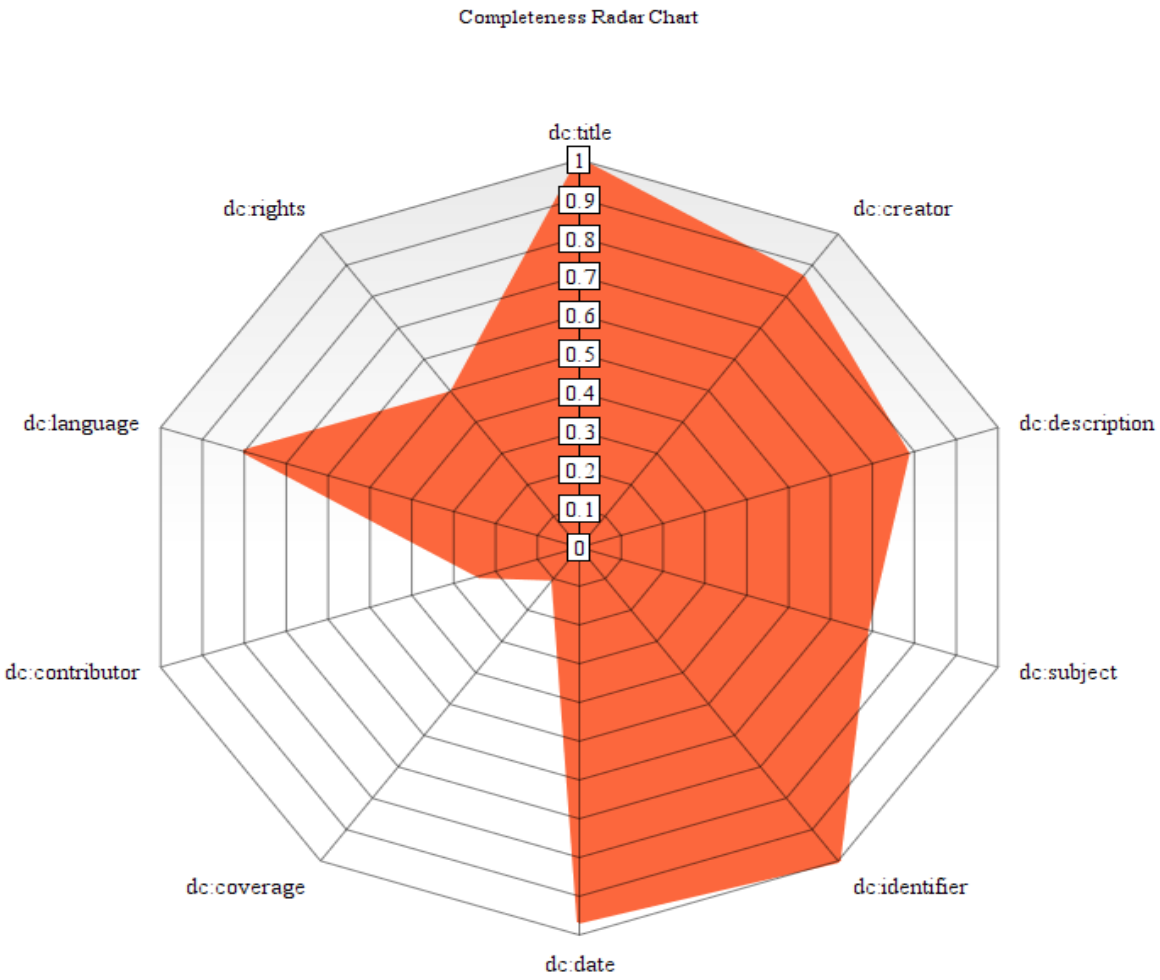
… Application

- **Repox:** http://repox.io
- 46 repositories





Completeness



Accuracy

dc:date	dc:language	dc:contributor	dc:rights
2011	English,Spanis	340: ICOM	4cc_by
Spring 2001	other	Animal Sciences	? The Authors
1917-1976 and updated	EN.	Digital Repository at the University of Maryland	Open
12-May			31/12/2100
1-May-12			This item is probably protected by Copyright Legislation

Data Quality in Europeana

- Data Quality Committee (DQC): to specify functional requirements that define the purpose of the metadata and guide data-quality evaluation.
- Multilinguality is an inherent aspect of these requirements.
  - Language of the object: access to objects in preferred language.
  - Language of the metadata: retrieval of items and determining their relevance.
- If several language tags in different languages exist, the multilingual value can be considered to be higher.

Data Quality in Europeana

- **Completeness**: The presence of fields with language tags or the presence of the dc:language field.
- **Consistency**: With regard to multilinguality, it assesses the variety of language values in the dc:language field.
  - define a standard for language notations and normalize the field in this regard.
- **Accessibility**: the degree to which multilingual information is present in the data.
  - how easy or hard it is for users with different language backgrounds to access information.

Results

- **Completeness**
  - Collection level: 904 out of 3548 collections have no value in the dc:language field
  - Record level: 58,03% of the records have a dc:language field.
- **Consistency**
  - Total values in the Europeana dataset: 33,070,941
  - Total values already normalized (ISO-639-1, 2 letter codes): 23,634,661
  - Total values already normalized (ISO-639-3, three letter codes): 4,831,534

Interoperability

- Real example.
- Don Quijote de la Mancha
  - multipart monograph
    - published in a single volume,
    - published as individual volumes too.

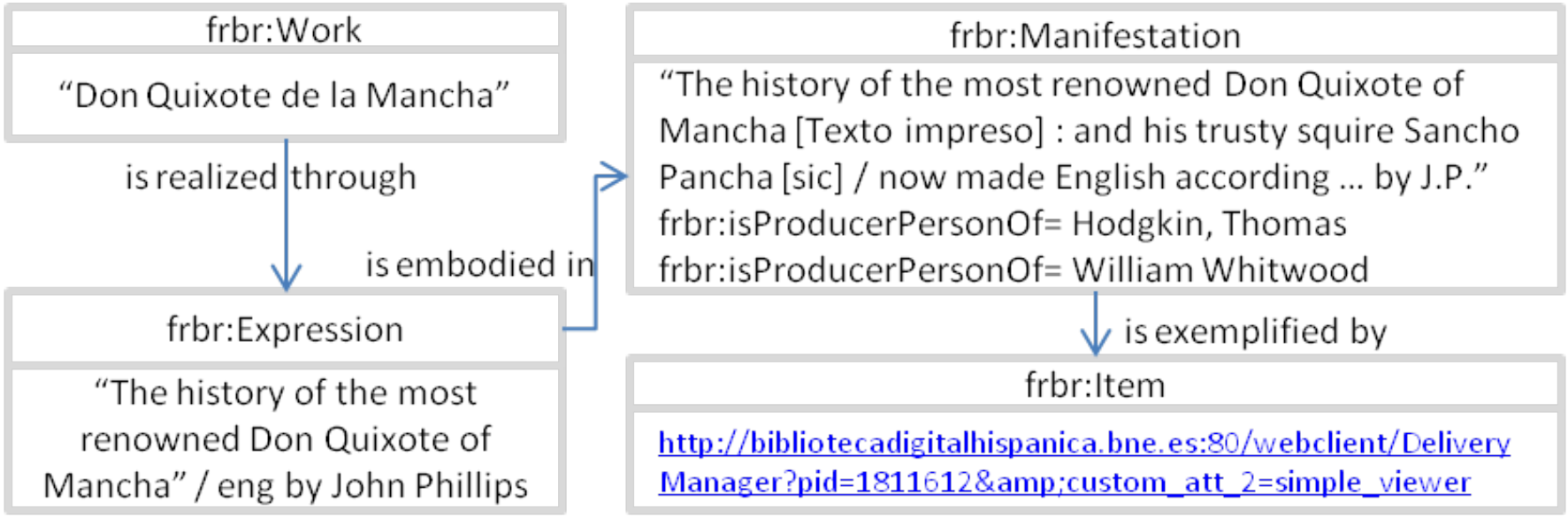
©César González via Flickr

Test case: Don Quijote de la Mancha

Table 1. Labeled version of the bima0000074081 MARCXML record

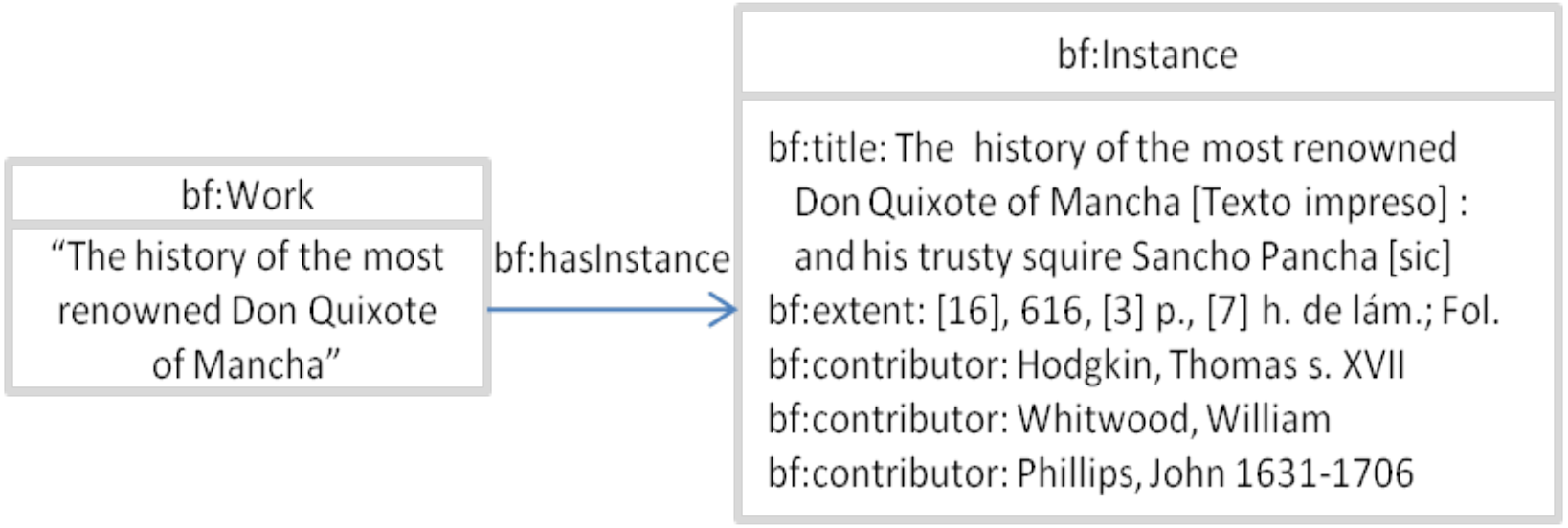
Personal name / Author	Cervantes Saavedra, Miguel de (1547-1616)
Uniform title	[Don Quijote de la Mancha. Inglés]
Title	The history of the most renowned Don Quixote of Mancha [Texto impreso]: and his trusty squire Sancho Pancha [sic] / now made English according to the humour of our modern language and adorned with seueral copper plates by J.P.
Publisher/Date	London : printed by Thomas Hodgkin and sold by William Whitwood ..., 1687
Physical description	[16], 616, [3] p., [7] h. de lám.; Fol.
Contents	Partes primera y segunda
Title note	Las iniciales J.P. corresponden al traductor, J. Philips, como consta en la dedicatoria
Added author	Hodgkin, Thomas (s. XVII), imp.; Whitwood, William, ed.; Phillips, John (1631-1706), tr.

FRBR representation



BIBFRAME representation

- Enable the transition of MARC21 data to the web of data



Interoperability: Explore

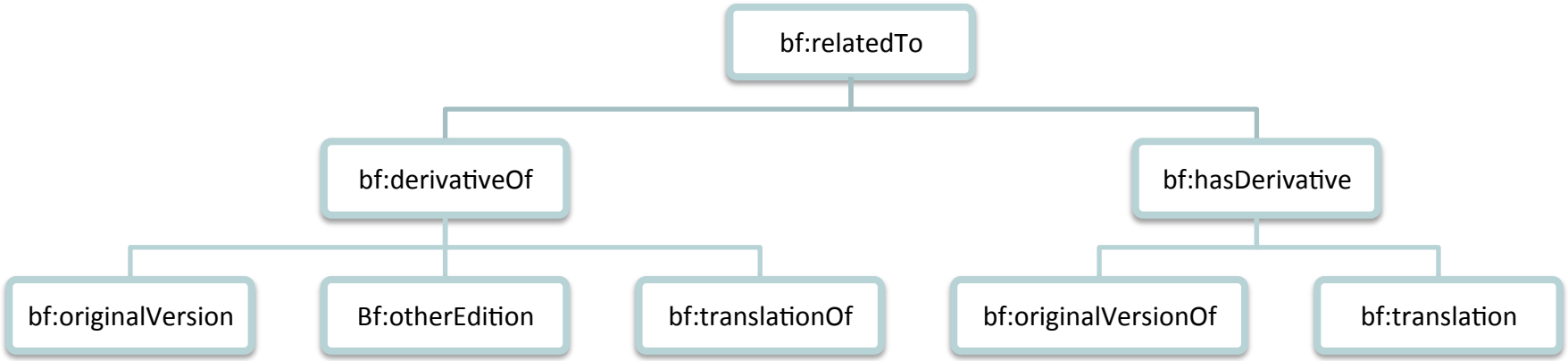
- To discover resources using the **relationships** between them and thus place the resources in a **context**.
- Content relationships
  - Equivalence
  - **Derivations**
  - Descriptive
  - Whole-part
  - Accompanying
  - Sequential
  - Shared characteristic

Complexity: Derivations in FRBR

Work to Work	Expression to Expression	Expression to Work
Summarization	Abridgement	Summarization
Adaptation	Revision	Adaptation
Transformation	Translation	Transformation
Imitation	Arrangement (music)	Imitation
	Summarization	
	Adaptation	
	Transformation	
	Imitation	

Complexity: Derivations in BIBFRAME

- Derivative relationships are mostly evolved in translations, adaptations, abridgements, dramatizations.



Interoperability

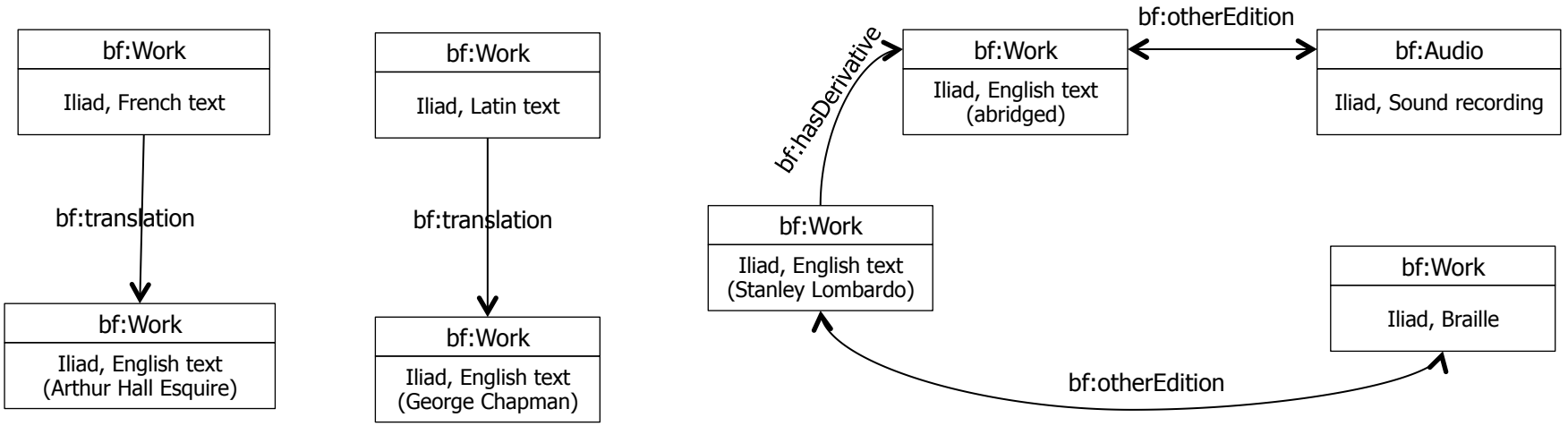
- Would be possible to map all these properties?
  - From FRBR or LRM to BIBFRAME
  - From BIBFRAME to FRBR and LRM
- Preserve Bibliographic families between models.

Bibliographic Family

- ‘***a set of related bibliographic works that are somehow derived from a common progenitor***’
  (coined by Prof. *Smiraglia)*
  - *Works or Expressions* within the same bibliographic family *may share* the *same intellectual content* and be related to the progenitor through different types of relationships
- Expresses how a Work (its ideas) is influenced by or influences other works in time

A bibliographic family in BIBFRAME

- Homer. Iliad



Mapping Evaluation Methodology

- Evaluate the efficiency of mappings using **Gold Standard datasets**
- Two Gold Standard Datasets
  - Gold FRBR
  - Gold BIBFRAME
- Mapping Gold FRBR to BIBFRAME (BIBF1 dataset)
- Compare BIBF1 against Gold BIBFRAME

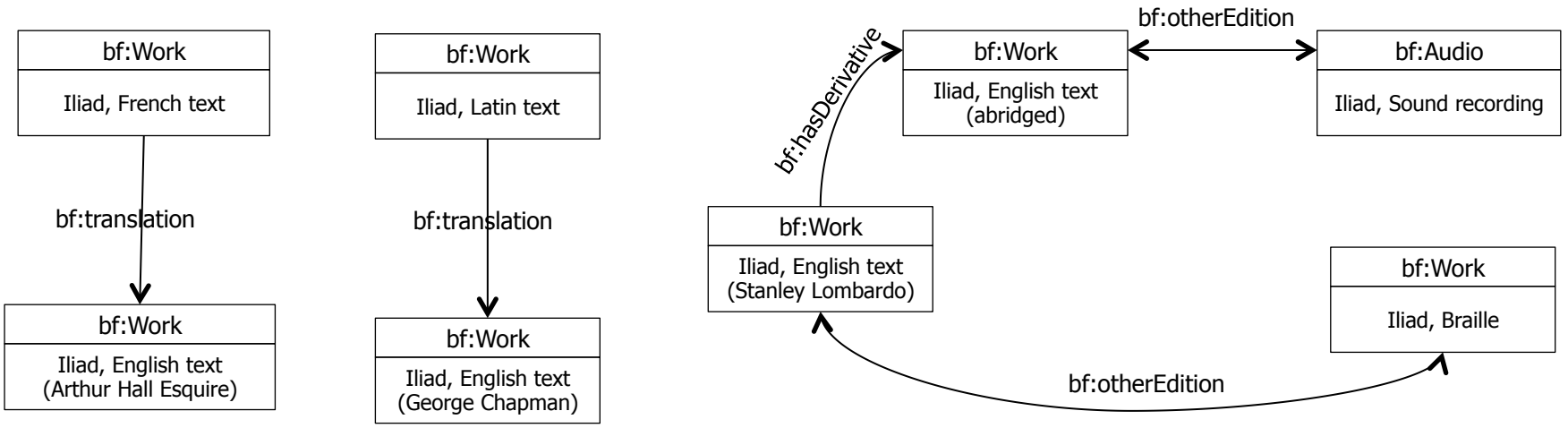
Records per family

- *Cien años de soledad* (15 records)
- *Crime and Punishment* (29 records)
- *Don Quijote* (11 records)
- *Faust* (28 records)
- *Iliad* (25 records)
- *Karamazov Brothers* (21 records)
- *Madame Bovary* (32 records)
- *Odyssey* (20 records)
- *The Scarlet letter* (24 records)
- *Tom Sawyer* (31 records)
- *Wuthering Heights* (20 records)

Total: 256 records

Development of datasets: Quality Issues

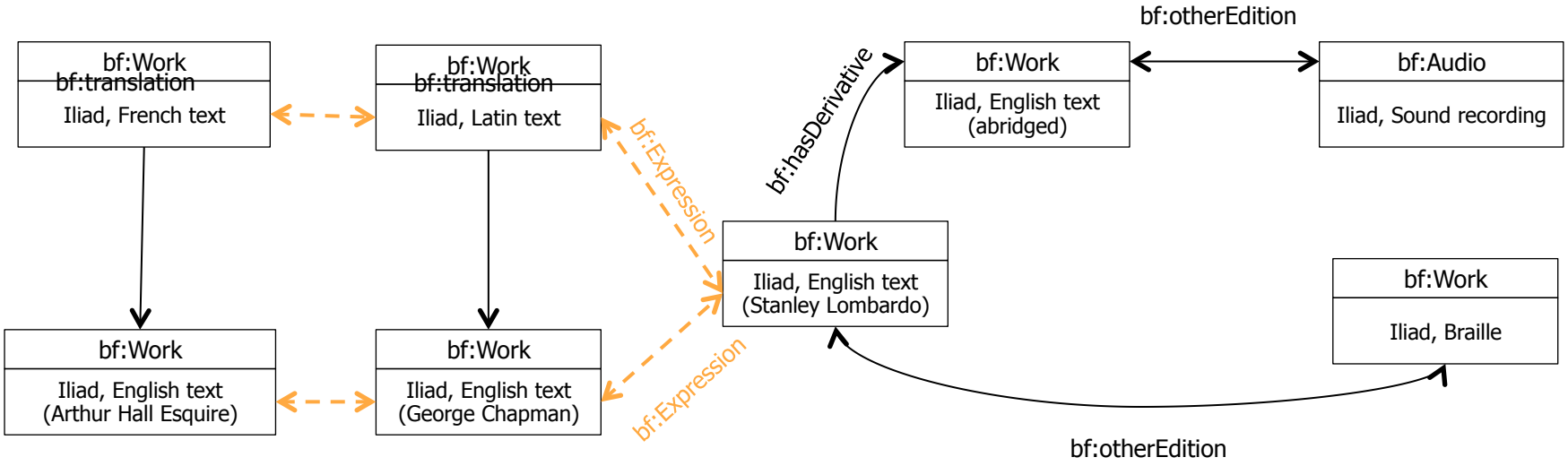
- Homer. Iliad



✘ … Some relationships are lost!

Development of datasets: Quality Issues

- Converting MARC records to BF
  - How effectively could MARC to BF conversion track down Content Relationships and Bibliographic Families?
- Assessment of conversion process



Mapping Results

- Core entities

Gold FRBR		Gold BIBFRAME		BIBF 1	
Expressions	Manifestations	Works	Instances	Works	Instances
229	257	230	257	229	257

Mapping Results

- Relationships

FRBR	BIBFRAME
Work-is realized through-Expression	bf:Work
Manifestation	bf:Instance
has a translation	bf:translation
has a revision	bf:hasDerivative
has an abridgement	
has adaptation	
has a transformation	
has an imitation	

Mapping Results

- Relationships
- Mappings with high accuracy from FRBR to BF
  - Core entities
  - Translation
  - NOT for other derivative relationships

Gold FRBR			Gold BIBFRAME		BIBF 1	
Translation (known source expression)	Literal Translation	Derivation	Translation	Derivation	Translation	Derivation
43	126	77	126	77	43	622

Conclusions

- Mappings with high accuracy from FRBR to BF
  - Core entities
  - Translation
  - NOT for other derivative relationships
- Mappings could be more accurate when preprocess MARC records or post process convertor results

Outlook

- The creation of a testbed for:
  - Hosting LLD datasets following different data models.
  - Algorithms for assessing the quality of legacy (MARC) data.
  - Algorithms for generating LLD.
  - Algorithms for mapping LLD.
  - Assessment of mapping quality: new metrics, new processes.

Bibliography

- Valentine Charles, Juliane Stiller, Péter Király, Wener Bailer, Nuno Freire, “Data Quality Assessment in Europeana: Metrics for Multilinguality”, Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Thessaloniki, Greece, September 21, 2017. CEUR Workshop Proceedings 2038, CEUR-WS.org 2018, http://ceur-ws.org/Vol-2038/paper6.pdf
- Dimitris Gavrilis, Dimitra-Nefeli Makri, Leonidas Papachristopoulos, Stavros Angelis, Konstantinos Kravvaritis, Christos Papatheodorou, Panos Constantopoulos, “Measuring Quality in Metadata Repositories”, TPDL 2015: 56-67
- Vangelis Nomikos, “Repolytics: Identifying Measurable Insights for Digital Repositories”, Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Thessaloniki, Greece, September 21, 2017. CEUR Workshop Proceedings 2038, CEUR-WS.org 2018, http://ceur-ws.org/Vol-2038/paper7.pdf
- Alireza Noruzi, “FRBR and Tillett’s, Taxonomy of Bibliographic Relationships”, https://core.ac.uk/download/pdf/33187095.pdf
- Kim Tallerås, “Metadata Structures of the Bibliographic Universe: Transformation, Interoperability, Conceptualizations, and Quality” PhD Thesis, Department of Archivistics, Library and Information Science, Oslo Metropolitan University, Norway
- Sofia Zapounidou, Michalis Sfakakis, Christos Papatheodorou, “Representing and integrating bibliographic information into the Semantic Web: A comparison of four conceptual models”, J. Information Science 43(4): 525-553 (2017)
- Sofia Zapounidou, Michalis Sfakakis, Christos Papatheodorou, “Preserving Bibliographic Relationships in Mappings from FRBR to BIBFRAME 2.0.”, TPDL 2017: 15-26

Thank you!

DBIS
database & information systems group
ionian university

12th MTSR
23 - 26 October 2018
Limassol, Cyprus