JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

JLIS.it

New ways of creating and sharing bibliographic information: an experiment of using the Wikibase Data Model for UNIMARC data

Giovanni Bergamin[(a)], Cristian Bacchi[(b)]

a) Independent scholar, http://orcid.org/0000-0001-6981-6188
b) Independent scholar, http://orcid.org/0000-0002-2912-5662

Contact: Giovanni Bergamin, giovanni.bergamin@gmail.com; Cristian Bacchi, cristian.bacchi@gmail.com
Received: 22 December 2017; **Accepted:** 1 January 2018; **First Published:** 15 September 2018

ABSTRACT

Starting from the consideration that UNIMARC (and in general the MARC) is in fact an ontology, this contribution proposes to make it explicit and to convert it – only at a syntactic level – Linked Data / RDF structures through the use of the Wikibase data model. The outcome could therefore become not only the publication of data as LOD, but also an environment for the production of bibliographic data that allows different ontological approaches.
We illustrate the possibility to achieve a restructuring of the UNIMARC record into distinct items by data type (potentially referred also to the different FRBR entities), retaining the possibility to recover all the information of the original format. Then we highlight the Wikibase solutions that become exploitable for the MARC: "usable version" of the record, with explicitation of the encoded values, and definitions connected to the data in the same system; identification of univocal data with URIs, as required in the context of the semantic web; source of the data recorded for each field; statistics on the presence of fields and subfields; new storage format natively designed for collaborative editing; export of all elements in standard RDF; support of modification via open API.

KEYWORDS

MARC; Linked Open Data; Wikidata; Wikibase Data Model; UNIMARC; Bibliographic ontologies.

CITATION

Bergamin, G., Bacchi, C. "New ways of creating and sharing bibliographic information: an experiment of using the Wikibase Data Model for UNIMARC data." *JLIS.it* 9, 3 (September 2018): 35-74. DOI: 10.4403/jlis.it-12458.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

Introduction

The objective of this work is to make a contribution to the ongoing initiatives on new ways of creating and sharing bibliographic records. The best known initiative in progress today is certainly BIBFRAME (Bibliographic Framework Initiative) which, in its objectives, also defines a possible direction of a transition path: it is not a matter of pursuing the goal of the *end of the MARC* (Tennant 2017), but of continuing to ensure in the *Web of data* the robustness of an exchange format that has assured us for over a half century (and continues to assure) the sharing of bibliographic resources (at creation stage and at access stage).[1]

The new context of transition initiatives is based on the technologies known as *Linked data* or *Linked open data*. These technologies (first defined in 2006) aim to implement in everyday life the vision of the *Semantic web,* proposed at the beginning of this century: to combine the *Web of documents* and the *Web of data*.[2]

The assumption of this contribution is that MARC ("machine readable catalog")[3] is well suited to play an important role in the *Semantic Web* that – in this context – could be defined as "machine readable and understandable web": obviously machines do not read and do not understand: they are only able to process information for practical purposes that we decide.

We should briefly recall here three terms used in the context of bibliographic information processing: *format*, *metadata schema* and *ontology*. In fact these terms from different points of view refer to the same concept: bibliographic information can be structured only by means of a given *format*;[4] *format* is defined by means of specific *metadata scheme* (i.e. a defined set of metadata); with the emergence of the *Semantic Web* and the *Linked data* the term *ontology* has replaced the term *metadata schema* (Coyle 2012, p 15). As we know the term *ontology* (in the information science context) refers to a representation model for a given domain of interest, mainly based on RDF (Resource Description Framework).[5] RDF – it is worth remembering – is a model (or a grammar[6]) for the data exchange on the web. Within these assumptions you can say for instance that MARC21 is a *de facto ontology* (beyond the fact that it is a *defined element set* or *metadata schema* that can be represented in different syntaxes: ex. g. ISO 2709, XML or JSON).

In fact, the transition paths – from MARC to the *Web of data* – have so far followed two ways: 1) *semantic restructuring of information* (not aiming to ensure round-trip format conversion); 2) *syntactic*

[1] "A major focus of the initiative will be to determine a transition path for the MARC 21 formats while preserving a robust data exchange that has supported resource sharing and cataloging cost savings in recent decades", http://loc.gov/bibframe.
[2] Seminal references: for *Linked data* (Berners-Lee 2006); for the *Semantic web* and the *Web of data* (Berners-Lee, Hendler, Lassila 2001).
[3] https://www.loc.gov/marc/faq.html#definition: "MARC is the acronym for MAchine-Readable Cataloging".
[4] Format as "structured information about an information resource" where "structured information" means that the information "must be recorded in accordance with some documented metadata scheme" (Caplan 2003. p. 3).
[5] Proposed by W3C in 1997: https://www.w3.org/TR/WD-rdf-syntax-971002/; here is the current version: https://www.w3.org/RDF.
[6] "Resource Description Framework (RDF) is the grammar for a language of data", "RDF statements follow a simple and consistent three-part grammar of subject, predicate, and object" (Baker 2013).

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

mapping of all the MARC elements (aiming to ensure round-trip format conversion). The first way has been the subject of many initiatives documented in literature (Hallo, Luján-Mora, Maté, Trujillo 2016) that have mainly dealt with the way of publishing bibliographic data on the web, with the exception of BIBFRAME which is systematically dealing also with the way to produce data. The second way has been followed by the complete mapping of MARC21 in RDF format,[7] but it does not seem that this mapping has been tested in any project. The first path starts from MARC, but programmatically pursues a change of the reference ontology, the second one is instead interested in the "translation" of MARC in the language of Web data and to maintain the starting ontology.

The contribution proposed here aims to pursue the second objective (complete mapping and ensure round-trip format conversion), taking as a starting point a *successful model* not coming from the bibliographic domain, but which has a strong impact on the way of producing, sharing and reusing information on today's web. We are referring to the Wikidata initiative and to the whole infrastructure (data models and reusable technological solutions) on which Wikidata was built.

Wikidata, Wikibase and data model

Wikidata is a *collaboratively edited knowledge base* (or *free linked database*) that can be read and edited by both humans and machines.[8] It is maintained by the Wikimedia Foundation and provides a service of centralized archiving for the "structured data" contained in the various Wikimedia projects such as Wikipedia, Wikiquote and Wikisource: it stores the information that can be represented with the triples of the semantic web (subject, property, object), such as the date of birth of a person.

The centralization of the common data is aimed at facilitating their maintenance independently of the specificity of the various Wikimedia projects (such as different languages) and facilitating their reuse in new contexts, with a Linked Open Data paradigm (Vrandečić, Krötzsch 2014).

The initiative was started by a team of researchers of Wikimedia Deutschland, with the support of the Paul Allen Institute for Artificial Intelligence, the Gordon and Betty Moore Foundation and Google.

From the point of view of user, the centralization of "structured data" can be verified by considering an example as the "Dante Alighieri" entry in Wikipedia: the pages in different languages corresponding to this Wikipedia entry, contain both discursive descriptions and "structured data";

[7] http://www.marc21rdf.info.
[8] Reference web site of the initiative: https://www.wikidata.org/wiki/Wikidata:Main_Page.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

Figure 1. "Dante Alighieri" entry in Wikipedia, in 2 different languages

these Wikipedia pages point to the same Wikidata "page" (called *Item*) of "Dante Alighieri",[9] in which all the "structured data" are collected.

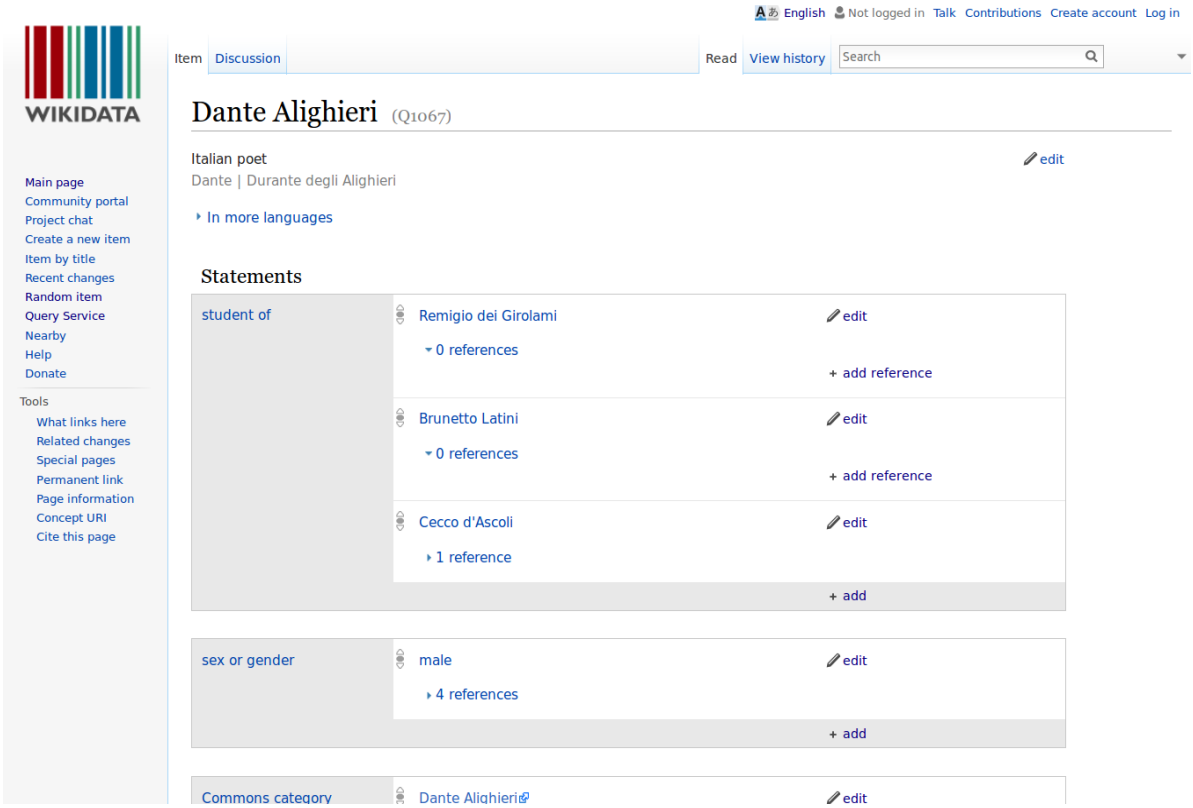


Figure 2. "Dante Alighieri" *Item* in Wikidata

[9] Wikidata *Item* about Dante Alighieri: https://www.wikidata.org/wiki/Q1067.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

The pointers to the different pages of Wikimedia projects are managed in the section *links of Item*; this section gives an idea of the role of "interconnection" played by Wikidata in the maintenance of the various Wikimedia projects: the links to pages in different languages are collected for each Wikimedia project (Wikipedia, Wikibooks, Wikinews, Wikiquote) in the same Wikidata *Item*.

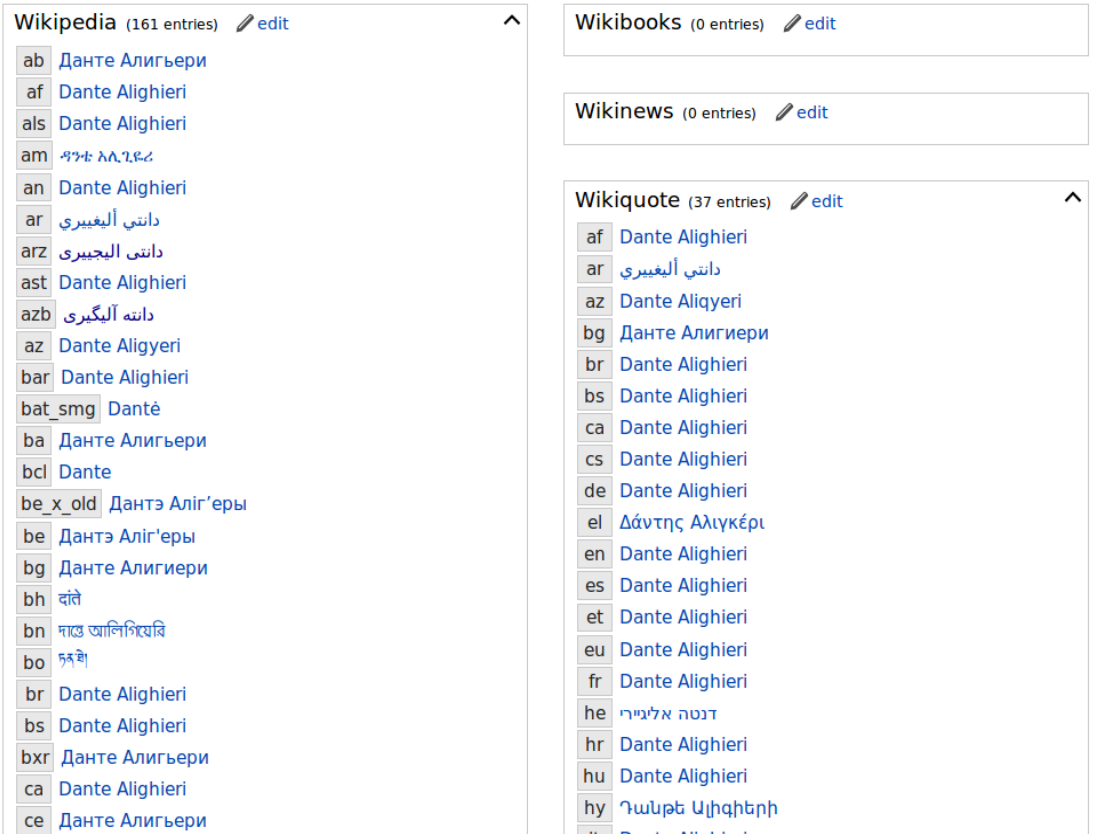


Figure 3. Link section of the *Item*

Wikibase Data Model

The Wikidata service is operated with the free software Wikibase[10] and it is based on a data model, the *Wikibase Data Model*,[11] which stores in a standard way all the information units (*Entity*), distinguishing the specific characteristics (*Property*) from the subjects described or correlated (the *Item*).

[10] Reference web site of the software: http://wikiba.se.
[11] For the study of the *Wikibase Data Model*, we used the two sources of documentation proposed by Wikimedia as "living document": https://www.mediawiki.org/wiki/Wikibase/DataModel and https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

Item and *Property* are identified by the identifier assigned by Wikibase (a progressive number preceded by a "Q" in the case of *Item,* and by a "P" in the case of *Property*) and by a different *Fingerprint* for each of the languages supported for that *Entity*: the *Fingerprint* consists of the *Label*, the *Description* and the alternative versions of *Label* (*Alias*, usually displayed with the definition "Also known as").

Dante Alighieri (Q1067)

Italian poet — edit

Dante | Durante degli Alighieri

In more languages Configure

Language	Label	Description	Also known as
English	Dante Alighieri	Italian poet	Dante Durante degli Alighieri
Italian	Dante Alighieri	poeta italiano autore della Divina Commedia	Durante degli Alighieri Dante
French	Dante Alighieri	poète, homme politique et écrivain italien	Dante
Sardinian	Dante Alighieri	No description defined	
Abkhazian	Данте Алигьери	No description defined	
Afrikaans	Dante Alighieri	No description defined	
Amharic	ዳንቴ አሊጊዬሪ	No description defined	
Aragonese	Dante Alighieri	No description defined	
Arabic	دانتي أليغييري	No description defined	
Egyptian Arabic	دانتى اليجييرى	No description defined	
Asturian	Dante Alighieri	poeta italianu	Dante

Figure 4. *Fingerprint* of the "Dante Alighieri" *Item* in Wikidata

The *Entity*, then, contains the *Statements* that host the semantic triples with which the "structured data" is recorded with *Property* and respective *Target*. The *Target* of the triple (the object) may contain a pointer to another *Entity* or a value (textual, numerical, a link, or other type of encoded data).

In the example of "Dante Alighieri" *Item*, it can be noticed that there is the reference to the place where he died: this is handled as a "structured data" (a semantic triple) that connects two *Entities*, Dante Alighieri as subject and the city of Ravenna as object, through the "place of death" *Property*.

place of death	Ravenna — edit
	2 references
	+ add value

Figure 5. *Statement* with reference to the place of death in the *Item* "Dante Alighieri" in Wikidata

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

The *Statements* are based on *Properties,* as usual in the semantic web triples. The *Properties*, like the *Items*, are identified with *Fingerprints* and "described" with *Statements*.

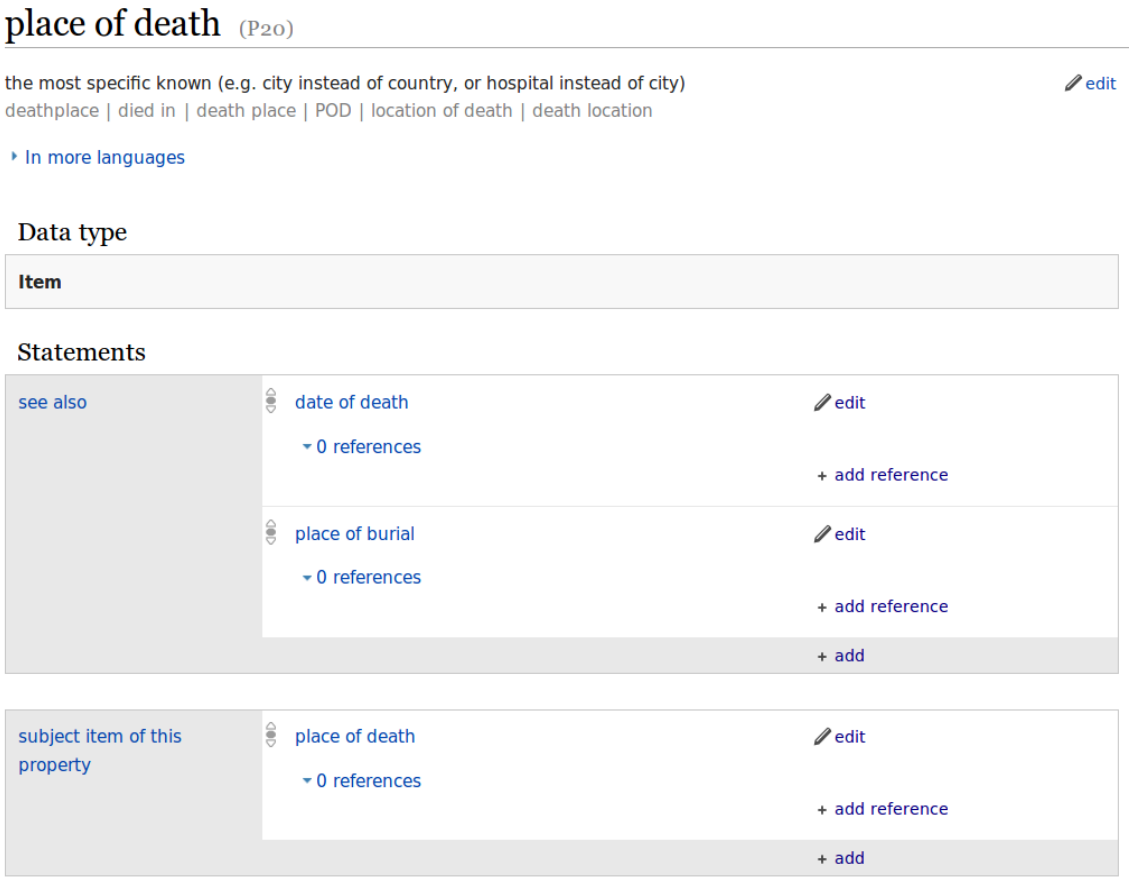


Figure 6. *Property* "Place of death" in Wikidata

The main characteristic of the *Wikibase Data Model* to be highlighted for the purposes of the idea here presented is that, within the *Statement*, the basic semantic triple (*Claim*) can be enriched with two types of information: the *Qualifier* which contains additional details about the *Statement* and the *Source* (or *Reference*) which contains the original source of the information.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

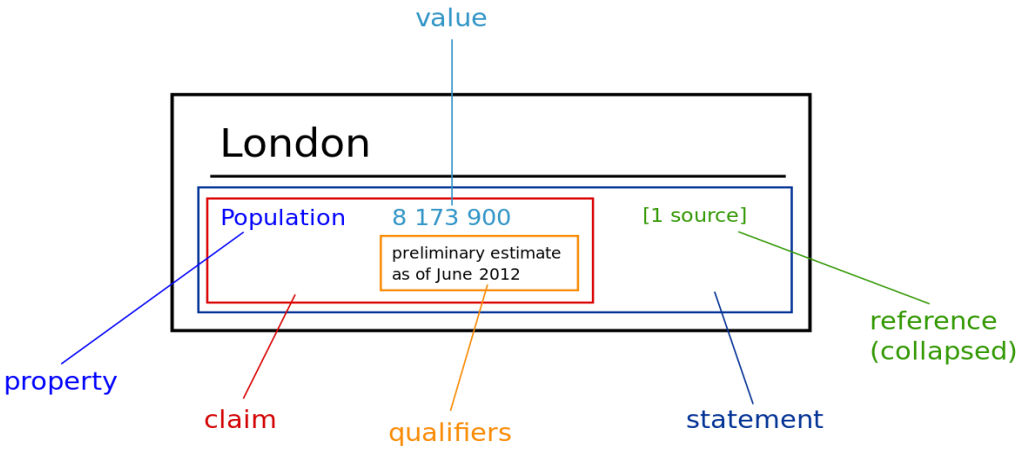


Figure 7. Graphic scheme of the *Wikibase Data Model*

In this sense, in the example of Dante's *Item,* we can note that the *Statement* with *Property* "date of birth" contains some additional specifications – the *Qualifiers* – also described with semantic triples, using specific *Properties:* "earliest date", the lowest limit for the date (*terminus post quem*) and "latest date", the maximum limit for the date (*terminus ante quem*).

Figure 8. *Statement* with reference to the date of birth in the *Item* "Dante Alighieri" in Wikidata

We can also note that this *Statement* is completed with the *Sources* from which the data is acquired (the Figure 8 shows the first two); each source is described again with a series of semantic triples: the first presents the reference to the source (with the link to the specific *Item*) and the date on which the information has been taken, the second also presents the URL of the website in which the information can be verified.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

Test of the use of the Wikibase Data Model for UNIMARC data

Taking advantage of the "enrichment" of the semantic triple implemented in the *Wikibase Data Model*, we hypothesized the possibility to replicate in this data model all the information details of the UNIMARC record, as it is exposed for example by the online catalogue of Italian library network *Servizio Bibliotecario Nazionale* (OPAC SBN).[12]

We therefore chose an SBN record of medium complexity (BID[13] IT\ICCU\CFI\0893220) that contains links to other bibliographic records and to headings (access points) for name, subject and classification, as well as terms from controlled vocabularies.

```
LEADER 05771nam0 2201813 i 4500
001 IT\ICCU\CFI\0893220
003 http://id.sbn.it/bid/CFI0893220
005 20150525130956.6
010 $a978-88-581-1170-3
020 $aIT$b2014-2498
100 $a20140514d2014 ||||0itac50 ba
101 $aita$cita
102 $aIT
181 1$6z01$ai $bxxxe
182 1$6z01$an
183 1$6z01$cnc$2rdacontent
200 1 $a<<Lo >>Stato innovatore$esfatare il mito del pubblico contro il privato$fMariana Mazzucato$gtraduzione di Fabio Galimberti
210 $aRoma$aBari$cGLF editori Laterza$d2014
215 $aXXVI, 351 p.$d21 cm
410 0$1001IT\ICCU\RAV\1797816$12001 $aAnticorpi$v43
500 10$a<<The >>entrepreneurial State$3IT\ICCU\CFI\0893222$9Mazzucato, Mariana
606 $aInnovazione tecnologica$xInterventi statali$2FN $3IT\ICCU\CFIC\227454
676 $a338.45$cINDUSTRIE SECONDARIE E SERVIZI. EFFICIENZA PRODUTTIVA$v23
700 1$aMazzucato$b, Mariana$3IT\ICCU\UFIV\143190
702 1$aGalimberti$b, Fabio$f <1972- >$3IT\ICCU\CFIV\224780
801 3$aIT$bICCU$c20160705
899 $1AL0114$2TO0 49$fP/G$qN
```

Figure 9. UNIMARC record, BID IT\ICCU\CFI\0893220

So we made an experiment with a data import in Wikibase and, consequently, with the format transformation from UNIMARC towards the *Wikibase Data Model*, using semiautomatic processes – input and manual corrections aided with batch procedures – in order to verify the following basic assumptions of the proposed idea:

- The *Properties* of Wikibase can be used to map all the structural elements of the MARC: field codes, respective indicators and subfield codes.

[12] http://opac.sbn.it/.

[13] The identifier of the UNIMARC record (label 001) will be indicated in the text with the name "BID" adopted in the SBN context, while the subfield 3 of the label block 7xx (Authority Record Number) will be indicated as "VID".

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

- The *Item* can host bibliographic records or authority records, and the entries of controlled lists (such as languages or publication countries), using
    - the *Target* of *Statement* to store the string (with its conventional punctuation, if present) as it appears in the MARC field;
    - the *Qualifiers* to store the details of the record (such as the subfields and the indicators);
    - and the *Source* to store, in each field, the source of the original record.
- The system of relations between the Wikibase *Items* can be used to support the relationships of the MARC bibliographic records with other descriptive elements:
    - other bibliographic MARC records (as in the case of multi-levels descriptions),
    - entries of controlled vocabularies,
    - authority records (such as names, subjects, classification or related titles).

UNIMARC within the Wikibase Data Model

The first operation required in the proposed perspective is the "mapping" of the semantics of the UNIMARC, which essentially means the implementation of an ontology of the standard within the *Wikibase Data Model*: the syntactic elements of the UNIMARC (record label, fields, indicators, subfields) and the coded data (language codes, countries) must be appropriately distributed between *Properties* and *Items* of Wikibase, with a detail and a relational logic that take into account on the one hand the need to fully preserve the original information details and on the other hand the possibility to reuse the data exploiting the functionalities of the new system.

Property: UNIMARC structure

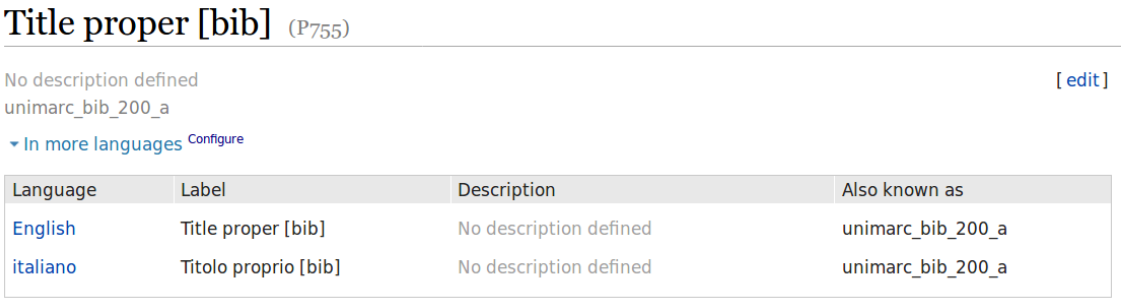
In the experiment we created the *Properties* for the structural elements of UNIMARC (fields, subfields and indicators) which can be found in the bibliographic and authority records we selected; the *Fingerprint* of the *Property* has been set according to the following criteria:

- the extended name of the UNIMARC element has been stored into the *Label* of the *Property*, translating it into multiple languages;
- the codes of the UNIMARC elements have been stored in the *Aliases*, according to a naming scheme that collects the identifying marks, from the most general to the most specific, separated by underscore.

For example: "unimarc_bib_200_a" for the subfield "a" (Title Proper) of the UNIMARC 200 field (Title and Statement of Responsibility)

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458



Title proper [bib] (P755)

No description defined [edit]
unimarc_bib_200_a
In more languages Configure

Language	Label	Description	Also known as
English	Title proper [bib]	No description defined	unimarc_bib_200_a
italiano	Titolo proprio [bib]	No description defined	unimarc_bib_200_a

Figure 10. *Property* corresponding to the "Title proper" subfield of the UNIMARC 200 field

In order to maintain the uniqueness of the *Labels*, we introduced the suffix [bib] in the *Property Labels* for the UNIMARC bibliographic record, and the suffix [auth] (or [aut] for Italian) in those for the UNIMARC record of the authority entries.

Date entered on file [bib] (P659)

No description defined [edit]
unimarc_bib_100_a_0-7
In more languages Configure

Language	Label	Description	Also known as
English	Date entered on file [bib]	No description defined	unimarc_bib_100_a_0-7
italiano	Data di immissione nella base dati [bib]	No description defined	unimarc_bib_100_a_0-7

Figure 11. *Property* corresponding to the characters positions 0 to 7 of the UNIMARC 100 field, subfield "a"

Flexibility of the Wikibase Data Model

The highly flexible nature of the *Wikibase Data Model* makes it possible (and also necessary) to avoid a translation of the UNIMARC syntax into a list of all possible combinations of its component parts (field, indicators, subfield):[14] we, therefore, exploited the opportunity offered by Wikibase to qualify the elements with a system of relations between the *Entities*.

For example, dealing with the field UNIMARC 200 (the title area), in order to account for the different cases of "significant title" or "not significant title" given by the value of its first indicator ("Title

[14] Here we refer to the mapping work "marc21rdf" – http://www.marc21rdf.info/ – created by Gordon Dunsire (as mentioned in the introduction) where, for example, for the 2xx fields of the standard "Properties [are created] representing every combination of MARC 21 tags 210 to 264, first and second indicators, and their subfield codes".

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

Significance Indicator"), we didn't map the entire field with multiple *Properties*: we separately mapped the first indicator, instead, with the specific "Title Significance Indicator [bib]" *Property*;

Title Significance Indicator [bib] (P1250)

No description defined [edit]
unimarc_bib_200_ind1
In more languages Configure

Language	Label	Description	Also known as
English	Title Significance Indicator [bib]	No description defined	unimarc_bib_200_ind1
italiano	Indicatore di titolo significativo [bib]	No description defined	unimarc_bib_200_ind1

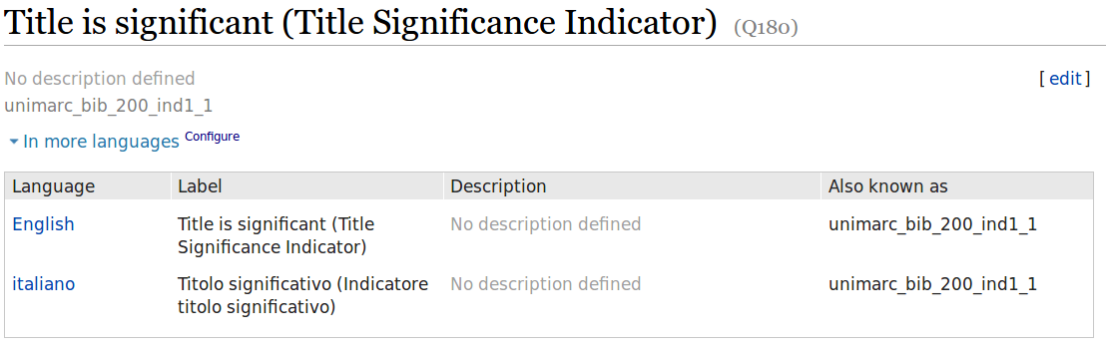
Figure 12. *Property* corresponding to the first indicator of the UNIMARC 200 field

the *Property* for the first indicator is used in a *Qualifier* of the *Statement* for the UNIMARC field 200, and its value allows us to know whether the title is significant or not.

Item: controlled vocabularies

To complete the mapping of UNIMARC semantics in the *Wikibase Data Model*, therefore, the terms of the controlled vocabularies of the standard have to be imported in apposite *Items*.

In the experiment we loaded the coded-data used within the selected UNIMARC records; continuing the example of the UNIMARC field 200, we imported the possible values for the first indicator: one of them is "Title is significant".



Title is significant (Title Significance Indicator) (Q180)

No description defined [edit]
unimarc_bib_200_ind1_1
In more languages Configure

Language	Label	Description	Also known as
English	Title is significant (Title Significance Indicator)	No description defined	unimarc_bib_200_ind1_1
italiano	Titolo significativo (Indicatore titolo significativo)	No description defined	unimarc_bib_200_ind1_1

Figure 13. *Item* corresponding to the value "Title is significant" for the first indicator of the UNIMARC 200 field

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

Item: UNIMARC record

Once we loaded into Wikibase all the *Entities* needed to represent the UNIMARC structure of the chosen sample of data, we experimented the loading of the bibliographic and authority records into Wikibase *Items*, using all the different kinds of semantic triples supported into the *Statements* of the *Wikibase Data Model*:

- the entire UNIMARC fields (with their conventional punctuation, if present) have been stored in the *Targets* of the entire *Statements*,
- while the indicators and the subfields have been stored in the *Qualifiers*.
- The *Sources* host, for each occurrence of the UNIMARC fields, the pointers to the sources of the original records.

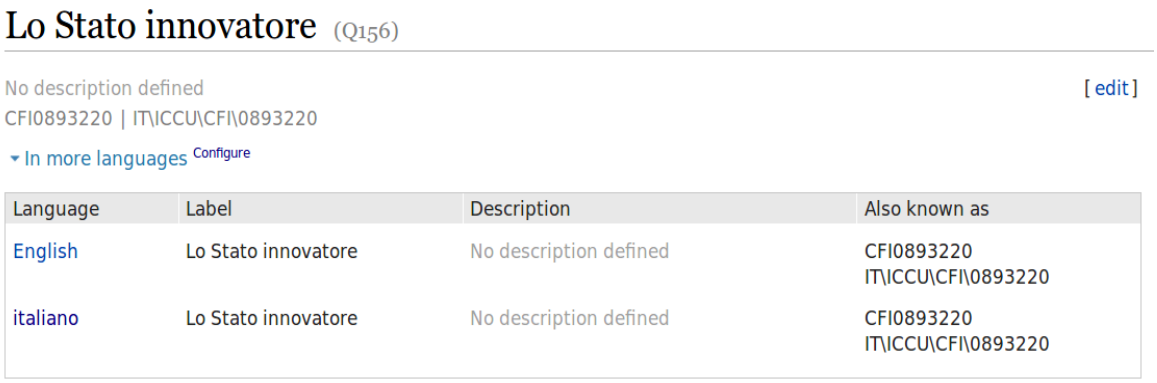
Figure 14. *Statement* corresponding to the UNIMARC field 200 of the record BID IT\ICCU\CFI\0893220

In the *Items* corresponding to UNIMARC bibliographic records, the identifier assigned in the original system (the BID of SBN) has been stored in the *Alias* of the *Item*; this way, even after uploading the record to Wikibase, a continuity in how to identify and retrieve the data can be guaranteed.

The distribution of this information between the *Label* and the *Aliases* (both for bibliographic records and for entries of controlled vocabularies or authority lists) follows the practices of the *Wikibase Data Model*, in which the "main label" is used to describe *Entity* in various languages, while the *Alias* is mainly used for research and to provide alternative access points: in this sense, the *Alias* is also more suitable to host multiple alphanumeric identifiers if necessary.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458



Lo Stato innovatore (Q156)

No description defined [edit]

CFI0893220 | IT\ICCU\CFI\0893220

In more languages Configure

Language	Label	Description	Also known as
English	Lo Stato innovatore	No description defined	CFI0893220 IT\ICCU\CFI\0893220
italiano	Lo Stato innovatore	No description defined	CFI0893220 IT\ICCU\CFI\0893220

Figure 15. *Fingerprint* of the *Item* corresponding to the entire UNIMARC record BID IT\ICCU\CFI\0893220. The SBN record identifier ("BID") is registered in the *Alias*

Each *Statement* corresponding to a UNIMARC field contains the *Source* from which data has been taken. The following details has been stored with appropriate *Properties*:

- the institution, with a pointer to the specific *Item* created in Wikibase,
- and the pointer to the original record, with a URL to the source

Publication, distribution, etc. [bib]

Roma ; Bari : GLF editori Laterza, 2014 [edit]

Place of publication, distribution, etc. [bib]	Roma Bari
Name of publisher, distributor, etc. [bib]	GLF editori Laterza
Date of publication, distribution, etc. [bib]	2014
MARC subfields order	a[1] a[2] c[1] d[1]

1 reference [edit]

Imported from	National Central Library of Florence
official web site	http://purl.org/bncf/marc/CFI0893220

[add reference]

[add]

Figure 16. *Statement* corresponding to the UNIMARC field 210 of the record BID IT\ICCU\CFI\0893220

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

The original order of subfields is stored in an apposite Qualifier

The UNIMARC standard establishes, among its fundamental characteristics, that fields are possibly repeated and that also the subfields may possibly be repeated within each occurrence of a field. Less obvious is the situation in which the subfields are repeated intermingled with other subfields, and still less obvious (and never explicitly permitted or prohibited by the standard) is the situation in which a subfield is repeated containing the same value.

An example[15] that presents both problems can be found in the title area (UNIMARC field 200) of the following record (BID IT\ICCU\URB\0620565), where we can see the presence of the subfield $f repeated twice with the same value and in different positions:

```
200 1 $aPour les valeurs bourgeoises$fpar Georges Hourdin$cContre les valeurs
bourgeoises$fpar Georges Hourdin
```

Both needs – to repeat the subfields (*Qualifiers*) in different positions and to store subfields with the same content (then *Qualifiers* with the same *Property* and the same *Target* in the same *Statement*) – would not be natively supported in the *Wikibase Data Model*; the Wikibase software does not accepts two *Claims* with identical *Property* and *Target* and, in fact, responds with an error to any loading attempt, both using the manual input interface of the software, or trying to send data through API using an automatic procedure (see below):

```
pywikibot.data.api.APIError: modification-failed: Claim has Already a
qualifier with hash d4d0bad3dc0bbd58a8c3c218fd135e73265a03e9
```

While, about the repetition of the subfields in different positions, the order would be lost in their representation by *Qualifiers*, because the *Qualifiers Claims* (within the same *Statement*) are grouped by the *Property* and, therefore, multiple UNIMARC subfields $f found in different positions of the field (as in the described case) would be collected into a single list under the *Property* with which the subfield $f has been mapped.

In the experiment, these two requirements were solved by introducing an additional *Qualifier* – the *Qualifier* with *Property* "MARC subfields order" – for storing the exact order of subfields (for each occurrence of the UNIMARC fields, and then the *Statements*).

In the case of the record (BID IT\ICCU\URB\0620565) described above, for example, the *Statement* corresponding to the title (UNIMARC field 200) will contain a *Qualifier* with *Property* "MARC subfields order" and *Target* with string "a[1] f[1] c[1] f[1]": these are the subfield codes, in their exact sequence (registered regardless of how Wikibase then stores the contents of each subfield); the number in brackets after the codes is the counter of repetition of each of them.

[15] In all likelihood (but we have not seen the book) it is a transcription error. A correct version of the bibliographic record can be found here: http://catalogue.bnf.fr/ark:/12148/cb33045411k.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

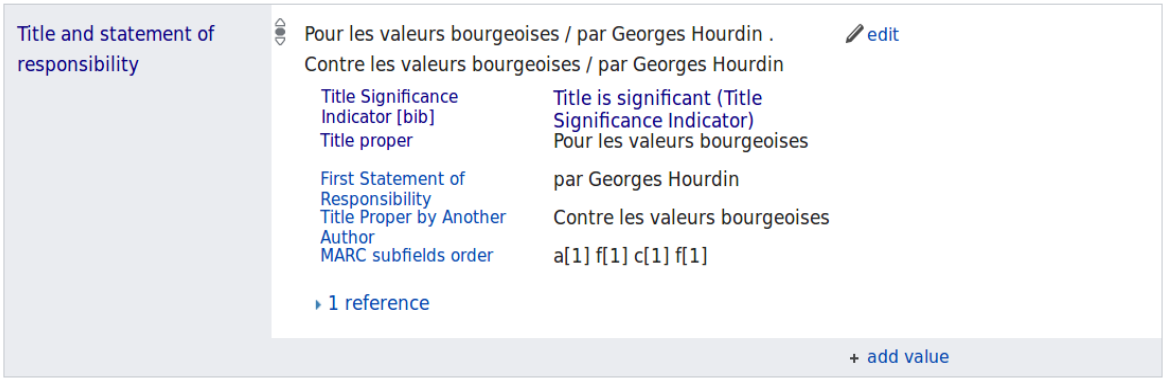


Figure 17. *Statement* corresponding to the UNIMARC field 200 of the record BID IT\ICCU\URB\0620565

In this way, the order of the subfields is correctly recorded, while the second occurrence of subfield (repeated identical) is not stored, it is only referenced.

With this solution, all the Wikibase indexing and data retrieval features are preserved: the "par Georges Hourdin" string is correctly connected to the specific occurrence of the *Statement* of the title area (UNIMARC 200), in the *Item* created for the record BID IT\ICCU\URB\0620565; while in the process of data extraction, it is always possible to reconstruct the exact sequence of the subfields, including their repetitions.

Relational model

In the experiment we necessarily addressed the "relational logic" of the *Wikibase Data Model* towards which we were transforming the UNIMARC format.

The UNIMARC record is known to be a "flat record", in which all common information is repeated identically; but it is equally well established that, with its support of unique identifiers and pointers between related data (as in the case of headings or multi-level descriptions handled in the Linking Entry Block of the standard), UNIMARC conveys in a consistent way the relational model of the elements that "organize" the bibliographic universe.[16]

The transformation from UNIMARC towards the *Wikibase Data Model*, therefore, required to break up all the UNIMARC elements that could be stored into unique Wikibase *Entities*, recording and qualifying all the necessary relations between them.

In an *Item* corresponding to a UNIMARC bibliographic record, therefore, the UNIMARC fields containing authority entries are managed with pointers to other *Items:* the name heading, for example, is stored in a *Statement* with the *Property* which maps the specific type of responsibility, and the

[16] "Data elements on bibliographic records might be classified into two categories: those that describe the entity in hand and those that relate the entities to other entities. Thus, in considering data elements to be included on bibliographic records, account needs to be taken not only of those that represent the format attributes of the entities described (descriptive elements) but also those whose purpose is to organize catalogues and by so doing to structure the bibliographic universe (organizing elements)" (Svenonius 1992).

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

content – *Target* – is replaced by a pointer to the *Item* of the authority record of that name (note that Wikibase automatically displays the *Label* of the linked *Item*)

Figure 18. 2 *Statements* corresponding to the names headers of the record BID IT\ICCU\CFI\0893220

Additionally, the *Qualifiers* of the *Statements* used for the "relations" between *Items*, can host the information (when present) that "describes" the relation: such as the relator-code of the name heading or the indication of the volume in the pointer to the series (from the Linking Entry Block of the UNIMARC).

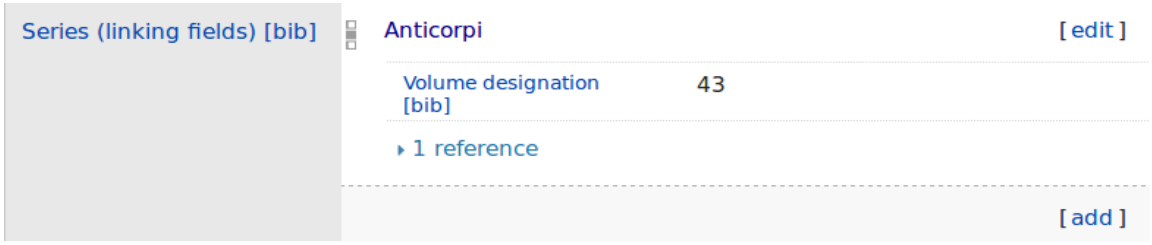


Figure 19. *Statement* corresponding to the pointer to the Series, UNIMARC field 410 of the record BID IT\ICCU\CFI\0893220

In the experiment, a distinction has been hypothesized between (a) pointers to the authority entries (which can have a separate record in the UNIMARC format), and (b) pointers to terms of closed lists of coded-data:

a) the *Statements* that link to authority entries with specific UNIMARC record exports (such as names, subjects, classification or related titles), have the pointers stored into the *Target* of the entire *Statement*.
In the bibliographic record, the details already recorded in the authority entry (stored in the separate *Item*) are discarded while "relationship attributes", if present, are maintained. (See the Figure 18 and 19).

b) The *Statements* containing coded-data related to "closed lists" (controlled vocabularies), have the pointers in the *Targets* of *Qualifiers*.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

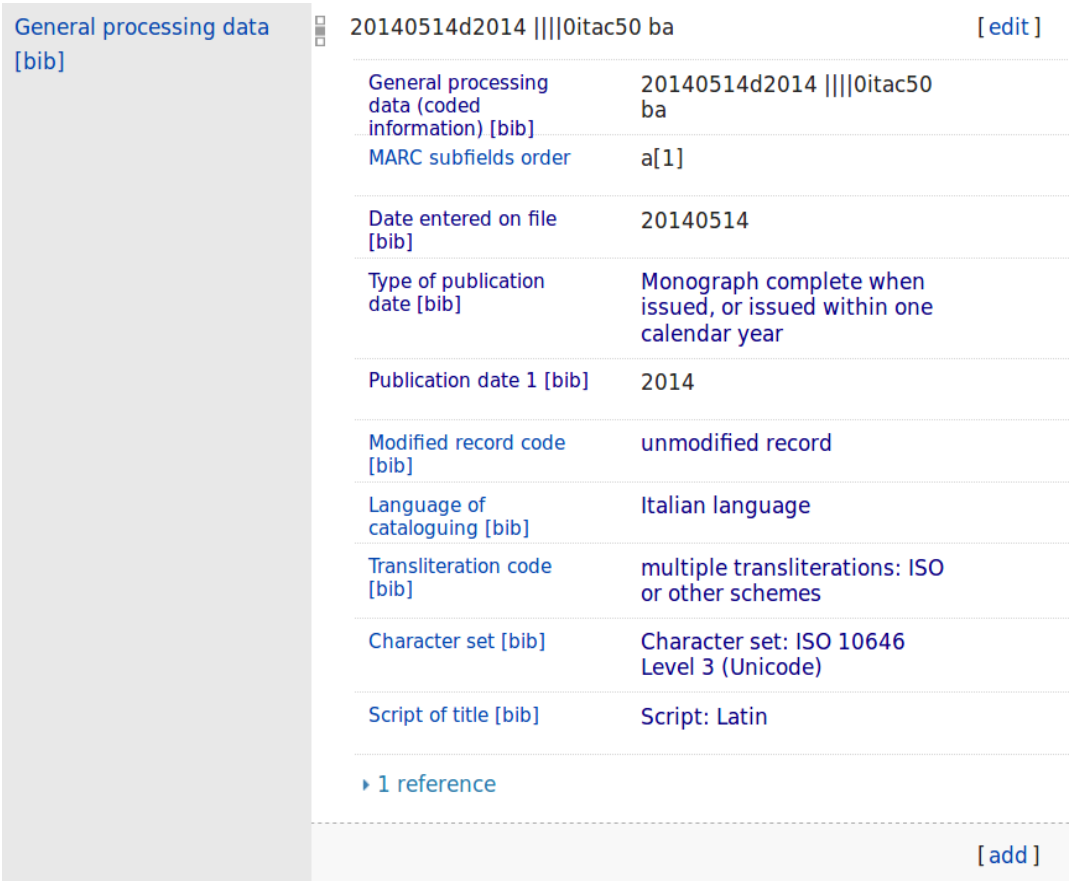


Figure 20. *Statement* corresponding to the UNIMARC field 100 of the record BID IT\ICCU\CFI\0893220

In the cases that require it, such as the fields "Language of the Item" (UNIMARC 101) and "Country of Publication" (UNIMARC 102), the pointers at subfield level can also support repeated values (i.e. different languages) in the same field occurrence.

The entry from the authority list, such as the entry for an author, becomes a separate *Item,* with all its *Statements* and specific *Properties* the same as the *Item* of the bibliographic record, and with the identifier of the authority-list (the SBN VID, in this case) stored in the *Alias* section of the *Fingerprint* (both for the Italian version and for the English one).

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

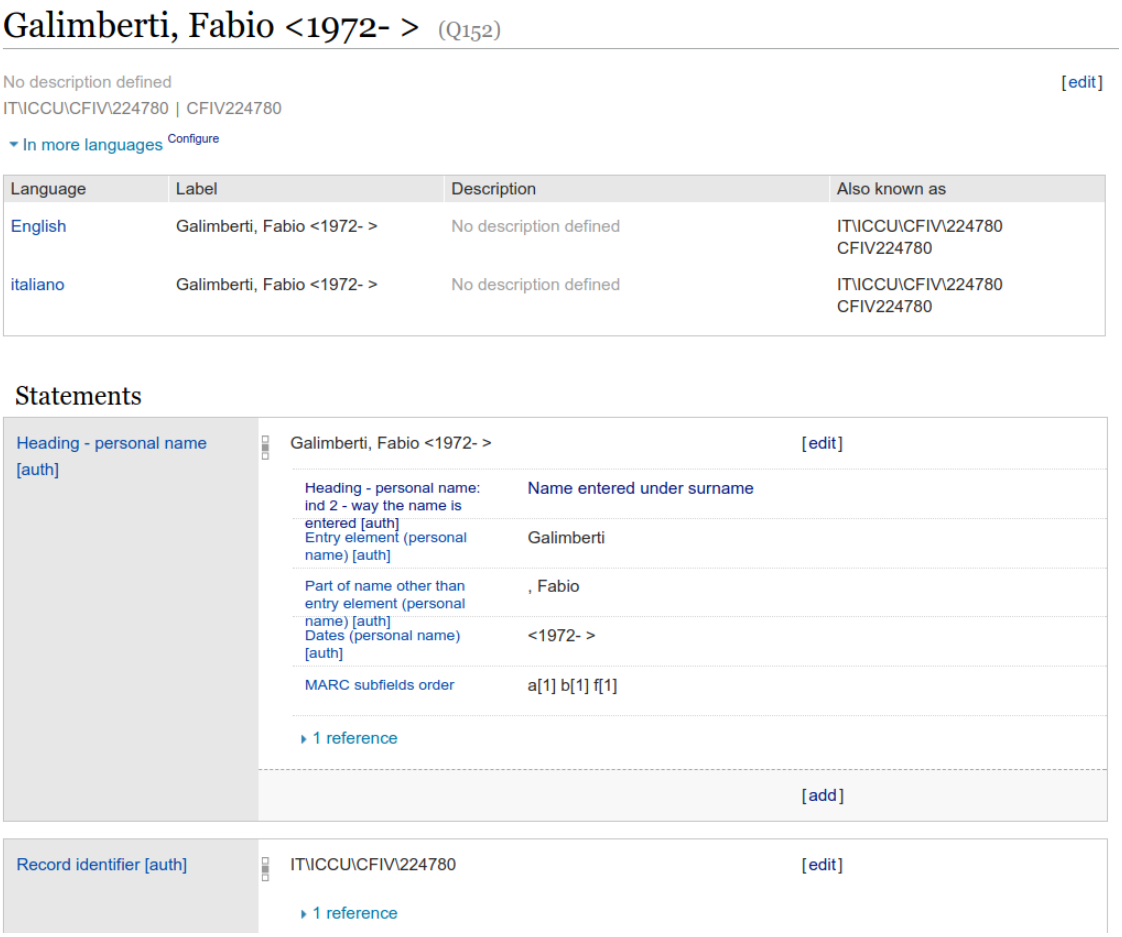


Figure 21. *Item* corresponding to the authority record for the author Galimberti, Fabio

Restructuring of UNIMARC records in separate Items

In this way, with the experiment it was possible to restructure the UNIMARC record into distinct *Items* by data type (potentially referred also to the different FRBR entities) and at the same time we tested the possibility to realize a mapping of all the UNIMARC elements, retaining the possibility to recover all the information of the original format.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

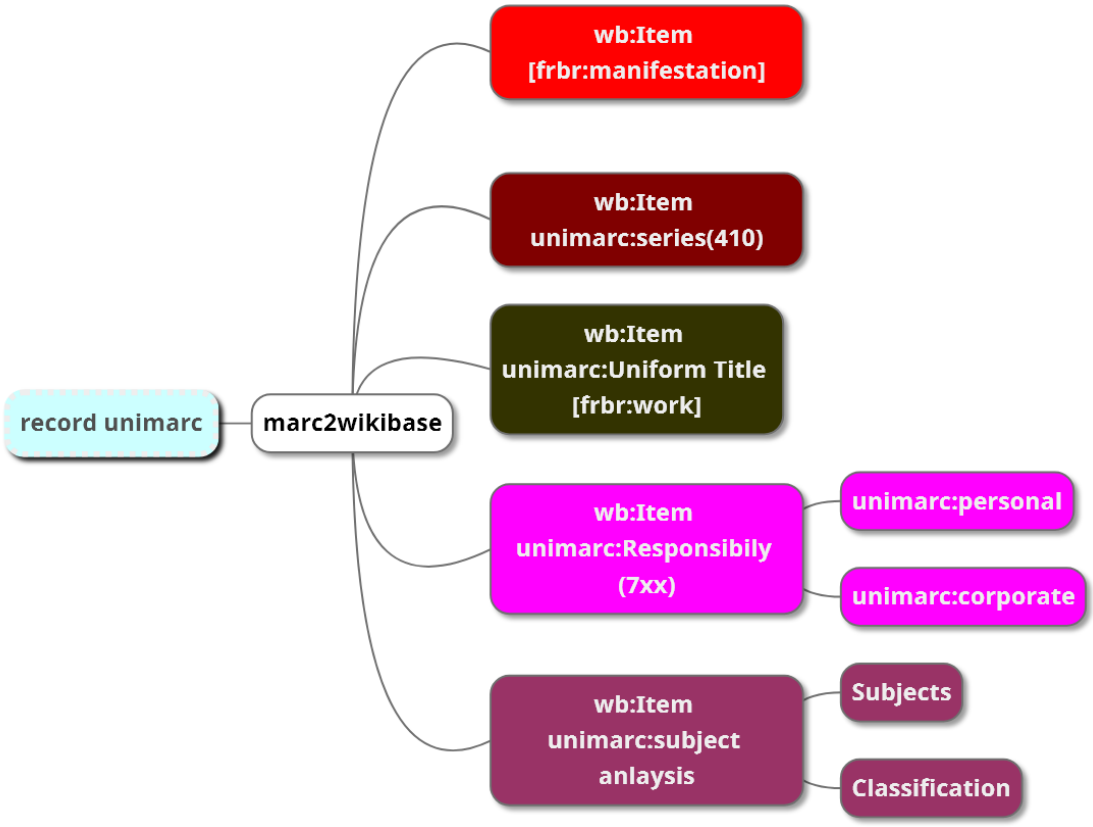


Figure 22. Graphical scheme of the restructuring of the UNIMARC record in *Items* distinct for given types

Quantitative data of the test

The quantitative data of the experiment give an idea of the articulation in different *Entities* which is required to represent one UNIMARC record and the elements it invokes within the *Wikibase Data Model*. A UNIMARC record of medium complexity – "Lo Stato innovatore" (BID IT\ICCU\CFI\0893220) – in Wikibase required one *Item* connected to more than 160 other *Entities* (*Items* and *Properties*): more than 100 *Properties* for the mapping of UNIMARC structural elements; about 40 *Items* for the encoded data (such as the codes for language or country); and then other *Items* for the related bibliographic records (as the series) or to the entries from authority lists (such as headings for the names, the subject, the classification or the uniform title).[17]

[17] The *Entities* were created in a temporary installation of Wikibase (set up only for the purposes of this test). The screenshot of the entire *Item* corresponding to the main UNIMARC record (as it is displayed in the Wikibase graphical interface) can be found at https://archive.org/details/Q156En; while a listing of all the *Entities* (encoded in the JSON format produced by Wikibase) created to reproduce the UNIMARC structure of the main record and the data connected to it, can be found at https://archive.org/details/list_of_used_Properties_and_Items.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

In details: the record has its own *Item*

- bibliographic record: Lo Stato innovatore (see Figure 15).

All the related *Entities* are meant to be reused in connection with other *Entities*.

Items for related records and authority entries:

- Bibliographic Record (series): Anticorpi (see Figure 19).
- Responsibility (authority entry): Galimberti, Fabio <1972-> (see Figure 21).
- Responsibility (authority entry): Mazzucato, Mariana (see Figure 23).
- Classification: 338.45 (ed. 23) - INDUSTRIE SECONDARIE E SERVIZI. EFFICIENZA PRODUTTIVA.
- Subject: Innovazione tecnologica - Interventi statali (see Figure 24).
- Uniform title: The entrepreneurial State.

Items and *Properties* for the mapping of UNIMARC syntax:

- About 40 *Items*, used to store the coded-data (see Figure 13).
- 116 *Properties* for the structural elements of UNIMARC (see Figure 10, 11 and 12).

Issues pointed out in the experiment

The experiment has also pointed out some issues that can be considered in the perspective of loading large amount of data, with fully (or almost fully) automated processes.

Acquisition of authority lists and controlled vocabularies

In the experiment, the acquisition of entries from authority lists and controlled vocabularies (in addition to bibliographic records) was made by testing different sources: the OPAC of the National Central Library of Florence for the bibliographic records, and the OPAC of Indice SBN for the authority records of names.

For the names entries, in particular, in addition to the case of names from UNIMARC authority record (such as "Galimberti, Fabio <1972->" already illustrated in the Figure 21), in the experiment we tested the case of names for which there is no authority record: in this case the *Item* for the name was generated in Wikibase only with the data derived from the 7xx heading of the bibliographical records.

This is the case of the name Mazzucato, Mariana:

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

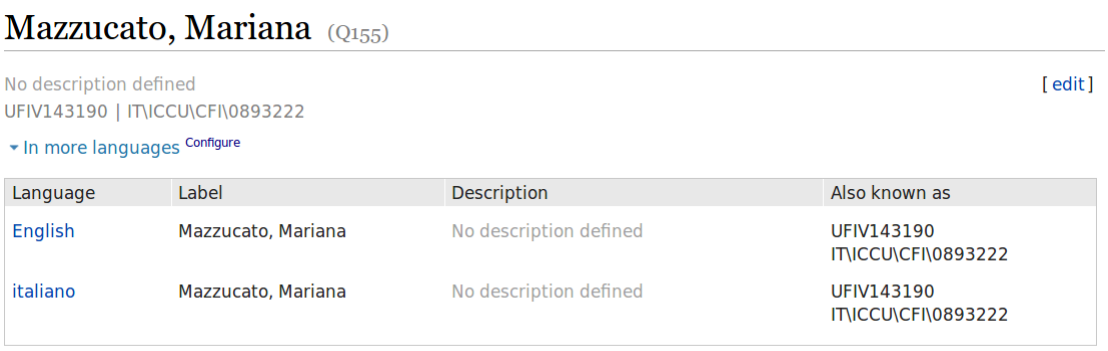


Figure 23. *Item* created for the author Mazzucato, Mariana

Similarly, we tested the generation of *Item* for authority entries such as uniform title ("The entrepreneurial State"), subject ("Innovazione tecnologica - Interventi statali") and classification ("338.45 (ed. 23) - INDUSTRIE SECONDARIE E SERVIZI. EFFICIENZA PRODUTTIVA") from the data recorded in the bibliographic record.

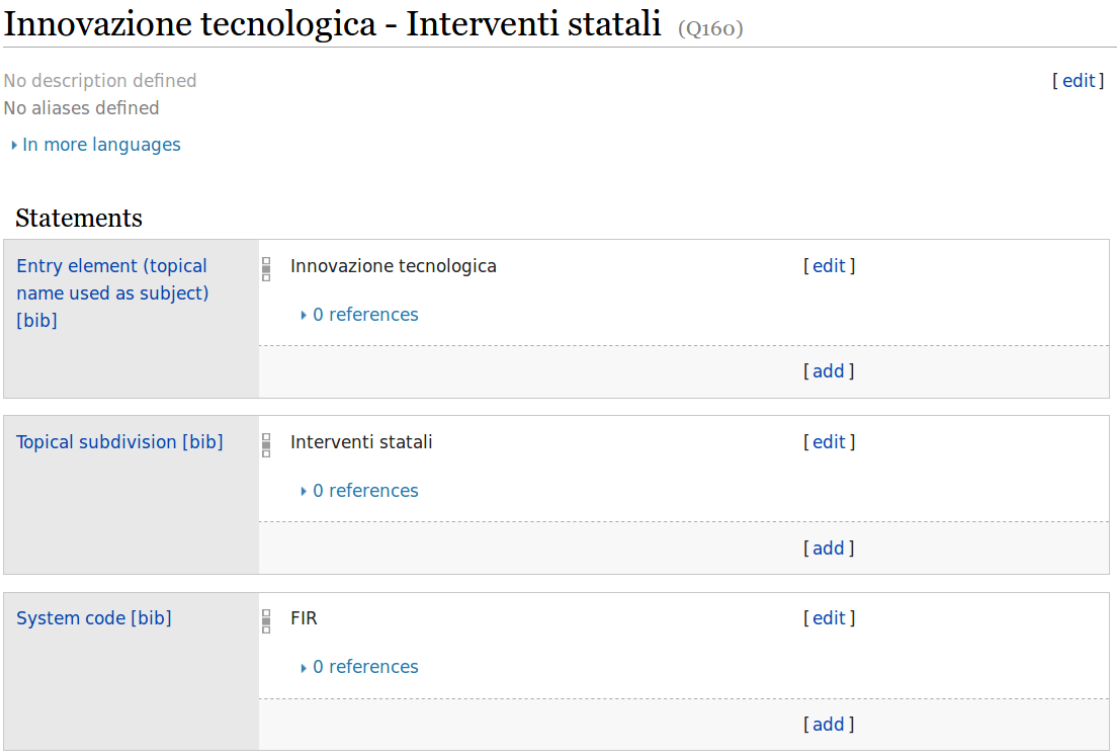


Figure 24. *Item* created for the subject entry

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

But certainly, in the perspective of loading large amounts of data, it should be preliminarily planned to acquire all the entries of controlled vocabularies and authority lists.

Transformation of the bibliographic format towards the Wikibase Data Model

In the experiment some choices have been made with respect to the textual *Target* of *Statements* corresponding to UNIMARC fields:

- when the treated field corresponds to an ISBD area, ISBD punctuation is reported in the *Target*.

Figure 25. *Statement* corresponding to the UNIMARC field 210 of the record BID IT\ICCU\CFI\0893220; the *Target* contains the original punctuation

- when the treated field does not correspond to an ISBD area with its prescribed punctuation, the *Target* has simply been filled with the list of all contents of the subfields (in their original order) separated by dashes ("-").

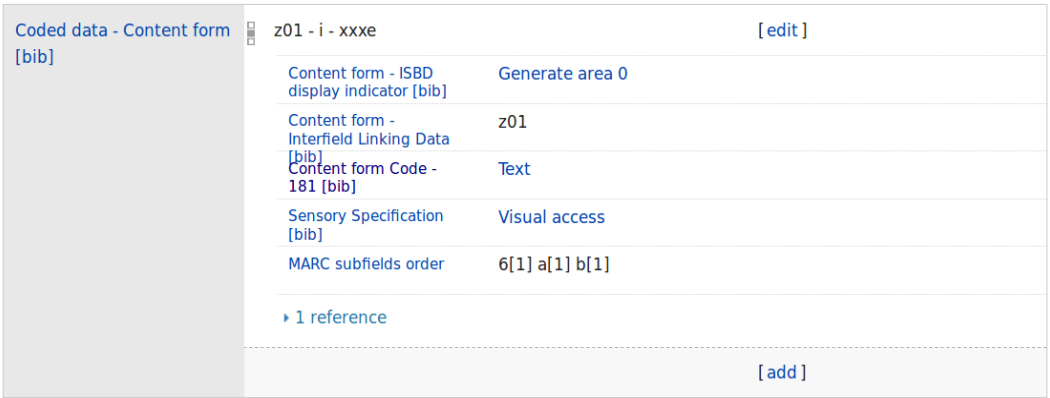


Figure 26. *Statement* corresponding to the UNIMARC field 181 of the record BID IT\ICCU\CFI\0893220

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

The *Label* of the entire *Item* corresponding to the bibliographic record contains only the title proper.

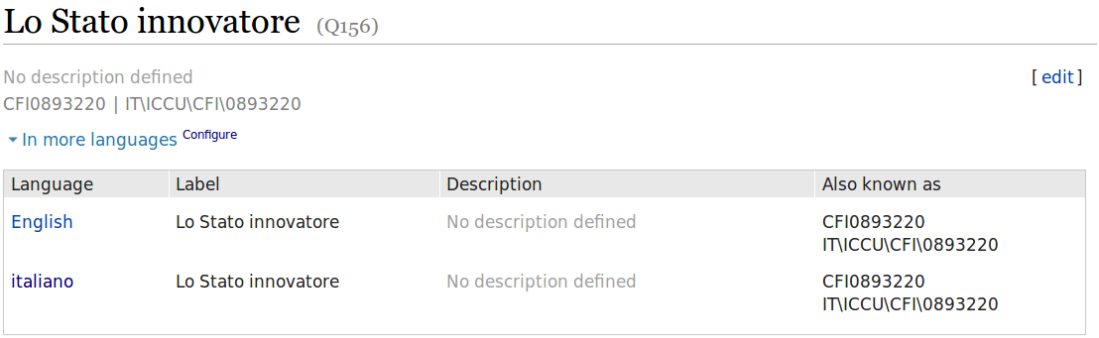


Figure 27. *Fingerprint* of the *Item* corresponding to the entire UNIMARC record BID IT\ICCU\CFI\0893220. The *Label* contains the title proper of the UNIMARC record

In any case, it is obvious that the *Fingerprint* of the *Item* will have to contain some elements or some interpolations of elements that the library catalog traditionally uses to identify the bibliographic record (title area, identifiers, etc); and, similarly, the *Targets* of the *Statements* which correspond to ISBD areas could host the sequence of the subfields with the predicted punctuation. To do this, planning to load large amounts of data should also include a process for re-generating the conventional punctuation used in bibliographic record display, thus supporting all the cases of punctuation according to combinations of fields and subfields.

Moreover, always in the perspective of managing numerous UNIMARC records as *Item* in Wikibase, the amount of record data to be included in the *Fingerprint* of the entire *Item* (*Label*, *Description*, *Alias*) should be carefully considered, bearing in mind that the title proper only is not sufficient either to identify a manifestation (or an edition) nor to ensure that you avoid, within the same instance of Wikibase, cases of equal *Label* in different *Items*.

Data-types of Wikibase

The experiment has highlighted the possibility of converting some UNIMARC data to specific data-types of the *Wikibase Data Model*: for example, the dates recorded in the Coded Information Block can be normalized to the data-type "Point in time".

In order to manage positional value code strings, specific storage methods can be studied: for example, in the experiment, the label of the UNIMARC record (in addition to being translated into all its component parts by specific *Qualifiers*) is stored in the *Target* of the *Statement* between double quotation marks, to save any spaces at the end of the character sequence.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

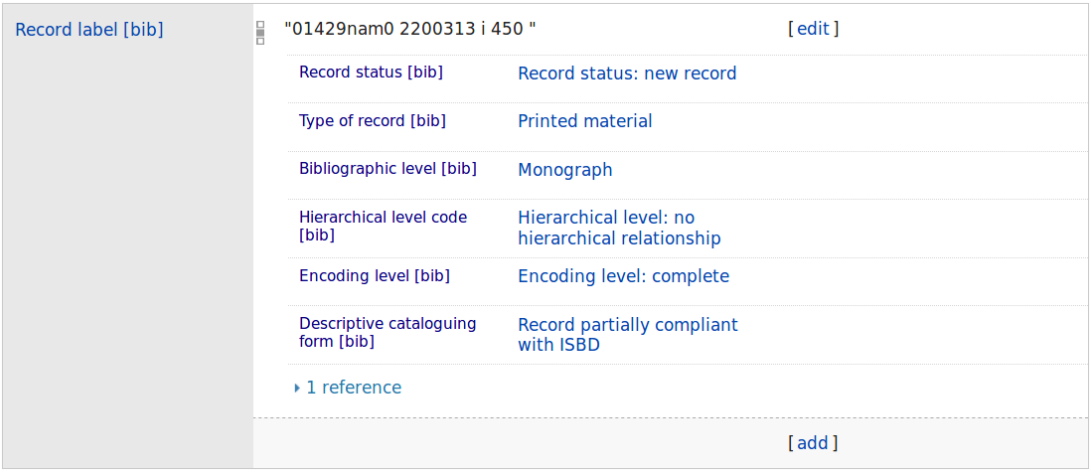


Figure 28. *Statement* corresponding to the "record label" of the UNIMARC record BID IT\ICCU\CFI\0893220

Implementation of automatic procedures

In the experiment, data loading was done by aiding the manual input of the various test solutions with the implementation of multiple automatic upload procedures – called *bot* (short for robot) – aimed in particular to loading portions of uniform data in a temporary installation of Wikibase.

Specifically, Python programming language was used for writing procedures, using the Pywikibot library[18] that supports the Application Programming Interface (API) of the Wikibase software,[19] which is an extension of MediaWiki API.[20]

Therefore the experiment, although it was focused on the data model and not on the development of a loader software, has allowed us to verify that for the import of data in Wikibase there is a large variety of open source tools maintained by the developers community surrounding Mediawiki systems. In this respect, subject to the tests of performance with large amounts of data, many tools can certainly be found to deal with the more usual issues in the data transfer process from one system to another.

For example: the replication of the relational setting of UNIMARC data within the *Wikibase Data Model* requires, of course, the creation of new pointers between the descriptive elements (for example, a bibliographic record pointing to an authority record of the author) that replaces the "old" identifiers of the original data (the BID of the bibliographic record and the VID of the name authority entry) with the "new" identifiers assigned to the *Entities* by Wikibase; but the "new" identifiers are created only after each *Entity* load and cannot be anticipated.

[18] The reference web site of the software library is https://www.mediawiki.org/wiki/Manual:Pywikibot, while the evaluation page in Mediawiki is https://www.mediawiki.org/wiki/API:Client_code/Evaluations/Pywikibot.
[19] The Wikibase APIs are documented in https://www.mediawiki.org/wiki/Wikibase/API.
[20] The Mediawiki APIs are documented in https://www.mediawiki.org/wiki/API:Main_page.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

This issue (solvable with the usual intermediate processing of the data to be loaded) can certainly be addressed by testing with large amounts of data the various options already offered by software libraries and API commands: in the case of loading pre-processed data organized in pre-ordered batches, the choice could be the full-record upload (with JSON encoding); in the case of an update of pointers between the *Entity* after their upload to Wikibase, the choice could be the use of the various editing options via APIs to modify the *Entities* in Wikibase.[21] In any case, the processing of the pointers (as well as the treatment of each element of the *Entity* managed in Wikibase) can take advantage of a complete set of response messages returned by the API for each intervention on the data: for each *Item* or *Property* created or modified, the identifier of the *Entity* and the complete detail of the modified data are returned; in this way it becomes possible to keep track of the new identifiers that replace the BID (or VID) of the records and plan the most effective method of updating the pointers among the *Entities* created.

New mappings of UNIMARC

The degree of flexibility with which the relationships between *Entities* are managed in the *Wikibase Data Model*, lets imagine that a "mapping" of the UNIMARC semantics can be connected with multiple other metadata schema (or ontologies) through references added in the *Aliases* of the *Properties* created to represent the UNIMARC itself, or through relationships (of equivalence, of inclusion or other) with the *Properties* created for the new scheme (the same could be done for the *Items* used for the controlled vocabulary entries).

For the connection between *Properties* (and between *Entities* in general) different "relationships" are already available in Wikidata such as "equivalent property in other ontologies" (*Property* P1628)[22] or "equivalent class in other ontologies" (*Property* P1709)[23] that can be used to "map" the *Entities* of one scheme with those of another.

An example of "property mapping" can be found in Wikidata in the "title" *Property* (P1476),[24] which is linked, through the "equivalent property" relationship (*Property* P1628), both to the "title" field of Dublin-Core[25] and to the "name" field of Schema.org.[26]

[21] See the *Entities* modification commands outlined in the documentation: https://www.wikidata.org/wiki/Wikidata:Creating_a_bot#API.
[22] https://www.wikidata.org/wiki/Property:P1628.
[23] https://www.wikidata.org/wiki/Property:P1709.
[24] https://www.wikidata.org/wiki/Property:P1476.
[25] http://purl.org/dc/terms/title.
[26] https://schema.org/name. For documentation on the scheme: https://schema.org.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

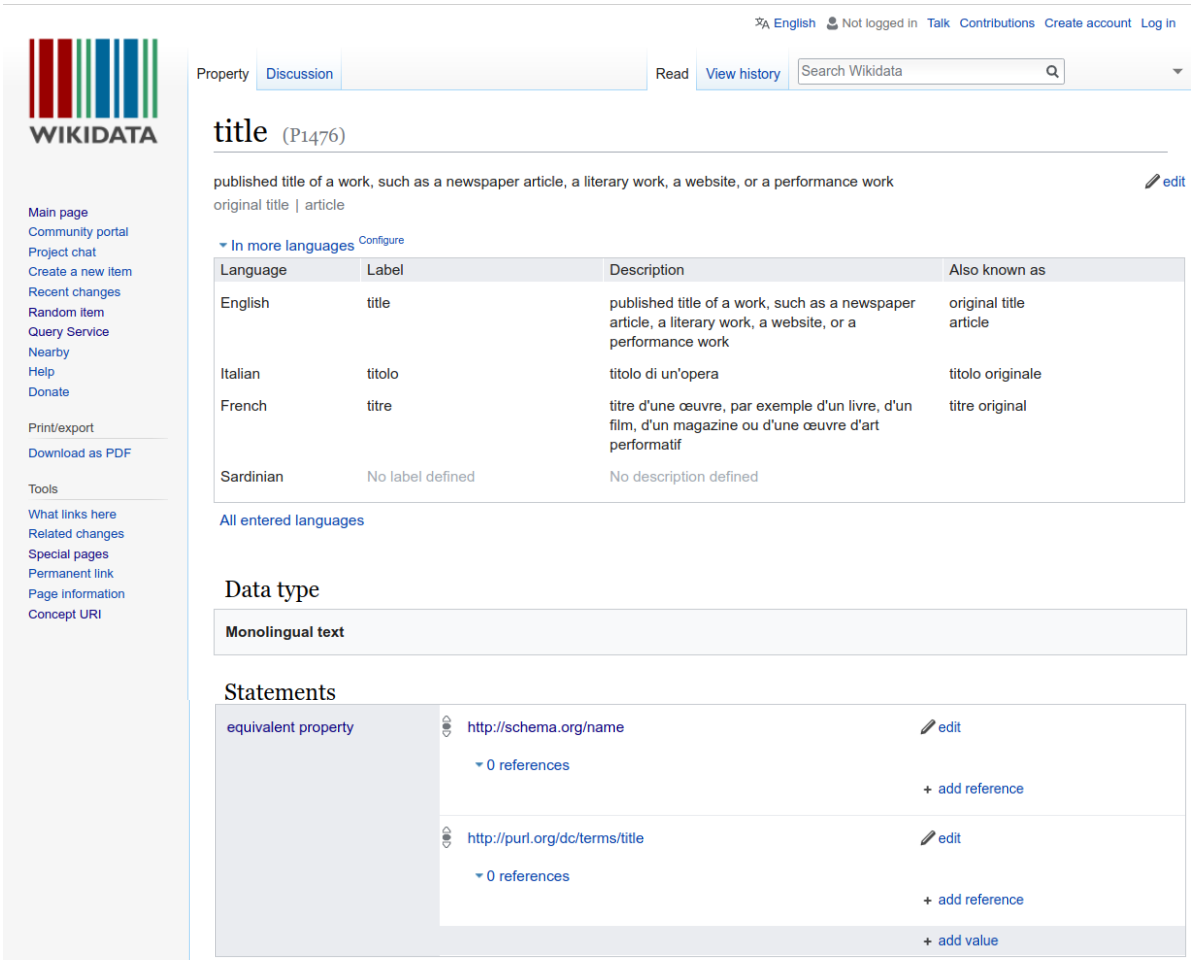


Figure 29. "Title" *Property* (P1476), linked to the "name" field of Schema.org and to the "title" field of Dublin-Core, through "equivalent property in other ontologies" (P1628)

An example of "class mapping" can be found in the "creative work" *Item* (Q17537576) of Wikidata[27] which is linked, through the "equivalent class" *Property* (P1709), to both the "CreativeWork" class of Schema.org[28] and to the "work" class of GND ontology.[29]

[27] https://www.wikidata.org/wiki/Q17537576.
[28] http://schema.org/CreativeWork.
[29] https://d-nb.info/standards/elementset/gnd#Work.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

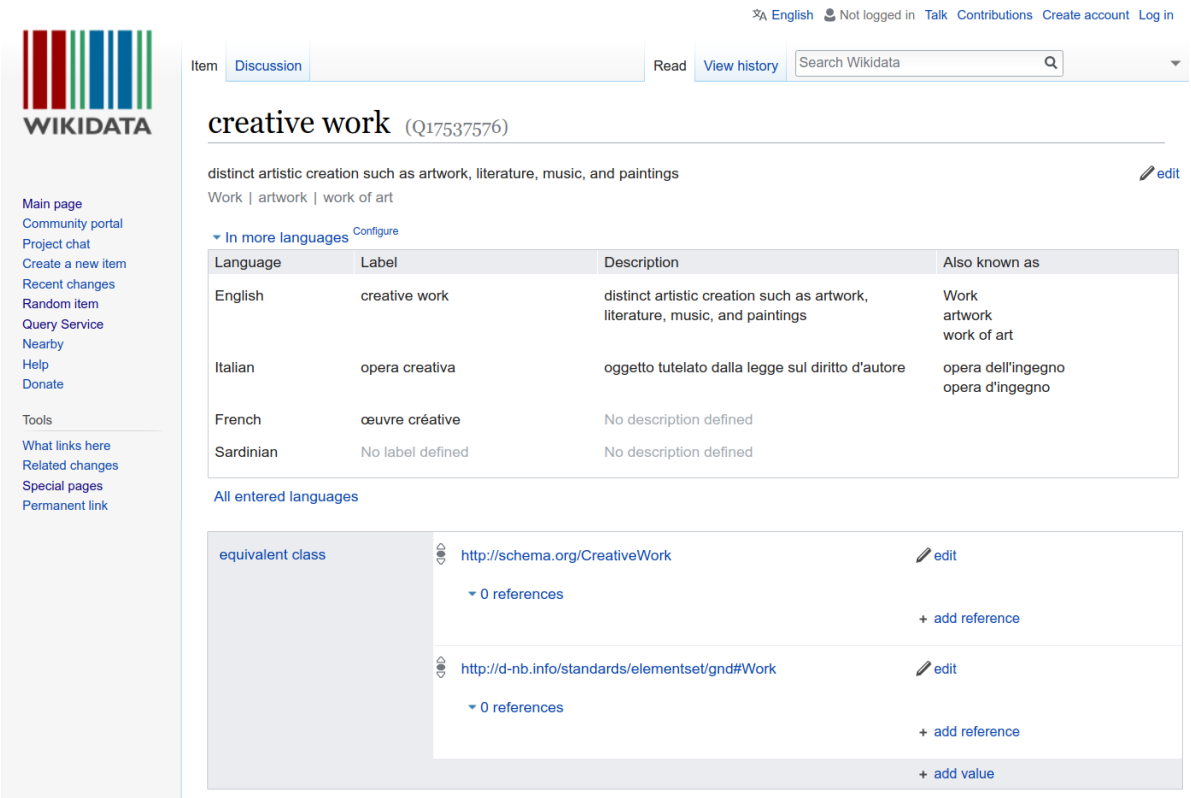


Figure 30. "Creative work" *Item* (Q17537576), linked to the "CreativeWork" class of Schema.org and to the "work" class of GND ontology, through the "equivalent class in other ontologies" *Property* (P1709)

Relations between *Entities* such as those described can then be exploited in the search stage to use any mapped scheme indifferently: remaining to the cited examples, with a simple query in the SPARQL endpoint of Wikidata[30] you can retrieve the "title" *Property* starting from the "title" field of Dublin-Core[31] or the "creative work" *Item* starting from both the "CreativeWork" class of Schema.org[32] or from the "work" class of GND ontology.[33]

Moreover, in the contents of Wikidata, and therefore also in the Wikibase mechanisms, the management of the hierarchical relations between *Properties* is envisaged too: among the examples of queries in the SPARQL documentation,[34] there is a query that collects all the "Subproperties" of the "location" *Property* (P276) even with multiple degrees of separation;[35] this type of hierarchical relation can be used, for example, to link MARC fields to the relative subfields or to single characters of fixed

[30] The "Query Service" of Wikidata is https://query.wikidata.org/, and the starting page of the documentation is https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Wikidata_Query_Help.
[31] See the SPARQL query http://tinyurl.com/ydfaen53.
[32] See the SPARQL query http://tinyurl.com/ydfmckyx.
[33] See the SPARQL query http://tinyurl.com/y9zjc5kx.
[34] See, in particular, the "Wikibase predicates" section https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples#Wikibase_predicates.
[35] See the SPARQL query http://tinyurl.com/ydyf2fu7.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

length fields, and also to link (in the case of mappings) multiple MARC fields to more generic concepts (such as "title" or "creator").

Linking mechanisms between *Entities* can be inserted into more complex queries to search for data based on the different mapped schemes: an immediate example can be the search for all Wikidata *Items* presenting the string "Alice's Adventures in Wonderland" as Dublin-Core "title".[36]

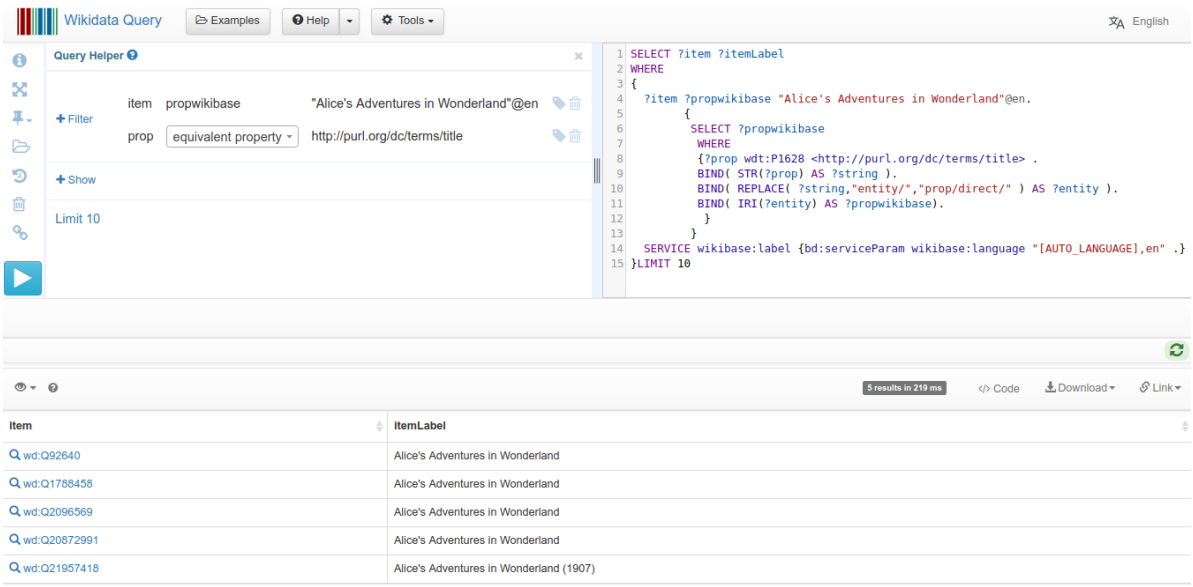


Figure 31. SPARQL query that searches the Dublin-Core "title" field for the string "Alice's Adventures in Wonderland"

In practice, the scenario that can be prefigured is that a large amounts of bibliographic data stored in Wikibase (with the informative detail verified in this experiment) can be accessed and searched both with the UNIMARC or other semantics, for example *schema.org* or *BIBFRAME*.

Main Wikibase technical potential for bibliographic information

Once all the MARC information detail is translated into the *Wikibase Data Model*, it becomes possible to exploit, for the bibliographic data, the technical solutions and the services implemented in Wikibase.

[36] See the SPARQL query http://tinyurl.com/y79l2sjt.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

The conversion of the UNIMARC record in Wikibase Data Model can facilitate the use of data

The degree of real "usability" of data structured in MARC, especially in this era of great innovation in storage formats and query protocols, is a widely debated topic in literature and is often a subject of criticism towards the standard.

Loading the full MARC detail in the *Wikibase Data Model* (starting with one of its dialects, such as UNIMARC) can probably respond to a good number of needs that are highlighted in the actual practice of using the standard. A quick list of most significant needs for which it is useful to seek answers, can certainly be drawn from the authoritative analysis already proposed for the problem, such as the article "MARC21 as Data: A Start" in which Karen Coyle highlights the need to overcome the limitations of the MARC format with a view to providing new bibliographic services in the context of the web, while underlining that it is necessary first to have a complete knowledge of all the informative detail that the MARC conveys. In her analysis, Karen Coyle lists a series of technical limitations of the MARC format that hinder the complete decoding of its content. These technical limitations can perhaps find a solution with the translation in the *Wikibase Data Model*: below are some significant examples taken from the article mentioned and from another by the same authoress.

A "usable version" of the MARC: with explicitation of the encoded values

The transformation into the *Wikibase Data Model* can provide a "usable version" of the MARC: it is known (and it is also obvious) that anyone or any system that needs to work on data structured in MARC must create its own elaboration of the format:[37] in Wikibase, definitions and descriptions of the data (*Labels*, *Descriptions* and *Aliases* of *Fingerprints*) coexist with data, in the same system by using *Properties* and relationships between *Entities*.[38] By exploiting this mechanism, the MARC coded data can be registered with pointers to the *Entities* containing their definitions (see above the paragraph "Relational model"); the same can be done for the data recorded in MARC as a placeholder in the fixed length fields: all the codes provided for the cases of ambiguity (the "fill character" for cases in which it is not possible to define the correct value, the "blank" to be used as a simple placeholder and the others "Unknown", "Combination", "Not applicable", "Not present" and "Other") can be managed with appropriate *Entities* containing the corresponding definitions.[39] Similarly, specific *Entities* can be created for all the values prescribed in the indicators of the MARC fields, including the value "Undefined".[40]

[37] "Anyone wishing to develop applications for MARC21 must create their own usable version" (Coyle 2011).
[38] "My own database still lacks definitions and descriptions, and those will probably need to be added by screen-scraping hundreds of screens from the LC web site" (Coyle 2011).
[39] "The use of values for 'unknown' and 'no attempt to code' are directly related to the characteristic of fixed length fields made up of positional data elements, where all positions must be filled in to retain the positioning" (Coyle 2011).
[40] "In addition to the fixed field values like 'Unknown', almost 60% of indicator positions (206 out of 350 in my database) have the value 'Undefined' These truly represent empty positions in the record format and they can be ignored" (Coyle 2011).

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

Identification of univocal data, and their respective sources

The relationship between *Entities* supported in the *Wikibase Data Model*, and the consequent articulation in unique *Entities* in the transformation from the MARC format, make it possible to ensure the identification of the univocal data.

This makes the translation of encoded data with their respective names and definitions in natural language completely transparent: the entries of authority lists and controlled vocabularies are recorded in separate *Items*, and are automatically replaced by Wikibase with the *Label* (in the graphical interface) and the identifier (in the data export) of the *Item* that hosts the unique data.[41]

All the *Entities* of the *Wikibase Data Model* are uniquely identified with URIs, as required in the context of the semantic web;[42] the source of the data can be recorded for each *Statement* (corresponding to UNIMARC field) using the *Source* and, with the described solution, it is also possible to reconstruct the exact sequence of the original details.[43]

Statistics of descriptive elements

Once loaded into Wikibase, all elements of MARC semantics (fields, subfields, indicators, positional values in fixed length fields and controlled vocabulary entries) are available not only in complete and searchable lists, but also with statistics of their use. The quantifications can be obtained through the queries supported by the system itself or through those supported by SPARQL endpoints: for example, among the queries proposed in the Wikidata Query Service documentation,[44] there is a query that gathers all the *Items* with a specific *Property* (the "Wikimedia database name" *Property*, P1800);[45] the same query can be easily adapted to return the count of these *Items*,[46] thus providing a statistic of the use of the *Property*.[47]

[41] "The fixed fields are fixed-length strings with positional data elements that take coded data that is presumably useful for machine processing. These are primarily in the form of controlled term lists with the terms represented by 1-2 character codes" (Coyle 2011).
[42] "All of the data elements in my study must be assigned an identifier, and I use http URIs for this under the registered domain name 'marc21.info'" (Coyle 2011).
[43] "It is also convenient if the description of the data elements contains information that would lead back to the original encoding of that data in MARC21" (Coyle 2011).
[44] See the documentation at https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples#All_items_with_a_property.
[45] See the SPARQL query http://tinyurl.com/yatx3324.
[46] See the SPARQL query http://tinyurl.com/ya8d52zr.
[47] "Having the MARC21 fixed fields, tags and subfields in a database first allowed me to do some quick statistics based on the names of fields and fixed field values" (Coyle 2011).

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

Sharing of bibliographic data between different systems

Wikibase can facilitate the sharing of component parts of the record, even starting from the most simple: it could host the shared parts of the record, also produced on the base of different standards.[48] The shared bibliographic data may reside in a shared central installation.[49]

Wikibase Data Model is designed for concurrent editing

Wikibase can handle the semantics of MARC into a new storage format natively designed for collaborative editing, even in individual parts of the record.[50] The *Wikibase Data Model* is an open format with a broad user community.[51] Wikibase (although it uses relational DBMS such as MySQL or MariaDB) is based on data storage in triples.[52]

Redirection between Item

Wikibase handles redirection in case of "replacements" of an *Item*: old *Item* will remain preserved, with its *Fingerprint* and all its information, but will contain the reference to the new *Item* that replaces it. This feature provides a technical solution to the problem of "replacement" between identifiers (BID) in SBN: in practice the need to merge two entities when it occurs that one of them was created by mistake.

[48] "Output rates of intellectual and cultural products is increasing. Libraries have already responded to this through shared cataloging and purchase of cataloging from product vendors. However, the records produced in this way are then loaded into thousands of individual catalogs in the MARC-using community" (Coyle 2017).

[49] "Those records are often edited for correctness and enhanced. Thus they are costing individual libraries a large amount of money, potentially as much or more than libraries save by receiving the catalog copy. [...] The repeated storing of the same data in thousands of catalogs means not being able to take advantage of true sharing. In a cloud solution, records would be stored once (or in a small number of mirrors), and a record enhancement would enhance the data for each participant without being downloaded to a separate system" (Coyle 2017).

[50] "'Sharing' in today's environment means exporting data and sending it as a file. Since MARC records can only be shared as whole records, updates and changes generally are done as a 'full record replace' which requires a fair amount of cycles" (Coyle 2017).

[51] "Moving more toward open source would be facilitated by moving away from a library-centric data standard and using at least a data structure that is commonly deployed in the information technology world" (Coyle 2017).

[52] "NoSQL databases and triple stores. The current batch of databases are open source, fast, and can natively process data in a variety of formats (although not MARC) [...] This makes deployment of systems easier and faster" (Coyle 2017).

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

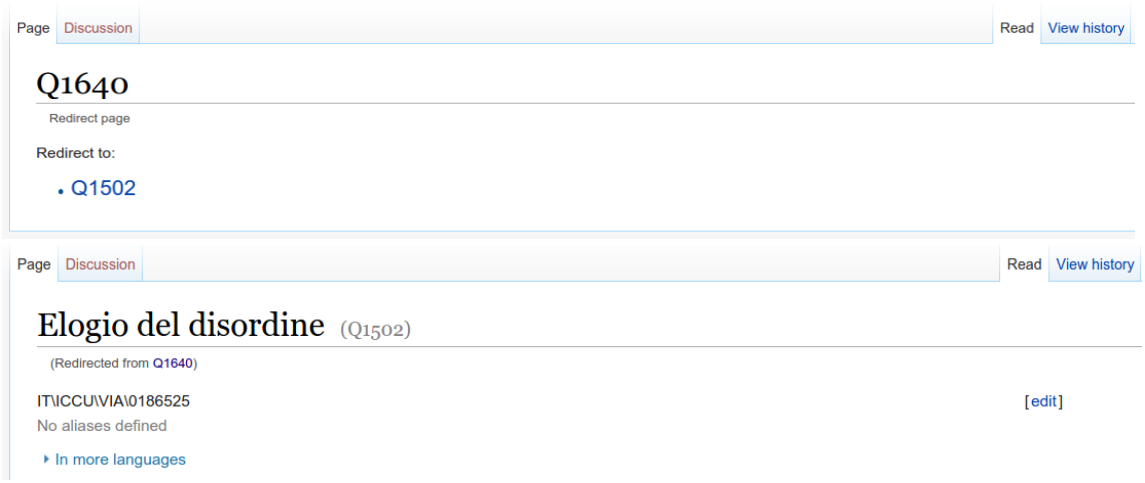


Figure 32. Example of redirection between two *Items* in Wikibase

History of all versions of Item

Wikibase allows to preserve the history of editing operations on all *Entities* (both *Items* and *Properties*) stored in the system: all the revisions of the *Entities* are kept in the history with date and author references of the changes and, in the best Wiki tradition, with the support of discussions about each version.

The versions can be individually restored at any time.

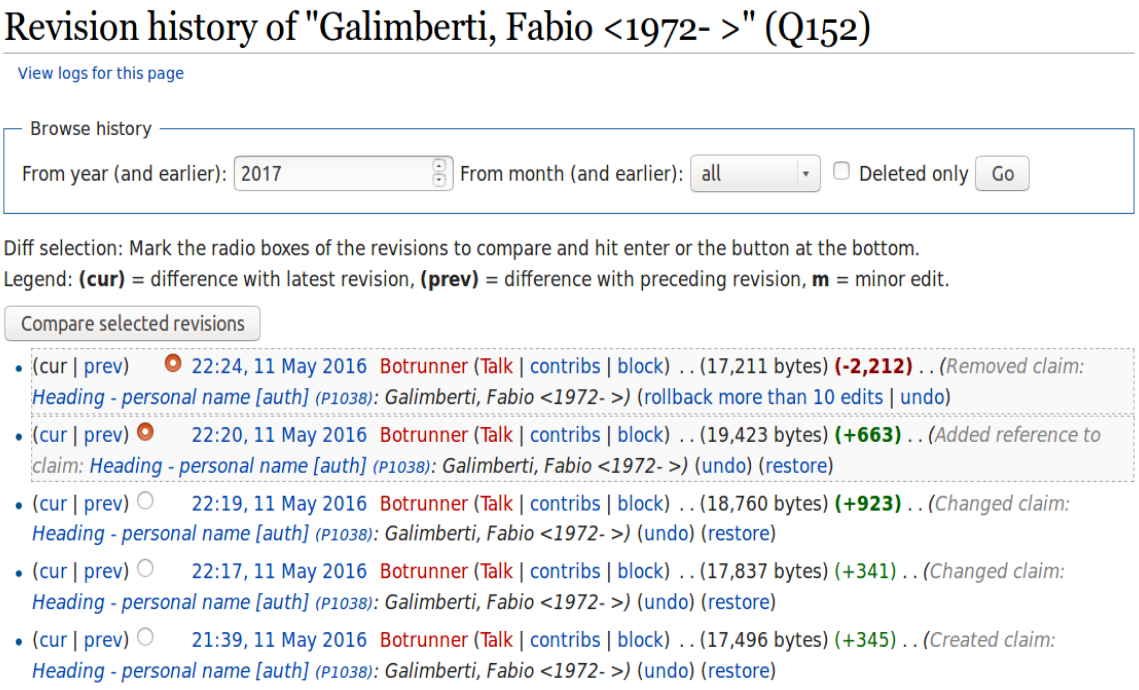


Figure 33. Revision history of the *Item* corresponding to the author Galimberti, Fabio

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

Export in RDF

Although an "enriched" version of the semantic triple is implemented in the *Wikibase Data Model*, Wikibase also provides an export of all *Entity* elements in standard RDF.

The solution is implemented in the Wikidata Toolkit (an additional library of tools connected to the Wikibase software),[53] and is based on the "reification" mechanism: the process by which, in the context of the Semantic Web, complex data structures (such as the *Statement* of *Wikibase Data Model* that also contain *Qualifiers* and *Sources*) are represented by introducing new elements that make it possible to use only elementary semantic triples (Erxleben, Günther, Krötzsch, Mendez, Vrandečić 2014).

Through this service already available in Wikibase, therefore, the upload of all the UNIMARC details tested in the experiment also obtains the effect of a standard mapping towards RDF and, consequently, the possibility to expose the data with the usual methods of the semantic web.

For example, looking again at the main bibliographic record tested in the experiment, the RDF Turtle export (in the section immediately following the initial prefix statements, @prefix) presents the contents of the *Fingerprint* of the *Item* and, subsequently, the list of *Targets* of all the *Statements* with which the record fields have been mapped: note, in the Figure 34, the entire string "Lo stato innovatore…" for the *Statement* with *Property* "P754" (title area), and the presence of pointers to other *Items* for heading fields, such as "Q152" (*Item* of Galimberti, Fabio) for *Property* "P1195" (Personal Name - Secondary Intellectual Responsibility).

[53] The reference site of the software library is https://www.mediawiki.org/wiki/Wikidata_Toolkit.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458


```
wdata:Q156 a schema:Dataset ;
    schema:about wd:Q156 ;
    cc:license <http://creativecommons.org/publicdomain/zero/1.0/> ;
    schema:softwareVersion "0.0.1" ;
    schema:version "2504"^^xsd:integer ;
    schema:dateModified "2016-08-02T22:58:11Z"^^xsd:dateTime .

wd:Q156 a wikibase:Item ;
    rdfs:label "Lo  Stato innovatore"@en ;
    skos:prefLabel "Lo  Stato innovatore"@en ;
    schema:name "Lo  Stato innovatore"@en ;
    rdfs:label "Lo  Stato innovatore"@it ;
    skos:prefLabel "Lo  Stato innovatore"@it ;
    schema:name "Lo  Stato innovatore"@it ;
    skos:altLabel "CFI0893220"@en,
        "IT\\ICCU\\CFI\\0893220"@en,
        "CFI0893220"@it,
        "IT\\ICCU\\CFI\\0893220"@it ;
    wdt:P644 "IT - 2014-2498" ;
    wdt:P779 "Roma ; Bari : GLF editori Laterza, 2014" ;
    wdt:P657 "20140514d2014    ||||0itac50      ba" ;
    wdt:P923 "IT - IT-FI0098 - 20140514" ;
    wdt:P754 "Lo Stato innovatore : sfatare il mito del pubblico contro il privato / Mariana Mazzucato ; traduzione di Fabio Galimberti" ;
    wdt:P793 "Anticorpi ; 43" ;
    wdt:P620 "9788858111703" ;
    wdt:P618 "CFI0893220" ;
    wdt:P673 "it" ;
    wdt:P619 "20160201124421.0" ;
    wdt:P671 "ita" ;
    wdt:P788 "XXVI, 351 p. ; 21 cm" ;
    wdt:P1195 wd:Q152 ;
    wdt:P1193 wd:Q155 ;
    wdt:P1218 wd:Q157 ;
    wdt:P1224 wd:Q159 ;
    wdt:P1226 wd:Q160 ;
    wdt:P1228 wd:Q161 ;
    wdt:P603 "\"01429nam0 2200313 i 450 \"" ;
    wdt:P1255 "z01 - i - xxxe" ;
    wdt:P1261 "z01 - n" ;
    p:P644 wds:Q156-d549bbc7-46cd-da5f-95f7-e7fc0d24eef8 .
```

Figure 34. Excerpt of the RDF Turtle export of the *Item* corresponding to the UNIMARC record BID IT\ICCU\CFI\0893220

Thanks to the choices made in the implementation of the *Wikibase data model*, the RDF export can manage a specific information content both as a whole (for example ISBD Area 1 or UNIMARC 200 label) and as an atomic element (for example Title proper or subfield a): this first part of the RDF export (Figure 34) can provide access (in querying or extracting data) to the contents of the "entire" UNIMARC fields (such as Area 1 ISBD, mapped with *Property* P754).

Later in the same RDF textual display (Figure 35), there are also the details of *Qualifiers* and *Sources* for each *Statement*: for example the one corresponding to UNIMARC 200 (ISBD Area 1), where we can found the contents of all the subfields (including subfield a, mapped with *Property* P755) and the *Source* "National Central Library of Florence" which was translated with wikibase:Reference "wdref: 872bd272cb87cb64b2cc120c5ec14406a1546efb".

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458


```
wds:Q156-43f2e678-4ac0-048e-003f-217c61585270 a wikibase:Statement,
        wikibase:BestRank ;
    wikibase:rank wikibase:NormalRank ;
    ps:P754 "Lo Stato innovatore : sfatare il mito del pubblico contro il privato / Mariana Mazzucato ; traduzione di Fabio Galimberti" ;
    pq:P1250 wd:Q180 ;
    pq:P755 "Lo Stato innovatore" ;
    pq:P759 "sfatare il mito del pubblico contro il privato" ;
    pq:P760 "Mariana Mazzucato" ;
    pq:P761 "traduzione di Fabio Galimberti" ;
    pq:P11 "a[1] e[1] f[1] g[1]" ;
    prov:wasDerivedFrom wdref:872bd272cb87cb64b2cc120c5ec14406a1546efb .

wd:Q156 p:P793 wds:Q156-79885b6a-46bf-c78e-f24b-8226b526ed69 .
```

Figure 35. Another excerpt of the RDF export of the *Item* corresponding to the UNIMARC record BID IT\ICCU\CFI\0893220

This more detailed level of RDF export can provide more targeted access to the content of subfields and indicators, or to coded informations (such as the data of the Coded Information Block of UNIMARC).

Wikibase can manage multiple descriptive models within the same Item

By exploiting the possibility to qualify each *Statement* and specify its source, it is also possible to record the standard used for each stored field: this makes it technically possible to manage, within the same *Item*, descriptions produced in accordance with multiple standards and from multiple sources, always maintaining the faculty to recognize and then filter the required fields.

The "coexistence" of multiple standards within the same *Item*, is a different and additional option than the mapping (described above) of the structural elements of the already loaded data and refers to the possibility of creating the same element following different standards or cataloging rules: for example, the date of publication can be recorded by following both the rules for transcribing the information provided by ISBD and those provided by the RDA.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

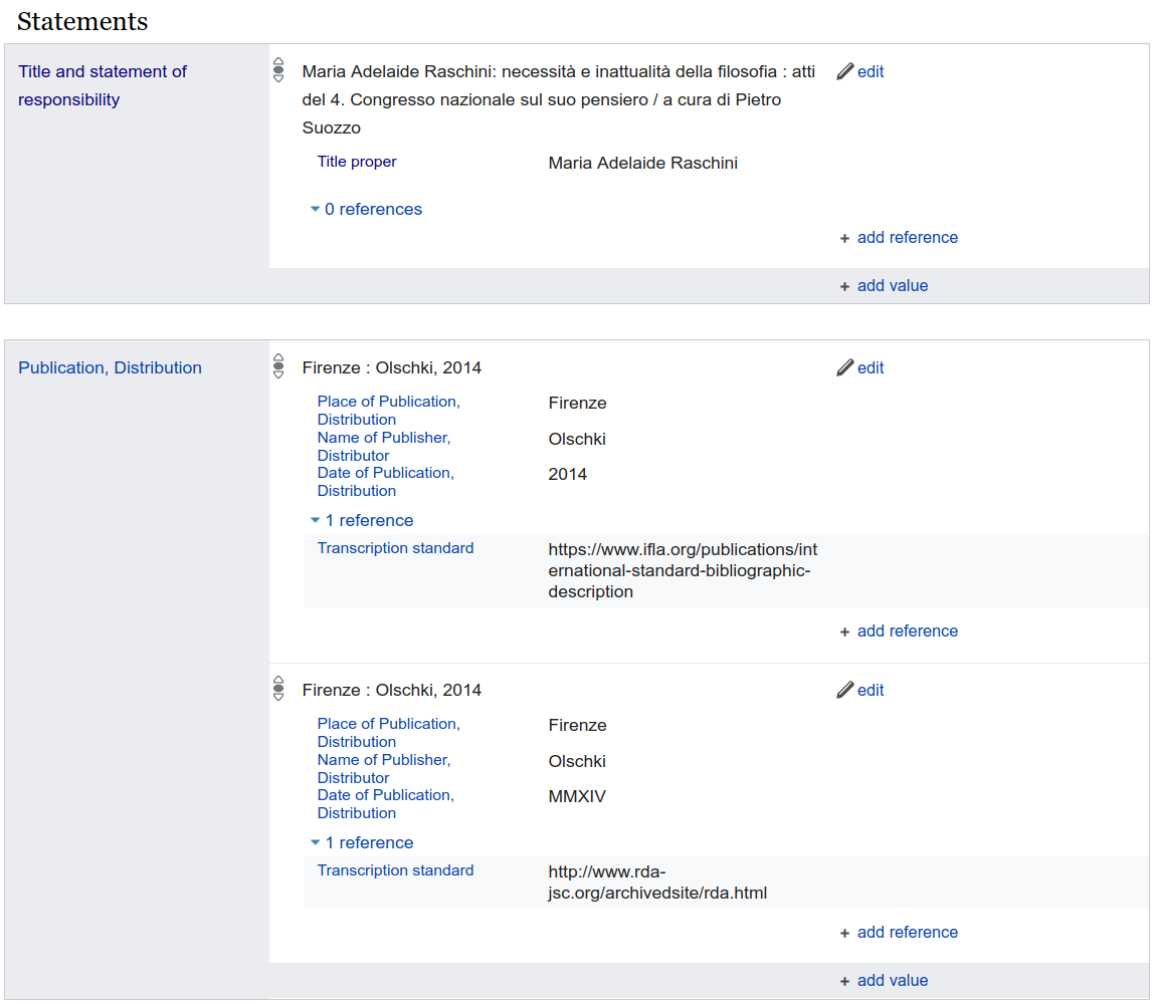


Figure 36. Example of 2 *Statements* corresponding to the publication data recorded following respectively the rules for transcribing the information provided by ISBD, and those provided by RDA (experimented on https://test.wikidata.org/wiki/Q166411)

Wikibase can handle multiple streams of modification

Wikibase offers open and documented interfaces (API) that allow external systems to query and modify managed data; such interfaces, along with the presence of a large amount of basic software libraries produced for Wikibase by the Wikimedia community, give the opportunity to expose the data to multiple editing flows from the outside.

Possible directions

As mentioned, in most of the existing initiatives the publication as LOD (Linked Open Data) of bibliographic data becomes an *additional* activity compared to the traditional flow of creation and use of the bibliographic record: in practice the existing bibliographic data are published as LOD and

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

their publication does not influence their creation. In these initiatives, the publication of data takes place through a process – usually automatic – of semantic conversion (not just syntactic): the ontology (or metadata scheme) traditionally used is converted into a new ontology considered most suitable for the *Web of data*. One of the reference points of these translations is the *FRBR conceptual model*.[54] An exception is BIBFRAME which – as mentioned – deals with both the problem of new strategies of managing bibliographic records, and the problems of the complete semantic conversion of the MARC21 format.

The proposal presented in this contribution starts from the assumption that UNIMARC (and in general the MARC format) is a *de facto* ontology and proposes to make it explicit and to convert it – only at a syntactic level – in Linked Data / RDF structures through the use of the *Wikibase data model*. The bibliographic data hosted in Wikibase could therefore become not only a way to publish data as LOD, but also an environment for the production of bibliographic data. Many could be the advantages of this new environment, among which:

- the collaboration potential of the Wiki model;
- the native ability to process data in a multilingual environment;
- the availability of data (in reading and writing) with open protocols (SPARQL, API Wiki, etc.);
- the availability of open software and utilities with a wide base of development and experimentation worldwide.

To these advantages must be added another: the offer of very agile instruments – in particular the ability to host different ontologies – that can also support the ongoing discussion on the bibliographic control (from FRBR to LRM[55]) without semantic conversions of legacy data that, in a production environment, can be really expensive and problematic.

In the course of over half a century of MARC application we have already witnessed different embodiments of the semantics of the standard (maintained substantially unaltered) from a "basic syntax" – ISO 2709 – to another – XML – with the according evolution of query protocols, from z39.50 to the different web service protocols (for example SRU[56] and SRW[57] based on XML over HTTP). On this path this contribution proposes to convey the semantics of MARC (preserving all the informative detail) through a new syntactic embodiment, the Wikibase Data Model, which allows to exploit a management system already widely used, the Wikibase software, complete with protocols and APIs for querying, creating and editing data.
The double aim of this "new embodiment" of the MARC intends to be, on the one hand, to bring MARC bibliographic data (retaining all the semantic details) towards the mainstream of the most widespread methods of data sharing and, on the other hand, to provide a management tool – or new

[54] The point on FRBR can be found in Karen, 2016.
[55] The reference site for IFLA RLM is: https://www.ifla.org/publications/node/11412.
[56] http://www.loc.gov/standards/sru/.
[57] http://www.loc.gov/standards/sru/companionSpecs/srw.html.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

production environment – that facilitates the MARC ontology mapping with new proposals of a semantic (or ontological) nature in the field of bibliographic data.

The Italian context

Focusing only on the specific case of Italy, using the *Wikibase data model* as a way to publish LODs and, above all, experimenting with the new production environment could be an interesting proposal also for the Italian library network *Servizio Bibliotecario Nazionale*. In particular, the SBNMARC protocol[58] (based on XML messages over HTTP), that is currently used by local library systems (named "poles") to update the central database "Indice SBN" (in fact the Italian union catalog), could read and write – with an appropriate interface – also on a "new Wikibase production environment" implemented in accordance with this contribution: in this way (for example to manage a transition phase) the software currently used by the local library systems could also update the data of the new production environment. This hypothesis would obviously require more investigations, specifications and testing, but from a technological point of view it is a viable solution. In a nutshell, the SBNMARC protocol essentially conveys two types of coded information. The first one is related to the application services that the "Indice SBN" makes available to local systems (or poles): it is an exchange of request messages (for example: Search, Create, Modify, Delete, Align) and of reply messages from the "Indice SBN". The second one refers to the bibliographic content exchanged in those messages: here SBNMARC is fully compliant with UNIMARC semantics (in XML syntax). In order to allow a local SBN library system to interact with a "new Wikibase production environment" it would therefore be necessary to map the application services of the SBNMARC protocol with the open and documented interfaces (API) of Wikibase.

In any case, given the lack in Italy of a large-scale initiative concerning bibliographic data as LOD, even the publication alone of SBN data using the *Wikibase Data Model* would be a significant step.

[58] The reference site of the SBNMARC protocol is http://www.iccu.sbn.it/opencms/opencms/it/main/sbn/evoluz_indice_sbn/pagina_144.html.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12458

References

Baker, Tom. 2013. "Designing data for the open world of the Web." *JLIS.it*, 4(2013), 1, https://www.jlis.it/article/view/6308/7865.

Berners-Lee, Tim. 2006. "Linked data." http://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, Tim, James Hendler and Ora Lassila. 2001. "The semantic web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities." *The scientific American*. http://www.scientificamerican.com/article.cfm?id=the-semantic-web.

Caplan, Priscilla. 2003. "Metadata fundamentals for all librarians." Chicago: ALA.

Coyle, Karen. 2011. "MARC21 as Data: A Start." Code4Lib Journal, 14.

Coyle, Karen. 2012. "Linked data tools: connecting on the Web." ALA Library Technology Reports, 48 (4).

Coyle, Karen. 2016. "FRBR, Before and After: A Look at Our Bibliographic Models." Chicago: ALA. http://www.kcoyle.net/beforeAndAfter/978-0-8389-1364-2.pdf.

Coyle, Karen. 2017. "Precipitating Forward." http://kcoyle.blogspot.it/2017/04/precipitating-forward.html.

Erxleben, Fredo, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. "Introducing Wikidata to the Linked Data Web." In *Proceedings of the 13th International Semantic Web Conference*: 50–65. Springer.

Hallo, María, Sergio Luján-Mora, Alejandro Maté, and Juan Trujillo. 2016. "Current state of Linked Data in digital libraries." *Journal of Information Science* 42:117–127, first published on July 21, 2015 Doi: 10.1177/0165551515594729.

Svenonius, Elaine. 1992. "Bibliographic Entities and Their Uses." in *Proceedings of the Seminar on Bibliographical Records*, held in Stockholm, 15-16 August 1990. Munchen: K.G. Saur: 3–18.

Tennant, Roy. 2017. "'MARC Must Die' 15 Years On." https://hangingtogether.org/?p=6221.

Vrandečić, Denny and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase." *Communications of the ACM,* 57 (10):78–85.