Still Waiting for That Funeral: the Challenges and Promises of a Next-Gen INTERMARC

Sébastien Peyrard
Metadata Department, Bibliothèque nationale de France, Paris, France.
sebastien.peyrard@bnf.fr

Mélanie Roche
Metadata Department, Bibliothèque nationale de France, Paris, France.
melanie.roche@bnf.fr

Abstract:

The National library of France has used its own MARC-flavoured INTERMARC as a production format for almost forty years, and is now revisiting it to make it compatible with Semantic Web technologies and recently developed bibliographic models such as FRBR and IFLA LRM. We believe that keeping MARC as a production format allows for professional continuity and enhances the expertise of bibliographers, while exporting it as Linked Data acknowledges and meets the needs of a community of mainly non bibliographers users. Our objective is to switch from records to linked, reusable, and trustworthy data by increasing the portability of MARC records to a finer-grained level — all the while fitting into the ISO 2709 formalism. The context is that of a national process called Bibliographic transition that aims at native LRMised cataloguing through a French declination of RDA. And the stakes are increased higher by the necessity to implement change management for all those concerned.

Keywords: Bibliographic transition, MARC, INTERMARC, Semantic Web, IFLA LRM

INTRODUCTION

Catalogues were the entry point for computers in libraries. In the late 1960s, the couple formed by library catalogues and computer technologies was made official by the advent of the first Machine-Readable cataloguing format (MARC), USMARC – which became MARC21 in 1999, the most commonly-used MARC flavour in the English-speaking world, used in OCLC's Worldcat. But the love story was soon endangered when it became clear that IT technology was evolving faster than bibliographic formats. To add insult to injury, at the turn of the 21st Century a new conceptual model for bibliographic data called FRBR (Functional Requirements for Bibliographic Records) came and sowed discord in a thirty-year-long idyll. A decade later, RDA (Resource Description and Access), the offspring of FRBR and Semantic Web ideals, consummated the divorce between bibliographic formats and IT by establishing a new set of cataloguing rules indifferent to encoding formats.

There is no denying that so far MARC formats have failed to embrace the manifold challenges posed by new cataloguing models and rules. Yet it seems that too often, speaking of "MARC formats" is a generalisation for MARC21, and pointing to MARC21's shortcomings should not condemn all MARC formats. This is the belief held at the National library of France (BnF, Bibliothèque nationale de France), who has engaged a radical overhaul of its production format, INTERMARC, so as to make it compatible with new standards for the Semantic Web, new bibliographic models, and new cataloguing rules.

Our paper will focus on methodology and results: how the idea of a "Next-Gen INTERMARC" emerged, how we turned it into an actual participative process, and what the format will eventually look like. Considering that we are halfway through the project, all we know for sure at that point is that the next generation of INTERMARC will be entity-oriented, rely heavily on controlled vocabularies, and include meta-metadata.

1. THE CONCLUSIONS THAT DAWNED ON US

1.1. MARC*21* must die

It has been more than fifteen years since the library world was shattered by MARC's obituary[1]. In a *Library Journal* column that has gone down on history, Roy Tennant predicted how MARC formats were sure to come to an end anytime soon, due to their technical obsolescence and overall inadequacy to users' and cataloguers' needs alike. The authors of this paper were not born when MARC formats came into being, nor were even librarians when Tennant's article went out, and therefore cannot be suspected of a personal or professional bias in favour of any such time or format. And yet, speaking as connected millennials, it is our belief that if we have been waiting so long for MARC's funeral, it may be because it is not altogether dead right now.

All of Tennant's arguments are absolutely correct – as far as MARC*21* is concerned. However, it would be unfair to dismiss thus harshly other MARC-based formats. UNIMARC, for one, has relied from the start on local identifiers to establish links between two records (either authority or bibliographic).

[1] https://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/

Tennant himself noted that the term “MARC” “conflat[es] several interrelated things. There are the MARC syntax, the MARC data elements, and the *Anglo-American Cataloging Rules* (*AACR*).” That too is very true for MARC21 but less so where UNIMARC is concerned, as is evidenced by the recent developments of the two formats. Since the Library of Congress decided to adopt RDA in 2012, MARC21 has evolved to accommodate the new rules, but has failed to undergo a structural revision in terms of modelling. However, compliancy with a model underlying one specific set of rules might eventually prove more efficient in terms of data interoperability, especially now that RDA has officially implemented IFLA-LRM – the conceptual model that succeeded to the FR- family of models[2]. This was the road chosen by UNIMARC a few years ago: the UNIMARC Authorities Format (UNIMARC/A) engaged a structural revision as early as 2010 so as to comply with FRAD[3], the new model for authority data that had only just been published. In the last four years, the UNIMARC Bibliographic Format (UNIMARC/B) has also evolved to subsume the concepts of Works and Expressions required for FRBRised cataloguing. Whilst UNIMARC is still far from fully implementing FRBR or IFLA-LRM, it paved a way we decided to take even further.

1.2. The stakes of a national bibliographic agency

Tennant believed that “There are only two kinds of people who believe themselves able to read a MARC record without referring to a stack of manuals: a handful of our top catalogers and those on serious drugs”. At the National library of France, we have a “handful” of about 300 people manipulating MARC data daily, not counting a whole bunch of metadata experts and trainers – and we can testify they are not (all) on “serious drugs”. Ever since the beginning of computer cataloguing in the late 1980s, BnF cataloguers have grown extremely familiar with MARC formats, and as French-speaking people are actually far more comfortable with numbered labels (as esoteric as they might seem to an outsider) than with an English-based format such as XML that Tennant envisioned as the future of cataloguing.

At BnF we have had our own in-house MARC-based cataloguing system for about twenty years, and our own MARC-flavoured format for over forty: INTERMARC was born in 1974[4] and presents a mix between MARC21 and UNIMARC. The labels and overall structure are closer to MARC21, but many of its fundamental features are reminiscent of UNIMARC, such as systematic links between authority and bibliographic records or between two bibliographic records; it is also structured at a finer-grained level than MARC21 and even UNIMARC. INTERMARC is a very versatile and progressive format. Twenty years ago, it already evolved from a series of formats tailored for specific kinds of ressources, into one consolidated format accommodating the description of all material types. The issue back then was one of granularity: we had to come up with a suitable format for all our collection departments, while providing the richness of information expected from a national bibliographic agency. Today, the issue is slightly different, since we are thinking in terms of data interoperability, yet one fundamental question remains the same: do we start again from scratch or do we rely on what we already know and master?

[2] http://www.rda-rsc.org/ImplementationLRMinRDA

[3] Functional Requirements for Authority Data, a conceptual model now superseded by the IFLA-LRM.

[4] http://www.bnf.fr/fr/professionnels/f_intermarc/s.format_intermarc_histoire.html?first_Art=non

The answer now is the same as it was then: for the institution, simply dismissing such a highly skilled workforce would be very dangerous, not only in terms of change management, but also because as the producer of the French National Bibliography, the National library is subjected to produce records within a reasonable time after Legal Deposit, and therefore we cannot risk expanding our cataloguing delays by getting used to a new format.

As a national library, we also serve a community of local libraries that are still using MARC — and will very likely continue to do so in the foreseeable future, because UNIMARC is still a widely-used exchange format, not only in France but also in Europe. It seemed unrealistic to ask them to wait indefinitely for the bibliographic revolution to come. Our responsibility towards those professional users is just as great as towards non professional users, which is why we incorporated our reflexion about INTERMARC into the wider context of a national Bibliographic transition[5].

1.3. Towards a Next-Gen INTERMARC

At BnF, we are not convinced by BIBFRAME developments as a replacement for a MARC-based structure, especially in terms of the interpretation of the FRBR — now IFLA-LRM, model. We do believe that exposing data in RDF is the key to being searchable and reusable in the Semantic Web (which is why we developed data.bnf.fr[6]), but we make a clear distinction between the way the data is produced and the way it is stored and disseminated. Storing and disseminating RDF triples doesn't necessarily mean producing them in current cataloguing.

Tennant believed that "If libraries cling to outdated standards, they will find it increasingly difficult to serve their clients as they expect and deserve." Since 2002 though, it is our contention that libraries *have* become "flexible, responsive organisations" that Tennant called for. Yet this has less to do with technology than with the way with which librarians have used technology. Technology in itself is neither necessary nor sufficient to bring forth meaningful changes, if it does not come with a paradigm shift in mentalities.

Reading Tennant's arguments actually convinced us that such a change of mindset mattered more than any technology shift: "To create standards that are both adequate for present needs and flexible enough to offer new opportunities, we should begin with the requirements of bibliographic description (see Functional Requirements for Bibliographic Records, for example) and devise an encoding standard that provides power and flexibility". With this we could not agree more (barring the fact that we are now speaking in terms of IFLA LRM), and this is exactly what we set up to do.

2. THE METHOD WE USED

2.1. Prefiguration

The concept of a next-gen INTERMARC needed to be tested among the core format experts of BnF. Two brainstorming sessions were set up to discuss the relevance of keeping a MARC-based format, its adequacy with the objectives of bibliographic data in the long run, and its compatibility with Linked Data principles. These brainstorming sessions confirmed

[5] https://www.transition-bibliographique.fr/enjeux/bibliographic-transition-in-france/

[6] http://data.bnf.fr/.

that keeping an internal format that would not be revolutionary to cataloguers was compatible with the very nature of the Bibliographic transition. It also uncovered areas that needed to be further explored, in a context that was as collaborative as possible for the core cataloguing experts to own the approach.

2.2. INTERMARC Camp

These were the topics that needed further exploration:

- Entities and Relationships. How can MARC effectively and efficiently express FRBR?
- Meta-metadata. How can we provide helpful data about the metadata so that users and metadata managers understand its quality in regards to a particular production context? At which levels of granularity?
- Reference information. In a context where the cornerstone for machine-ready, interoperable data is the web of data, how can we upgrade MARC to achieve full linked data compatibility?
- Syntax and format conversions. Is the current syntax (ISO-2709) still relevant? What format mechanisms can change or simplify? What impact does it have on the up-and-running conversions, to keep BnF data interoperable with the rest of the library world?

These topics needed to be discussed and shared across a wide group of experts, as representative as possible of the various cataloguing contexts at BnF (cataloguing workflows, material types…), and in a collaborative setting. Each participant should be allowed to express their opinion on all questions at some point, so as to actively contribute to the evolution of the format and produce a deliverable that the group as a whole would own. The discussions should not remain abstract but take into as much account as possible concrete data encoding and feasibility considerations. For efficiency purposes, the workshop should also take place within a short period of time — but not *too* short though, considering the complexity of the subject at hand.

Obviously, a traditional meeting setting was not adapted to such an approach, which is why we opted for a “World Café”[7] setting. The “INTERMARC Camp” took place within a 2-day period (2017, Feb. 23rd-24th), so as to be efficient, focused and to avoid distraction by other daily tasks.

- The 4 areas identified in the brainstorming sessions were discussed in 4 dedicated workshops that were run as parallel sessions, with a general introduction setting up the rules; each workshop consisted in 3 meetings on day One and 2 meetings on day Two;
- Each workshop had a permanent chair and reporter across the 2 days;
- Each participant would attend a given workshop on day One, and a second workshop on day Two, with the following missions:
    - On day One, imagine which features the next-gen INTERMARC should have and draft a prioritised list of the desired features[8];
    - On day Two, challenge the list drafted the day before by the previous group.

[7] https://en.wikipedia.org/wiki/World_caf%C3%A9

[8] The requirements were prioritised using the MoSCoW method : https://en.wikipedia.org/wiki/MoSCoW_method

- To avoid too much information cluttering during the workshops, 3 *liberos*[9] were nominated with the mission to attend a different workshop at each session and circulate information between workshops;
- The late afternoon of day Two was dedicated to a plenary session gathering all participants. This session was prepared by the chairs, reporters, and *liberos* in the early afternoon, resulting in 4 lightning talks that summarised the conclusions (and open questions) of each workshop. It also allowed each participant to contribute to the two workshops s.he did not attend in a final discussion.
- Last but not least, an idea box circulated among the audience, to receive feedback on what name this next-generation INTERMARC should be given, now that we had a rough idea of what it would look like.

We think this workshop was a great success: INTERMARC being well-known to the audience, they were the right people to discuss its mutation; it allowed sharing a common vision; it also allowed us to get a collective sense of the different needs at stake, with the different cataloguing tracks and/or cataloguing cultures inside the BnF. Eventually, it laid out the foundation for the leading principles that would dictate the evolution of INTERMARC.

2.3. Working Groups

The INTERMARC Camp did not only lay out the foundations, it also needed the ideas to be tested, refined, cataloguing-proof and technically feasible. This work was undertaken by a subset of the INTERMARC Campers in a series of traditional meetings tackling the following topics:

- Entity-driven meetings: Works and Expressions; Manifestations; Items; Agents; Topics and subject indexing; Nomen and Nomen attributes. These meetings were to delineate each entity, the list of attributes and relationships needed, and perform a rough gap analysis with the existing INTERMARC (fields and subfields to keep, to repurpose, to create);
- Bibliographic treatment-driven meetings: collections, performances;
- Mechanism-driven meetings: meta-metadata, controlled values.

These meetings took place between May and December 2017. Then 4 discrete meetings were held between the IT and Metadata Departments to assess the feasibility of the conclusions that we had reached, uncovering areas to further refine. The time was therefore ripe for our Metadata Steering Committee (COMET) to validate the principles behind the next-Generation INTERMARC — which they did on 2018, Feb. 12th. The next step was to draft a Magna Carta for the new format, stating its core principles and providing guidance to the experts who would be tasked with the rejuvenation of good old INTERMARC.

3. THE FUNCTIONALITIES OF THE NEW FORMAT

The features of this next-generation format are now outlined in a Magna Carta, and can be declined into 3 core principles: implementing the entity-relationship design of the IFLA LRM model; expressing relationships at a finer-grained level than that of a “record”; expressing controlled values through individual entities.

[9] https://en.wikipedia.org/wiki/Defender_%28association_football%29

3.1. Entities/relationships

Cataloguers had a linked data mindset before the Web was even born, since for a long time they had been describing things through documents — records — and had linked them through identifiers. The FRBR era defined a core model that reconciled, through core entities, what cataloguers describe and what users are looking for on the Web. One core idea behind INTERMARC-NG is to reconcile these abstract entities with the records in which they are described, which leads us to the following principles:

- Each entity described in INTERMARC-NG must have its own record. This means that a Place should have a single record, when today the same place can be described in two distinct records: one in RAMEAU[10], for subject indexing, and another one in our Geographic Authority file, used to identify places mentioned in our cartographic ressources.
- Each INTERMARC-NG record must describe a single entity. This means de-assembling information that is now mixed up in a single record. Typically, a traditional MARC bibliographic record contains information that pertains to Works, Expressions and Manifestations alike. Work- and Expression-level elements will therefore be moved from the bibliographic record into Expression and Work records, resulting in a Manifestation record that will look like a slimmer bibliographic record. Expression records will have to be created from scratch, whereas Work records will be based on our Uniform Titles Authority format, and enriched with new attributes and relationships.
- Links between entities must be expressed through links between records describing these entities. A link between a Manifestation and an Expression must be expressed by a link from the Manifestation record to the Expression record through the means of its identifier. The main idea is to use the same mechanisms throughout the format, whereas they are now different in the Authority and Bibliographic formats. All in all it boils down to three core subfields: the identifier of the linked entity; the nature of the relationship; the reliability of the link in case of dubious ones (e.g. for *supposed* or *alleged* creators of Works)

We nevertheless decided that the LRM E9-Nomen and LRM-E11Time Span entities would not be treated as distinct records, but as attributes of the record. The relationship between an entity and a Nomen or Time-Span will be expressed directly in the entities that they describe, with the tag or subtag that contains them expressing how they are used. This has the advantage of staying as close as possible to the existing structure and avoiding entity cluttering.

Specific subclasses of Res will have to be created locally to express specific entities, such as fictional agents and places, concepts, events, cultural events, Dewey concepts and controlled values.

The resulting data model is summarized in figure 1.

[10] RAMEAU is the French equivalent of LCSH. See http://rameau.bnf.fr/index_en.htm.

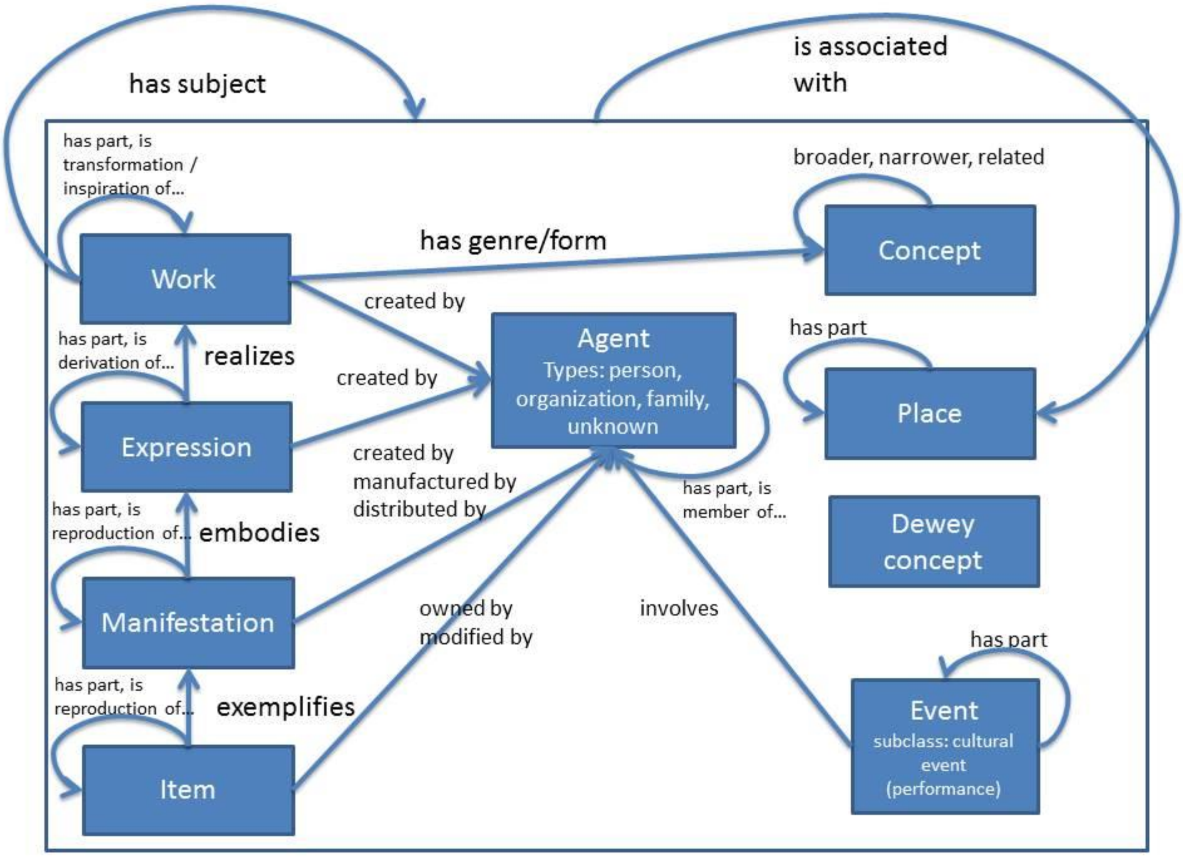
has subject
is associated with
has part, is transformation / inspiration of...
broader, narrower, related
Work
has genre/form
Concept
created by
Agent
Types: person, organization, family, unknown
has part
has part, is derivation of...
realizes
created by
Place
Expression
created by manufactured by distributed by
has part, is member of...
has part, is reproduction of...
embodies
Dewey concept
Manifestation
owned by modified by
involves
has part
has part, is reproduction of...
exemplifies
Event
subclass: cultural event (performance)
Item

Figure 1: Simplified Intermarc-NG data model, with one record per entity and one MARC field per relationship.

We also addressed the question of subject indexing, at a time when the French subject indexing language, Rameau, is currently being remodelled. The main challenge is to allow for relationships between entities that are relevant in the context of a given work: France has no consubstantial relationship with Cooking (☺) nor with the 19th century, but such a relationship between a concept, a place and a time-span becomes true in the context of a Work about French cooking in the 19th century. The role of MARC subfields in the Subject indexing tag is to be repurposed from expressing the structure of the subject headings to expressing the way the subject of a Work is composed of entities that are grouped or linked.

3.2. Meta-metadata

Since 2014, BnF opened up its bibliographic data as freely reusable data[11]. From then on, BnF data could be freely reused and integrated to external datasets, provided that the source (BnF) is explicitly stated. However, BnF also uses its own sources to establish information (birth dates and places, preferred names and variants…). It would be ideal to provide such information to assess the quality of an individual fact in our data that one could explicitly quote in its own dataset. For instance, in Wikidata, the language of the French songwriter Georges Brassens was imported from the corresponding information field in the BnF data.

[11] The “Etalab” licence is a French open license, roughly equivalent to CC-BY. See http://www.etalab.gouv.fr/licence-ouverte-open-licence.

BnF is stated as a source. It would be great to provide the source actually used by BnF (*Who's who, 20th century*) to establish this individual fact.

This means that a record, but also an individual information field in the source data, should be sourced. The former already exists in the INTERMARC structure (sources for the authority record), but not the latter. It was decided that 4 meta-metadata elements would be created at the level of a MARC tag:

- source of a manifestation statement;
- link to an internal source (if the BnF has the corresponding Manifestation);
- link to an external web resource through its URL;
- for an external web resource, date the URL was accessed.

However, most of the time, a given source provides information for all the information fields in the record. To avoid repetition, it was decided to have two levels of information for the source when cataloguing: one generic record-level source information, and one tag-level source information, for specific cases. In other terms, a tag without source information has the general sources statement as a source. This makes a good trade-off between efficient cataloguing and fine-grained information, when required. Upon export, each MARC information field will be provided with source subfields, which will be retrieved from record level source fields when they are missing in the original record.

The attributes of the LRM-E9 Nomen (that is, a given access point or a given character string) are also defined as meta-metadata elements. It was therefore decided that they should have their own information subfields, such as: encoding scheme (for authorised access points), intended audience, context of use, language, script and script conversion. Since such meta-metadata elements are relevant for virtually any of the MARC information tags, each meta-metadata element has to be assigned a generic subfield code that will always be the same, no matter which field uses it. To meet that end, we pushed MARC syntax a step further by using capital letters as subfield codes and reserving them to meta-metadata elements.

Other meta-metadata elements were key to assess metadata quality and inform decisions, such as:

- history of a record and its successive corrections and enrichments;
- origin of an information field;
- license for the metadata.

Such information was deemed irrelevant in the MARC structure, because it is not useful in day-to-day cataloguing workflows, but only for retrospective cataloguing (the history of the information record, the origin of an information field, and so forth) or for external reusers (licensing information). Such information is still vital, but will be expressed in a different event-oriented format that still has to be defined.

3.3. Controlled vocabularies

INTERMARC data should be compliant with the aims of the Semantic Web, that is a Web in which "machines become capable of analysing all the data on the Web"[12]. This means that each data point that is not a label should be shaped as a controlled vocabulary composed of

[12] Berners-Lee, Tim, *Weaving the Web: the Original design and Ultimate Destiny of the World Wide Web*, HarperBusiness: 2000, p. 157.

individual values that are all expressed through the same mechanism. This matched an internal need to allow each controlled value to be attached the following attributes whenever needed:

- preferred and alternate labels;
- links to narrower, broader or related values;
- history and usage notes for the statements;
- a code, in case such values have a standard code (e.g. ISO language codes);
- the set of values (controlled vocabulary) it belongs to.

Such information fields can already be expressed in INTERMARC, especially for RAMEAU subject headings. It was decided to try and reuse such mechanisms, and apply them to controlled values.

Many subfields and fixed length coded fields already use controlled values in the INTERMARC format, but such values only associate codes with labels. In INTERMARC-NG, it is therefore necessary to:

- consider each controlled value as a distinct entity that should have its distinct record (principle 1) and identifier;
- reuse the existing mechanisms in RAMEAU to express the different information fields on a given controlled value.

3.4. Impact on the overall syntax of the format

The aforementioned principles and generic mechanism conflict with certain specifics of the MARC structure that had to be abolished:

- Meta-metadata could not be expressed on fixed length codes fields in the INTERMARC structure, unless using complex internal links mechanisms for which MARC was not shaped. Neither could it easily be expressed for indicators for the very same reasons. However, it could be well expressed on non-fixed fields, with meta-metadata dedicated subfields;
- Controlled values could only be expressed through codes, not identifiers, in fixed length coded fields and indicators. They could be expressed in subfields, by storing the identifier for the value instead of the existing code in the subfield. Furthermore, the fixed fields and indicators have a limited number of values.

For such reasons, it was decided to **abandon indicators and fixed length coded fields,** and replace both with standard subfields within a MARC tag, which will store the record identifier for the individual controlled value. This will avoid defining 3 different mechanisms in the format to handle controlled values and meta-metadata. The resulting format will also have a homogenized syntax.

CONCLUSION

In conclusion, what Tennant provided sixteen years ago was a very powerful analysis of one local context and one single MARC format, but such an analysis proves weaker when applied to other machine-readable formats. The authors of this paper couldn't agree more with all of Tennant's premises, but instead of RSVPing to his hasty invitation to a funeral, we at BnF chose rather to explore the possibilities of rising MARC from the dead.

Defining a Next-Generation INTERMARC proved a challenge in terms of organisation, but we soon realised that experts generally agreed with the leading principles the new format needed to implement. We are still in the early stages of the development, which explains why so far we have only been able to delineate the scope of our Next-Gen INTERMARC. How it will be drafted by the format specialists, interpreted by our IT department, and integrated by our "handful" of cataloguers, is an altogether different story.

Acknowledgments

The authors of this paper wish to thank Françoise Leresche, whose in-depth knowledge of INTERMARC, ISO 2709 and all things related to bibliographic formats proved invaluable. We also wish to address many thanks to all INTERMARC Campers for their enthusiasm, dedication and expertise.

References

BERNERS-LEE, Tim. *Weaving the Web: the Original design and Ultimate Destiny of the World Wide Web*, HarperBusiness, 2000.

LE PAPE, Philippe. "La vie quotidienne d'UNIMARC au temps de la Transition bibliographique", 2017. Available online: http://www.bnf.fr/documents/jsyd2017_le-pape.pdf

TENNANT, Roy. "MARC must die" *Library Journal*, 2002. Available online:
https://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/

BnF website:
http://www.bnf.fr/fr/professionnels/f_intermarc/s.format_intermarc_histoire.html?first_Art=non