How FAIR is MARC?: FAIR Data Principles applied to a bibliographic data standard

Brian Dobreski
University of Tennessee-Knoxville, USA
bdobreski@utk.edu

Heather Moulaison-Sandy
University of Missouri, USA
moulaisonhe@missouri.edu

Bradley Wade Bishop
University of Tennessee-Knoxville, USA
wade.bishop@utk.edu

Abstract

FAIR Data Principles provide a framework for considering how best to make data available in a way that is 1) findable, 2) accessible, 3) interoperable, and 4) reusable. Designed to be simple to understand and machine-actionable, FAIR principles support data use and reuse. This conceptual paper investigates the application of FAIR principles to bibliographic data through an examination of the current standard for encoding library records, MARC. To this end, this paper begins by describing the FAIR principles. It then looks to understand the MARC standard and applies the FAIR principles to the data affordances provided by the MARC encoding itself. In doing so, it probes the question of the extent to which MARC, as a standard, is FAIR. Ultimately, MARC is historically designed for machine-readability, not machine-actionability; although it is well suited to the description of bibliographic materials and is widely used, it does not adhere fully to any of the four FAIR principles. Even so, this examination suggests that FAIR principles could be useful in assessing specific MARC record datasets, particularly as bibliographic data is more widely shared and reused.

Keywords: FAIR Data Principles; MARC standard; library (meta)data: machine-actionable data; bibliographic data

1. Introduction

The concept of data sharing to advance science by permitting others to verify results, replicate experiments, and lead to new application and discovery through data re-use, are long-standing goals of all open data movements. Whereas humans can attempt to overcome (or ignore) incomplete data and metadata for reuse, machines lack those representative heuristics to assume, and only compute (Pryor, 2012; Bishop et al., 2019). To address the crisis in findability and ultimately in usability, the research data community has developed principles to support access. The FAIR Data Principles provide all data-intensive fields of science with brief, easily read and understood aspirational attributes for data to be *machine-actionable data*. Machines not only process data, but also, with machine learning, make discoveries humans cannot. From this perspective, machine-actionable data must be findable, accessible, interoperable, and reusable without human intervention.

One concern is the nebulous concept of data (Furner, 2017); is data created or recorded, or both? The combination of the terms 'metadata' and 'data' into "(meta)data" exists because in some natural and life sciences there are no clear distinctions between the two concepts (e.g., specimen lists are data and metadata). One researcher's data could be another's metadata. Similarly, bibliographic data collected by libraries is in fact metadata that describes library resources. To alleviate some of the ongoing confusion and misuses of FAIR, some of the original authors re-emphasized machine-actionable as the central reason for the principles (Mons et al., 2017).

This paper presents a conceptual exercise considering the FAIR Data Principles in relationship to a well-known standard for bibliographic data, MARC (MAchine-Readable Cataloging). MARC is currently the most commonly used machine-readable bibliographic data format, designed by a

particular community (i.e., the library community) for a specific purpose (i.e., cataloging library materials).

Despite the importance of the FAIR principles to the task of making data available, the metadata community has not, to our knowledge, studied the way that FAIR principles might apply to the standards in use in libraries, including those like MARC that encode bibliographic data. Although some work has been done in measuring the FAIR Data Principles for data repositories and science research data (e.g., Wilkinson et al., 2018), no previous research has looked at the MARC standard or individual records' FAIRness. To begin to address this gap in the field's understanding of FAIR's applicability to the stores of bibliographic data maintained in libraries, this paper investigates the ways in which FAIR principles could be used to assess the MARC bibliographic standard. Using the standard itself as the unit of analysis, rather than any specific records or sets, the authors compared the FAIR framework to the official documentation for the MARC standard (Library of Congress, 2022a), looking for correlates between FAIR requirements and MARC affordances. Another way to think of this question is to ask "How FAIR is MARC?" – and to consider the best case scenario in terms of the possibilities for FAIR bibliographic data that the MARC standard permits. This can be seen as an initial exploration that opens the door for assessment of specific bibliographic datasets using the FAIR principles.

2. FAIR Data Principles

In 2016, the Future of Research Communication and e-Scholarship (FORCE11) created the FAIR Data Principles as guidelines to describe aspirational attributes that any meta(data) should address to be machine-actionable. The original paper, since cited more than ,7,300 times as of August 2022, provides fifteen principles to enable machine-actionable data reuse (Wilkinson et al., 2016). Table 1 presents the FAIR Data Principles.

TABLE 1: FAIR Data Principles (Wilkinson et al, 2016).

Principles	Defined
To be findable:	F1. (meta)data are assigned a globally unique and eternally persistent identifier. F2. data are described with rich metadata. F3. (meta)data are registered or indexed in a searchable resource. F4. metadata specify the data identifier.
To be accessible:	A1. (meta)data are retrievable by their identifier using a standardized communications protocol. A1.1 the protocol is open, free, and universally implementable. A1.2 the protocol allows for an authentication and authorization procedure, where necessary. A2 metadata are accessible, even when the data are no longer available.
To be interoperable:	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles. I3. (meta)data include qualified references to other (meta)data.
To be reusable:	R1. meta(data) have a plurality of accurate and relevant attributes. R1.1.(meta)data are released with a clear and accessible data usage license. R1.2.(meta)data are associated with their provenance. R1.3.(meta)data meet domain-relevant community standards.

By retaining the everyday meaning of *fair* for all involved parties, including decision makers, FAIR caught fire and has been widely adopted across many organizations and disciplines (Wolstencroft *et al*., 2017; Rodríguez-Iglesias *et al*., 2016; Diepenbroek *et al*., 2017; Lightsom et al., 2022).

3. MARC and FAIR

Bibliographic data is metadata used to represent and provide access to library resources, and is typically stored in records and compiled into discovery tools such as catalogs. The content of these records is governed by content standards, which vary according to linguistic and cultural setting.

A separate encoding standard is used to make this data machine operable, and in most modern library settings, this standard is MARC.

Developed in the early 1960s, the MARC format was designed to enable to conversion of card catalogs into electronic databases. In the past 60 years it has become the de facto standard for encoding bibliographic descriptions in electronic catalogs worldwide (Joudrey, Taylor, & Miller, 2015, p. 796). Today, MARC continues to serve as the primary electronic and communication format for data about library resources. This aging standard has been cited for its lack of extensibility, its incompatibility with web technologies, and its general marginalization among information technology standards (Tennant, 2002). MARC is, however, commonly used in libraries and throughout the cultural heritage sector. It is also, to an extent, customizable. Local practice can and does affect how MARC records look, including the standards used in the MARC records to record data, the way that data is entered, and even the kinds of resources selected for treatment.

Another advantage of the MARC standard is that its records are machine-readable. In 1967, machine-readable meant that records could be stored on magnetic tape, and, when that tape was fed to a machine, the machine could display the record and print a physical copy of it (Avram, 1968). In this way, "machine-readable" denoted a limited range of action, as opposed to "machine-actionable," which allows other functions. As electronic and online catalogs emerged in the ensuing years, the MARC format would eventually enable a wider array of actions, including sending and receiving, searching, retrieving, editing, and linking.

In 2022, an era of FAIR principles, how does MARC's current machine-actionability stand up? Table 2 gives an overview of the potential for FAIRness for MARC-encoded bibliographic data, by comparing FAIR requirements to affordances of the MARC standard itself. Overall, MARC data has the potential to be FAIR, even if the MARC standard was not conceived to address FAIR concerns.

TABLE 2: Potential for FAIRness of MARC.

FAIR Principle	Implications of MARC
To be findable:	F1. MARC does not prescribe a globally unique and eternally persistent identifier; the resources described often have one, however, which permits duplicate records to be matched on the back end. F2. The metadata included in MARC records is rich and is well-adapted to describing bibliographic materials. F3. This varies by institution. In the case of OCLC member libraries, MARC records will be uploaded and the *contents* will be made publicly searchable through Open WorldCat (though, not the records themselves). F4. MARC records can encode various identifiers such as ISBN, UPC, and CODEN.
To be accessible:	A1. MARC records are retrievable using standardized protocols such as Z39.50 and SRU A1.1 These protocols are open, but often require specialized software. A1.2 MARC retrieval protocols can be configured to require authorization. A2. This varies by institution, though MARC records are usually not retained for deaccessioned materials.
To be interoperable:	I1. MARC records are widely used in libraries; outside of the cultural heritage sector, however, MARC is essentially unknown. I2. Best practices for MARC require the use of vocabularies or standards, though not all are FAIR because they are not freely accessible (e.g., Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) have fee-based access). I3. MARC allows for the recording of URIs in various fields, but there is room for improvement in the provision of qualified references to other (meta)data.
To be reusable:	R1. MARC bibliographic records are ostensibly accurate (often community-vetted) and reusable; because the standard is community-developed, attributes of MARC records are relevant to the description of bibliographic resources. R1.1. MARC 038 field allows the recording of data usage license information about the metadata itself. R1.2. MARC fields such as 040 allow some provenance information indicating where records were created and possibly where edited or enhanced. R1.3. MARC is a community-developed and maintained standard. Applications of local practice in the choice of data provided, the input standards used, and the quality of the data (e.g., accuracy, correctness) (Margaritopoulos et al., 2008) affect the consistency (and reusability) of MARC records.

4. Implications for Bibliographic Data

The brief comparison in Table 2 reveals that the MARC standard enables, in part, each of the four FAIR Data Principles. As a standard, however, MARC does not adhere fully to any of them, indicating there is likely room for improving the FAIRness of bibliographic data, either through additional practices or encoding via alternate formats.

MARC's strengths seem to be in its long history of use in the library community and its wide adoption, which has led to a standard that is incredibly well-adapted to the materials it describes. This strength, however, is conceivably related to a number of the weaknesses in terms of MARC's FAIRness. Perhaps because of the close-knit community of cultural heritage institutions, creating documentation that would allow for broader access to MARC records in a FAIR manner has not been a priority. Many of these insufficiencies would be challenging to address. For example, URIs for MARC records would be easily created, yet who would mint and coordinate them? Only permitting FAIR vocabularies or standards to be included would also help with the FAIRness of MARC bibliographic data, yet curtailing libraries' vocabulary choices and expecting them to abandon non-FAIR vocabularies such as DDC seems unreasonable.

While some challenges to FAIRness are likely to persist due to the nature of bibliographic records and current cataloging practices, other encoding formats may better afford FAIRNness and alleviate at least some issues. MARCXML serves as a currently available alternative to standard MARC formats, and has been used by Library Congress as well as OCLC (Library of Congress, 2022b). While it offers the benefits of XML, no data on its actual adoption throughout the library community is available. Emerging, linked data formats likely hold more promise for increasing the FAIRness of bibliographic data, for example, Library of Congress's RDF-based BIBFRAME standard. Further opportunities to evaluate BIBFRAME in relation to FAIR will arise as more data and data practices become established. In either case, the FAIRness of actual bibliographic data will reflect the resources, infrastructure, and training available at any given institution, which can vary widely.

For now, the current work opens up new discussion around FAIR, library data, and library practices. While this exploration focused on affordances of the MARC standard itself, it must be noted that how MARC is used can vary from institution to institution, and even from collection to collection. This raises the question, at what level should we consider FAIR? Standard/schema? Community? Institution? Dataset? At a broader level, further exploration and discussion is warranted around if libraries can, and should, better adhere to FAIR principles with their data, including what are the benefits of FAIRer, machine-actionable bibliographic data, and who may benefit from it.

5. Conclusion and Future Study

This conceptual paper reveals that bibliographic data adhering to the MARC standard has the potential to adhere to many of the FAIR principles. Unsurprisingly, given its history and the culture in which it has evolved, the MARC standard is not, however, FAIR-compliant, nor was it intended to be any more than a means of sharing machine-readable bibliographic data among libraries. The present study opens up new questions around the FAIR principles and library standards and data. Libraries should not find themselves (and their data) excluded from the open science movement over the use of an encoding scheme, no matter how well-adapted and common it may be. As libraries look to share their data for reuse more widely on the web, additional machine-actions must also be taken into consideration, including potentials for use with visualizations, machine learning, and AI, a range of actions which MARC was never designed to enable.

At the same time, in practice the MARC standard may be used differently from institution to institution, and an investigation of actual MARC records which accounts for local practices emerges as a possible next step to this initial investigation. Indeed, some MARC records are likely FAIRer than others. Additionally, FAIR may be more appropriate to apply to specific datasets or types of materials (e.g., digital) rather than the entire encoding standard.

As libraries look to leave MARC behind for standards capable of accommodating linked data infrastructures, emerging alternatives such as BIBFRAME should also be investigated for their potential for FAIRness and their ability to support open science and open access. A comparison of the same bibliographic dataset, encoded in both MARC and BIBFRAME, could also offer insight while further delving into questions around what level FAIR data principles should be applied and evaluated at. Regardless of encoding format, with greater consideration of FAIR principles bibliographic data will be better positioned to benefit from advances in machine operability, enabling improved organization, access, use, and reuse.

References

Avram, H. D. (1968). *The MARC pilot project: Final report*. Library of Congress.

Bishop, B. W., Hank, C. F., Webster, J., & Howard, R. A. (2019). Scientists' data discovery and reuse behavior: (Meta)data fitness for use and the FAIR Data Principles. *Proceedings of the Association for Information Science and Technology, 56*(1). https://doi.org/10.1002/pra2.4

Diepenbroek, M., Schindler, U., Huber, R., Pesant, S., Stocker, M., Felden, J., … Weinrebe, M. (2017). Terminology supported archiving and publication of environmental science data in PANGAEA. *Journal of Biotechnology, 261*, 177–186. https://doi.org/10.1016/j.jbiotec.2017.07.016

Furner, J. (2017). Philosophy of data: Why? *Education for Information, 33*(1), 55–70. DOI:10.3233/EFI-170986

Joudrey, D. N., Taylor, A. G., & Miller, D. P. (2015). *Introduction to cataloging and classification*. Libraries Unlimited.

Library of Congress. (2022a). *MARC21 format for bibliographic data*. https://www.loc.gov/marc/bibliographic/

Library of Congress. (2022b). *MARCXML*. https://www.loc.gov/standards/marcxml/

Lightsom, F.L., Hutchison, V.B., Bishop, B., Debrewer, L.M., Govoni, D.L., Latysh, N., and Stall, S. (2022). Opportunities to improve alignment with the FAIR Principles for U.S. Geological Survey data. *U.S. Geological Survey Open-File Report 2022–1043*, https://doi.org/10.3133/ofr20221043.

Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., & Manitsaris, A. (2008, September). A Conceptual Framework for Metadata Quality Assessment. In *Dublin Core Conference* (pp. 104-113). https://library.oapen.org/bitstream/handle/20.500.12657/32535/610315.pdf?sequence=1#page=120

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use, 37*(1), 49–56.

Pryor, G. (2012). Why manage research data? In G. Pryor (Ed.), *Managing research data* (pp. 1–16). Facet Publishing.

Rodríguez-Iglesias, A., Rodríguez-González, A., Irvine, A. G., Sesma, A., Urban, M., Hammond-Kosack, K. E., & Wilkinson, M. D. (2016). Publishing FAIR data: An exemplar methodology utilizing PHI-base. *Frontiers in Plant Science, 7*, 641. https://doi.org/10.3389/fpls.2016.00641

Tennant, R. (2002). MARC must die. *Library Journal, 127*(17), 26–27.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*(160018). https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. *Scientific Data, 5*, 180118.

Wolstencroft, K., Krebs, O., Snoep, J. L., Stanford, N. J., Bacall, F., Golebiewski, M., … Goble, C. (2017). FAIRDOMHub: A repository and collaboration environment for sharing systems biology research. *Nucleic Acids Research, 45*(D1), D404–D407. https://doi.org/10.1093/nar/gkw1032