Data Dictionary: Metadata for Phonograph Records

Catherine Lai
Schulich School of Music
McGill University
Montreal, QC Canada H3A 1E3
lai@music.mcgill.ca

Ichiro Fujinaga
Schulich School of Music
McGill University
Montreal, QC Canada H3A 1E3
ich@music.mcgill.ca

Abstract

The creation and maintenance of a metadata data dictionary is essential to large-scale digital repositories. It assists the process of data entry, ensures consistency of records, facilitates semantic compatibility and interoperability between systems, and, most importantly, forms the foundation for efficient and effective information retrieval infrastructure.

In this paper we explain in detail the necessity of metadata data dictionaries to digitization projects and digital library retrieval services. We also describe the development process of our Data Dictionary for phonograph records. We then present the underlying data model of our Data Dictionary and provide information about the meaning and use of semantic units defined in the Data Dictionary. We stress the usefulness of the generation and maintenance of our Data Dictionary for MIR as it provides a means to ensure accurate, consistent, and comprehensive metadata annotation. For maximum interoperability between systems, digital repositories not only need to agree on the same metadata fields, but also the meanings of the fields. To this end, we believe our Data Dictionary is the cornerstone of optimal retrieval of music information about phonograph records.

Keywords: Metadata, Data Dictionary, Phonograph Records, Digitization, Standardization, Management

1. Introduction

Libraries have a history of managing large collections of information and using technology to carry out associated services such as delivering bibliographic information in the form of electronic records to library patrons. With the advent of the Internet, patrons' expectations for access to information have increased dramatically. Libraries, archives, and cultural institutions worldwide are thus reinventing their roles in today's networked society to meet patrons' new demands.

As digitization technology, metadata standards, data management techniques, and digital preservation awareness have evolved and advanced, libraries, archives, cultural institutions, and other organizations have initiated digitization programs for preservation of their unique and precious analogue holdings, including analogue sound recordings such as 78 rpm and long-playing phonograph (LP) records. Development of digitization programs lowers the barriers to physical access. With the availability of digital collections, users can gain direct access to the digitized versions of artifacts as well as retrieve the metadata stored with the digital objects through automated library services such as online catalogues.

Metadata issues are central to discussions about the evolution of digital library information retrieval services. There are many definitions of metadata. In fact, the Task Force on Metadata compiled more than 25 different definitions of the term in the appendix section of its final report. The formal working definition of the term "metadata" by the Task Force is "metadata are structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities" [1].

Another phrase that is often heard in discussions of metadata is the term "metadata schema." There are also many definitions of the term. The definition by Murtha Baca in her *Introduction to Metadata* edited for the Getty Research Institute defines metadata schema as "a set of rules for encoding information that supports specific communities of users" [2].

Metadata is indeed important in this digital era where digital objects need to have information attached to them simply in order to be found. However, if digital collections are to be effectively retrieved and shared among digital repositories, an agreement on adoption and use of standards needs to be established. Metadata that is randomly or arbitrarily added to a digital object without any overarching principles or established framework will lack interoperability with other resources. It will consequently be difficult to locate, and therefore be underused.

The ultimate purpose of the implementation of this metadata Data Dictionary is to facilitate retrieval of digital collections and promote interoperability between different systems by defining the semantics (i.e., what each metadata element means) of the comprehensive list of metadata elements developed for digitized representations of phonograph records. Metadata permits efficient and effective information retrieval only if digital repositories agree not only on the same metadata fields, but also the meanings of the fields. Our Data Dictionary clarifies the scope and type of metadata associated with the digital representations of the original recordings (e.g., label issue number vs. matrix number). It helps to prevent duplicate handling of data, inconsistencies, and lack of integrity during metadata entry (e.g., as part of the digitization process) and data management.

The next section describes the development of metadata in the context of the history of cataloguing, and shows why metadata and data dictionary have become so important in recent years for retrieval of digital objects. Section 3 then explains the need for a data dictionary in order to assist projects digitizing phonograph records and digital library information retrieval services. Section 4 describes the development process of our Data Dictionary. The paper then presents the underlying data model of our Data Dictionary and provides information about the meaning and use of semantic units defined in the Data Dictionary.

2. Historical Background

2.1 Library Catalogues

The concepts and techniques of metadata creation have been around since the first library catalogue was created more than 2000 years ago. The first appearance of the term metadata dates back to the 1960s and became established in the context of database management systems in the 1970s. The first national cataloguing code, arranged by author entry, was traced back to the French code of 1791, which used catalogue cards and rules of accessioning and cataloguing. Various cataloguing rules were developed and published in different countries, for example, Sir Anthony Panizzi's cataloguing rules for the British Museum Library in 1841 and Charles A. Cutter's *Rules of a Dictionary Catalog* in the U.S.A. in 1876. Library associations in the two countries continuously worked to develop and improve cataloguing rules. In 1904, the American Library Association and the Library Association in the UK co-operated to produce an international cataloging code [3].

At the International Conference on Cataloguing Principles in Paris in 1961, international participants drafted twelve "Paris Principles" to draw a common basis for the assignment and form of access points (e.g., subject headings). The American and the British library associations cooperated again and published in 1967 the first edition of the *Anglo-American Cataloguing Rules (AACR)*, based on the "Paris Principles." As a means for the international exchange and sharing of bibliographic information, in 1974, the International Federation of Library Associations issued the *International Standard Bibliographic Description (ISBD)*. The second edition of *AACR*, *AACR2*, published in 1978, incorporated *ISBD* and brought cataloguing of non-book materials into the mainstream [4]. Libraries worldwide have gradually come to adopt and use the interpretation and implementation of *AACR2* for cataloguing of bibliographic records. The adoption of a common standard has enabled an authority control of bibliographic records and facilitated transparency in the exchange and sharing of catalogue records. The most recent revised edition of *AACR2* was published in 2002 (*AACR2r*) [5].

2.2 Electronic Catalogues

Since the late 1960s, the sharing of catalogue records has become computerized. The fundamental tool that enables the exchanges of catalogues in computer-based systems is the MARC (Machine-Readable Cataloging) format developed by the Library of Congress. The availability of MARC records has benefited library patrons to a wider access of searchable catalogues. The implementations of *AACR2* and MARC have made universal bibliographic control and interoperability between different systems possible and have since then provided an efficient method of retrieving items from library holdings.

2.3 Metadata Initiatives

The idea of cataloguing the resources on the World Wide Web bourgeoned in the mid-1990s in response to the tremendous growth of the Internet. The idea became known as the Dublin Core Metadata Initiative (DCMI), which provides a minimum set of 15 metadata elements designed to describe document-like web objects to facilitate resource discovery [6]. DCMI is an ongoing initiative. Its goal of annotating metadata to help manage and retrieve electronic resources has led to a widespread use and application of metadata across different disciplines. Some of the most well-known metadata standards are EAD (Encoded Archival Description), TEI (Text Encoding Initiative), VRA (Visual Resource Association) Core, and MPEG-7 (Motion Picture Expert Group).

3. Rationale for the Data Dictionary

A new set of metadata standards is necessary for describing digitized representations of phonograph records because traditional formats and standards for cataloguing sound recordings such as described in [7, 8, 9] are inadequate for the encoding and retrieval of digital representations of phonograph records. Specifically, the traditional practices of cataloguing sound recordings are generally limited to bibliographic description of relatively few elements [10]. Information about artwork or photographs in the album packaging of a LP, for example, is usually not included. A few recognized authorities have begun to contribute and expand the utility of metadata to sounds. IASA (International Association of Sound and

Audiovisual Archives), an association working to support the professional information exchange between audiovisual archives in all fields, provides cataloguing rules and guidelines on the production and preservation of digital audio objects. However, the cataloguing rules only cover descriptions of sound recordings in general [11]. The rules are not refined enough to provide a foundation for information retrieval of digital representations of phonograph records. MPEG-7, a formal system for describing multimedia content, defines elements for description of audio and video content [12]. However, MPEG-7 does not have characteristics tailored to the structural complexity that is necessary for the full description of sound recordings. For example, significant types of important information at the individual track or song level are missing. Unspecified information such as recording location, recording session date, recording engineers, or recording equipment used, if available, are potentially useful and convenient retrieval points. MPEG-7, instead, places most emphasis on low-level perceptual features of multimedia data which are meant to be extracted automatically and the data is typically not human-readable. We therefore, created a set of metadata elements specifically designed for phonograph records for the encoding of metadata as part of a large phonograph records digitization management system [13].

A metadata data dictionary, which provides semantic meanings to metadata elements, is perhaps even more important than the metadata element set. Unlike in the library community where cataloguers have well-established rules for data content description in the form of the *AACR*, the lack of a data dictionary can cause problems within and across digital collections. Digitizers may encode the same data element using different metadata elements, or they may encode different data elements using the same metadata element. As a result, digital repositories may not be able to combine or map data across systems because the definitions used are not consistent or identical. A data dictionary (similar to *AACR2*) that systematically defines the semantic meaning, descriptive requirement, and formatting principles of metadata is therefore necessary to create semantic compatibility (i.e., consistent assignment and form of content data to metadata entries), facilitate efficient interoperability between systems, and provide effective searching mechanism.

4. Development Process

4.1 Methodology

Taking the set of metadata elements previously developed for phonograph records, we conducted a comparative study of existing data dictionaries of well-established metadata standards in closely related fields and used it as a guideline to assist in building our Data Dictionary. The data dictionaries studied included:

- AACR2r
- California Digital Library (CDL) for administrative and structural metadata
- CDWA for arts metadata
- Dublin Core Metadata Initiative
- EAD
- IASA cataloging rules for digital audio objects
- MARC 21
- NISO Z39.87 for technical metadata for digital still images
- PREMIS for preservation metadata
- Variations2 for musical work metadata
- VRA for cultural work and images metadata

4.2 Implementation Considerations

Depending on practicability and suitability, our Data Dictionary partially incorporates features from these established standards. Types of metadata and their functions sometimes overlap because different types of metadata do not always have well-defined boundaries. For example, the metadata “image bit depth” is categorized as administrative metadata in CDL and technical metadata in NISO. In our Data Dictionary, types and functions of metadata are classified based on practicality that best matches the structural complexity of digitized representations of phonograph records (i.e., the data model). We used the following guidelines in defining our Data Dictionary to maximize interoperability with the existing systems.

4.2.1 Flexibility

The Data Dictionary should be constructed to handle fine metadata granularity and directly use or modify existing establishments whenever possible.

4.2.2 Extensibility

The Data Dictionary should allow extensions to be developed later as necessary. The design should be flexible to support changing user needs and new technologies.

4.2.3 Effectiveness

The creation and maintenance of the Data Dictionary should be as economical and efficient as possible without sacrificing the usability.

4.2.4 Openness

The creation and ongoing management of the Data Dictionary should be as cooperative and inclusive as practicably possible.

4.2.5 Unambiguity

The Data Dictionary should provide for consistent and unique interpretation and employ controlled vocabularies to increase accuracy, precision and recall of records.

4.2.6 User Needs

The Data Dictionary should be designed to meet the needs of users, and not based on ease of implementation.

5. Limits to the Scope of the Data Dictionary

Metadata elements in this Data Dictionary focus solely on digitized representations of phonograph records of music. Sound recordings in other formats such as CDs, tapes, cylinders or phonograph records of speech or animal sound, which may include metadata elements such as species name, habitat, weather conditions, are beyond the scope of this dictionary.

Our Data Dictionary currently includes five types of metadata: description metadata to enable discovery and identification of resources; administration metadata to support management of resources; structure metadata to describe font and layout characteristics of texts and images; legal rights metadata to protect intellectual property rights; and technical information metadata to record the capture process and technical characteristics of digital objects. The metadata pertains to the recording itself as well as to the content of the recording.

6. Fields Reference Guide

Each entry in our Data Dictionary offers these attributes of a semantic unit:

6.1 Field Name

The field name is devised to be descriptive and unique within the Data Dictionary. This name is different from the display name (see the Label field below) presented on a computer display interface.

6.2 Definition

The definition field provides semantic meaning of metadata elements with the intention of disambiguating metadata terms, thereby clarifying the conceptual intent of each metadata term.

6.3 Multiplicity

The multiplicity field [14] indicates the possible number of data entries for each metadata element. This field combines two commonly used fields in other data dictionaries: requirement and repeatability. For example, the value “0…many” indicates that the metadata is not required and may be repeated, the value “0…1” indicates that the metadata is not required and may not be repeated, and the value “1” indicates that the metadata is required and may not be repeated.

6.4 Data Type

The data type field indicates the format in which the data is to be entered into the system and indicates any vocabulary controls (e.g., Library of Congress Subject Headings) enforced.

6.5 Label

Label refers to the contextual instance of a metadata element. This field dictates the way that a metadata element is presented on a computer display interface. The field provides an easy mechanism to change its display name to a variant name while still pertaining to the same semantic meaning. Although the goal of the data dictionary is standardization, this flexibility allows terminologies to be renamed so vocabularies that are special to the field appear natural and logical to users.

6.6 Provenance

The provenance field pertains to the derivation history of a metadata element from its original source(s), including its unique identification. This information can help to validate and determine quality of the Data Dictionary.

6.7 Examples or Notes

The examples or notes field suggests encoding guidelines by providing examples of usage and describing the nature and properties of the metadata element.

6.8 Data Constraints

The data constraint field indicates the scope of metadata elements. Examples include constraints applied to content data at the collection level, artwork level, or track level.

6.9 Version Tracking

The version-tracking field supports changes in data management, semantic evolution, and new technical requirements. The issued date, modified dates, replacement, or the deprecation of fields are provided when applicable.

7. Data Model

In addition to describing individual metadata elements, we also defined the relationships among metadata elements in our Data Dictionary. The data model adopted in this design facilitates the management of the wide variety of objects (e.g., tracks, discs, performers) that comprise phonograph records. The metadata elements belong to one or more hierarchical classes: Collection, Album, Image, Disc, and Track [13].

8. The Data Dictionary

Our Data Dictionary currently contains 103 description metadata, 4 administration metadata, 3 structure metadata, 11 legal rights metadata, and 52 technical information metadata, totaling more than 170 metadata elements. Due to space constraints, we show in this paper the metadata we created as well as a few examples of metadata that we incorporated from existing well-established standards.

8.1 Descriptive Metadata (examples)

Field Name	grooveCharacteristics
Definition	Give the groove characteristic of a

	phonograph record if it is not standard
Multiplicity	0…1
Data type	String
Label	Groove Characteristics
Provenance	AACR2r
Example or notes	microgroove
Data constraints	Disc level
Version tracking	Issued on 2004-07-13

Field Name	trackDuration
Definition	Give the duration of a track
Multiplicity	0…1
Data type	Record in the representations of ISO 8601: hh:mm:ss
Label	Track Duration
Example or notes	00:02:59
Data constraints	Track level
Version tracking	Issued on 2004-07-13

Field Name	dateRecording
Definition	Give the date the recording was made
Multiplicity	0…1
Data type	Record in the representations of ISO 8601: YYYY-MM-DD, YYYY-MM, or YYYY
Label	Date of Recording
Example or notes	1938-06-24
Data constraints	Track level
Version tracking	Issued on 2004-07-13 Modified on 2005-02-03

Field Name	trackPeculiarityNote
Definition	Make notes of any peculiarities associated with the track
Multiplicity	0…Many
Data type	String
Label	Track Peculiarity Note
Example or notes	Track number typo: this track should be track 4, following track 3, but the track number is printed track 3 again on the disc label.
Data constraints	Track level
Version tracking	Issued on 2004-07-13

8.2 Administrative Metadata (examples)

Field Name	provenance
Definition	Specify the source the data is derived, for example, the disc label, a discography website, or legacy metadata from the library's primary catalogue
Multiplicity	0…Many
Data type	String
Label	Provenance of Data
Example or notes	MARC record from McGill Library's MUSE.
Data constraints	
Version tracking	Issued on 2005-02-03

Field Name	provenanceComment
Definition	Give comments about the provenance of data
Multiplicity	0…Many
Data type	String
Label	Comment about the Provenance of Data
Example or notes	The date information provided here may not necessarily be the date of recording. The event associated with this date was not explicitly stated on the MARC record.
Version tracking	Issued on 2005-02-03

8.3 Technical Information Metadata (examples)

Field Name	stylusDimension
Definition	Give the radius of the tip of the stylus
Multiplicity	0…1
Data type	String
Label	Stylus Dimension
Example or notes	2.5 mil
Data constraints	Disc level
Version tracking	Issued on 2005-02-03

Field Name	stylusShape
Definition	Give the shape of the stylus
Multiplicity	0…1
Data type	String
Label	Stylus Shape
Example or notes	elliptical
Data constraints	Disc level
Version tracking	Issued on 2005-02-03

Field Name	trackingForce
Definition	Give the total force holding the stylus in place in the record groove
Multiplicity	0…1
Data type	String
Label	Tracking Force
Example or notes	1.5 g
Data constraints	Disc level
Version tracking	Issued on 2005-02-03

Field Name	Turnover
Definition	Give the equalizer settings used to accomplish the playback adjustment for bass turnover
Multiplicity	0…1

Data type	String
Label	Turnover
Example or notes	500 Hz RIAA
Data constraints	Disc level
Version tracking	Issued on 2005-02-03

Field Name	playbackSpeed
Definition	Give the playing speed of an analog disc in revolutions per minute (rpm).
Multiplicity	1
Data type	String
Label	Playback Speed
Provenance	AACR2r
Example or notes	76.6-rpm
Data constraints	Disc level
Version tracking	Issued on 2004-07-13

Field Name	imageProducer
Definition	Identify the organization-level producer(s) of the image.
Multiplicity	1…Many
Data type	String
Label	Image Producer
Provenance	ANSI/NISO Z39.87 proposed draft, 2005
Example or notes	Music Technology Area, Schulich School of Music of McGill University
Version tracking	Issued on 2004-07-13 Modified on 2005-02-03

9. Conclusion and Future Work

The creation and maintenance of our Data Dictionary is important to MIR as it provides a means to ensure accurate, consistent, and comprehensive metadata annotation and promotes interoperability between different systems. Specifically, the Data Dictionary assists the process of data entry, ensures consistency of records, facilitates semantic compatibility and interoperability between systems and, most importantly, forms the foundation for an efficient and effective information retrieval infrastructure.

The Data Dictionary has been used and tested for McGill University's audio preservation digitization pilot projects [15]. Our Data Dictionary, however, still needs to be used and evaluated in future phonograph records digitization projects. It is not intended to be fixed or final, but rather provides a starting point for improvements and enhancements to digital library services based on community experience and feedback.

The most immediate future research to be performed involves the implementation of the Data Dictionary in an XML-based format that facilitates the exchange of metadata information between digital repositories.

References

[1] Association for Library Collections & Technical Services, Committee on Cataloging: Description and Access, Task Force on Metadata, "Final Report," pp. 16, [Web site] 2000, [2006 April 4], Available: http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html

[2] Bacca, M., ed., *Introduction to Metadata*, Los Angeles: Getty Research Institute, [Web site] 1998, [2006 April 4], Available: http://www.getty.edu/research/conducting_research/standards/intrometadata/

[3] D. Haynes, *Metadata for information management and retrieval,* London: Facet publishing, 2004.

[4] A. Taylor, *The organization of information*, Englewood, Colo.: Libraries Unlimited, 1999.

[5] Joint Steering Committee for Revision of AACR, American Library Association, et al., *Anglo-American cataloging rules,* 2nd ed., 2002 revision, 2005 update. Chicago: American Library Association; Ottawa: Canadian Library Association; London: Chartered Institute of Library and Information Professionals, 2005.

[6] Dublin Core Metadata Initiative, [Web site] 2006, [2006 April 4], Available: http://dublincore.org/

[7] S. Mudge, and D. Hoek, "Describing jazz, blues, and popular 78 rpm sound recordings: Guidelines and suggestions," *Cataloging & Classification Quarterly,* vol. 29, no. 3, pp. 21–48. 2000.

[8] T. Simpkins, "Cataloging popular music recordings," *Cataloging & Classification Quarterly,* vol. 31, no. 2, pp. 1–35. 2001.

[9] R. Smiraglia, *Describing music material*, 3rd ed., Lake Crystal, MN.: Soldier Creek Press, 1997.

[10] H. Hemmasi. "Why not MARC?" in *Proceedings of the International Conference on Music Information Retrieval*, 2002, pp. 242–248.

[11] M. Miliano, and IASA. *The IASA cataloguing rules: a manual for the description of sound recordings and related audiovisual media*, IASA Cataloguing Rules Editorial Group, Eds. Stockholm: International Association of Sound and Audiovisual Archives, 1999.

[12] R. Koenen, and F. Pereira, "MPEG-7: A standardised description of audiovisual content," *Signal Processing: Image Communication,* vol. 16, no. 1, pp. 5–13, 2000.

[13] C. Lai, I. Fujinaga, and C. Leive, "Metadata for phonograph records: Facilitating new forms of use and access to analog sound recordings," in *Proceedings of the Joint Conference on Digital Libraries*, 2005, pp. 385.

[14] Variations2 Version 1 Data Dictionary. [Web site] 2003, [2006 April 19], Available http://variations2.indiana.edu/pdf/DML-metadata-elements-v1.pdf

[15] C. Lai, B. Li, and I. Fujinaga. "Preservation digitization of David Edelberg's Handel LP collection: A pilot project," in *Proceedings of the International Conference on Music Information Retrieval*, 2005, pp. 570–575.