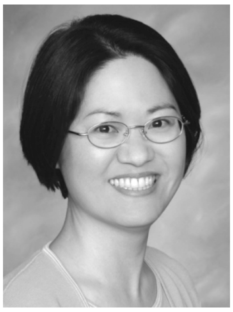
Metadata Object Description Schema (MODS) in Digital Repositories: An Exploratory Study of Metadata Use and Quality

Jung-ran Park* and Susan Maszaros**

*College of Information Science and Technology, Drexel University,
3141 Chestnut Street, Philadelphia, PA 19104 USA
< jung-ran.park@ischool.drexel.edu>

**Tennessee State Library & Archives, 403 7th Avenue North, Nashville TN 37243 USA
<sue.maszaros@state.tn.us>



Jung-ran Park is assistant professor at the College of Information Science and Technology, Drexel University. Prior to joining Drexel University, she worked as a cataloger at Indiana State University. She received a Ph.D. in linguistics and an MLIS from the University of Hawaii. She has applied her linguistics background to her primary research areas in knowledge organization and representation, computer-mediated communication and cross-lingual information access. Dr. Park is currently the principal investigator of a grant project "Metadata Creation and Metadata Quality Control across Digital Collections: Evaluation of Current Practices" from the Institute of Museum and Library Services.

Susan Maszaros serves as the Digital Documents Coordinator for the Tennessee State Library and Archives in Nashville, Tennessee. She is a graduate of Drexel University's College of Information Science and Technology where she received her MLIS. In her current position, Ms. Maszaros is co-coordinator for TSLA's microfilm digitization pilot project and overseas the development and implementation of local metadata practices for this program. In addition to her duties at the TSLA, she has served as webmaster for the TENN-SHARE library consortium and is currently chair of the TENN-SHARE Web Committee.

Park, Jung-ran, and Maszaros, Susan. **Metadata Object Description Schema (MODS) in Digital Repositories: An Exploratory Study of Metadata Use and Quality.** *Knowledge Organization, 36(1)*, 46-59. 25 references.

ABSTRACT: This study examines the use of the Metadata Object Description Schema (MODS) within three digital collections. It identifies the MODS metadata elements that evidence the most frequently occurring inconsistent and inaccurate application. For this, a total of sixty metadata records (twenty from each collection) were collected. The surveyed collections cover a wide range of material from digitized sound recordings and monographs, pre-1800 imprints to born-digital web resources. As a means of comparison in evaluating the quality of the metadata, local guidelines for the MODS metadata application are also consulted in order to determine the usage of MODS metadata elements in local collections against the guidelines. Analysis of the surveyed data drawn from the three collections shows that the five most frequently used elements (titleInfo, originInfo, recordInfo, physicalDescription and subject) appeared in 86 percent of the records. The total number of MODS elements represented in each collection ranged from twelve to fifteen (out of 20 MODS top-elements). Results of this study indicate that the MODS metadata scheme is suitable for describing a wide range of materials and resource types. The results also indicate that easily accessible local guidelines for metadata creation contribute significantly to the consistent and accurate application of the MODS metadata scheme.

1.0 Introduction

As digital collections within libraries continue to grow, so too does the need for a metadata schema that offers compatibility with existing library data and interoperability across different metadata schemas as well as the ability to provide rich description of resources (Guenther 2003). The Metadata Object Description Schema (MODS) is a standard that offers the potential to meet these needs. This project is a study of the MODS schema as it has been implemented within three digital collections.

The impact of metadata quality on resource discovery is significant. However, the critical issues affecting metadata quality evaluation have been relatively unexplored (Moen et al. 2003; Barton et al. 2003). There is a growing awareness of the essential role played by metadata quality assurance for successful resource access and sharing across distributed digital collections. Through an examination of learning objects and e-prints of communities of practice, Barton et al. (2003) discuss the importance of quality assurance for metadata creation while pointing out the lack of formal investigation into the metadata creation processes. Problems inherent in the metadata creation process, such as inaccurate data entry (e.g., spelling, abbreviations, format of date [date of creation or date of publication] and consistency of subject vocabularies) that result in adverse effects on resource discovery are examined.

Studies dealing with metadata quality issues mostly concern digital repositories using the Dublin Core metadata scheme (Park 2005; Park 2006). To the best knowledge of the authors, there are no studies evaluating MODS metadata records. We speculate that this is in part due to the fact that the MODS metadata scheme is relatively new and there are to date few projects fully implementing the scheme.

The goal of this exploratory study is to examine how the MODS metadata scheme is being used across three digital repositories that cover a wide range of material from digitized resources to born-digital resources such as websites. This study also identifies MODS metadata elements that evidence the most frequently occurring inconsistent, inaccurate and incomplete application. The use of controlled vocabularies for subject element description is also examined. Implications drawn from evaluation of the current status of MODS metadata application in relation to the issue of metadata semantics are also discussed. For the project, a randomly collected sample of MODS metadata records (n = 60) from three digital repositories is analyzed. In conjunction with looking at MODS user guidelines, local guidelines for MODS metadata application are secured in order to determine how MODS metadata elements are utilized vis-à-vis the local guidelines.

2.0 Metadata Object Description Schema (MODS) and metadata quality: An overview

The Metadata Object Description Schema (MODS) were developed by the Library of Congress' Network Development and MARC (Machine-Readable Cataloging) Standards Office and first implemented in 2002 (Library of Congress 2007). Derived from MARC 21, MODS is a descriptive metadata standard envisioned as an abbreviated version to MARC built to be "more compatible with library data than either the Dublin Core or ONIX (Online Information Exchange) applications (Guenther 2003, 139). Online Information Exchange is an international metadata standard developed by publishers and used within the book industry for the creation of basic bibliographic records for purposes of describing resources intended for sale (NISO 2004, 7).

A derivative of MARC 21, MODS offers much of the richness and granularity of the MARC standard but is expressed using the Extensible Markup Language (XML) schema language, which offers greater flexibility, especially in describing electronic resources (NISO 2004). Even though the "richness" of the MODS metadata schema is often touted as one of its advantages (NISO 2004), it is the fact that MODS offers a simpler structure than MARC—a reduced number of "fields" and "language-based tags"—that has made it more appealing for users (Guenther 2003, 139).

The MODS metadata scheme is comprised of twenty top-level elements, all of which are repeatable. These elements are a "repackaging" of sorts of the MARC fields, which either were "combined with other elements to form a single element" or "dropped altogether" (Guenther 2003, p. 140). According to the *MODS User Guidelines* (Library of Congress 2007), there is no one specific element that is mandatory.

The top-level elements are shown in Table 1.

Each of these top-level elements can be refined using attributes and subelements that can be applied throughout the entire schema. For instance, the titleInfo element can be refined using some of the following attributes and subelements: Attributes—type, displayLabel, xlink, ID, lang, xml:lang, script, trans-

titleInfo	*note*
name	*subject*
typeOfResource	*classification*
genre	*relatedItem*
originInfo	*identifier*
language	*location*
physicalDescription	*accessCondition*
abstract	*part*
tableOfContents	*extension*
targetAudience	*recordInfo*

Table 1. Top-level elements of MODS

literation; Subelements—title, subTitle, partNumber, partName, nonSort. The following illustrates this:

```
<titleInfo>
  <nonSort>The</nonSort>
  <title>winter mind</title>
  <subtitle>William Bonk and American letters</subtitle>
</titleInfo>
```

As stated, MODS offer more than just flexibility and compatibility with library data. McCallum (2004) highlights both the sophistication and usability of MODS by examining key features in its design. As McCallum (2004) notes, some of the benefits of MODS can be found in its ability to operate in an XML environment using fewer tags than MARC. It also offers better linking capabilities and is designed to accommodate the description of digital resources.

Guenther (2004) examines the capabilities of MODS through the lens of the *MODS User Guidelines* (Library of Congress 2007). By detailing the use of MODS elements within the Library of Congress' MINERVA project, she demonstrates not only how MODS can be used as a tool for the creation of records for digitally born objects, but also how the success (and accuracy) of its application is directly related to the "extensive guidelines" that accompany this schema (Guenther 2004, 93).

As well, Missingham (2004) illustrates the potential uses of MODS in her study of the National Library of Australia's Kinetica service, a web-based service that allows libraries in Australia to contribute records to the national union catalog of resources—the Australian National Bibliographic Database. The study demonstrates the effectiveness of MODS as an "intermediary format" in the conversion of records from one metadata schema to another. Missingham (2004, 7) presents the potential that MODS offers in its ability to "reuse…descriptive records" as well as to support greater interoperability between standards. Furthermore, the use of MODS records have the potential to allow for greater access to a wider variety of materials, "regardless of their format" as well as increased opportunities for discovery via the web by opening up such collections to OAI-PHM harvesting services (Missingham 2004, 8).

While the implementation of MODS is not yet widespread, the successful use of MODS in such projects as the Australian National Bibliographic Database as well as the MINVERA project demonstrate the paths on which MODS will continue to advance, especially within the library community. The full bibliographic descriptions within a framework that provides greater flexibility and interoperability are key to the metadata schema of the future.

Let us now present an overview on metadata quality. The critical issues affecting metadata quality evaluation have been relatively unexplored and very few studies have attempted to define "metadata quality" (Moen et al. 1997, Barton et al. 2003). As mentioned, studies dealing with metadata quality issues mostly concern digital repositories using the Dublin Core metadata scheme (Park 2005; Park 2006).

While examining metadata in e-print archives, Guy et al. (2004) state that "high quality metadata supports the functional requirements of the system it is designed to support, which can be summarized as quality is about fitness for purpose." They suggest that functional requirements be established by defining internal and external requirements; that is, define the internal functional requirements relating to the archive's Web user interface. The internal functional requirement can be defined in relation to end-users' needs in a local archive. The external functional requirements can be defined in relation to disclosed and exposed local metadata relating to external service providers such as the Open Archives Initiative.

Metadata quality can be assessed based on the above mentioned functional requirements (Guy et al. 2004). For example, if searching and browsing by date is listed as a functional requirement, then it is necessary to have content rules specifying the format of dates (e.g., 05-06-2007) to meet the functional requirements. Otherwise, different formats (e.g., 05/06/2007 or 05-06-07) can be used, which will interfere with sorting of the documents. This will in turn hamper users as they search and browse documents by dates. In line with this functional perspective, Hillmann et al. (2004) point out that "the utility of metadata can best be evaluated in the context of services provided to end-users."

According to NISO (2004), "good" metadata supports interoperability, the qualities of archivability, persistence, unique identification and the long-term management of objects. As well, it uses controlled vocabularies to reflect the what, where, when and who of the content and includes conditions and terms of use. It is also authoritative and appropriate to the collection and its users. The criteria and principles articulated by NISO (2004) function to provide a framework of guidance for building good digital collections.

In the aforementioned definitions (Guy et al. 2004; NISO 2004; Hillmann et al. 2004), the quality of metadata reflects the degree to which the metadata in question perform the core bibliographic functions of discovery, use, provenance, currency, authentication and administration. In other words, the principal purpose of metadata is to a large degree related to that of the traditional online library catalogs and databases in finding, identifying, selecting and obtaining items (IFLA 1998).

Even though there is no established framework for measuring metadata quality, studies have identified major criteria that can be used for assessing metadata quality. Such functional perspectives are closely tied with the criteria and measurements that are used for assessing metadata quality.

Statistics Canada's Quality Assurance Framework (2002) presents six dimensions of information quality: relevance, accuracy, timeliness, accessibility, interpretability and coherence. Bruce and Hillmann (2004) further refine these six principles by modifying them for the library community. The suggested criteria concern completeness, accuracy, provenance, conformance to expectation, logical consistency, coherence, timeliness and accessibility. These criteria are particularly developed in the context of aggregated collections.

Moen et al. (1997) measured accuracy, consistency, completeness and currency for their analysis of Government Information Locator Services (GILS) metadata records. Park (2005 2006) and Bui and Park (2006) also measured accuracy, completeness and consistency in relation to interoperability. Zeng (2006) utilized similar criteria in examining the metadata quality of the National Science Digital Library.

Below we will briefly touch on three criteria, as we will utilize these for examining MODS metadata records: completeness, accuracy and consistency. The completeness of metadata description is conditioned by the access capacity to individual local objects and connection to the parent local collection(s) (Bruce and Hillmann 2004). This reflects the functional purpose of metadata in resource discovery and use (Guy et al. 2004; NISO 2004). The completeness of metadata description is also conditioned by characteristics of the resource type within a given domain and specifically by local application profiles such as guidelines and best practices (Duval et al. 2002). The local application profiles are further modulated by the functional purpose (e.g., information access/service). In this sense, the characteristics of local communities (e.g., collections, agency creating the metadata) as well as the resource itself seem to modulate the completeness of the metadata description. Thus, the completeness of metadata description entails several factors: resource type (i.e., object), its relation to the local collection(s) and the metadata creation guides.

Accuracy can be measured in terms of precise data input such as the elimination of typographical errors, conforming expression of personal names and place names and use of standard abbreviations (Bruce and Hillmann 2004). Several studies report problems in the DC metadata description of data content on this level. For instance, Currier, et al. (2004) report problems inherent in the metadata creation process, such as inaccurate data entry, which covers inaccuracy in spelling, abbreviations and formatting of date (e.g., date of creation or date of publication).

On another level, the accuracy of metadata can be measured by taking into account its context. On this level, Zeng describes accuracy in terms of the correctness of data element's content, intellectual property and instantiation. She also characterizes the accuracy of metadata in terms of the accurate representation of the original resources. There are several studies reporting accuracy problems in applying the DC metadata scheme (see Park 2006 for details).

Consistency (also known as comparability or coherence) can be measured in terms of both semantic consistency in data value and structural consistency in data format (i.e., syntax) (Stvilia et al. 2004). Consistency issues in digital repositories stem especially from the heterogeneous nature of resource types and of federated repositories. Metadata creation guidelines vary institution by institution and remain somewhat open to interpretation. This also affects consistency. For instance, the DC identifier can be used for a variety of data elements such as call number (e.g., LCC, *DDC*), image number, negative number, accession number, serial number and photographer's reference number. Park's studies also (2005; 2006) look at problems in relation to consistency. For instance, there is great confusion in employing some of the DC elements Type and Format and they are interchangeably used in the surveyed three collections. The DC elements Source and Relation are also inconsistently employed across these collections.

3.0 Data and research methods

For this exploratory study, MODS metadata records were collected from March through April 2007 from three digital collections: Election 2002 Web Archive, the Library of Congress (n=20 records), African-American Band Music and Recordings, Music, Theater, and Dance collection, the Library of Congress (AABMR, n=20 records), and the Copac Academic and National Library Catalogue (n=20 records) (See Resources). These three sample collections are selected based on the following criteria: 1) the repositories employ the same MODS metadata scheme; 2) the XML tags of the collections are viewable; 3) the collections cover a wide range of material from digitized sound recordings (e.g., AABMR), monographs such as pre-1800 imprints and periodicals (e.g., Copac), and born-digital items such as websites (e.g., Election 2002 Web Archive).

A sampling of metadata records was derived from numbers produced using the Research Randomizer tool (http://www.randomizer.org/). A randomized sampling of ten sets of six unique numbers based on unsorted numbers between 1 and 1000 was generated from this tool. For this exploratory study, a total of sixty metadata records (twenty from each collection) were collected by using these random numbers. While records were selected at random, there was an effort within the Copac catalogue to choose a variety of the distribution of the records. Of the twenty records from this collection, four are pre-1800 imprints, five are pre-1970 imprints, ten are post-1970 and one periodical has no date listed.

Let us briefly present an overview of the surveyed collections. The Election 2002 Web Archive is described as a project developed by the Library of Congress in collaboration with the State University of New York Institute of Technology's WebArchivist.org and the Internet Archive. Election 2002 is a selective collection of nearly 4,000 sites archived between July 1 2002 and November 30 2002. It includes "Web sites associated with United States 2002 mid-term Congressional elections, and mayoral elections in 15 major United States cities." (United States Election 2002 Web Archive, see Resources). The websites archived in this collection are searchable by eight different categories: Candidates, Citizen, Civic & Advocacy, Government, Political Party, Press, Public Opinion and Miscellaneous. However, there is no search feature with this collection.

The African-American Band Music and Recordings, Music, Theater, and Dance collection by the Library of Congress (AABMR) comprises digitized sound recordings, images of musical scores and arrangements, as well as articles and biographies related to the musicians and music represented in this collection. It contains approximately 300 digitized items. The collection is searchable by formats such as "all formats," "Instrumental parts" or "Recordings," (see Resources, African-American Band Music Recordings). There is a keyword search feature with this collection.

The Copac Academic and National Library Catalogue contains records for materials found within the catalogs of "all major university and National libraries in the UK and Northern Ireland" (see Resources, About Copac). Copac is essentially a union catalog. It consists of a collection of over 32 million records supplied by the CURL - Consortium of Research Libraries (see Resources, About Copac). Monographs mostly represent the Copac database; periodicals represent 6% of records and conferences 3%. There are three types of searches conducted within this catalog: a quick search, a main search and a map search. Users have the option of searching a variety of fields; however, it does not appear that this collection contains a mechanism for browsing.

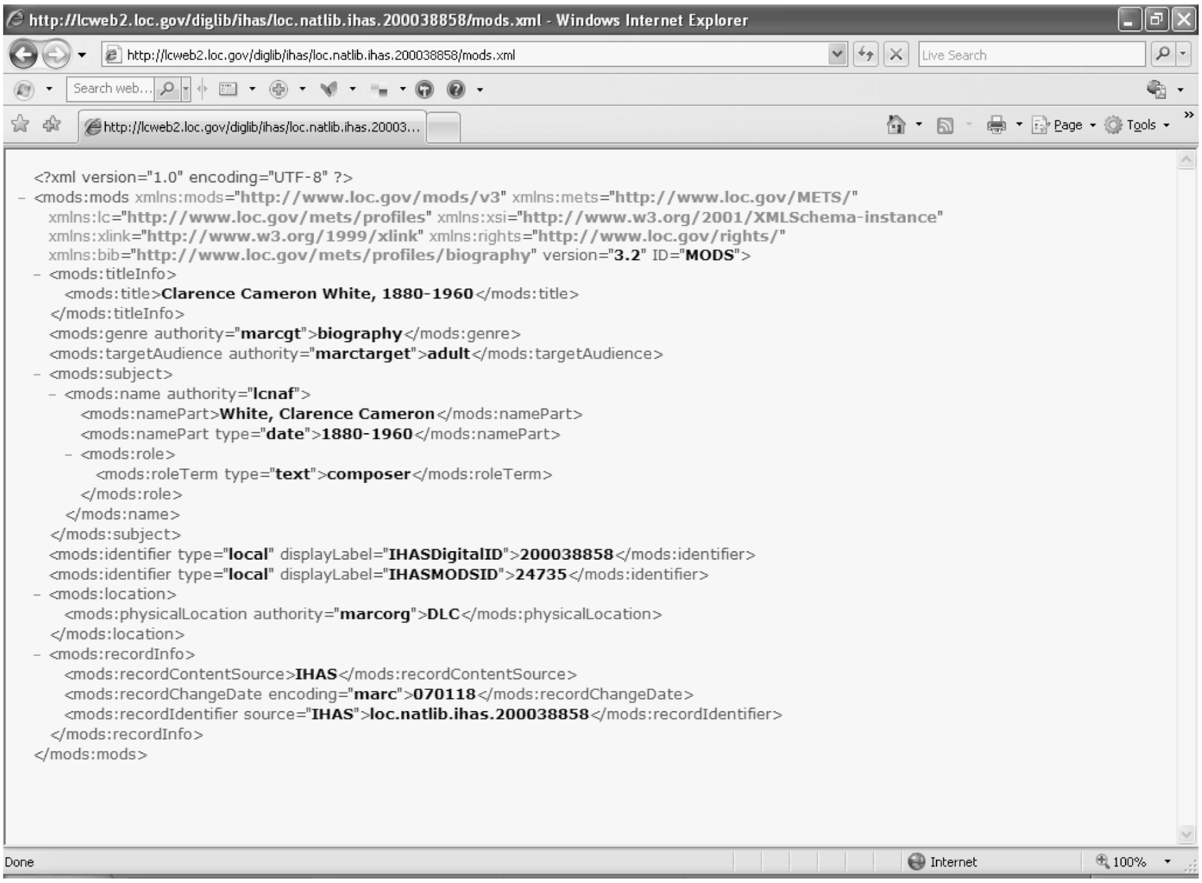
To be able to evaluate the quality of the metadata, it is critical to see the source code in XML format. Both AABMR and the Copac catalogue provide direct links to MODS XML records as part of their display records within the collection. The Election 2002 site does not provide XML records within the display record. As such, MODS XML records were retrieved by manually inserting the .xml extension in the display record.

Figure 1 below illustrates a sample XML version of the MODS record.

This study has been conducted in three phases. The first phase consisted of retrieving MODS XML records (as shown in the above figure) from each collection and importing them into an Excel spreadsheet. Frequency of use was documented for each of the MODS elements, attributes and sub-elements. Display records were also imported into an Excel spreadsheet alongside the original XML record for comparison.

Excel spreadsheets offer ready-to-view visual inspection. This also allowed us to read a record with all its elements across a page. Scrolling up and down was also helpful in the identification of any anomaly. An Excel file can have as many worksheets as the sys-

https://doi.org/10.5771/0943-7444-2009-1-46 https://www.inlibra.com/de/agb - Open Access -


http://lcweb2.loc.gov/diglib/ihas/loc.natlib.ihas.200038858/mods.xml - Windows Internet Explorer
http://lcweb2.loc.gov/diglib/ihas/loc.natlib.ihas.200038858/mods.xml
Live Search
Search web...
http://lcweb2.loc.gov/diglib/ihas/loc.natlib.ihas.20003...
Page
Tools
Done
Internet
100%

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <mods:mods xmlns:mods="http://www.loc.gov/mods/v3" xmlns:mets="http://www.loc.gov/METS/"
    xmlns:lc="http://www.loc.gov/mets/profiles" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:rights="http://www.loc.gov/rights/"
    xmlns:bib="http://www.loc.gov/mets/profiles/biography" version="3.2" ID="MODS">
  - <mods:titleInfo>
      <mods:title>Clarence Cameron White, 1880-1960</mods:title>
    </mods:titleInfo>
    <mods:genre authority="marcgt">biography</mods:genre>
    <mods:targetAudience authority="marctarget">adult</mods:targetAudience>
  - <mods:subject>
    - <mods:name authority="lcnaf">
        <mods:namePart>White, Clarence Cameron</mods:namePart>
        <mods:namePart type="date">1880-1960</mods:namePart>
      - <mods:role>
          <mods:roleTerm type="text">composer</mods:roleTerm>
        </mods:role>
      </mods:name>
    </mods:subject>
    <mods:identifier type="local" displayLabel="IHASDigitalID">200038858</mods:identifier>
    <mods:identifier type="local" displayLabel="IHASMODSID">24735</mods:identifier>
  - <mods:location>
      <mods:physicalLocation authority="marcorg">DLC</mods:physicalLocation>
    </mods:location>
  - <mods:recordInfo>
      <mods:recordContentSource>IHAS</mods:recordContentSource>
      <mods:recordChangeDate encoding="marc">070118</mods:recordChangeDate>
      <mods:recordIdentifier source="IHAS">loc.natlib.ihas.200038858</mods:recordIdentifier>
    </mods:recordInfo>
  </mods:mods>
```

Figure 1. A sample MODS XML record

tem's capacity allows. This break-up method did not cause any negative effect on data analysis.

For the second phase of this study, the subject elements were examined. In this phase the types of controlled vocabularies in the MODS records and local guidelines were identified. The third phase of this project entailed analyzing the metadata quality of the collected data. The data analysis was formulated based on both a qualitative and quantitative examination of the usage of the MODS metadata elements. In order to examine usage and completeness of MODS metadata elements, the frequency of metadata elements of a total of 60 metadata item records are calculated. In conjunction with *MODS user guidelines* (Library of Congress 2007), local guidelines for the MODS metadata application were secured. Through utilizing qualitative analysis, we examined the manner in which MODS metadata elements are used vis-à-vis local guidelines (see Resources). The semantics of the MODS metadata element name and its corresponding definition are examined through utilization of linguistic semantic analysis.

Prior to data analysis, we will briefly outline the surveyed collections' local guidelines.

Election 2002 Web Archive and African-American Band Music and Recordings (AABMR) are taken from of the Library of Congress' collections, providing the most easily accessible information pertaining to local practices. The guidelines (*MODS: Description of Elements; Subject Terms for Use in Cataloging*, see Resources) specify only eleven of the nineteen top-level elements: Title, Name, Abstract, Date Captured, Genre, Physical Description/Format, Related Item, Identifier, Language, Access Condition and Subject. The element recordInfo is also part of the MODS XML record, although no information pertaining to this element was addressed in the guidelines.

The Copac catalogue does not provide information on the use of the MODS metadata scheme. However, it must be taken into consideration that Copac is a union catalogue of which records are supplied through the CURL (Consortium of Research Libraries) database. MARC records are converted to MODS XML records through an automated process using a crosswalk developed by software technicians (S. Cousins, personal communication). The CURL provides a document entitled *Technical Requirements for Data supplied to Copac* as well as documentation related to

CURL Minimum Standards for Bibliographic Records (see Resources). As these documents address the creation of MARC records only, it was necessary to construct a crosswalk in order to compare MODS elements to the list of requirements for CURL and Copac records.

By utilizing the *MARC Mapping to MODS* (Library of Congress 2006) as well as the above mentioned documentation, we created a basic table of mandatory and desirable but non-mandatory elements that should be represented in each Copac record (see Appendix A).

4.0 Discussion

The following sections discuss the analysis of the surveyed MODS metadata item records and controlled vocabularies for subject element description within the context of the surveyed collections. Implications drawn from analysis of metadata quality of MODS records are also briefly discussed in relation to the issue of metadata semantics.

4.1 Analysis and findings

As stated, the completeness of metadata description entails several factors such as resource type (i.e., individual local object), its relation to the local collection(s) and the metadata creation guidelines. As well, the assessment of the completeness demands examining the "size and distribution of elements among the records" as well as "the degree to which the general metadata functions of resource discovery authentication, and administration are fulfilled" (Zeng 2006; Bruce and Hillmann 2004; Guy et al. 2004; NISO 2004; Duval et al. 2002).

Table 2 below illustrates the frequency of use of MODS metadata elements and the total number of elements used in the surveyed collections.

As illustrated in Table 2, the most frequently and commonly used elements among the three collections are the following (elements listed in descending order):

- titleInfo, originInfo, recordInfo, physicalDescription, subject, name, identifier, and language

In the case of the Election 2002 Web Archive, the analysis of the data informs us that, with the exception of the accessCondition, all the elements listed in the local guidelines are well-represented in the surveyed record (n=20). Subject elements are also accurately represented in conformance to the guidelines. In the case of Copac, the titleInfo, name, originInfo and extension elements are used in all the surveyed records and at a noticeably higher rate compared to that of Election 2002. The physicalDescription element is absent in only one surveyed metadata record and the note element is applied to half the records mostly due to inaccurate mapping.

When attempting to examine the presence of the other mandatory elements within the Copac records, it is important to point out the publication dates of items the records represent, inasmuch as the CURL recommendations for the construction of bibliographic records differ according to imprint date. As well, identifiers are more likely to occur in records for post-1970 publications. Of the twenty records from this collection, four are pre-1800 imprints, five are pre-1970 imprints, ten are post-1970 and one periodical has no date listed. The distribution of the records by date may account for the low number of subject and identifier elements. Overall, the mandatory elements are fairly well represented in the Copac collection.

According to Zeng (2006), the performance of resource discovery can be measured by the presence of primary elements such as title, author and creator, together with subject and keywords. However, as shown in Table 2 above, within the Copac records nearly half the surveyed records do not contain the subject element. However, as stated earlier, the distribution of the records by date needs to be taken into account regarding the low number of the subject element in the case of Copac. The AABMR provides names in nine of the surveyed records, indicating that the resource discovery function is hindered.

The total number of top-level elements used at least once within the collection ranges from twelve to fifteen. Nineteen of the twenty elements are used at least once in the sixty records studied. The only MODS element that does not appear in any of the surveyed collections is the part element.

As illustrated in Table 2 above, the surveyed collections as a whole are strong in providing authentication metadata through the use of the recordInfo element. As well, the recordIdentifer subelement is employed in all sixty records. The recordContentSource, recordChangeDate and recordCreationDate sub-elements are also represented in two of the three collections. While the MODS metadata scheme presents the element accessCondition, which can be employed for specifying use and restriction notes, this element is not fully employed; it is used in only fifteen re-

MODS element	Election 2002 (n/20)	% of total number of elements used (n/235)	Copac (n/20)	% of total number of elements used (n/193)	AABMR (n/20)	% of total number of elements used (n/257)	Total (n/60)	% of total element usage
titleInfo	20	8.51	20	10.36	20	7.78	60	100
Name	20	8.51	20	10.36	9	3.50	49	81.7
typeOfResource	0	0.00	11	5.70	20	7.78	31	52.0
Genre	20	8.51	6	3.11	0	0.00	26	43.3
originInfo	20	8.51	20	10.36	20	7.78	60	100
Language	20	8.51	19	9.84	0	0.00	39	65.0
physicalDescription	20	8.51	19	9.84	20	7.78	59	98.3
Abstract	20	8.51	0	0.00	0	0.00	20	33.3
tableOfContents	0	0.00	1	0.52	0	0.00	1	1.6
targetAudience	0	0.00	0	0.00	20	7.78	20	33.3
Note	0	0.00	11	5.70	20	7.78	31	51.6
Subject	20	8.51	12	6.22	20	7.78	52	86.6
classification	0	0.00	3	1.55	20	7.78	23	38.3
relatedItem	20	8.51	6	3.11	8	3.11	34	56.6
Identifier	20	8.51	5	2.59	20	7.78	45	75.0
Location	0	0.00	0	0.00	20	7.78	20	33.3
accessCondition	15	6.38	0	0.00	0	0.00	15	25.0
Part	0	0.00	0	0.00	0	0.00	0	0.0
Extension	0	0.00	20	10.36	0	0.00	20	33.3
recordInfo	20	8.51	20	10.36	20	7.78	60	100.0
locally added element	0	0.00	0	0.00	20	7.78	20	33.3
Total number of elements used	235	100.00	193	100.00	257	100.00	685	1141.5

Table 2. Frequency of use

cords and they all appear within the Election 2002 collection.

Accuracy can be measured in terms of accurate data input and data content (Bruce and Hillmann 2004). It is not feasible to assess if the Copac records accurately represent the original item owing to the absence of digital objects next to the surveyed records. However, in the case of the other two collections, nearly all the records surveyed in this study seem to accurately represent data content from the original source. The only exception occurs within the titleInfo element of Election 2002. The collection guidelines indicate that all values are to be extracted from the "Base URL" or HTML source code. However, titles are in some cases not representative of the html source code <Title>.

Table 3 below illustrates the inaccurate description in data content, format, input and mapping found in the records surveyed. Inaccurate coding mostly occurs collection-specifically and not across the surveyed collections.

The primary sources of inaccurate description derive from the use of attributes. There are inaccurate descriptions in the encoding type listed and the value provided. There are missing values for mandatory attributes. In addition, there are invalid attributes given and inaccurate mappings of the encoding type to the value of the element (rather than having it listed as an attribute). The following from AABMR illustrates this:

```
<mods:recordChangeDate encoding="marc">07030616
</mods:recordChangeDate>
```

While the encoding is listed as "marc," the format of the date is not correct.

In part owing to an aggregated system, Copac records present the greatest distribution and types of in-

Element Name	Description of Inaccuracy	Number of occurrences
titleInfo	Includes data that are part of contents notes/tableOfContents	4
	No inclusion of "nonSort" attribute	5
	Title not transcribed as it appeared within <Title> tag of web page as per local guidelines	1
Name	Inaccurate mapping : part of title included	1
	Data value entered incorrectly	1
	Lack of attribute "type" against local guidelines	20
originInfo	Inaccurate mapping: publisher name and publication date mapped onto <placeTerm> subelement	1
note	Includes inaccurate data information such as RelatedItem	1
	Includes attribute information (e.g., system requirements) as data value	1
	Includes inaccurate data value for attribute "type" (copyright)	4
	Inaccurate mapping of "contents note" (e.g. list of songs included in a collection of sheet music is used in the *note* element rather than in the *tableOfContents*.)	3
physicalDescription	Error in data input (uses inaccurate measurement mechanism—a degree sign instead of centimeter)	3
	No data value entered	1
subject	Missing <namePart> subelement	1
	Error in capitalization	1
	Subject entered not part of local guidelines	1
identifier	Includes inaccurate data value	20
language	Inaccurate data input (e.g., language listed as undetermined for "English")	3
	Lists invalid attribute ("text")	19
recordInfo	Date entered for <recordChangeDate> subelement does not match format for encoding type listed	20

Table 3. Inaccurate description

accuracies among the surveyed collections. These inaccuracies might be derived from the automated conversions from MARC to MODS. Because the Copac catalogue is so large, human intervention, other than "setting up of the processes involved in creating it" is next to impossible--(S. Cousins, personal communication). Thus, there are inaccuracies such those found in the language element; for instance, when the value is noted as undetermined but the language of the item would have been obvious to a human cataloger. In the case of the display records, the Copac catalogue effectively shows how the use of the role sub-element can be translated to the main entry field.

As stated, consistency can be measured in terms of both semantic consistency in data value and structural consistency in data format (Stivilia et al. 2004). In this study, we looked at both the semantic and syntactic consistency of the metadata elements within individual collections and also across the three collections when the same element was present in two or more of the collections. Title and name elements are the most consistently applied elements in all MODS records. Genre across the Election 2002 and the Copac records, when applied, is consistent in application. Even though the attributes and respective element values were different, they were still used consistently. The example below is illustrative:

<genre authority="**marcgt**">**Bibliography**</genre>
<genre authority="**local**">**Electronic books**</genre>

However, the placeTerm subelements within originInfo are inconsistently applied. For instance, both "city and state" values are used in some MODS records; in other cases, only "city" is applied. Encoding attributes are not always present for recordCreationDate subelements (i.e., recordInfo); when encoding attributes are present, values are not consistently applied. For instance, a record from one collection encodes the value at "20060814" while a record from another collection, citing the same encoding type, encodes the value as "20039150000."

Turning to specific collections, there are inconsistencies in the use of the accessCondition element in Election 2002. In the case of the AABMR collection, for sheet music images, the different instrumental parts are noted as relatedItem. While data in this element are not searchable, they are available for selection through a drop-down box (see Appendix B). However, this practice is not consistently applied in the sense that there are inconsistencies in the addition of the instrumental parts to the note element.

4.2 Controlled vocabulary use and metadata semantics

There are a variety of controlled vocabularies used in these surveyed collections. In particular, Copac catalogs utilize the widest variety of controlled vocabularies resulting owing to the fact that Copac is a union catalogue the records of which are supplied through the CURL (Consortium of Research Libraries).

For subject description all surveyed records (n=60) utilize the *Library of Congress Subject Headings*. The Election 2002 collection uses a slight variation of the *LCSH*. Guidelines and recommended terms and subject strings are provided for this collection. Both the Copac and the AABMR list "lcsh" as the authority within their XML records, as shown below:

```
<mods:subject authority="lcsh">
<mods:topic>African Americans--
  Music</mods:topic>
</mods:subject>
<mods:subject authority="lcsh">
<mods:topic>Popular music--United Sta-
  tes</mods:topic>
</mods:subject>
<mods:subject authority="lcsh">
<mods:topic>Band music</mods:topic>
</mods:subject>
```

The Election 2002 collection does not embed the subject authority used within the records; however, it provides the most extensive account in the local guidelines regarding the derivation and use of subject terms for cataloging. This collection utilizes the simplified Library of Congress-style subject headings and such modifications to *LCSH* can be observed in the sample records. For instance, a heading such as Third Parties without the gloss (United States politics) is used. Even though such variations seem relatively minor, this may dwarf the performance of "cross-database searching" due to the vocabulary compatibility issue.

Controlled vocabularies are consistently applied in the surveyed three collections. Subject vocabularies such as *LCSH* represent the "aboutness" of the document being described. Since these surveyed collections contained little illustrative matter, *LCSH* seemed to be the most suitable choice of controlled vocabulary for subject descriptions. However, other controlled vocabulary schemes such as the *Art and Architecture Thesaurus* (*AAT*) might also be suitable candidates for describing the AABMR collection, inasmuch as musical genres are represented in *AAT*.

Metadata semantics affect consistency as well (see Park 2006 for details). For instance, there are semantic overlaps among DC metadata Type and Format as well as Physical Description in the sense that the semantic boundaries among these elements are fuzzy and not clear cut; consequently, they may be used interchangeably with resulting confusion and inconsistency.

The analysis of this study also brings to light the area of metadata semantics within the MODS framework. Some of the MODS metadata elements engender difficulty and confusion during the metadata creation process. For instance, the MODS framework has a note element as well as a tableOfContents element. The confusion lies within the *MODS User Guidelines* (Library of Congress 2007), which specify that the tableOfContents element "contains contents notes for a resource. It is roughly equivalent to MARC 21 field 505." However, the surveyed data inform us that the content notes tend to be used in the note element rather than in the tableOfContents. This indicates that semantic clarity is needed between these two elements within the MODS framework.

5.0 Conclusion

The MODS records examined in this study encompass digitized sound recordings, monographs, periodicals and born-digital web resources. The results of this exploratory study indicate that the MODS metadata scheme is suitable for describing such a wide range of materials and resource types. Metadata quality assessment of the MODS records surveyed for this study evinces that the top five most frequently used elements (titleInfo, originInfo, recordInfo, physicalDescription and subject) appeared in

86 percent of the records. The total number of MODS elements represented in each collection ranged from twelve to fifteen (out of 20 MODS top-elements), with at least ten of these elements present in over 50 percent of the total number of records. The results of this study show that MODS elements, sub-elements and attributes are underutilized.

The analysis of metadata quality for this study also indicates that easily accessible local guidelines for metadata creation contribute to the consistent and accurate application of the MODS metadata scheme. For instance, the Election 2002 Web Archive provides the most easily accessible information pertaining to local practices. As such, this collection shows the greatest consistency in terms of the required elements per local guidelines that appear in each record. Furthermore, this collection contains the least number of inaccurate descriptions. The metadata quality of this collection demonstrates the effectiveness of local guidelines in improving the quality of metadata. However, the effect of local guidelines for improving metadata quality needs to be examined further by comparing a larger number of local guidelines and their usage. This will be dealt in a new project which is currently underway.

Metadata semantics greatly affects consistency. Some of the MODS metadata elements (i.e., note and tableOfContents) engender particular difficulty and confusion during the metadata creation process. As reflected in this empirical study of metadata quality analysis, conceptual ambiguities and semantic overlaps among some MODS metadata elements affect the accurate and consistent application of MODS metadata. Further studies are needed to examine the semantics of MODS metadata elements relative to their impact on the application of the scheme and on semantic interoperability across MODS repositories.

The Copac catalogs seem to be good candidate for further examination, as they have been converted from MARC to MODS through an automated process. The surveyed sample records (n =20) from this collection illustrate that there are a variety of inaccuracies within the Copac records. In many cases, such inaccuracies seem to occur because of "miscues" in the MARC to MODS conversion. St. Pierre and LaPlant (1998) note that "if the metadata and crosswalk transformations could be captured in a formal way that is consistent throughout the many metadata standards, the implementation of the standards and their crosswalks would be vastly simplified." Further examination of the Copac records might provide a better opportunity for gaining an insight into possible improvements in the creation and application of crosswalks between metadata schemas.

Despite the findings of this study, there are several limitations. The major limitation stems from the sample size. We examined MODS metadata use and metadata quality by using only three digital repositories with small number of sample records from each repository. It is possible that a sizable number of metadata records in conjunction with a larger pool of MODS repositories may bring forth different results on metadata use and quality. In this sense, further examination with a larger sample size will enable us to have better understanding on the current state of MODS metadata use and quality. Another limitation stems from the comparability issue of surveyed repositories: Copac catalogs are different from the other two repositories in the sense that they have been converted from MARC to MODS through an automated process. Future research also lies in further examination of metadata quality by examining digital repositories similar in contextual matters such as history, subject, resource type, and target users.

References

Barton, J., Currier, S., and Hey, J.M.N. 2003. Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. *2003 Dublin Core Conference*. http://purl.oclc.org/dc2003/03barton.pdf

Bruce, Thomas R. and Hillmann, Diane I. The continuum of metadata quality: defining, expressing, exploiting. In Diane Hillmann & Elaine L. Westbrooks (Eds.). *Metadata in practice.* Chicago: American Library Association.

Bui, Yen and Park, Jung-ran. 2006. An Assessment of Metadata Quality: A Case Study of the National Science Digital Library Metadata Repository. In Haidar Moukdad (Ed.). *Information science revisited: approaches to innovation*, CAIS/ACSI 2006 Proceedings of the 2006 annual conference of the Canadian Association for Information Science held with the Congress of the Social Sciences and Humanities of Canada at the York University, Toronto, Ontario. June 1 – 3 2006. http://www.cais-acsi.ca/proceedings/2006/bui_2006.pdf

Currier, S., Barton, J., O'Beirne, R., and Ryan, B. 2004. Quality assurance for digital learning object repositories: issues for the metadata creation process. *ALT-J research in learning technology* 12: 5-20.

Duval, E., Hodgins, W., Sutton, S., and Weibel, S.L. 2002. Metadata principles and practicalities. *D-lib* *magazine* 8(4). http://www.dlib.org/dlib/april02/weibel/04weibel.html

Guenther, R. S. 2003. MODS: the metadata object description schema. *Portal: libraries and the academy 3*: 137-50.

Guenther, R. S. 2004. Using the metadata object description schema (MODS) for resource description: guidelines and applications. *Library hi tech 22*: 89-98.

Guy, M., Powell, A., and Day, M. 2004. Improving the quality of metadata in eprint archives. *Ariadne* 38.

Hillmann, D., Dusshay, N. and Phipps, J. 2004. Improving metadata quality: augmentation and recombination. Paper presented at the International Conference on Dublin Core and Metadata Applications (DC-2004), Shanghai, China, October 2004. http://lcweb2.loc.gov/cocoon/minerva/html/elec2002/about-metadata.html

International Federation of Library Associations and Institutions. 1998. *Functional Requirements for Bibliographic Records: Final Report*. http://www.ifla.org/VII/s13/frbr/frbr.pdf

Library of Congress. 2007. Metadata Object Description Schema (MODS). http://www.loc.gov/standards/mods/

Library of Congress. 2007. MODS User Guidelines. http://www.loc.gov/standards/mods/v3/mods-userguide.html

Library of Congress. 2006. MARC Mapping to MODS Version 3.2. http://www.loc.gov/standards/mods/v3/mods-mapping.html#mapping.

McCallum, Sally H. 2004. An introduction to the Metadata Object Description Schema (MODS). *Library hi tech 22*: 82-88.

Missingham, R. 2004. Reengineering a national resource discovery service: MODS down under. *D-lib magazine* 10(9). http://www.dlib.org/dlib/september04/missingham/09missingham.html

Moen, W.E., Steward, E.L., and McClure, C.R. 1997. The role of content analysis in evaluating metadata for the U.S. Government Information Locator Service: results from an exploratory study. http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm.

National Information Standards Organization. 2004. *Understanding metadata*. Bethesda: MD: NISO Press. http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

Park, Jung-ran. 2005. Semantic interoperability across digital image collections: a pilot study on metadata mapping. In Liwen Vaughan (Ed.). *Data, information, and knowledge in a networked world*, Proceedings of the 2005 annual conference of the Canadian Association for Information Science held with the Congress of the Social Sciences and Humanities of Canada at the University of Western Ontario, London, Ontario, June 2 - 4 2005. http://www.cais-acsi.ca/proceedings/2005/park_J_2005.pdf.

Park, Jung-ran. 2006. Semantic interoperability and metadata quality: an analysis of metadata item records of digital image collections. *Knowledge organization* 33: 20-34.

Statistics Canada, Minister of Industry. 2002. *Statistics Canada's Quality Assurance Framework*. http://www.statcan.ca/english/freepub/12-586-XIE/12-586-XIE02001.pdf.

St. Pierre, M., & LaPlant, W. P. 1998. Issues in crosswalking content metadata standards. http://www.niso.org/press/whitepapers/crsswalk.html.

Stvilia, B., Gasser, L., Twidale, M., Shreeves, S., and Cole. T. 2004. Metadata quality for federated collections. Metadata quality for federated collections. In *Proceedings of ICIQ004—9th International Conference on Information Quality*. pp: 111-125.

Svenonius, Elaine. 2000. *The intellectual foundation of information organization.* Cambridge, MA: MIT Press.

Urbaniak, Geoffrey C. & Plous, Scott. 2008. Research Randomizer. http://www.randomizer.org/.

Zeng, Marcia. 2006. Metadata quality study for the National Science Digital Library (NSDL) Metadata Repository. Presented paper at the *Research and teaching talk series* in Information Science and Technology at Drexel University.

Resources

Copac Academic and National Library Catalogue. http://copac.ac.uk/

About Copac: http://copac.ac.uk/about/

Technical Requirements for Data supplied to Copac. http://www.curl.ac.uk/projects/challengefundFAQ.htm

CURL Minimum Standards for Bibliographic Records. http://www.curl.ac.uk/database/bibstandards.htm

United States Election 2002 Web Archive, Archived in the Library of Congress Web Archives. http://lcweb2.loc.gov/cocoon/minerva/html/elec2002/elec2002-about.html

MODS: Description of Elements (Election 2002 Web Archive). http://lcweb2.loc.gov/cocoon/minerva/html/elec2002/about-metadata.html

Subject Terms for Use in Cataloging (Election 2002 Web Archive). http://lcweb2.loc.gov/cocoon/minerva/html/elec2002/about-metasubjects.html

African-American Band Music and Recordings, 1883-1923. http://memory.loc.gov/cocoon/ihas/html/stocks/stocks-home.html

Appendix A: Mapping MARC to MODS for Copac Catalogs

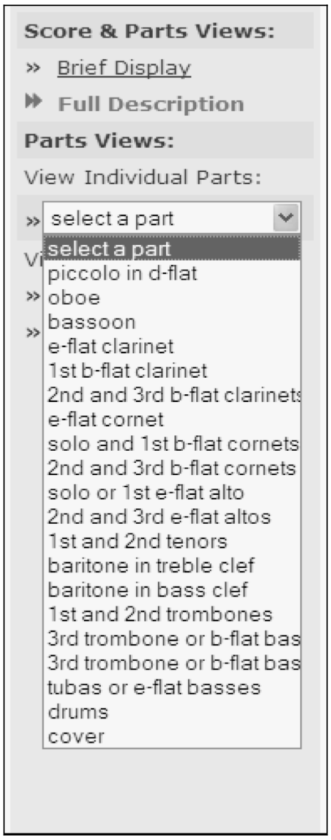
MARC field	MODS element
Mandatory data elements	
130 240 $a, $d, $f, $k, $l, $m, $o, $r 245 $a, $b, $n, $p 246 740	titleInfo
100, 110, 111 700, 710, 711	name
Leader/06*	typeOfResource
008/07-14 250 $a 260 $a, $b, $c	orginInfo
300 $a, $b, $c, $e 130, 240, 242, 245, 246, 730 $h (text only)	physicalDescription
505	tableOfContents
245 $c 500 534 $p, $a, $b, $c, $e, $f, $k, $l, $m, $n, $t	note
600, 610, 611 630 650, 651(post-1800 imprints)	subject
440 490 770 800-830	relatedItem
506 $a, $b, $c $d $3 $5 540 $a $b $c $d $3 $5	accessCondition
020 022	identifier
No mapping elements; local identifier	extension
Non-mandatory data elements	
240 $a, $d, $f, $k, $l, $m, $o, $r	titleInfo
041	language
510 546 561 562	note
245 $h (other than text) 300 $b, $c	physicalDescription
650 - LCSH 651 - LCSH 655 752	subject

Appendix B:
XML and Display Record—*relatedItem*

```
      - <mods:relatedItem
ID="DMD_p0001">
      - <mods:titleInfo>
        <mods:title>Piccolo in D-
flat</mods:title>
    </mods:titleInfo>
     </mods:relatedItem>
    - <mods:relatedItem
ID="DMD_p0002">
      - <mods:titleInfo>

<mods:title>Oboe</mods:title>
     </mods:titleInfo>
     </mods:relatedItem>
    - <mods:relatedItem
ID="DMD_p0003">
      - <mods:titleInfo>

<mods:title>Bassoon</mods:title>
     </mods:titleInfo>
     </mods:relatedItem>
```


Score & Parts Views:
» Brief Display
Full Description
Parts Views:
View Individual Parts:
» select a part
select a part
piccolo in d-flat
oboe
bassoon
e-flat clarinet
1st b-flat clarinet
2nd and 3rd b-flat clarinets
e-flat cornet
solo and 1st b-flat cornets
2nd and 3rd b-flat cornets
solo or 1st e-flat alto
2nd and 3rd e-flat altos
1st and 2nd tenors
baritone in treble clef
baritone in bass clef
1st and 2nd trombones
3rd trombone or b-flat bas
3rd trombone or b-flat bas
tubas or e-flat basses
drums
cover