Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data Providers of Cultural Heritage Materials

Arwen Hutt
University of Tennessee Digital Library Center
1015 Volunteer Blvd.
Knoxville, TN 37757
+1-859-865-6910
ahutt@utk.edu

Jenn Riley
Indiana University Digital Library Program
1320 E. 10th St, E170
Bloomington, IN 47404
+1-812-856-5759
jenlrile@indiana.edu

ABSTRACT
This study analyzes metadata shared by cultural heritage institutions via the Open Archives Initiative Protocol for Metadata Harvesting. The syntax and semantics of metadata appearing in the Dublin Core fields creator, contributor, and date are examined. Preliminary conclusions are drawn regarding the effectiveness of Dublin Core in the Open Archives Initiative environment for cultural heritage materials.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, standards, systems issues, user issues.*

General Terms
Documentation, Reliability, Standardization

Keywords
Open Archives Initiative, Dublin Core, metadata quality, interoperability, digital libraries

1. INTRODUCTION
The creation of digital resources is increasing at a rapid pace among organizations charged with organizing and preserving information. In cultural heritage institutions such as libraries, archives, museums, and historical societies, this creation is largely taking the form of digitization of existing analog materials. In addition to large well-established digitization programs, there are now an increasing number of smaller organizations creating digital objects. The growing prevalence of these small projects has naturally led to a desire to gather the resulting disparate collections into more centralized repositories. This trend is manifested in the current heightened interest in the creation and planning of aggregated collections [1] [8] [9] [26] [31]. Many of these new aggregated collections are being built upon the foundation of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [24].

JCDL '05, June 7–11, 2005, Denver, Colorado, USA

The OAI-PMH, hereafter referred to as OAI, was originally developed as a low-barrier method for the sharing of metadata about e-prints, electronically published research papers [17]. There are two types of participants in the framework: data providers and service providers. Data providers expose metadata, which can then be harvested by service providers. Service providers commonly add value by aggregating metadata from multiple repositories into larger searchable collections [24]. Although OAI was originally developed primarily for textual materials (e-prints), it is now widely used with many other types of resources and in many different knowledge communities.

OAI data providers are required to expose an unqualified, or "simple," Dublin Core (DC) metadata record for every item represented in the repository. DC is a simple yet flexible metadata standard, intended to describe a wide variety of resources [14]. As DC favors "document-like objects" [21] it provided a good match for describing the e-prints that were the original focus of the OAI protocol. Although richer secondary metadata records are permitted in addition to the required DC record, the majority of OAI data providers make only this basic record available [30]. However, given a choice, most cultural heritage institutions do not implement strict simple DC in their local environments. One survey of DC usage in libraries indicated that 9% used strict simple DC, 18% used strict qualified DC, and 73% added local qualifiers to the base qualified DC set [18].

Dublin Core prescribes very few exact mechanisms for its metadata, yet it does include numerous indications of best practice. One of the guidelines for creating DC records most relevant to this study is the One-to-One (1:1) principle: "In general Dublin Core metadata describes one manifestation or version of a resource, rather than assuming that manifestations stand in for one another." [21] Therefore, the metadata in a DC record should when possible describe one and only one manifestation of a resource. Because of DC's inherent flexibility and limited content guidelines, there is a great deal of variability in how metadata is represented in OAI records. As a result, the aggregation of widely varied materials into collections via OAI has stimulated discussions of what makes truly valuable "shareable metadata." [7]

2. GOALS
This study builds upon previous work in this area by conducting data analysis of Dublin Core usage in OAI data providers of

cultural heritage materials. This focus on DC use via OAI within a single community of practice allows us to conduct an in-depth investigation of the specific descriptive needs of that community. We examine the usage of the DC creator, contributor, and date fields looking at the semantic content and syntactic form of their values. We selected these fields for initial analysis because we hypothesized they would demonstrate descriptive needs of cultural heritage institutions. Analysis of other DC fields will be performed at a later stage of this project. This study should be beneficial to the OAI and cultural heritage communities by helping to inform best practices for consistent metadata and aid data normalization and indexing practices of service providers. One of the initial motivations for this study was the desire to better understand current practices in the cultural heritage community to inform our personal work as OAI data providers.

Previous studies on metadata shared via OAI have generally concluded with a discussion of the great deal of investigation still necessary; this study is yet another step along the way toward a more thorough understanding of how different communities of knowledge use the OAI protocol for sharing metadata within their community and beyond.

3. RELATED RESEARCH

A number of studies of the usage of DC fields have been conducted on OAI records. Jewel Ward performed a thorough study of DC element use across 82 data providers. She found that data providers used "an average of eight DC elements … per record," and "the top five DC elements accounted for 71 per cent of all element usage." She concludes that, despite its simplicity, "...the DCMES [Dublin Core Metadata Element Set] is not used to the fullest extent possible." [33] The University of Illinois at Urbana-Champaign (UIUC) Digital Gateway to Cultural Heritage Materials compiled similar statistics [29].

The Arc service provider, which harvests data providers from various subject domains, has focused much of their analysis on the semantics of various DC fields within OAI records. They have calculated usage of controlled values in the DC subject, type, format, language, and date elements, and used this data to construct browse interfaces for language, type, and format [25].

Several studies have also been published focusing on DC OAI usage within a specific knowledge domain. Staff from the National Science Digital Library (NSDL) categorized errors observed in metadata they harvested into four groups [16]:

1. Missing data, where information deemed essential to the service provider was not present in records from a data provider;
2. Incorrect data, where an element's content did not fit the definition of the element in the metadata schema;
3. Confusing data, where multiple values were placed within a single element, often with inconsistent indications of where one value ends and another begins, or extraneous data such as HTML tagging within an element's content; and
4. Insufficient data, where the metadata, particularly when presented in simple DC format, was not robust enough for good indexing and display of search results.

Analysis of the presence of each of the fifteen DC elements in metadata harvested by the UIUC cultural heritage materials service provider "showed wide disparities in how and for what purpose various DC elements were utilized. Encodings used in even standard elements such as 'date,' 'coverage,' 'format,' and 'type' varied significantly." The authors of this study concluded that these differences "in part … relate to whether the metadata author chose to focus on describing a given work itself or on describing the digital surrogate of a given work. This in turn appears to be traceable to the nature of the original environment or project for which the metadata was created." [4] The NSDL and UIUC studies show that metadata consistency and quality are significant problems in the OAI environment.

4. METHOD

In this study, we harvested Dublin Core metadata via OAI from data providers included in the UIUC Digital Gateway to Cultural Heritage Materials [32]. We were able to successfully harvest DC records from 44 data providers. A total of 750,945 records were harvested once deleted records and duplicates were removed. We took a 5% sample, yielding 37,564 records for analysis.

From the sampled records we extracted our target elements (creator, contributor and date) to create three element silos. Each silo was an individual XML file containing records from the sample that had one or more instances of the target element. Each of the three silos was then processed to separate multiple instances of each element into a single record and to aggregate repeated values, providing us with multiple views of the metadata. For each element, we recorded data about certain characteristics of the enclosed value. Each of these characteristics, listed in Table 1, was coded as an attribute in the element tag.

Table 1. Characteristics recorded

All elements (Creator, Contributor & Date):	The presence of multiple discrete values in a single element
	The presence of pseudo-qualifiers within the value that refined the meaning of the element
	Whether the value was appropriate within the specified element based on DC rules and usage guidelines
Creator and Contributor:	The semantic type of the value (personal name, corporate name or other)
	Whether the entity is known, unknown or ambiguous
	Whether the value is inverted or in direct order
Date:	The semantic type of the value (creation, copyright or digitization)
	The general specificity of the date (single date, range or period)
	Indication that a date is not definitive (that it is estimated or approximate)
	Whether the value is purely numeric or contains non-numeric text

Categorization was done in several stages, starting with automated processing based on patterns present in an element (e.g., the string "Inc" indicated a corporate name in the creator file). These categorizations were iteratively reviewed and revised to improve their accuracy. We then performed a basic manual review of each silo in order to further sort values that were not matched in the automated review. The result was a rough categorization of element values, providing a big-picture view of trends in DC element content, rather than a precise categorization of every element value. After each silo had been manually reviewed, the occurrence of each attribute and certain combinations of attributes were tabulated providing the numerical data discussed in the remainder of this paper.

At several points in the data manipulation process, count scripts were run to provide data for analysis and to verify data consistency through the project lifecycle. In addition, we performed data cleanup at several stages of processing to remove anomalies such as empty elements and leading spaces from element values. However, other content variations were not normalized, such as trailing punctuation, HTML tagging within element values, and quotes around element content.

5. FINDINGS

5.1 Dates

5.1.1 Appropriateness of Date Values

Of the elements analyzed, date strings had the highest level of conformance to a strict definition of the Dublin Core element, with only 0.61% (n=80) of unique values and 2.08% (n=701) of total values categorized as inappropriate for <dc:date>. Some of the inappropriate values seem to be a result of processing errors by data providers (e.g., "unknown," "(ca.)" with no date following), and this sort of value frequently appeared in a large number of records, creating a very high average of 8.76 uses per inappropriate value. Most of the inappropriate values, however, were long strings mixing several types of information into a single element, and could have been marked as a valid date value according to a looser interpretation of the <dc:date> element definition. The low occurrence of inappropriate values in the date element could be attributed to the fact that date of creation is an important access point for many cultural heritage materials, and as such might be recorded in a consistent manner in local metadata systems.

5.1.2 What the Record Describes

Dates in DC records are among the elements that most clearly show the effect of the 1:1 principle. Different versions of a resource, especially a cultural heritage resource, can be created at vastly different times. For example, one record in our sample described a photograph taken around 1920, while the digitized version was created in 2001, and both dates were provided in the OAI DC record. To partially analyze the effect of the 1:1 principle on DC date usage in OAI records, we sorted dates from records in our sample into three classes: creation date of the original version of resource, copyright date of the resource, and date the resource was digitized or otherwise put online. This sorting, like others in this study, was a rough classification intended to provide a high-level picture of DC usage. High-level assumptions were made to facilitate sorting; we did not individually verify all dates in the sample.

Table 2. Types of dates appearing in sample

	Unique Values		All Values	
Digitization date	5286	40.68%	8026	24.33%
Creation date	7585	58.37%	24624	74.64%
Copyright date	124	0.95%	339	1.03%
TOTAL	**12995**	**100.00%**	**32989**	**100.00%**

As seen in Table 2, creation dates made up the majority of <dc:date> values in our sample, representing 58.37% (n=7585) of the unique values and 74.64% (n=24624) of the total values in the date silo. Each date string representing the date a resource was created, therefore, was used an average of 3.25 times. Digitization dates also appeared frequently in OAI DC records, making up 40.68% (n=5286) of the unique values and 24.33% (n=8026) of the total values in the date silo. Digitization dates were re-used much less frequently (averaging 1.5 occurrences per date string) than creation dates. This is probably because digitization dates tended to be more granular than creation dates, often including timestamps. Copyright dates made up a small percentage of the dates present in the sample, 0.95% (n=124) of unique date values and 1.03% (n=339) of total date values. This small number of copyright dates in our sample may be due to the fact that only dates explicitly indicated as copyright dates were counted; many other dates for published materials may have been copyright dates but were not indicated as such within the <dc:date> value. The distinction between a copyright date and a creation date is only considered important in certain communities, e.g., libraries, and thus would only likely appear in records created within that community.

Many of the dates classified in this study as digitization dates were from OAI data providers run by the ContentDM digital asset management system [5]. Turnkey digital asset management systems not only provide search and display of digital content, but also provide users with some administrative functions for their digital content. Thus the system must store metadata about multiple versions of a resource, and users of systems such as these often may not have complete control over how these multiple versions are represented in OAI DC records.

5.1.3 Pseudo-qualifiers

Few date elements in our sample included "pseudo-qualifiers," extra text within the element intended to further refine the element's meaning within simple DC (e.g., "Digitized: 2005-01-19"). These occurred only 0.83% of the time for unique values and 0.64% of the time for all values in our sample. However, 6.44% of the records in the sample contained more than one <dc:date> element, indicating that creators of OAI DC records commonly left ambiguities in their records instead of attempting to add qualifiers to date element values.

5.1.4 Format

As seen in Table 3, the overwhelming majority of <dc:date> values appeared in numeric form, comprising 74.24% (n=9649) of unique values and 82.82% (n=28485) of total values in the sample. The remaining values were coded as textual representations of dates. Some of these were named periods (e.g., "20th Century"), but most were written-out month names (e.g.,

"Jan." or "January"). Month names appeared at the beginnings of date strings (e.g., "January 31, 1948"), following the day (e.g., "10 Jan 1944)", and in several other variant forms. Many descriptive standards for cultural heritage materials prescribe some version of a written-out month name when this is known. Despite the wide variety of formats used for named months, it might be possible to write parsing routines that go a long way towards normalizing these values into standard formats.

The Dublin Core Element Set lists as a best practice for the date element use of the WC3DTF profile of ISO8601 [15]. W3CDTF defines six levels of granularity for dates [6]:

Year:
YYYY (eg 1997)
Year and month:
YYYY-MM (eg 1997-07)
Complete date:
YYYY-MM-DD (eg 1997-07-16)
Complete date plus hours and minutes:
YYYY-MM-DDThh:mmTZD (eg 1997-07-16T19:20+01:00)
Complete date plus hours, minutes and seconds:
YYYY-MM-DDThh:mm:ssTZD (eg 1997-07-16T19:20:30+01:00)
Complete date plus hours, minutes, seconds and a decimal fraction of a second
YYYY-MM-DDThh:mm:ss.sTZD (eg 1997-07-16T19:20:30.45+01:00)

The <dc:date> values in our sample conformed to one of these six W3CDTF formats just over 17% of the time (n=2200). The dates coded as textual, described above, by definition do not conform to the W3CDTF format. However, even within the numeric date values, most did not conform to W3CDTF. Many of these included timestamps separated from a preceding date in YYYY-MM-DD format by a space (e.g., "2000-10-31 00:00:00" or "2003-03-13 13:44:34"), rather than with the W3CDTF-prescribed "T."

Table 3. Format of dates appearing in sample

	Unique Values		All Values	
Numeric	9650	74.26%	27321	82.82%
Textual	3345	25.74%	5668	17.18%
TOTAL	**12995**	**100.00%**	**32989**	**100.00%**

5.1.5 Needs Not Addressed by W3CDTF

Many date values in our sample were not expressible in W3CDTF form. While 89.34% (n=11610) of unique values and 83.33% (n=27489) of total values were coded as representing a single date, not all of these are expressible in W3CDTF. Cultural heritage materials, by their nature, frequently do not have known precise dates of creation, although this is assumed by W3CDTF.

Some dates in the sample were at a less granular level than any of the W3CDTF formats, with only a decade or a century known, instead of the exact year. Others are a named time period, for example, "Summer 1957." Many cultural heritage institutions locally use data content standards that provide for less granular date levels than a known year. These standards use a variety of formats to represent this lack of granularity, including "19--" and "1930s."

Descriptive standards in use by many cultural heritage institutions provide mechanisms for indicating whether a date is known or estimated. Definitions of "known" may differ between standards, however. The Anglo-American Cataloging Rules, 2nd edition (AACR2) [2] describe the concept of a "chief source of information" from which cataloging data of various sorts should be obtained. In some cases, data can be obtained from sources other than the chief source, but this must be indicated by enclosing the value obtained elsewhere in square brackets. Other conventions to indicate an uncertain or estimated date include "ca.," "or," "between," and "?." These qualifiers in AACR2 all have slightly different meanings, none of which are currently available in W3CDTF formats. The cultural heritage community would benefit from the development of normalization algorithms for the conventions in AACR2, Archives, Personal Papers, and Manuscript (APPM) [20], and Describing Archives: A Content Standard (DACS) [28], for improved date searching from OAI service providers including cultural heritage materials.

Table 4. Specificity of dates appearing in sample

	Unique Values		All Values	
Date range	1309	10.07%	4815	14.60%
Single date	11610	89.34%	27489	83.33%
Period	76	0.58%	685	2.08%
TOTAL	**12995**	**100.00%**	**32989**	**100.00%**

Date ranges, representing 10.07% (n=1309) of unique values and 14.60% (n=4815) of date values in the sample, are also not expressible in W3CDTF form or the much more extensive ISO 8601 date standard [22]. While date ranges appeared in our sample much less frequently than single dates, it is clear that the best practice for DC dates should include some mechanism for expressing them. The DC Date Working Group [10] is addressing these issues. This group's 2004/2005 work plan includes the following action item [11]:

> Investigate options to provide for the interoperable representation of commonly-recorded dates which cannot be satisfactorily represented using *ISO 8601 Data elements and interchange formats–Information interchange–Representation of dates and times*, including the following categories of dates:
>
> - B.C.E. dates
> - Questionable dates
> - Approximate dates
> - Open-ended date ranges
> - Non-Gregorian dates
> - Large dates (e.g., geologic periods, astronomical time)
> - Soft termini (i.e. the outer bounds for one or more termini is known or can be associated with a known period, but one or both of the exact boundaries of the event referenced are not known)
> - Elapsed time less than date range interval (i.e. the duration is less than the complete interval between two termini, as in an intermittent activity)

5.2 Creator

5.2.1 Appropriateness of Creator values

The analysis of Dublin Core creator elements showed a low occurrence of values inappropriate to the DC definition of the element, 1.14% (n=168) of unique values and 2.43% (n=683) of total values. This high level of conformance can be partially attributed to the consistency of the concept of "creator" in a generic sense across varied disciplines. Our assumption of appropriateness in the absence of obvious contrary evidence probably also contributed to this result.

Although constituting a small percentage of total elements, a significant proportion of inappropriate values in creator fields contained data relating to time periods (e.g. Byzantine: Venice) which might more appropriately belong in a Coverage element. This may point to a difficulty faced by parts of the cultural heritage community, especially museums of cultural history, in describing ancient and unattributed cultural materials.

5.2.2 Types of Creators

As part of the analysis of the creator element, the content was sorted into three general classes: personal names, corporate names, and other values. This primary distinction between corporate (or group) names and personal names is drawn from the library community, and although not necessarily explicitly articulated by other groups, it can be applied to cultural heritage data in general. In most cases it was clear to which category a value belonged, but when there was doubt (e.g., a one-word name with no initials or corporate indicators), personal name was considered the default. The majority of creators were personal names, making up 90.56% (n=10645) of unique values and 82.95% (n=22790) of the total values in the creator silo. Only 14.82% (n=4071) of unique creator values were group names. This is roughly consistent with the distribution of name records within the Library of Congress Authority Files where 78.84% are personal names.

Table 5. Types of creators appearing in sample

	Unique Values		All Values	
Personal Name	10645	90.56%	22790	82.95%
Corporate Name	1064	9.05%	4071	14.82%
Other	46	0.39%	615	2.24%
TOTAL	**11755**	**100.00%**	**27476**	**100.00%**

There was a small percentage of "other" values, accounting for 0.39% (n=46) of unique values and 2.24% (n=615) total values. Roughly a third of these are values that mix personal and corporate names. The majority of the values typed as "other" are those that indicate the creator is unknown (e.g., "not known," "unknown"), composing 75.93% (n=467) of total type "other" values.

5.2.3 Pseudo-qualifiers

In contrast to the other elements examined, a very large proportion of creator elements contained pseudo-qualifiers. In analysis, text within the element value indicating the role the creator played in production of the object was considered a pseudo-qualifier. These were present in 28.01% (n=3292) of unique creator values and 29.44% (n=8090) of total creator values. The presence of such a high proportion of qualification, in spite of the lack of support for it within simple DC, is suggestive. It implies that data providers of cultural heritage materials feel role qualification is of significant importance either within a local collection interface or within an aggregated search environment.

Table 6. Psuedo-qualifiers in Corporate & Personal Name Creator Elements

	Unique Values		All Values	
Personal Names	2980	27.99%	5906	25.91%
Corporate Names	308	28.95%	2112	51.88%

Rudimentary analysis implies that records with multiple creator elements could account for the majority of elements with role pseudo-qualifiers. Further study could analyze the frequency of qualification as a function of the number of creators within a single record, to determine if role qualification primarily occurs in cases where multiple creators contribute to a single resource.

The value of role qualification in general seems to be supported by the work of the DCMI Libraries Working Group on the Library Application Profile [13]. Early versions of the profile included role as a refinement for both creator and contributor. These are absent in the September of 2002 version, but the following comment is included: "Refinements for Creator are needed to express role, as well as structured values to express further information about the creator. They are not included in the application profile, awaiting approval by DCMI of a mechanism to express these." [12] Although the most recent version of the application profile, from September 2004 [13], retains the prohibition on role refinement of <dc:creator>, the efforts the DC-Lib group made to find some mechanism for communicating this information supports the view that role qualification is considered important.

Not surprisingly, there was very little consistency among data providers on the syntax of role pseudo-qualifiers. Although service providers could parse this data from a list of possible role qualifiers, the lack of a consistent vocabulary usage and syntax again places an increased burden on the service providers.

5.2.4 What (else) is in a Name?

In addition to the presence of role pseudo-qualifiers, elements also contained data that qualified or refined the identity of the creator as an individual. Such content was considered different from role qualifiers in that it did not refine the element itself, but instead provided further description of the identity of the person named. Although not explicitly tracked in this study, these qualifications were almost exclusively present in personal name values.

Dates are probably the most common form of this type of qualification, as they help to uniquely distinguish individuals with similar names. Other types of information seen in the creator field included: occupations (governor, student teacher, mathematician), geographic origins, honoraries (sir, knight, baron) and institutional affiliations. As with role pseudo-qualifiers, this content was present in a wide variety of syntactical formats. This variation in form was further compounded by the frequent occurrence of more than one type of qualification in a value. Without consistent values or syntax, these qualifiers would prove

a significant obstacle to the processing and collocation of creator data.

5.2.5 Form and Usage of Controlled Vocabularies

Examination of the form of creator values showed that 85% of all personal names appeared in inverted order. This could indicate conformance to DC best practice, or that cultural heritage institutions frequently inverted names in their local metadata. There is also a fair amount of repetition of personal name values (on average 2.14 uses per value). This and the high occurrence of inverted form may imply the use of a controlled vocabulary for value assignment, even if this control is only internal to the repository.

Table 7. Inversion in Corporate & Personal Name Creator Elements

	Unique Values		All Values	
Personal Names	9097	85.46%	19641	86.18%
Corporate Names	15	1.41%	63	1.55%

In contrast, during manual review it was clear that there were a number of occurrences of the same name not being collocated because of variations in spelling, formatting or punctuation. Slight variations could represent inconsistencies in the parsing or export of data (e.g., presence of ending period on a name) rather than variations in the name itself. But differences in spelling and role or identity qualification weaken the possibility that controlled vocabularies or authority control are being used frequently.

Corporate names were repeated more often then personal names, each unique value being used 3.82 times on average. Although this may indicate a higher level of authority control, it may also be attributable to the greater simplicity (and thus smaller room for variation) in corporate names present in the creator silo. Corporate names seemed to lack the pseudo-qualifiers, inversion and other variations that occurred in personal names.

5.3 Contributor

5.3.1 Appropriateness of Contributor Values

Of the elements surveyed, contributor was used the least, being present in only 6.98% of the records harvested. In addition, contributor had the highest occurrence of values inappropriate to the element, 4.00% (n=41) of unique values and 30.05% (n=1072) of total values. These two characteristics seem to indicate that there is confusion as to how <dc:contributor> should be used in the description of cultural heritage materials. The DC definition of contributor, "an entity responsible for making contributions to the content of the resource," [15] is sufficiently vague to allow almost any agent information to be included.

As contributor is described as "the most general of the elements used for 'agents'" [21], we made the presumption that more specific elements, when appropriate, would be preferred. In determining inappropriate values, we included those that, although possibly correct based on a loose interpretation of the contributor definition, fit more naturally within another DC element. The primary example of this was the inclusion of information relating to the institution responsible for the digital object and collection available. This information would be more accurately placed within <dc:publisher> as "the entity that provides access to the resource" [21]. This use accounted for the great majority of inappropriate values, and explains their high reuse, averaging 26.14 uses per value.

This use of contributor instead of publisher may be in part due to the highly specific connotation "publisher" has within libraries, a major part of the cultural heritage community. The placement of digital publisher data within <dc:contributor> may indicate a reluctance to broaden the traditional concept of publisher to contain digital publication information, but may also indicate confusion as to element usage. More research is needed to determine whether the low and varied usage of contributor is rooted in a disconnect between the DC notion of the element and standard descriptive practices in the cultural heritage community, the absence of clear discussion of how to use the element, or a simple lack of need. What seems clear is that contributor is not being used frequently and consistently enough to make it a useful source of information for service providers. Perhaps because of this, the OAIster OAI service provider went so far as to disregard data included in <dc:contributor> elements [19].

5.3.2 Types of Contributors

The same basic classes were applied to the contributor element as were applied to creator: personal name, corporate name and other. In contrast to the distribution of types seen in creator, 21.87% (n=215) of unique values and 41.96% (n=1047) of total contributor values were of the corporate type. Although personal names still account for the majority of values, 77.31% of unique values and 57.52% of total values, there is a substantial decrease in the proportion of personal to corporate names from creator to contributor. This shift towards corporate name types would be even stronger if the values relating to digital publication, marked as inappropriate to the contributor element, were included in this count. There was a very low occurrence (less than 1% of unique and total values) of values with the type "other" in the contributor element. The higher prevalence of corporate names within <dc:contributor> compared to <dc:creator> supports the impression that institutions and groups more often play a supporting role in creative endeavors.

5.3.3 Pseudo-qualifiers

Psuedo-qualifiers were present in few contributor elements, accounting for only 4.58% (n=45) of unique values and 4.21% (n=105) of total values. This might be attributed to a perceived lower importance of contributor both in general and specifically within an aggregated searching environment, or to a lack of clarity of use of the contributor element in general. Interestingly, the most recent version of the DC Library Application Profile allows for role qualification of <dc:contributor>, although not <dc:creator> [13]. As might be expected, based on the lower proportion of personal names and perception that contributors are less important than creators, there did not seem to be a large amount of non-role qualification in the contributor elements.

6. CONCLUSIONS

6.1 The OAI DC Record Context

Discussion of what makes a useful and shareable metadata record requires the consideration of two relationships. The first of these is between the OAI DC metadata record and the intellectual object it describes. The second is the relationship between the OAI DC record and the aggregated search environment the record

is being brought into. These two areas of interaction provide a framework for discussion of the results of this survey.

6.1.1 The OAI DC Record & the Intellectual Object

The 1:1 principle, which requires DC records to describe exclusively one version of a resource, is particularly problematic for cultural heritage institutions where the majority of digital objects are not born digital but are instead created from the digitization of existing analog materials. Therefore, it is common for multiple versions of an intellectual object to exist within a single institution, often including the original analog materials, the master digital file and at least one derivative digital file. Representing this complexity in the OAI DC environment obviously presents a challenge. In some advanced local environments, mechanisms can exist to relate simple DC records for multiple manifestations of a work to one another and thus follow the 1:1 principle. In such a system the full description of the work is formed from the combination of multiple records, and as a result, a complete picture of the work does not appear in any single DC record. However, in an OAI environment, each record must stand on its own; the external semantics necessary to make complete sense of the relationships between records are not shared.

This leaves data providers with two choices, create records that adhere to the 1:1 rule and omit pertinent information, or violate the rule. We observed many cases in which data providers chose to violate the rule and combine data about the original intellectual object as well as its digital manifestation. This can be seen in the high presence of dates associated with both the original and digital object, and the inclusion of digital publication data. This sort of departure from DC best practice may be due in some cases to lack of knowledge of the best practice, but seems to be often caused by the difference in focus between DC as a "core" set of descriptive practices for any and all resources and the specific descriptive needs of the cultural heritage community.

6.1.2 The OAI DC Record & the Aggregated Search Environment

A common suggestion among metadata implementers is that "...it is helpful to think of metadata as multiple views that can be projected from a single information object." [23] Applying this principle in an OAI environment, an OAI DC record should represent one view of a more complete metadata record for a specific resource. A defining characteristic of the OAI DC record context is the removal of the individual metadata records from their original collection for eventual aggregation with records from other collections. In most cases a metadata "view" useful in this new environment will be different from that useful in local systems. This primarily involves the addition of contextual information unnecessary in a local environment and the removal of information *only* relevant locally.

DC records in this study frequently showed evidence that data providers had not created them with their combination with metadata from other data providers in mind. Many of the DC elements examined in this study showed evidence of "hacks" to achieve a specific result in a data provider's local context. For example, one repository often represented date ranges with every year in the range listed within a single <dc:date> element (e.g.., <dc:date>1901 1911 1902 1903 1904 1905 1906 1907 1908 1909 1970 1911</dc:date>), presumably for a local search engine to retrieve the record for a search on any of the included years. Administrative metadata that is primarily of local value similarly appears in harvested records. Refraining from exposing this sort of local information would make DC practice more consistent across data providers and thus allow for better aggregation of metadata by service providers.

6.2 Needs of Cultural Heritage Data Providers

Records in this study clearly showed a disconnect between the structure and goals of simple DC and the descriptive needs of cultural heritage institutions. There were two main areas where this seemed especially pronounced; role qualifications and date granularity and uncertainty.

Data in this study demonstrated the need (or desire) of the cultural heritage community to record the relationship between a creator and a resource. As discussed previously, such qualifications are more important when describing the varied material types that cultural heritage institutions typically produce, than when dealing with collections of a fairly consistent type (e-prints). Similarly, this study demonstrated the need of the cultural heritage community to support in their metadata records a wide variety of types and granularities of dates that are not currently provided for within stated best practice for the DC date element.

6.3 Moving Towards Better Metadata

As noted by the managers of the Arc OAI service provider, "the effort of maintaining a quality federation service is highly dependent on the quality of the data providers." [25] The results of this study suggest that metadata quality problems are widespread. Problems of this sort have no easy solutions, but instead can be approached using a variety of strategies. Although there are many options worth further investigation, each has its own obstacles and drawbacks. The most drastic change would be to remove the OAI requirement for a simple Dublin Core record. While this change has been discussed within the OAI community, no consensus was reached and therefore no change was made [27]. The total effect this change in policy would have across different knowledge domains is unclear, but some other mechanism to achieve basic metadata interoperability would be required to take the place of the use of simple Dublin Core. This could take the form of community specific required metadata formats or perhaps requiring a qualified Dublin Core record. If alternatives are not provided, are unclear, or focus too strongly on one subset of the OAI community, such a change seems likely to exacerbate current problems rather than fix them.

A slightly less drastic, but probably no less controversial, strategy would be the development of best practice documentation for cultural heritage materials in Dublin Core that deviate from current DC best practice. The development of such guidelines within the cultural heritage community could bring greater content standardization and consistency to metadata created in this format, allow expression of important concepts not currently supported, and possibly increase search and retrieval functionality by service providers. One of the major problems with such a strategy would be the threat of weakening the nature of Dublin Core as a digital "pidgin" [21] by sanctioning deviation from its rules.

As is often the case, the more palatable strategies are often the least dramatic options. One such strategy is to continue to educate metadata providers on how to create quality shareable metadata. Many significant advancements have recently been made in this area, including a chapter by Thomas R. Bruce and Diane I. Hillmann in the book *Metadata in Practice* [3]. Another is to strongly encourage cultural heritage data providers to make use of the capability within OAI to expose other metadata formats in addition to OAI DC. While these data providers must still make difficult decisions about how to implement simple DC, exposing metadata records in richer standard formats will increase their ability to effectively communicate information about their resources to other members of their community.

While it is more desirable for metadata quality issues to be addressed at the data provider level, service providers will always be required to do some data normalization. Another strategy for improving metadata quality is for service providers to share the data normalization tools and strategies they have developed for their own use. A common theme among these options for improving metadata quality in the current OAI environment is one of communication. Any combination of strategies resulting in an improvement of metadata quality will necessarily be a result of discussion among both service and data providers, and the many knowledge communities that utilize OAI.

7. FURTHER RESEARCH

The research presented in this paper is the first stage of a larger project. We plan to continue analysis on the date, creator and contributor data, and also expand to other DC elements, including subject, coverage, and publisher. Later stages of the project will focus on analyzing temporal information across the date and coverage elements, geographic information across the subject and coverage elements, and name information across the creator, contributor, and publisher elements. Data from these analyses could be used to evaluate potential strategies for improving metadata consistency among data providers of cultural heritage materials.

8. ACKNOWLEDGMENTS

The authors would like to thank the University of Tennessee Libraries' for financial support of this project through their Faculty Research Incentive Program, Sarah Shreeves of the University of Illinois at Urbana-Champaign for generously sharing the list of cultural heritage data providers harvested by UIUC, and Michelle Dalmau, Sarah Shreeves, John A. Walsh, and Jon W. Dunn for their helpful reviews of drafts of this paper.

9. REFERENCES

[1] American West Project, http://www.cdlib.org/inside/projects/amwest/

[2] *Anglo-American Cataloging Rules* (AACR2), 2nd ed., 2002 revision. Chicago: American Library Association; Ottawa: Canadian Library Association; London: Chartered Institute of Library and Information Professionals, 2002.

[3] Bruce, Thomas R. and Diane I. Hillmann. "The Continuum of Metadata Quality: Defining, Expressing, Exploiting." In Diane I. Hillmann and Elaine L. Westbrooks, eds. *Metadata in Practice*. Chicago: American Library Association, 2004.

[4] Cole, Timothy W. and Sarah L. Shreeves. "Lessons Learned from the Illinois OAI Metadata Harvesting Project." In Diane I. Hillmann and Elaine L. Westbrooks, eds. *Metadata in Practice*. Chicago: American Library Association, 2004.

[5] ContentDM Digital Collection Management Software, www.contentdm.com

[6] Date and Time Formats, W3C Note, http://www.w3.org/TR/NOTE-datetime

[7] Digital Library Federation (DLF) OAI Best Practice Working Group, http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl

[8] Digital Library Federation (DLF) Project Aquifer, http://www.diglib.org/aquifer/

[9] Documenting the American South, http://docsouth.unc.edu/

[10] Dublin Core Date Working Group, http://www.dublincore.org/groups/date

[11] Dublin Core Date Working Group 2004-2005 Workplan, http://www.dublincore.org/groups/date/workplan_2004-2005.shtml

[12] Dublin Core Library Application Profile, 2002-09-24, http://dublincore.org/documents/2002/09/24/library-application-profile/

[13] Dublin Core Library Application Profile, 2004-09-10, http://dublincore.org/documents/2004/09/10/library-application-profile/

[14] Dublin Core Metadata Initiative, http://dublincore.org/

[15] Dublin Core Metadata Terms, 2005-01-10, http://dublincore.org/documents/dcmi-terms/

[16] Dushay, Naomi and Diane I. Hillmann. Analyzing Metadata for Effective Use and Re-Use. In *2003 Dublin Core Conference: Supporting Communities of Discourse and Practice--Metadata Research & Applications*. (September 28 - October 2, 2003, Seattle, Washington). Information Institute of Syracuse, Syracuse, NY, 2003.

[17] Eprints Glossary, http://www.eprints.org/glossary/

[18] Guinchard, Carolyn. Dublin Core use in libraries: a survey. *OCLC Systems & Services, 18, 1* (2002), 40-50.

[19] Hagedorn, Kat. OAIster: A 'no dead ends' OAI service provider. *Library Hi Tech, 21, 2* (2003), 170-181.

[20] Hensen, Steven, comp., *Archives, Personal Papers, and Manuscripts*, 2nd ed. Society of American Archivists, Chicago, 1989.

[21] Hillmann, Diane. Dublin Core usage guide, 2003-8-26. http://dublincore.org/documents/usageguide

[22] International Standards Organization. ISO 8601:2004, Data elements and interchange formats - Information interchange - Representation of dates and times, 2004. http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=40874

[23] Lagoze, Carl. Keeping Dublin Core simple: cross-domain discovery or resource description?.*D-Lib Magazine, 7, 1,* (January 2001). http://www.dlib.org/dlib/january01/lagoze/01lagoze.html

[24] Lagoze, Carl and Herbert Van de Sompel. The making of the Open Archives Initiative Protocol for Metadata Harvesting. *Library Hi Tech*, *21, 2* (2003), 118-128.

[25] Liu, X. et al. Federated searching of interface techniques for heterogeneous OAI repositories. *Journal of Digital Information, 2, 4,* (2002). http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/

[26] National Science Digital Library (NSDL) http://www.nsdl.org/

[27] OAI Implementers email thread, starting with http://www.openarchives.org/pipermail/oai-implementers/2003-August/000945.html

[28] Society of American Archivists. *Describing Archives: A Content Standard* (DACS). Society of American Archivists, Chicago, 2004.

[29] University of Illinois at Urbana-Champaign. Analysis of Dublin Core use by data provider, http://oai.grainger.uiuc.edu/Analysis_DCuse_Providers.doc

[30] University of Illinois at Urbana-Champaign. Distinct Metadata Schemas found by UIUC Experimental OAI Registry, http://gita.grainger.uiuc.edu/registry/ListSchemas.asp

[31] University of Illinois at Urbana-Champaign. OAI Metadata Harvesting Project, http://oai.grainger.uiuc.edu/

[32] University of Illinois at Urbana-Champaign. UIUC Digital Gateway to Cultural Heritage Material, Descriptions of Collections, http://oai.grainger.uiuc.edu/AboutCollections.htm

[33] Ward, Jewel. Unqualified Dublin Core usage in OAI-PMH data providers. *OCLC Systems & Services, 20, 1*, (2004), 40-47.