**Yen Bui, Ph.D. Student**
**Jung-ran Park, Ph.D., Assistant Professor**
**College of Information Science and Technology, Drexel University**
**3141 Chestnut Street, Philadelphia 19104, USA**

# An Assessment of Metadata Quality: A Case Study of the National Science Digital Library Metadata Repository

**Abstract:** The goal of this study is to assess the quality of current metadata records in the NSDL repository. For this, we harvested over one million Dublin Core metadata records submitted through November 2005 to the repository using the Open Archives Initiative Protocol (OAIP). This study reports on the preliminary results of the tabulations and assessment of metadata quality.

**Résumé :** Le but de cette étude est d'évaluer la qualité des enregistrements de métadonnées actuels à partir du référentiel d'entrepôt de la NSDL. À cette fin, nous avons recueilli plus d'un million d'enregistrements de métadonnées Dublin Core soumis au référentiel jusqu'en novembre 2005, en utilisant le Open Archives Initiative Protocol (OAIP). Cette étude présente les résultats préliminaires des statistiques et de l'évaluation de la qualité des métadonnées.

## 1. Introduction

The Metadata Repository is a major part of the NSDL. The NSDL supports the open source approach and the reuse of metadata. This means that the repository will accept metadata records contributed by various external organizations and is also open for the public to harvest its metadata records for their own use. The Open Archives Initiative Protocol (OAIP) is used for both submitting and harvesting. The Repository is composed of over 100 collections; each collection generally represents the contribution from an external organization. When submitting, the metadata records need to conform to the standard format used by the NSDL repository, which is the Enhanced Dublin Core. The metadata records are used in the search engine (Search and Discovery by UMASS) to return results for a search. When the entire text of a resource cannot be accessed freely due to licensing issues, the metadata is likely the main source of information about this resource. Since incoming records do not go through a standardization process, the metadata submitted by the different organizations can vary greatly in quality.

The goal of this study is to assess the quality of the current metadata records in the NSDL Repository in general and also the quality of each collection in the holding. For this, we harvested over one million Dublin Core metadata records submitted through November 2005 to the repository using the Open Archives Initiative Protocol (OAIP). The data harvested was loaded into an Excel database and exhaustive tabulations of all the Dublin Core metadata fields were performed. The data was also broken down in order to enable the evaluation of the metadata quality in each collection and subject matter. In other words, the richness or sparseness of the contents of records in each area of subject (e.g., Chemistry, Mathematics, etc.) was examined. The criteria of quality assessment are based on metadata uses in the following areas: frequency, consistency, completeness, accuracy and local additions of data providers (Park 2005). This study reports on the preliminary results of the tabulations and assessment of metadata quality.

## 2. Semantic Interoperability and Metadata Quality

The critical issues affecting metadata quality evaluation have been relatively unexplored (Moen et al. 2003, Barton, J., Currier, S., & J.M.N. Hey 2003). However, there is a growing awareness of the essential role of metadata quality assurance for successful resource access and sharing across distributed digital collections. Through examining learning objects and e-prints of communities of practice, Currier, et al. (2003) discuss the importance of quality assurance for metadata creation while pointing out the lack of formal investigation of the metadata creation processes. The 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], consistency of subject vocabularies) that result in adverse effects on resource discovery are examined. Moen, et al. (2003) also discuss problems of metadata quality through examination of 80 metadata records from the Government Information Locator Service (GILS) using a set of criteria such as completeness, accuracy and currency.

With the objective of enhancing semantic interoperability and requiring metadata quality assurance, Heery (2004) points out the increasingly rising number of local additions and variants to metadata standards. She emphasizes the necessity of building a mediation mechanism that can be sharable across libraries. Currier, et al. (2003) also point out the necessity of guidelines for metadata creation and quality control. Bruce, T.R. and D. I. Hillmann (2004) address challenges in approaching questions of quality by stating "quality standards and measures are sorely missed." In reaction to improving metadata quality, the study suggests examination of documentation practices and standards documents accompanying best practice guidelines and examples.

Challenges in enhancing access to digital collections have been reported by various studies (Heflin, J. and J. Hendler 2000, Doerr 2001, Park 2002, Vizine-Goetz, D., et al. 2004, Hegg and Knab 2003). Park 2002 presents an overview from a linguistic perspective of the characteristics of natural language, focusing on issues of synonymy and polysemy that pose particular challenges in semantic interoperability across heterogeneous knowledge organization schemes. Heflin, J. and J. Hendler (2000) report hindrances in integrating DTDs (Document Type Definition) by way of addressing the problems of polysemy and synonymy. The study stresses the critical importance of metadata creation by cataloging professionals and human indexers: "it is difficult for machines to make determinations of this nature, even if they have access to a complete automated dictionary and thesaurus."

McClelland, M., et al. (2002) discuss issues and challenges stemming from iLumina project experiences of mismatches of imported metadata from data providers, such as missing and incorrect data value: "metadata will be incomplete and contain errors, don't count on accuracy in data." Likewise, according to the analysis of Godby, et al. (2003) of 400 Dublin Core records, the incorrect and inconsistent metadata uses occur in the following way:

> Subject and Description both contain subject headings and free-text descriptions; Format and Type both contain names of media types such as photograph; and the data in the Language of the metadata record and the language of the content. Without extensive human-mediated correction, or training that promotes more consistent application of the Dublin Core element semantics when the records are created, even the goal of limited interoperability is compromised. (Underlined emphasis by the authors)

As the results of the above mentioned studies indicate, extensive research efforts focused on the identification of inaccurate, incomplete and inconsistent metadata creation and the factors behind such, together with adequate training of cataloging professionals, are critically needed for enhancing metadata quality and accordingly enhancing resource sharing and access across digital collections. Development of common data models that are sharable across libraries demands assessment of the current practice of metadata creation and locally defined metadata element sets; i.e., application profiles.

Park's recent studies (Park 2005 a & b) of 659 Dublin Core (DC) metadata records also points out critical metadata quality problems that inevitably hinder resource sharing and access across digital collections. The analysis of 659 metadata item records shows evidence of frequent inaccurate, incomplete and inconsistent metadata element uses. Some examples: the 'physical description' field is either inaccurately used as DC 'format' or 'description'; there is great confusion in employing the DC elements 'type' and 'format' and they are interchangeably used; the DC elements 'source' and 'relation' are inconsistently used; the DC element 'relation' is interchangeably used with cataloger-defined field names such as 'digital collection'. Some of the most frequently identified locally added field names are: 'contact information', 'ordering information', 'acquisition.'

Table 1 below represents the usage of Dublin Core metadata elements by three digital image collections. The total of 659 metadata item records was collected thus: from digital collection A (n/203 records), B (n/215 records) and C (n/241 records).

**Percentage of the Total Number of DC Metadata Elements Used by Three Collections (A, B, C)**

| DC Elements | A n/203 | % of total number of DC elements used n/3476 | B n/215 | % of total number of DC elements used n/2721 | C n/241 | % of total number of DC elements used n/2606 | Total n/659 | % of total usage of DC |
|---|---|---|---|---|---|---|---|---|
| Title | 203 | 5.8 | 217 | 8.0 | 241 | 9.2 | 661 | 100.3 |
| Creator | 196 | 5.6 | 148 | 5.4 | 30 | 1.2 | 374 | 56.8 |
| Subject | 580 | 16.7 | 416 | 15.3 | 448 | 17.2 | 1444 | 219.1 |
| Description | 203 | 5.8 | 210 | 7.7 | 263 | 10.1 | 676 | 102.6 |
| Publisher | 203 | 5.8 | 231 | 8.5 | 0 | 0.0 | 434 | 65.9 |
| Contributor | 289 | 8.3 | 100 | 3.7 | 19 | 0.7 | 408 | 61.9 |
| Date | 201 | 5.8 | 113 | 4.2 | 236 | 9.1 | 550 | 83.5 |
| Type | 0 | 0.0 | 150 | 5.5 | 235 | 9.0 | 385 | 58.4 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Format | 384 | 11.0 | 139 | 5.1 | 417 | 16.0 | 940 | 142.6 |
| Identifier | 265 | 7.6 | 107 | 3.9 | 7 | 0.3 | 379 | 57.5 |
| Source | 362 | 10.4 | 0 | 0.0 | 0 | 0.0 | 362 | 54.9 |
| Language | 63 | 1.8 | 0 | 0.0 | 5 | 0.2 | 68 | 10.3 |
| Relation | 121 | 3.5 | 98 | 3.6 | 4 | 0.2 | 223 | 33.8 |
| Coverage | 203 | 5.8 | 281 | 10.3 | 241 | 9.2 | 725 | 110.0 |
| Rights | 203 | 5.8 | 215 | 7.9 | 241 | 9.2 | 659 | 100.0 |
| Locally added elements | 0 | 0.0 | 296 | 10.9 | 219 | 8.4 | 515 | 78.1 |
| Total | 3476 | 100.00 | 2721 | 100.0 | 2606 | 100.0 | 8803 | 1335.8 |

**Table 1. Dublin Core Metadata Usage in Three Digital Image Collections**

Among the three collections, the following metadata elements are the most frequently employed, in descending order: subject, description, title, format and coverage. Usage of these five metadata elements constitutes over 50% of all the DC metadata elements.

With specific regard to the NSDL collections, many agree that there is currently no method for evaluating and integrating the results from the more than 100 submitting projects (Dushay, Hillmann, 2003 and Silva et al., 2004). Some efforts have been made to do some evaluation of the metadata, but they were either preliminary at the outset where there were few metadata records and were used more as a test of the OAIP and the collections or to explore techniques that could be used for evaluation purposes (Dushay, Hillmann, 2003). Shin (2004) proposes the use of a testbed by identifying the multiple classification dimensions that users need and by listing the key testbed components that are required. The latest metadata quality study was done by Zeng et al. (2004) in a NSF funded project. This study examined 186,237 records to measure for completeness, correctness, consistency and duplication. Only a research poster has been published; we could not find the complete project report.

The consensus seems to be that while the metadata repository plays a key role in the NSDL architecture and supports providers of services such as the NSDL Search function, it is impossible to impose detailed requirements for standards that every collection must follow. A compromise solution was the oai_dc format, which enhances the Dublin Core native metadata records submitted by the collections (Arms et al., 2003). When no native metadata is provided, very basic information is generated comprising the URL and whatever can be generated automatically from textual materials on the website. The trade-off between collection size and metadata quality is a typical challenge.

Recently, movement towards Library 2.0 (the importation of the Web 2.0 concepts into the library environment) has been evident. Fulker (2003) describes the effort to make the NSDL more user-participatory through a distributed model intended to engage and be more responsive to the needs of users. As a part of this strategy, the NSDL is working on defining and specifying the details to allow for direct user involvement in collection building and resource description. Along the same lines, the NSDL is also making use of the Fedora architecture, a tool to represent complex content, data, metadata and semantic relationships through an information network overlay (INO). The effort is a step to try to present contextualized information in the NSDL repository (Lagoze et al., 2005). These efforts are predicated on the move towards allowing users to participate in the indexing of metadata.

This paper evaluates the current state of the metadata repository at the NSDL. The repository is now at a much more mature stage and has built up a significant number of metadata records. We conducted an evaluation of all the metadata records in the NSDL holdings and utilized some of the methods and ideas suggested by previous metadata quality studies to explore ways to improve the quality of the repository.

## 3. Data and Research Methods

Metadata collections held by the NSDL were harvested using the OAI Metadata Harvesting Protocol 2.0 (OAI-PMH 2). The records harvested were submitted to the NSDL Repository from 9/8/03 through 1/23/05, as indicated by their datestamp values. The repository has no records prior to 9/8/03. A program written in the script language PHP was used to systematically harvest metadata records chronologically. Each time a request is made, approximately 200 to 250 records were returned along with a "resumption Token" value so that the request for the next batch of records could be made by including this value in the request. The task of extracting the resumption Token value and continuing the harvesting was handled by the PHP program. Records retrieved were saved in multiple XML files.

Only records with the oai_dc format were harvested. The repository also contains 3 other metadata formats: nsdl_dc (schema at http://ns.nsdl.org/schemas/nsdl_dc/nsdl_dc_v1.02.xsd), nsdl_all (schema at http://ns.nsdl.org/schemas/nsdl_all/nsdl_all_v1.02.xsd) , nsdl_links (schema at http://ns.nsdl.org/schemas/nsdl_links/nsdl_links_v1.00.xsd). These formats allow for additional elements not included in the oai_dc format. However, Dublin Core (DC) is mandated for lowest common denominator interoperability. Further research examining the records stored in these three other formats should yield more insight into the repository.

The records retrieved were actually collection records. A collection record has, besides the metadata elements, also top level information for the primary purpose of administration (See Figure 1 for an example of a collection record). A PHP program was used to parse the retrieved XML formatted records and to extract and export the metadata portions of the records to Excel files. In addition to the metadata elements, we also extracted the *NSDL ID*, the *datestamp*, and the *setSpec* elements. The *NSDL ID* element identifies each record uniquely; *datestamp* indicates the date when the record was submitted to the NSDL Repository; *setSpec* represents the collection set to which the record belongs. *SetSpec* is the ID given to each collection (i.e., organization) submitting records.

For each metadata element, if there was more than one value, they were all stringed together. For example, if a record's *Creator* element has "Smith" and "Brown", they would be extracted as "Smith/Brown". The only exception was the *Subject* element, for which we saved each value separately. For example, if a *Subject* element was listed as "Math", "Earth Science", "Computer", they would all be recorded separately. The reason for this lies in our intention to analyze the richness of the various subject areas in the repository.

The NSDL Repository has 111 record sets; the list was obtained using the ***ListSets*** command from the OAI protocol. For the time period from which we retrieved the records (9/8/03 to 11/23/05), six sets were not included because there were no records belonging to these records submitted during this period. The six sets were: *433248, 464485, 476579, 604607, 1468455,* and *alsos.*

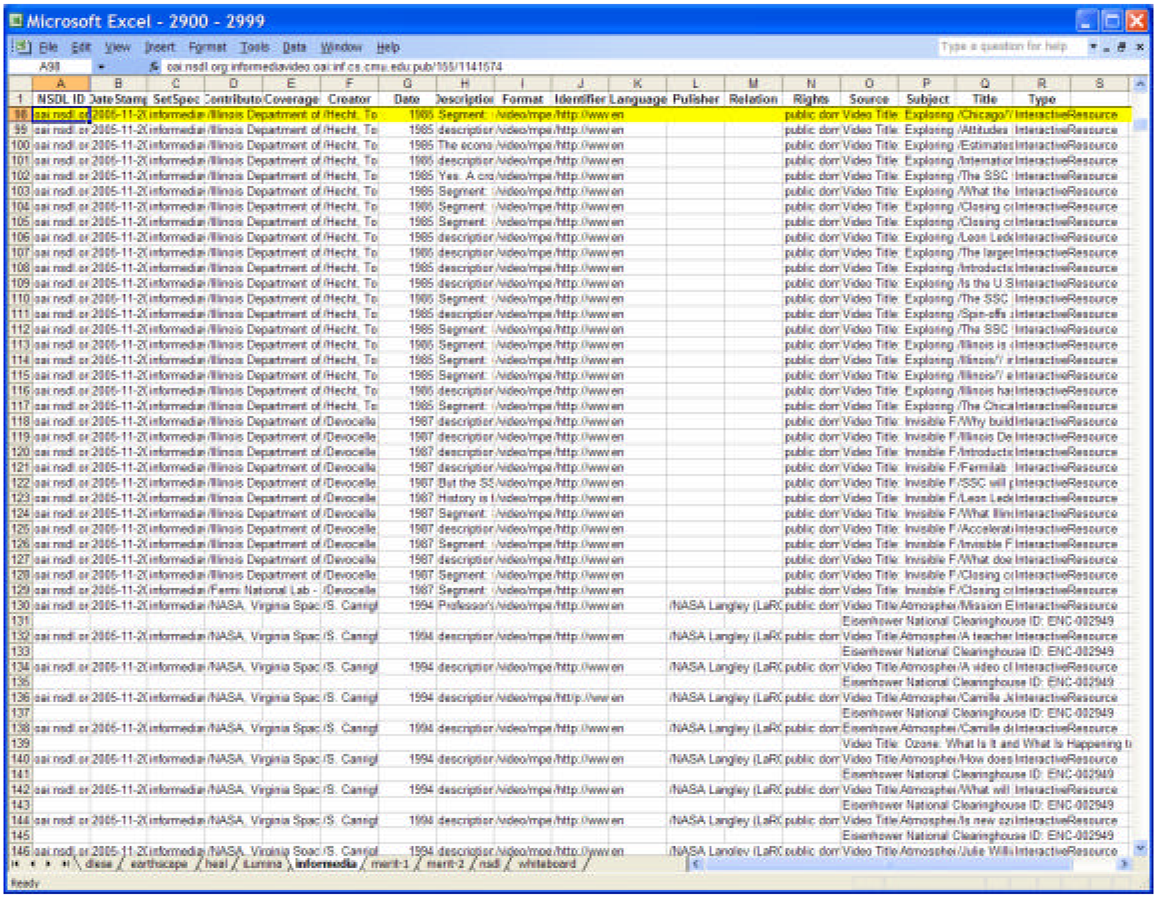
Once the records were parsed and exported to Excel, the records needed to be cleaned. For example the NSDL ID appears twice in a collection record. When exported to Excel, both would show up in the Excel worksheet; one needed to be deleted. The tasks in Excel were done using some simple Visual Basic modules embedded in macros. The records in the Excel files were then rearranged by record collections.

We chose to use Excel as a database for analysis because Excel spreadsheets offer ready-to-view visual inspection. We could read a record with all its elements across a page. Scrolling up and down was also helpful to identify any anomaly. One inconvenience with Excel is that each worksheet can only have up to 65,000 rows, but an Excel file can have as many worksheets as the system's capacity allows. We handled this limitation by storing the records in multiple worksheets in several files. This break-up method did not cause any negative effect on our analysis efforts. Another methodology consideration was to work with samples and statistics instead of the entire record database. We opted to work with all the records instead because once the PHP and Visual Basic programs were written, there was not much difference in working with samples or working with all the records. The computer automation took care of most of the work, with the advantage that we would have a more complete picture without the need to do statistical extrapolations.

```
<record>
-         <header>
-         <identifier>
oai:nsdl.org:informediavideo:oai:inf.cs.cmu.edu:pub/155/1141574
</identifier>
<datestamp>2005-12-21T03:31:10Z</datestamp>
<setSpec>informediavideo</setSpec>
</header>
-         <metadata>
-         <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>Chicago's Tunnel And Reservoir Plan (TARP)</dc:title>
<dc:creator>Hecht, Tom</dc:creator>
-         <dc:description>
Segment: #15 of 21, start 0:19:1.574, duration = 0:0:49.183
</dc:description>
-         <dc:description>
Another factor in favor of Illinois as the SSC site is the expertise that has been gained nearby
with Chicago's TARP project, the Tunnel and Reservoir Plan. More than seventy-two miles of
tunnels have been built hundreds of feet beneath the city by the Metropolitan Sanitary District
to help alleviate pollution and flooding. The project has been called the eighth wonder of civil
engineering. Illinois has become a world leader in tunnel projects because of TARP. And the
knowledge gained from that construction would certainly prove beneficial in building the
super collider, since it too would be located in bedrock hundreds of feet underground.
</dc:description>
<dc:contributor>Informedia at Carnegie Mellon University</dc:contributor>
<dc:contributor>Fermi National Lab - DOE</dc:contributor>
```

```
<dc:contributor>Illinois Department of Energy and Natural Resource</dc:contributor>
<dc:date>1985</dc:date>
<dc:type>InteractiveResource</dc:type>
<dc:format>video/mpeg</dc:format>
-         <dc:identifier>
http://www.infsearch.cs.cmu.edu/cgi-bin/stream100/FRM/FRM33/1141574.wmx
</dc:identifier>
<dc:source>Video Title: Exploring the Universe in Illinois</dc:source>
<dc:language>en</dc:language>
<dc:rights>public domain</dc:rights>
</oai_dc:dc>
</metadata>
-         <about>
-         <nsdl_about schemaVersion="1.00.000"
xsi:schemaLocation="http://ns.nsdl.org/nsdl_about_v1.00
http://ns.nsdl.org/schemas/nsdl_about/nsdl_about_v1.00.xsd">
-         <brand>
<iconURL>http://crs.nsdl.org/brands/informediavideo.gif</iconURL>
<title>Informedia</title>
<width>74</width>
<height>30</height>
</brand>
<category>item</category>
<firstCreated>2003-07-14T10:51:53Z</firstCreated>
-         <primaryIdentifier>
http://www.infsearch.cs.cmu.edu/cgi-bin/stream100/FRM/FRM33/1141574.wmx
</primaryIdentifier>
<link linkType="primaryCollection">oai:nsdl.org:nsdl.nsdl:00164</link>
</nsdl_about>
</about>
-         <about>
-         <provenance schemaVersion="1.00.000"
xsi:schemaLocation="http://ns.nsdl.org/provenance_about_v1.00
http://ns.nsdl.org/schemas/provenance_about/provenance_about_v1.00.xsd">
-         <originDescription altered="true" harvestDate="2005-12-21T05:11:58Z">
<harvestType>OAI2.0</harvestType>
<dataSource public="true">http://infsearch.cs.cmu.edu/cgi-bin/oai.pl</dataSource>
<identifier>oai:inf.cs.cmu.edu:pub/155/1141574</identifier>
<datestamp>1997-01-17T00:00:00Z</datestamp>
<metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
</originDescription>
</provenance>
</about>
</record>
```

**Figure 1. An Example of a Collection Record**



oai:nsdl.org:informedia/video.oai.inf.cs.cmu.edu:pub/155/1141674

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | NSDL ID | DateStamp | SetSpec | Contributo | Coverage | Creator | Date | Descriptio | Format | Identifier | Language | Pulisher | Relation | Rights | Source | Subject | Title | Type | |
| 98 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1986 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Chicago/'/ | InteractiveResource | |
| 99 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1986 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Attitudes | InteractiveResource | |
| 100 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | The econo | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Estimates | InteractiveResource | |
| 101 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1986 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Internatior | InteractiveResource | |
| 102 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | Yes. A cra | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /The SSC | InteractiveResource | |
| 103 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /What the | InteractiveResource | |
| 104 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1986 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Closing c | InteractiveResource | |
| 105 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Closing c | InteractiveResource | |
| 106 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Leon Lede | InteractiveResource | |
| 107 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1986 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /The larges | InteractiveResource | |
| 108 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Introductio | InteractiveResource | |
| 109 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Is the U.S | InteractiveResource | |
| 110 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1986 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /The SSC | InteractiveResource | |
| 111 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Spin-offs | InteractiveResource | |
| 112 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /The SSC | InteractiveResource | |
| 113 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1986 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Illinois is | InteractiveResource | |
| 114 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Illinois/'/ a | InteractiveResource | |
| 115 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Illinois/'/ a | InteractiveResource | |
| 116 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1986 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /Illinois ha | InteractiveResource | |
| 117 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Hecht, To | 1985 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Exploring | /The Chica | InteractiveResource | |
| 118 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /Why build | InteractiveResource | |
| 119 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /Illinois De | InteractiveResource | |
| 120 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /Introductio | InteractiveResource | |
| 121 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /Fermilab | InteractiveResource | |
| 122 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | But the SS | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /SSC will p | InteractiveResource | |
| 123 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | History is t | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /Leon Lede | InteractiveResource | |
| 124 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /What Illin | InteractiveResource | |
| 125 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /Accelerat | InteractiveResource | |
| 126 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /Invisible F | InteractiveResource | |
| 127 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | descriptior | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /What doe | InteractiveResource | |
| 128 | oai:nsdl.or | 2005-11-2 | informedia | /Illinois Department of | | /Devocelle | 1987 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /Closing c | InteractiveResource | |
| 129 | oai:nsdl.or | 2005-11-2 | informedia | /Fermi National Lab - | | /Devocelle | 1987 | Segment: | /video/mpe | /http://www | en | | | public dom | Video Title: | Invisible F | /Closing c | InteractiveResource | |
| 130 | oai:nsdl.or | 2005-11-2 | informedia | /NASA, Virginia Spac | | /S. Carrigl | 1994 | Professor's | /video/mpe | /http://www | en | /NASA Langley (LaRC | | public dom | Video Title | Atmosphe | /Mission E | InteractiveResource | |
| 131 | | | | | | | | | | | | | | | Eisenhower National Clearinghouse ID: ENC-002949 | | | | |
| 132 | oai:nsdl.or | 2005-11-2 | informedia | /NASA, Virginia Spac | | /S. Carrigl | 1994 | descriptior | /video/mpe | /http://www | en | /NASA Langley (LaRC | | public dom | Video Title | Atmosphe | /A teacher | InteractiveResource | |
| 133 | | | | | | | | | | | | | | | Eisenhower National Clearinghouse ID: ENC-002949 | | | | |
| 134 | oai:nsdl.or | 2005-11-2 | informedia | /NASA, Virginia Spac | | /S. Carrigl | 1994 | descriptior | /video/mpe | /http://www | en | /NASA Langley (LaRC | | public dom | Video Title | Atmosphe | /A video c | InteractiveResource | |
| 135 | | | | | | | | | | | | | | | Eisenhower National Clearinghouse ID: ENC-002949 | | | | |
| 136 | oai:nsdl.or | 2005-11-2 | informedia | /NASA, Virginia Spac | | /S. Carrigl | 1994 | descriptior | /video/mpe | /http://www | en | /NASA Langley (LaRC | | public dom | Video Title | Atmosphe | /Camille Je | InteractiveResource | |
| 137 | | | | | | | | | | | | | | | Eisenhower National Clearinghouse ID: ENC-002949 | | | | |
| 138 | oai:nsdl.or | 2005-11-2 | informedia | /NASA, Virginia Spac | | /S. Carrigl | 1994 | descriptior | /video/mpe | /http://www | en | /NASA Langley (LaRC | | public dom | Eisenhowe | Atmosphe | /Camille d | InteractiveResource | |
| 139 | | | | | | | | | | | | | | | Video Title: Ozone: What Is It and What Is Happening t | | | | |
| 140 | oai:nsdl.or | 2005-11-2 | informedia | /NASA, Virginia Spac | | /S. Carrigl | 1994 | descriptior | /video/mpe | /http://www | en | /NASA Langley (LaRC | | public dom | Video Title | Atmosphe | /How does | InteractiveResource | |
| 141 | | | | | | | | | | | | | | | Eisenhower National Clearinghouse ID: ENC-002949 | | | | |
| 142 | oai:nsdl.or | 2005-11-2 | informedia | /NASA, Virginia Spac | | /S. Carrigl | 1994 | descriptior | /video/mpe | /http://www | en | /NASA Langley (LaRC | | public dom | Video Title | Atmosphe | /What will | InteractiveResource | |
| 143 | | | | | | | | | | | | | | | Eisenhower National Clearinghouse ID: ENC-002949 | | | | |
| 144 | oai:nsdl.or | 2005-11-2 | informedia | /NASA, Virginia Spac | | /S. Carrigl | 1994 | descriptior | /video/mpe | /http://www | en | /NASA Langley (LaRC | | public dom | Video Title | Atmosphe | /Is new oz | InteractiveResource | |
| 145 | | | | | | | | | | | | | | | Eisenhower National Clearinghouse ID: ENC-002949 | | | | |
| 146 | oai:nsdl.or | 2005-11-2 | informedia | /NASA, Virginia Spac | | /S. Carrigl | 1994 | descriptior | /video/mpe | /http://www | en | /NASA Langley (LaRC | | public dom | Video Title | Atmosphe | /Julie Willi | InteractiveResource | |

**Figure 2. Screen Shot of an Example of the Excel file.**

In the above figure, the first 3 columns are NSDL ID, DateStamp, SetSpec. Columns D through R represent the metadata elements. Blank cells indicate that the records do not have data for this element. Note that the columns are sized so that all columns can fit on the screen; hence the contents of the cells are not shown in full. The corresponding record in Figure 1 is highlighted.

## 4. Discussion

We retrieved a total of 1,311,169 collection records. Removing the empty records (usually this is because a record has been deleted), the total non-empty records came to 1,040,034. We tabulated the population of each Dublin Core element for the entire retrieved set and found that overall the records are rather well populated, at least for those elements that are more critical. For searching and retrieval purposes, we believe that the six most important DC elements, not in any particular order, are *descriptor, subject, title, identifier, type* and *creator*. Search queries most often use terms that are embedded in these elements. For example, a searcher would often like to look for a document written by some author (creator), about some topic (subject, title, descriptor), of some type such as map or text (type), together with how to access this document (identifier).

The title and identifier elements come close to 100%, which is to be expected. However, we expected to see the *creator* element higher than 83.34%. This lower than expected percentage is due to a few collection sets where we suspect that the *creator* element is not so easily identified because the document is created by either an organization or group. These collection sets usually have their entire collection with none of the creator fields filled in. The *subject* element is also very critical but at the same time is not an element that can be automatically identified. We found that while most collection sets are diligent in filling in this element with only basic subjects, some collection sets are very thorough and populated this element with many subject terms. One collection that stands out is the Wolfram Research, Inc. where the subject element is very well populated. Another characteristic that we noticed is the fact that the provision of some DC element in a collection set is usually consistently very good or very sparse. That is, it is often a case of either a very high percentage or close to or equaling 0%.

Table 2 summarizing the results for all the retrieved records is presented below.

| **Total number of non-empty records in the NSDL repository: 1,040,034**<br>**(9/2003 – 11/2005)** | | | | |
|---|---|---|---|---|
| **Descriptor:**<br>867,423<br>(83.40%) | **Subject:**<br>805,432<br>(77.44%) | **Title:**<br>1,039,168<br>(99.92%) | **Identifier:**<br>1,033,271<br>(99.35%) | **Type:**<br>782,998<br>(75.29%) |
| **Creator:**<br>866,754<br>(83.34%) | **Date:**<br>894,823<br>(86.04%) | **Format:**<br>455,239<br>(43.77%) | **Language:**<br>395,126<br>(37.99%) | **Contributor:**<br>88,412<br>(8.50%) |
| **Coverage:**<br>18,972<br>(1.82%} | **Publisher:**<br>333,275<br>(32.04%) | **Relation:**<br>69,165<br>(6.65%) | **Rights:**<br>160,583<br>(15.44%) | **Source:**<br>152,094<br>(14.62%) |

**Table 2. Summary of Metadata Elements for All Records Retrieved**

Another characteristic that stands out is the distribution of records in the NSDL repository. The majority of the records come from only a few contributors. The top three contributors (Arsiv, Osti, Citadel) constitute 53.38% of the repository. The next two contributors (458940/Wofram and 491770) add another 15.99% for a total of 69.37%. In short, five out of the 105 collection sets makes up more than two-thirds of the repository. This is evident of a power-law distribution wherein a few members account for most of the repository holdings. Table 3 shows the detail of this core and scatter distribution.

| Collection Set | # of records | Percentage | Cumulative |
|---|---|---|---|
| **Arxiv** | 339,369 | 32.63% | Top 3 sets<br>53.38% |
| **Osti** | 113,629 | 10.93% | |
| **CITIDEL** | 102,084 | 9.82% | |
| **458940** | 87,616 | 8.42% | Top 5 sets<br>69.37% |
| **4917704** | 78,771 | 7.57% | |
| **498820** | 46,984 | 4.52% | Top 10 sets<br>82.57% |
| **Euclid** | 32,086 | 3.09% | |
| **BioMedCentral** | 20,820 | 2.00% | |
| **MathForum** | 18,757 | 1.80% | |
| **316881** | 18,614 | 1.79% | |
| **The rest (95 sets)** | 181,304 | 17.43 | 100% |

**Table 3. The Power-Law Distribution of the NSDL Repository Holdings**

A graph is also provided (Figure 3) to illustrate this core and scatter distribution. The spike at the left represents the core; the long tail represents the rest of the collections. The labels in the figure (names of the collections) are selected at certain intervals and are not consecutive; that is, they do not go from the collection with the highest rank to the second rank, and so on.

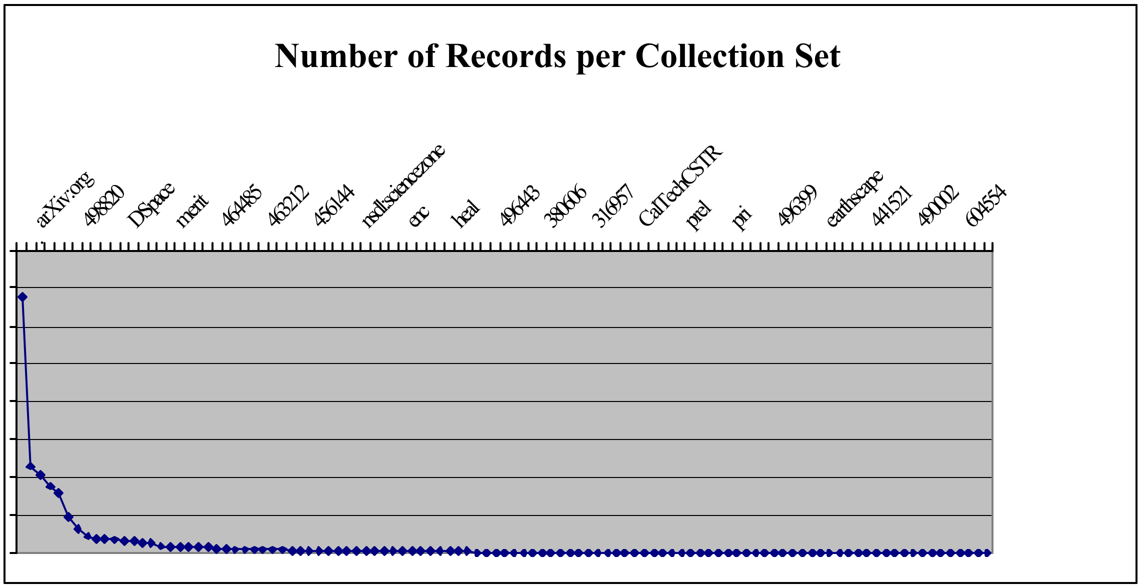


**Figure 3. Graph showing the power-law distribution of the records in the repository**

## 5. Future Study

The NSDL repository is made up of records from over 100 contributing organizations, but most of the records come from a few major contributors. Overall, the more important

DC elements are fairly well populated. Currently, we are still working on identifying the distribution of the subject areas as well as on a detailed analysis of how the metadata elements are used. Recommendations for enhancements and improvements in this area will depend on the results of this detailed analysis. One possibility that seems promising is for the NSDL to allow the professional catalogers or contributors to edit online the metadata of the records after they have been submitted. This would appear to reduce the cumbersome requirement that the contributor submit the records again.

**References**

Barton, J., Currier, S., and J.M.N. Hey. 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, T.R., and D.I. Hillmann. 2004. The continuum of metadata quality: defining, expressing, exploiting. In Diane Hillman & Elaine L. Westbrooks (eds.) *Metadata in Practice*. Chicago: American Library Association.

Lagoze, C., Krafft, D., Jesuroga, S., Cornwell, T., Cramer, E., and E. Shin. 2005. An information network overlay architecture for the NSDL (poster), presented at 2005 *JCDL*.

Dushay, N. and D. Hillmann. 2003. Analyzing metadata for effective use and re-use. *Dublin Core Conference: Supporting Communities of Discourse and Practice-Metadata Research & Applications*.

Fulker, D. 2003. Metadata strategies to address NSDL objectives. *Proceedings of the 5th Russian Conference on digital libraries RCDL2003*.

Godby, C. J., Smith, D., and E. Childress. 2003. Two paths to interoperable metadata. *DC-2003: Supporting Communities of Discourse and Practice—Metadata Research & Applications*. http://www.oclc.org/research/publications/archive/2003/godby-dc2003.pdf

Heery, Rachel. 2004. Metadata future: Steps towards semantic interoperability. In Diane Hillman & Elaine L. Westbrooks (eds.) *Metadata in Practice*. Chicago: American Library Association.

Heflin, J. and J. Hendler. 2000. Semantic interoperability on the Web. In *Proceedings of Extreme Markup Languages*. Graphic Communications. http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf

Hegg, K.J. and A.R. Knab. 2003. Using Dublin Core to facilitate cross-collection searches in an enterprise image repository. In *Dublin Core Conference: Supporting Communities of Discourse and Practice--Metadata Research & Applications*. http://dc2003.ischool.washington.edu/Archive-03/03hegg.pdf.

McClelland, M. et al. 2002. Challenges for service providers when importing metadata in digital libraries. In *D-Lib Magazine* 8(4). http://www.dlib.org/dlib/april02/mcclelland/04mcclelland.html

Moen, W.E., Steward, E.L., and C.R. McClure. 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.

Park, Jung-ran. 2002. Hindrances in semantic mapping among metadata schemes: A linguistic perspective. *Journal of Internet Cataloging*, Vol. 5(3): 59-79.

Park, Jung-ran. 2005a. Semantic interoperability across digital image collections: A pilot study on metadata mapping. *CAIS/ACSI 2005 Data, Information, and Knowledge in a Networked World*, Liwen Vaughan (ed.). 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 and Sang-joon Park. 2005b. Digital collection management software employed by libraries and museums: Evaluation of metadata semantic mapping functionality. Presented at the poster session at ALISE (Association/Library and

Information Science Education) annual conference, January 11-14, 2005 in Boston, Massachusetts.

Shin, P. 2004. Towards making the NSDL collection more accessible though a testbed. *Report from the Annual NSDL Meeting, November 14-17, 2004*. http://nsdl.comm.nsdl.org/meeting/session_docs/2004/2443_NSDL-Position-Peter-*Shin.doc.*

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

W. Arms, W., Dushay, N., Fulker, D., and C. Lagoze. 2003. A case study in metadata harvesting: The NSDL. *Library Hi Tech* 21: 228-237.

Zeng, Marcia Lei, Bhagirathi Subrahmanyam, and Gregory M. Shreve. 2004. Metadata quality study for the National Science Digital Library (NSDL) metadata repository. In *Digital Libraries: International Collaboration and Cross-Fertilization. 7th International Conference on Asian Digital Libraries,* ICADL 2004 Shanghai, China, December 2004.