A QUANTITATIVE ANALYSIS OF DUBLIN CORE METADATA ELEMENT SET (DCMES) USAGE IN DATA PROVIDERS REGISTERED WITH THE OPEN ARCHIVES INITIATIVE (OAI)

by
Jewel Hope Ward

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

November, 2002

Approved By:

____
Advisor:

Jewel Ward. A Quantitative Analysis of Dublin Core Metadata Element Set (DCMES) Usage in Data Providers Registered with the Open Archives Initiative (OAI). A Master's paper for the M.S. in I.S. degree. November, 2002. 68 pages. Advisor: Gregory B. Newby

This research describes an empirical study of how the Dublin Core Metadata Element Set (DCMES) is used by 100 Data Providers (DPs) registered with the Open Archives Initiative (OAI). The research was conducted to determine whether or not the DCMES is used to its full capabilities.

Eighty-two of 100 DPs have metadata records available for analysis. DCMES usage varies by type of DP. The average number of Dublin Core elements per record is eight, with an average of 91,785 Dublin Core elements used per DP. Five of the 15 elements of the DCMES are used 71% of the time. The results show the DCMES is not used to its fullest extent within DPs registered with OAI.

Headings:

Electronic data archives – Standards.

Virtual Library.

Metadata.

Dublin core.

Preprints.

Science and Technology – Databases.

Acknowledgements

I would like to thank the following people, without whose support and advice this paper would not have been possible. I would first like to thank Gregory B. Newby, for being a great employer, instructor, and an excellent and patient advisor on this paper. I would next like to thank Michael L. Nelson for suggesting the initial area that evolved into the actual research topic, as well as for his support throughout this process and his great patience in answering my many questions about digital libraries and OAI-PMH.

I would like to thank the following people whose support ensured I made it through the master's program. I would like to thank Jonathan Magid, Paul Jones, and the rest of the Ibiblio folks for providing much advice, laughter, sanity checks and excellent conversations. I would like to thank my former managers and coworkers from Interpath Communications, Inc., for their support and encouragement while I worked full-time and went to graduate school part-time. In particular, I would like to single out Greg Cox, i.e., "ratnix", with whom I also worked at MCI, for his excellent advice and ability to make technical "stuff" understandable to me. I also would like to thank Jennifer Manning, Beth Yongue, Karin Krchnak, Ann Owens, Liz Boyle, Jenn Dubernas, and the rest of the RTP crew for their support and advice throughout this process. The actives and alumni of the Fraternity of Delta Psi, St. Anthony Hall, Xi Chapter, also deserve multiple snaps.

Table of Contents

Introduction 7
Literature Review 12
  Metadata in the Electronic Environment 12
  DC Metadata Records and the OAI-PMH 15
  OAI-PMH: Foundations and Current Status 17
Method 23
Results 25
  Data Providers 25
  Metadata Records 26
  DC Metadata Elements 26
Discussion 29
Summary 33
References 34
Appendix A 37
  The DCMES, July 1995 37
  The DCMES v. 1.0 38
  The DCMES v. 1.1 42
Appendix B 47
  Example of an OAI-PMH Metadata Record 47
Appendix C 49
Appendix D 54
Appendix E 62
Appendix F 64

List of Tables

Table 1 - Summary of Chi-Square Values 28
Table 2 – Number of DPs by Domain Name 52
Table 3 - % of DC Elements by All DC Elements, All Records, and All DPs 63

List of Figures

Figure 1 - Number of DPs by Top-Level Domain 49
Figure 2 - % of Top-Level Domains Across 82 DPs 50
Figure 3 - % of Top-Level Domains Across 82 DPs (Adjusted for Duplicate Domain Names) 51
Figure 4 - Summary of Records, All DPs and by Type 54
Figure 5 - Number of DPs per Type of Archive 55
Figure 6 - % of DPs by Type of Archive, Not Adjusted for Duplicate Domain Names .. 56
Figure 7 - % of DPs by Type of Archive, Adjusted for Duplicate Domain Names 57
Figure 8 - % of Records Held by All DPs 58
Figure 9 - % of Records Held by Each STI DP 59
Figure 10 - % of Records Held by Each Humanities DP 60
Figure 11 - % of Records Held by Each Combo (STI-Humanities) DP 61
Figure 12 - Total DP Records vs. the Number of DC Elements Used by All DPs 62
Figure 13 - Comparison of Total DC Elements Used, Overall and by Type 64
Figure 14 - DC Metadata Element Usage by % Across All 82 DPs 65
Figure 15 - DC Metadata Element Usage by % Across STI DPs 66
Figure 16 - DC Metadata Element Usage by % Across Humanities DPs 67
Figure 17 - DC Metadata Element Usage by % Across Combo (STI-Humanities) DPs.. 68

Introduction

As the World Wide Web grows in size and the amount of available information continues to expand, researchers are making efforts to increase the relevance and precision of the results returned to searching Internet users. Currently, web crawlers index most web pages for search engines but only index an estimated 16% of the vast numbers of text and non-text digital objects available (Lawrence & Giles, 1999). Those objects that are indexed often cannot be found, because webmasters change, delete or move links. Users end up frustrated, because they cannot find the information they are seeking.

One method information providers use to solve the information indexing and retrieval problem is to create data about the digital objects and to make that data searchable. The set of descriptions about the resource itself is called metadata. "Metadata is structured data about data that supports the discovery, use, authentication, and administration of information objects" (Greenberg, 2001, p. 918). The resource to be discovered may be a text or non-text digital object, but the text that describes it is metadata. During the 1990s, researchers developed standards for metadata so that it could be used to index, store, and retrieve digital objects from electronic resources using World Wide Web standards and protocols in the hopes of improving resource discovery by improving resource description.

Some information communities have developed their own metadata schemas to support their particular information needs. Bioinformatics, for example, has many different metadata schemas of varying complexity that match the discipline's need for resource description at the physical, object, network, and general/ontological level (J. MacMullen, INLS 252 presentation, September 25, 2001). However, heterogeneous schemas are difficult to search, because the retrieval tool must "understand" each of the different schemas in order to return relevant results to a user.

One way for information retrieval tools to search heterogeneous metadata is for each group to provide a second set of metadata in a common format. That is, each group may keep their own rich metadata format for more precise resource discovery within the sub-group, but must "translate" that metadata into a common schema that provides for interoperability within a larger information community (i.e., "cross-domain searching"). One common standard in use is the Dublin Core Metadata Element Set (DCMES or DC), which offers 15 main elements "expressed as simple attribute-value pairs without any 'qualifiers' (such as encoding schemes, enumerated lists of values, or other processing clues)" ("Dublin Core Metadata Initiative", 2002, FAQ, para. 19).

By accessing a common metadata format, networked search engines of varying types and purposes can discover a resource based on a small set of information rather than the current model of searching and indexing the entire text and ignoring non-text objects. Searching text documents based on metadata rather than the entire text saves both time and system resources, as well as improving resource discovery, assuming that the digital object has been cataloged correctly. As well, non-text digitized objects can be retrieved easily. The gathering of metadata about a digital resource by an information retrieval tool is called metadata "harvesting".

One protocol that facilitates metadata harvesting is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), formally released in January 2001. OAI-PMH is "an application-independent interoperability framework that can be used by a variety of communities who are engaged in publishing content on the Web" (Van de Sompel & Lagoze, 2001, para. 1). The authors of the protocol use the DCMES as the common metadata format and the protocol is "designed as a simple, low-barrier way to achieve interoperability through metadata harvesting" (Warner, 2001, para. 3). The foundation of the protocol's framework lies in the definition of metadata providers (Data Providers, or DPs) and metadata harvesters (Service Providers, or SPs). A DP exposes metadata records for harvesting by an SP, which processes the records and provides a variety of value-added end-user services such as searching.

For a digital library to be OAI-PMH-compliant, it must expose its metadata using the DCMES and OAI-PMH. For example, a digital library (DL) community may use their own richer metadata set for intra-community search and retrieval, but must expose a second set of metadata with the DCMES. Alternately, the community is not required to have a second schema, and may elect only to use the DCMES. While a group of DLs could use the OAI-PMH with a metadata schema that is not the DCMES, they would not be OAI-PMH-compliant and could only interoperate with each other. That is, the OAI-PMH can be used with any metadata schema, but the DCMES must be used to be OAI-PMH-compliant and thus, interoperable across domains. Other metadata "formats currently in use include MARC (33), RFC-1807 (34), Open Languages Archives

Community Metadata Set (35), and the Electronic Theses and Dissertation Metadata Set (36)" (Nelson, 2001, p. 143).

Although the DCMES is an accepted standard that provides for interoperability between disparate information communities, it is not without its critics. Lagoze (2001) defines two groups within the Dublin Core Metadata Initiative (DCMI) community – minimalists and structuralists. Minimalists "saw the value of Dublin Core as an agreed set of broad categories usable for simple, unadorned attribute-value metadata. The latter, in contrast, saw Dublin Core as the foundation of a richer and monolithic descriptive language" (Lagoze, 2001, section 3, para. 1). The debate over whether or not the DCMES should remain either a simple schema that focuses primarily on resource discovery or should evolve into the basis of a more detailed description of resources continues, but Lagoze argues in favor of keeping the DCMES simple as the most effective method for ensuring resource discovery and DL interoperability.

Critics of the DCMES within the DL community would like to see the OAI-PMH use a metadata format that provides more detail about resources than the unqualified DCMES currently does. The criticism stems from the fact that when administrators of online collections decide to use the OAI-PMH and then convert their richer, domain-specific metadata to the DCMES to provide for interoperability and resource discovery across domains, some information about the individual resource may be lost, depending on the complexity of the original metadata schema.

Clarke (1997), who is one critic of the DCMES' simplicity, argues that simplicity of use and rich metadata are not mutually exclusive. He believes that if the lessons learned from data modeling are applied to the metadata interoperability problem, users will have a system with a complex underlying data structure that provides ease-of-use, rich metadata description and resource discovery across disparate information communities.

As a result of reading about this debate, I decided to analyze DCMES usage by registered OAI-PMH-compliant DPs. My hypothesis is that DCMES is not used to as full an extent as possible by the DPs. My research aims to answer the following questions.

- Which individual elements of the DCMES are used or not used?
- Which individual elements of the DCMES are used the most? Which are used the least?
- Are there different "types" of DPs? If so, does usage of individual elements of the DCMES vary by type?

The answers to these questions are applicable to the debate over whether or not the unqualified DCMES is an appropriate metadata schema for the OAI-PMH.

Literature Review

In the first part of this section I will discuss metadata as it is used in the electronic environment. A thorough discussion of the history of metadata and the many schemas available for the discovery and description of both digital and non-digital objects would be a research paper in and of itself. Therefore, for the purposes of this paper, I will discuss metadata within the scope of the OAI-PMH only.

In the second part of this section I will describe simple DC metadata records and OAI-PMH metadata records. In the third part of this section I will discuss the literature that forms the foundation of and describes the OAI-PMH.

Metadata in the Electronic Environment

Weibel (1995) provided one of the first major introductions to the basic concepts of metadata usage in the DL environment. He summarized the purpose of and the agreements from the first metadata conference held in Dublin, Ohio, in March 1995, which provided a framework for future work on the topic. Weibel provided a background to the development of the first DCMES and presented the need for a simple format that would encourage resource description of digital objects by non-information professionals, in order to provide for resource discovery of digital objects and interoperability amongst varying retrieval tools.

The July 1995 version of the DCMES contained 13 descriptors: Subject, Title, Author, Publisher, OtherAgent, Date, ObjectType, Form, Identifier, Relation, Source, Language, and Coverage. The architects of the element set focused primarily on describing intrinsic, not extrinsic data (i.e., the focus is on describing the object itself, not on the context in which the object is used). They designed the DCMES to be extensible and syntax independent with all elements optional, repeatable and modifiable. The DCMI released DCMES v 1.0 in September 1998 and v. 1.1 in July 1999. The DCMI defined 15 elements in v. 1.x: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights (for further information on the DCMES versions, see Appendix A).

The DCMES has been referred to as “simple” or “unqualified” DC, as it does not provide any qualifiers to increase the precision of the metadata (i.e., “DC.Creator”). “Qualified” DC has provided mechanisms for refining the metadata content of a resource (i.e., “DC.Creator.PersonalName”). The OAI-PMH has used unqualified DC.

Lagoze (1996) discussed the results of the second metadata workshop held in Warwick, U.K. The stated purpose of the workshop was to review the DCMES a year later, and to address any issues that had developed. Attendees agreed that a higher-level context for DCMES was needed, and developed a framework for metadata interoperability that organized metadata into aggregated logical collections (the container) for exchange from distinct metadata items (packages). The authors named it “The Warwick Framework”, and it has provided an architecture that has facilitated metadata interoperability while allowing information communities to use their own vocabularies and to opt in for either minimal or more structured metadata.

Metadata in the electronic environment has evolved into much more than a mechanism for resource discovery, description and identification, however. In addition to the three functions just mentioned, the attendees of the second DCMES workshop described six types of metadata that have been required in the digitized library work environment: terms and conditions, administrative data, content ratings, provenance, linkage or relationship data, and structural data.

Digital libraries developed from the melding of library science methodology with computer science systems. Metadata also evolved out of these two traditions. Burnett, Ng and Park (1999) described and compared the contributions each field has made to the evolution of metadata. The library science field contributed the bibliographic approach and the computer science field contributed data management. "Both approaches use metadata schemes to locate, identify, retrieve and manipulate information" (Burnett, Ng & Park, 1999).

The authors stated that librarians have been less concerned with the information itself than in managing the information object, while computer scientists have been using metadata to enhance use of the information held by the object. They argued that different contexts proscribe different functions for metadata use. Therefore, the type of metadata schema used may vary depending on the function of the electronic environment: data management, data access, or data analysis, and whether or not the implementers and users of a particular schema are more interested in the properties of the object (the intrinsic information) or the context in which the information object is used (the extrinsic information). The authors defined one other division by function: that between system level and end-user. Burnett, Ng and Park defined system level metadata as the function of metadata that facilitates interoperability and resource discovery, while end-user level metadata aids in determining what resources are available, whether or not the object fits the information need, how the resource can be acquired, and how to access the resource.

The authors compared the elements of the DCMES, URC (Uniform Resource Characteristics), Semantic header, USMARC (United States MAchine-Readable Cataloging), IAFA (Internet Anonymous FTP Archives) Templates, and TEI (Text Encoding Initiative) header in order to determine functions common across the major metadata schemas. They noted that all formats use the same three elements: title, author, and identifier, and that five of the six include place and date. All formats contained intrinsic elements, but both the DCMES and URC lacked extrinsic elements. As one result of their study, Burnett, Ng & Park (1999) combined the functional use of metadata by both traditions to provide an integrated definition of metadata: "metadata is data about data that characterizes source data, describes their relationships and supports the discovery and effective use of source data."

DC Metadata Records and the OAI-PMH

The DCMI has defined a metadata record as "some structured metadata about a *resource* [the digital or non-digital object being described] comprising one or more *properties* [or, individual DC metadata elements] and their associated *values* [a literal string]" ("Dublin Core Metadata Initiative", 2002). To form a record, the DCMES has been encoded with Extensible Markup Language (XML). In other words, the DCMES has formed the building block for resource description, while XML has provided the framework for resource discovery across multiple networked systems.

The W3C ISO standard for SGML (Standard Generalized Markup Language) has defined it as a system for creating a document markup language or tag set. W3C developed XML from SGML, and it "is a pared-down version of SGML, designed especially for Web documents. It allows designers to create their own customized tags, enabling the definition, transmission, validation, and interpretation of data between applications and between organizations” (“Webopedia”, 2002). Both SGML and XML have provided a standard for tagging elements, but neither has provided rules for formatting the elements themselves.

A simple example of a DCMES record is:

```
<?xml version="1.0"?>
<metadata
  xmlns="http://foo.net/bar/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://foo.net/bar/ http://foo.net/bar/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
        <dc:title>
        A Quantitative Analysis of Dublin Core Metadata Element Set (DCMES)
        Usage in Data Providers Registered with the Open Archives Initiative
        (OAI)
        </dc:title>
        <dc:creator>
        Jewel Hope Ward
        </dc:creator>
        <dc:description>
        A Master’s paper submitted to the faculty of the School of Information
        and Library Science of the University of North Carolina at Chapel Hill in
        partial fulfillment of the requirements for the degree of Master of Science
        in Information Science.
        </dc:description>
        <dc:publisher>
        The University of North Carolina at Chapel Hill
        </dc:publisher>
</metadata>
```

The Resource Description Framework (RDF) has been one application of XML with the DCMES that has provided interoperability between applications based on the Warwick Framework. The W3C developed RDF as an infrastructure that “enables the encoding, exchange and reuse of structured metadata [by a variety of] disparate information communities” (Miller, 2001, Abstract section, para. 1). The creators of the OAI-PMH used common encoding standards such as XML and RDF to provide interoperability between SPs and DPs.

Lagoze, Van de Sompel, Nelson and Warner (2002) defined an OAI-PMH record and the process for an SP to obtain it from the DP as "metadata expressed in a single format. A record is returned in an XML-encoded byte stream in response to an OAI-PMH request for metadata from an item" (2.5 Record section, para.1). The authors have required that an OAI-PMH record contain a unique identifier, a metadata format description, and a datestamp as part of two of three main sections: header, metadata, and about. They defined header as containing the unique identifier, the datestamp, and the `setSpec` element, with an optional status attribute for deleted records; metadata as holding a single metadata format that describes the object; and about as an optional section for information about the metadata in the record. I have provided an example of an OAI-PMH record in Appendix B.

OAI-PMH: Foundations and Current Status

The OAI-PMH has its technical roots in the Universal Preprint Service (UPS) and the Dienst protocol. In turn, Dienst is based on the Kahn-Wilensky Framework (KWF). Thus, KWF led to Dienst, Dienst to UPS, and UPS to OAI-PMH.

Kahn and Wilensky (1995) defined a high level framework of the basic entities and structure of a DL service: digital objects, handles, metadata, repositories, handle generators, originators, users, naming authorities, and a repository access control. A drawback to the framework has been that the authors did not provide a working prototype of the specified design, although it has succeeded in stimulating other researchers to develop working models based on it.

Davis and Lagoze (2000) built on KWF and developed the Dienst protocol, an open architecture for federated distributed document libraries. The authors defined an architecture with four core services: repository, index, a user interface, and a collection service. The Networked Computer Science Technical Reference Library (NCSTRL) has used the Dienst model to create a working, searchable, interoperable network of digital libraries for computer science reports. Overall, the architecture has proved successful, but distributed searching does not scale well and NCSTRL eventually moved to a more centralized search service. Another drawback to Dienst has been that it takes a lot of effort to install and maintain it. This high-barrier aspect of the protocol has prevented its wider adoption and continued maintenance within the digital archive community.

The purpose of the UPS convention, held in Santa Fe, NM in October 1999, was to identify the key issues preventing the implementation of services such as linking and searching across large, diverse, distributed e-print archives. Participants have wanted to transform scholarly publishing by taking it out of the hands of the publishers and giving it back to the authors in the form of e-print archives, and they have believed that interoperability is the first step in achieving this transformation. The convention attendees also hoped to reach a consensus on the technical and organizational solutions needed to overcome the identified issues that prevented true interoperability amongst dissimilar DLs. Van de Sompel, et al. (2000) wanted to prove to the convention participants the feasibility of cross-archive value-added services built on data pulled from dissimilar e-print archives. The authors prepared the UPS Prototype as the demonstration model.

The authors of the UPS Prototype refined Dienst in order to provide access to the content of multiple e-print archives by searching the archives' metadata as a way to overcome the limits of distributed searching. The researchers gathered metadata from dissimilar cross-discipline archives, converted the metadata to a common format, and then ran services such as searching, buckets and linking on that metadata. The authors extended Dienst in the form of NCSTRL+ (Nelson, Maly, Shen & Zubair, 1998) and the software functioned as the search facility service on the metadata.

In spite of several difficulties, most of which involved extracting heterogeneous metadata from dissimilar archives, the authors of the UPS Prototype successfully demonstrated the feasibility of cross-archive end-user services based on a metadata harvester/data provider model. Consequently, they submitted several recommendations to the UPS meeting attendees. Van de Sompel, et al. (2000) proposed that:

- a distinction be made between the harvesting service and the metadata provider;
- distributed searching be set aside in lieu of searching harvested metadata;
- a framework be developed to outline the terms and conditions, technical characteristics, and administrative characteristics of a metadata harvester/data provider model;
- a universal, uniform, unique identifier namespace for e-prints be created; and,
- DPs adopt a common metadata format.

Attendees of the Santa Fe convention developed a consensus to adopt a UPS Prototype-based metadata harvesting model as a workable technical and organizational framework for delivering digital archive content and services to end users. The harvesting model allowed “E-print (content) providers to expose their metadata via an open interface, with the intent that this metadata be used as the basis for value-added service development” (Lagoze & Van de Sompel, 2001, p. 55).

Meeting participants also agreed upon the basic definitions, concepts, technical components and organizational aspects of interoperable e-print archives (Van de Sompel & Lagoze, 2000). These agreements became known as the “Santa Fe Convention”. Shortly after the meeting in Santa Fe, members of UPS changed the name to the Open Archives Initiative (OAI) to refer to the overall group of people and its philosophy, and named the protocol itself, the “OAI-PMH”. They also expanded the cross-archive interoperability framework demonstrated by the UPS Prototype beyond the e-print community to any academic and government organizations involved in scholarly publishing.

Van de Sompel and Lagoze (2001), along with the members of the OAI Technical Committee, released version 1.0 of the OAI-PMH in January 2001, after a period of beta testing. As a result of the consensus built at the Santa Fe Convention, in addition to the change in focus from e-prints to “document like objects”, two other major changes occurred in the version 1.0 protocol: the authors dropped the Dienst verb set in lieu of an OAI-PMH-specific six verb set, and dropped the Open Archives Metadata Set (OAMS) in lieu of unqualified Dublin Core. The authors did not plan to make changes to the version 1.0 protocol for a period of 12 to 18 months after the initial release, but they adopted the then newly released World Wide Web Consortium (W3C) XML standards, and upgraded the OAI-PMH in July 2001. The authors considered the 1.1 version of the protocol to be experimental, and the 12 to 18 month observation phase provided a static time period during which problems with the protocol were identified and evaluated without forcing early adopters to take on the cost of multiple rewrites (Warner, 2001).

In June 2002, Lagoze, Van de Sompel, Nelson and Warner (2002), along with the members of the new OAI Technical Committee, released version 2.0 of the OAI-PMH. The authors considered this to be a stable, non-experimental version. Changes from 1.1 to 2.0 included referring to "resources" rather than "document like objects". Other improvements included: the use of a single XML schema, DCMI XML; removing ambiguities in usage and the definitions of terms from the written protocol; more expressive options for the OAI-PMH six verb set; and a cleaner separation of roles and responsibilities between http and the OAI-PMH (Nelson, Van de Sompel & Warner, 2002). One drawback to the new release for current implementers of the protocol is that 2.0 was not designed to be backwards compatible with version 1.1.

In summary, the Kahn-Wilensky Framework has provided a model that defines the basic entities and structure of a digital library (DL) service. Lagoze and Davis refined KWF to create the Dienst protocol, and applied the latter to create NCSTRL. In turn, Dienst led to a new framework for digital library services, UPS. The authors of the UPS Prototype demonstrated the feasibility of cross-archive value-added services at the Santa Fe Convention in October 1999, but the researchers developed the service and the prototype with only the e-print community in mind. After the convention, the members of UPS changed the name to the OAI-PMH and extended it to a wider range of scholarly digital archives. The OAI-PMH has since expanded the interoperability concepts agreed upon at the Santa Fe Convention. It is no longer an experiment, but a working protocol.

The OAI-PMH “has emerged as a practical foundation for digital library interoperability” (Van de Sompel & Lagoze, 2002, p. 144).

Method

The 77 DPs registered on the OAI web site as of 8 May 2002 provided the initial metadata harvest. On 28 July, I added an additional 23 DPs that registered between 9 May and 28 July to the initial harvest group, for a combined total of 100 DPs. I did not include any DPs that registered after 28 July in this analysis. I harvested from v. 1.1 repositories, with the exception of arXiv. I harvested arXiv's metadata in November 2002 from an aggregator, Celestial, which used version 2.0 of the OAI-PMH. Otherwise, I stopped harvesting metadata as of 12 October 2002.

I used the Perl OAI Harvester v. 1.1 (Suleman & Fox, 2001) as the SP with which to harvest the metadata. After the software performed the initial harvest, I set up cron jobs to initialize the software to harvest each DP on a weekly basis. The harvester software ran on a Dell Precision 530, with dual 1.7 Ghz Xeon processors, 2 GB RAM, and a U160 SCSI with 10,000 RPM drives.

I could not use the Perl OAI Harvester to harvest metadata from arXiv, although I tried many times. The Perl OAI Harvester cannot harvest metadata from arXiv. This problem is now a known bug of the software. Between May and October 2002, instead of using software to harvest arXiv, I manually harvested a partial set weekly via a WWW browser. In November 2002 I harvested a full set of arXiv's metadata from the aggregator.

In order to analyze usage of the DCMES, I wrote a Perl program to count the number of records harvested from each DP, and parsed the individual elements from the content of each record in order to count the number of times a record contained each of the 15 elements. To determine the number of active and deleted records, since each file harvested contains one metadata record, I approximated the number of deleted records by counting the total number of files harvested from each DP and subtracting the total count of <dc></dc> or <oai_dc></oai_dc> tags. I did not use the `setSpec` option in the header to find the deleted records, as the harvester software dropped the header from all files.

I determined the type of repository by using a web browser to issue an `Identify` request combined with the `baseURL` of a DP registered on the OAI web site. If that did not provide sufficient information, I reviewed the records held by a DP (`baseURL` + `?verb=ListRecords&metadataPrefix=oai_dc`). If that still did not provide enough information, I pushed back to the domain name and explored the library and university web sites to determine the category of a DP's holdings. To determine the type of holdings of a non-English language repository, I either consulted someone who could read the language or made a "best guess".

Results

Data Providers

I was unable to harvest 12 of the 100 DPs registered as of 28 July 2002, even after several retries over a 2 ½ month period. Of the remaining 88 DPs, 6 provided no records at all, and an additional 11 of the 88 could only be harvested for part of the time period between 8 May and 12 October. Therefore, I harvested records from 82 of the 100 DPs, but only 76 of the 82 DPs could be harvested consistently.

The `baseURL`s for the 82 DPs represented 16 different top-level domains. .edu was the most common, followed by .org, with .com and .de tied for third place. Those four top-level domains represented more than two-thirds of the top-level domains used, even after I adjusted for duplicate domain names. The language-archives.org domain name was present in 10 of the 82 `baseURL`s registered. Ten domain names represented 40% of the names used by each of the 82 individually registered repositories. I consolidated registered repositories by domain name, which reduced the number of repositories to 59, but this did not affect which DPs were the largest repositories by number of records. (Please see Appendix C for the figures and table presenting the details.)

Based on the information I gathered from reviewing the metadata records and/or web sites of the 82 DPs, I divided them into three broad categories: STI (Scientific and Technical Information), Humanities, and Combo (both STI and Humanities).

Metadata Records

The total number of records harvested from the 82 DPs was 910,919, with active records making up 99.9% of the total. The average number of records per DP was 11,109. When I divided the number of records as a percentage among the three types of repositories, Humanities repositories held 43% of the records, STI repositories 31%, and Combo (STI-Humanities), 26%. By number of repositories, 33 DPs fell into the Humanities category, 27 into STI, and 22 into Combo (STI-Humanities), not adjusting for duplicate domain names. The trend across all types was for no more than three DPs to provide 89% or more of the records within each type. The eight largest of the 82 DPs I analyzed provided 92% of the total number of records across all DPs, regardless of type. At the other extreme, the bottom 10% provided less than 1% of all records. (Please see Appendix D for the figures presenting the details.)

DC Metadata Elements

The total number of DC elements across all DPs was 7,526,331. Thus, as a ratio against the total number of records, there was an average of eight DC elements used per record, with an average of 91,785 DC elements used per DP. The top five DC elements used, taken as a proportion of either the total number of DC elements or the total number of records were, from most- to least-used: creator, identifier, title, date, and type. The top five DC elements accounted for 71% of all element usage. The least-used five elements were, from most- to least-used – language, format, relation, contributor and source – and accounted for 6% of usage.

When the 15 DC elements were cross tabulated as a percentage within each DP, the top five elements used remained the same, but the order, from most- to least-used, changed to: title, creator, date, identifier, and type. When calculated as a percentage within each DP, almost 99% of DPs used the title tag. Calculating the least-used DC elements as a percentage within a DP changed two of the five least-used elements, compared to the previous order. They were, in order of most- to least-used: rights, contributor, source, coverage, and relation. (Please see Appendix E for the figures and table presenting the details.)

STI, Humanities, and Combo (STI-Humanities) DPs each used creator, title, identifier, and type as their top four most-used elements. STI and Combo (STI-Humanities) DPs both used date to round out the top five, matching the trend across all DPs, while Humanities DPs used rights as the fifth most-used element. The creator and identifier elements accounted for more than half of the DC elements used by STI DPs, while in Humanities DPs, title, identifier, creator and type accounted for 48% of element usage. The creator, identifier and date elements accounted for almost 60% of the total number of DC tags used by Combo (STI-Humanities) DPs. The results showed that just over half of the 82 DPs used only the creator and identifier elements for approximately half of their overall usage. (Please see Appendix F for the figures presenting the details.)

I reviewed my record numbers and DP data against the information posted at the "Celestial Hall of Shame" (http://celestial.eprints.org/cgi-bin/status) on November 24th 2002, and found that my data set was within range, taking into account that I stopped adding new DPs to harvest at the end of July, and that I stopped harvesting metadata records in mid-October.

I ran tests for Frequencies and Statistics using SPSS, but the results did not show anything particularly significant. As shown in Table 1, the chi-square values test showed no significant difference in the observed versus the expected results for the eight most-used DC elements within a DP, but p < .05 for the 7 least-used DC elements. I ran three Independent Samples t-tests, and paired the 3 types of DPs as three sets against each of the 15 DC elements. The results did not produce p < .05 in any of the three tests.

Table 1 - Summary of Chi-Square Values

DC Element	Pearson Chi-Square Test Results			
	Value	df	Asymp. Sig (2-sided)	p≤ 0.05?
title	1.503	2	0.472	N
creator	0.185	2	0.912	N
date	5.243	2	0.073	N
identifier	0.066	2	0.968	N
type	2.336	2	0.311	N
subject	3.677	2	0.159	N
description	1.518	2	0.468	N
language	3.841	2	0.147	N
publisher	9.441	2	0.009	Y
format	15.363	2	0.000	Y
rights	10.581	2	0.005	Y
contributor	9.609	2	0.008	Y
source	6.970	2	0.031	Y
coverage	10.007	2	0.007	Y
relation	14.124	2	0.001	Y

I discussed the Crosstabs results earlier in this chapter; a summary of the full results may be viewed in Appendix E, Table 3.

Discussion

The results support my hypothesis – DC, as simple as it is, is not used to the fullest extent possible. I did not expect every author or cataloguer who submits metadata to use each element at least once, but neither did I expect that two elements out of fifteen would make up half the element usage in over half of the DPs.

One area for future work would be an examination of how and why authors and cataloguers choose to use or not use certain elements. Although some DPs are author self-archiving, others contain metadata prepared by information professionals. I think a user analysis would be useful for two reasons. The first is to determine further the appropriateness of the current incarnation of DC as the foundation for cross-domain resource discovery for the OAI-PMH. The second is to provide administrators of DPs with a basis for building better tools with which authors and cataloguers can describe the information object(s). For example, is source the least-used element because it is not relevant to most resources, or is it because the box on the GUI that an author uses to enter his or her metadata requires the user to scroll down?

The trend I see across all of the results is for a very small number, whether it is DPs or DC elements, to dominate. Out of 82 DPs, five (citebase, arXiv, dlpscoll, lcoa1, and uiLib) hold 85% of the metadata records. Users have a choice of 15 DC elements, but five (creator, identifier, title, date, and type) are used 71% of the time. Of the 16 top-level domains represented, three top-level domains are used by 64% of DPs.

The fact that approximately a quarter of the DPs could not be harvested either with the Perl OAI Harvester or via a WWW browser, could not be harvested regularly, or just plain did not provide any records would be unacceptable if version 1.1 of the OAI-PMH was not the experimental phase. My criticism is not of the protocol itself, but aimed more towards the administrators of DPs that register a repository without validating it or maintaining it. I realize that using the OAI-PMH is voluntary, but as the protocol matures and establishes a production environment, some form of quality control should be considered and implemented at the DP level.

I was not always able to completely determine why I could not harvest the twelve repositories. Eleven of the twelve returned one of the following error messages: 301 (moved permanently), 302 (moved temporarily), 404 (not found), 500 (server error), and 502 (bad gateway). Again, I will criticize the administrators of the DPs, several of whom registered a DP and then either failed to update the `baseURL` he or she registered with OAI or to validate his or her repository.

I found the predominance of Humanities DPs to STI DPs interesting, given that OAI came out of the e-print community. I expected STI repositories to dominate. The number of "Combination" DPs also surprises me, as I expected DPs to be predominately STI or Humanities, with a handful that would qualify as both. The high number of Humanities and "Combination" DPs supports the belief in the information community that OAI has long since extended beyond its e-print roots.

The results of the study of DPs by domain are useful in a general sense, but are limited in relevance since I did not trace the country of origin for the top-level domains .gov, .com., .edu, and .org. The results from aggregating DPs based on domain names has limited usefulness, as from a system standpoint it does not matter if a repository is one-of-one or is one-of-many. Organizing the 10 repositories that use the language-archives.org domain name under one umbrella URL would not make harvesting the repositories easier or faster, the searching more accurate or improve the resource description.

Studying the domain names, does, however, show how administrators are using the OAI-PMH. The arXiv administrator, for example, has one URL and distinguishes among the DP's different collections by organizing them into groups that can be discovered by issuing the `ListSets` request defined in the protocol. Administrators of other organizations choose to list their collections as individual repositories. These different patterns of usage may or may not have implications for the scalability of the protocol.

Another interesting figure to me is the high number of active records – 99.9% – out of the total number of records, even taking into consideration the youthfulness of the OAI community. I think that a figure in the 98-99% range would be more accurate. There could be many reasons for this figure, ranging from the deleted records policies of individual DPs to my Perl coding, so I would be the first to attest to the limits of both claiming and not claiming that 99.9% of the records held by OAI-registered DPs are actives, not deletes.

Burnett, Ng, & Park (1999) studied six metadata standards and found that title, author, and identifier are common to all the schemes, and that two others – place and date – are common to five of the six schemes. Each of the five elements common to the six schemes is a metadata type for identification and resource discovery. The top five elements used in the OAI-PMH DPs are: title, creator, date, identifier, and type, whether viewed as a proportion of total elements, total records, or total DPs. Thus, the results correlate with the results of previous studies of metadata elements, but support the results at the system level, rather than the schema level.

Summary

This research describes an empirical study of how the Dublin Core Metadata Element Set (DCMES) is used by 100 Data Providers (DPs) registered with the Open Archives Initiative (OAI). The research was conducted to determine whether or not the DCMES is used to its full capabilities.

Eighty-two of 100 DPs have metadata records available for analysis. DCMES usage varies by type of DP. The average number of Dublin Core elements per record is eight, with an average of 91,785 Dublin Core elements used per DP. Five of the 15 elements of the DCMES are used 71% of the time. The results show the DCMES is not used to its fullest extent within DPs registered with OAI.

References

Burnett, K., Ng, K., & Park, S. (1999). A Comparison of the Two Traditions of Metadata Development. Journal of the American Society for Information Science, 50(13), 1209-1217.

Davis, J. R., & Lagoze, C. (2000). NCSTRL: Design and Deployment of a Globally Distributed Digital Library. Journal of the American Society for Information Science, 51(3), 273-280.

Clarke, R. (1997). Beyond the Dublin Core: Rich Meta-Data and Convenience-of-Use Are Compatible After All. Retrieved November 24, 2002, from the Australian National University, Faculty of Engineering and Information Technology Web site: http://www.anu.edu.au/people/Roger.Clarke/II/DublinCore.html.

Dublin Core Metadata Initiative. (1998). Dublin Core Metadata Element Set, Version 1.0: Reference Description. Retrieved November 24, 2002, from http://www.dublincore.org/documents/1998/09/dces/.

Dublin Core Metadata Initiative. (1999). Dublin Core Metadata Element Set, Version 1.1: Reference Description. Retrieved November 24, 2002, from http://www.dublincore.org/documents/dces/.

Dublin Core Metadata Initiative. (2002). Guidelines for implementing Dublin Core in XML. Retrieved November 24, 2002, from http://dublincore.org/documents/2002/09/09/dc-xml-guidelines/.

Dublin Core Metadata Initiative. (n.d.). What is the difference between "Simple" ("unqualified") and "Qualified" Dublin Core? Retrieved September 21, 2002, from http://www.dublincore.org/resources/faq/#whatisthedifference.

Greenberg, J. (2001). A Quantitative Categorical Analysis of Metadata Elements in Image-Applicable Metadata Schemas. Journal of the American Society for Information Science, 52(11), 917-924.

Kahn, R., & Wilensky, R. (1995). A Framework for Distributed Digital Object Services. Retrieved February 2, 2001, from http://www.cnri.reston.va.us/home/cstr/arch/k-w.html.

Lagoze, C. (1996, July). The Warwick Framework A Container Architecture for Diverse Sets of Metadata. D-Lib Magazine, 2(7). Retrieved May 4, 2001 from http://www.dlib.org/dlib/july96/lagoze/07lagoze.html.

Lagoze, C. (2001, January). Keeping Dublin Core Simple: Cross-Domain Discovery or Resource Description? D-Lib Magazine, 7(1). Retrieved January 27, 2001 from http://www.dlib.org/dlib/january01/lagoze/01lagoze.html.

Lagoze, C., & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries, 54-62.

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-6-14, document version 2002/09/13T11:34:00Z. Retrieved July 12, 2002, available from http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html.

Lawrence, S., Giles, C. (1999). Accessibility of Information on the Web. Nature, 400, 107-109.

Miller, E. (1998, May). An Introduction to the Resource Description Framework. D-Lib Magazine, 4(5). Retrieved October 2, 2001, from http://www.dlib.org/dlib/may98/miller/05miller.html.

Nelson, M. (2001). Better Interoperability through the Open Archives Initiative. The New Review of Information Networking, 7, 133-146.

Nelson, M., Maly, K., Shen, S., & Zubair, M. (1998). NCSTRL+: Adding Multi-Discipline and Multi-Genre Support to the Dienst Protocol Using Clusters and Buckets. Proceedings of the Advances in Digital Libraries 1998, 128-136.

Nelson, M., Van de Sompel, H., & Warner, S. (2002, July). Advanced Overview of Version 2.0 of the Open Archives Initiative Protocol for Metadata Harvesting. Tutorial 5 presented at the Second ACM/IEEE-CS Joint Conference on Digital Libraries, Portland, OR.

Suleman, H., & Fox, E. (2001, December). A Framework for Building Open Digital Libraries. D-Lib Magazine, 7(12). Retrieved November 26, 2002, from http://www.dlib.org/dlib/december01/suleman/12suleman.html.

Van de Sompel, H., & Lagoze, C. (2000, February). The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, 6(2). Retrieved March 20, 2001, from http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html.

Van de Sompel, H., & Lagoze, C. (2001). The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 1.1 of 2001-7-02, document version 2001-06-20. Retrieved September 9, 2001, available from http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html.

Van de Sompel, H., & Lagoze, C. (2002). Notes from the Interoperability Front: A Progress Report on the Open Archives Initiative. Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, 144-157.

Van de Sompel, H., Krichel, T., Nelson, M., Hochstenbach, P., Lyapunov, V., Maly, K., Zubair, M., Kholief, M., Liu, X., & O'Connell, H. (2000, February). The UPS Prototype An Experimental End-User Service across E-print Archives. D-Lib Magazine, 6(2). Retrieved March 20, 2001 from http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html.

Warner, S. (2001). Exposing and Harvesting Metadata Using the OAI Metadata Harvesting Protocol: A Tutorial. High Energy Physics Libraries Webzine, 4. Retrieved October 14, 2001 from http://library.cern.ch/HEPLW/4/papers/3/.

Webopedia. (2002). XML. Retrieved November 24, 2002, from http://www.webopedia.com/TERM/X/XML.html.

Weibel, S. (1995, July). Metadata: The Foundations of Resource Discovery. D-Lib Magazine, 1(7). Retrieved May 3, 2001 from http://www.dlib.org/dlib/July95/07weibel.html.

Appendix A

The DCMES, July 1995

Below is a brief description of the elements in the Dublin Core **Dublin Core Element Description**

- **Subject:** The topic addressed by the work
- **Title:** The name of the object
- **Author:** The person(s) primarily responsible for the intellectual content of the object
- **Publisher:** The agent or agency responsible for making the object available
- **OtherAgent:** The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work
- **Date:** The date of publication
- **ObjectType:** The genre of the object, such as novel, poem, or dictionary
- **Form:** The physical manifestation of the object, such as Postscript file or Windows executable file
- **Identifier:** String or number used to uniquely identify the object
- **Relation:** Relationship to other objects
- **Source:** Objects, either print or electronic, from which this object is derived, if applicable
- **Language:** Language of the intellectual content
- **Coverage:** The spatial locations and temporal durations characteristic of the object (Weibel, 1995, Scope section, para. 6).

The DCMES v. 1.0

Title:	**Dublin Core Metadata Element Set, Version 1.0: Reference Description**
Date Issued:	1998-09
Identifier:	http://dublincore.org/documents/1998/09/dces/
Supersedes:	Not Applicable
Is Superseded By:	http://dublincore.org/documents/1999/07/02/dces/
Latest version:	http://dublincore.org/documents/dces/
Translations:	http://dublincore.org/resources/translations/
Status of document:	This is a DCMI Recommendation.
Description of document:	This document is the reference description, version 1.0 of the Dublin Core Metadata Element Set. See the DCMI Home Page (http://dublincore.org) for further information about the workshops, reports, working group papers, projects, and new developments concerning the Dublin Core Metadata Element set. Note: This document has also been published as: Weibel, S.; Kunze, J.; Lagoze, C.; Wolf, M. 1998. Dublin Core Metadata for Resource Discovery. IETF #2413. The Internet Society, September 1998.

Introduction

This document is the reference description of the Dublin Core Metadata Element Set. See the Dublin Core Home Page (http://dublincore.org) for further information about the workshops, reports, working group papers, projects, and new developments concerning the Dublin Core Metadata Element set.

The current list of elements and their general definitions were finalized in December 1996. The elements and their names are not expected to change substantively from this list, though the application of some of them is currently experimental and subject to varying interpretation from implementation to implementation.

Note that elements have a descriptive name intended to convey a common semantic understanding of the element. To promote global interoperability, a number of the element descriptions may be associated with a controlled vocabulary for the respective element values. It is assumed that other controlled vocabularies will be developed for interoperability within certain local domains. In the element descriptions below, a formal single-word label (expressed in all upper case) is specified to make the syntactic specification of elements simpler for encoding schemes. Each element is optional and repeatable.

Questions or comments regarding the Dublin Core Element Set may be addressed to dcmi-feedback@dublincore.org.

Element Descriptions

1. Title
   Label: Title
   The name given to the resource, usually by the Creator or Publisher..
2. Author or Creator
   Label: Creator
   The person or organization primarily responsible for creating the intellectual content of the resource. For example, authors in the case of written documents, artists, photographers, or illustrators in the case of visual resources.
3. Subject and Keywords
   Label: Subject
   The topic of the resource. Typically, subject will be expressed as keywords or phrases that describe the subject or content of the resource. The use of controlled vocabularies and formal classification schemas is encouraged.
4. Description
   Label: Description
   A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources.
5. Publisher
   Label: Publisher
   The entity responsible for making the resource available in its present form, such as a publishing house, a university department, or a corporate entity.
6. Other Contributor
   Label: Contributor
   A person or organization not specified in a Creator element who has made significant intellectual contributions to the resource but whose contribution is secondary to any person or organization specified in a Creator element (for example, editor, transcriber, and illustrator).
7. Date
   Label: Date
   A date associated with the creation or availability of the resource.

Recommended best practice is defined in a profile of ISO 8601 (http://www.w3.org/TR/NOTE-datetime) that includes (among others) dates of the forms YYYY and YYYY-MM-DD. In this scheme, the date 1994-11-05 corresponds to November 5, 1994.

8. Resource Type
   Label: Type
   The category of the resource, such as home page, novel, poem, working paper, technical report, essay, dictionary. For the sake of interoperability, Type should be selected from an enumerated list that is under development in the workshop series.
9. Format
   Label: Format
   The data format and, optionally, dimensions (e.g., size, duration) of the resource. The format is used to identify the software and possibly hardware that might be needed to display or operate the resource. For the sake of interoperability, the format should be selected from an enumerated list that is currently under development in the workshop series.
10. Resource Identifier
    Label: Identifier
    A string or number used to uniquely identify the resource. Examples for networked resources include URLs and URNs (when implemented). Other globally-unique identifiers, such as International Standard Book Numbers (ISBN) or other formal names would also be candidates for this element.
11. Source
    Label: Source
    Information about a second resource from which the present resource is derived. While it is generally recommended that elements contain information about the present resource only, this element may contain metadata for the second resource when it is considered important for discovery of the present resource.
12. Language
    Label: Language
    The language of the intellectual content of the resource. Recommended best practice is defined in RFC 1766 http://info.internet.isi.edu/in-notes/rfc/files/rfc1766.txt
13. Relation
    Label: Relation
    An identifier of a second resource and its relationship to the present resource. This element is used to express linkages among related resources. For the sake of interoperability, relationships should be selected from an enumerated list that is currently under development in the workshop series.
14. Coverage
    Label: Coverage
    The spatial and/or temporal characteristics of the intellectual content of the resource. Spatial coverage refers to a physical region (e.g., celestial sector) using place names or coordinates (e.g., longitude and latitude). Temporal

coverage refers to what the resource is about rather than when it was created or made available (the latter belonging in the Date element). Temporal coverage is typically specified using named time periods (e.g., Neolithic) or the same date/time format (http://www.w3.org/TR/NOTE-datetime) as recommended for the Date element.

15. Rights Management
    Label: Rights
    A rights management statement, an identifier that links to a rights management statement, or an identifier that links to a service providing information about rights management for the resource ("Dublin Core Metadata Initiative", 1998).

The DCMES v. 1.1

Title:	**Dublin Core Metadata Element Set, Version 1.1: Reference Description**
Date Issued:	1999-07-02
Identifier:	http://dublincore.org/documents/1999/07/02/dces/
Supersedes:	http://dublincore.org/documents/1998/09/dces/
Is Superseded By:	Not Applicable
Replaces:	Not Applicable
Is Replaced By:	Not Applicable
Latest version:	http://dublincore.org/documents/dces/
Translations:	http://dublincore.org/resources/translations/
Status of document:	This is a DCMI Recommendation.
Description of document:	This document is the reference description, version 1.1 of the Dublin Core Metadata Element Set. This document supersedes the Dublin Core Metadata Element Set, version 1.0. See the Dublin Core Home Page (http://dublincore.org) for further information about the workshops, reports, working group papers, projects, and new developments concerning the Dublin Core Metadata Element set.

Introduction

The document summarizes the updated definitions for the Dublin Core metadata elements as originally defined in [RFC2413]. These new definitions will be officially known as Version 1.1.

The definitions utilise a formal standard for the description of metadata elements. This formalisation helps to improve consistency with other metadata communities and enhances the clarity, scope, and internal consistency of the Dublin Core metadata element definitions.

Each Dublin Core element is defined using a set of ten attributes from the ISO/IEC 11179 [ISO11179] standard for the description of data elements. These include:

- **Name** - The label assigned to the data element
- **Identifier** - The unique identifier assigned to the data element
- **Version** - The version of the data element
- **Registration Authority** - The entity authorised to register the data element
- **Language** - The language in which the data element is specified
- **Definition** - A statement that clearly represents the concept and essential nature of the data element
- **Obligation** - Indicates if the data element is required to always or sometimes be present (contain a value)
- **Datatype** - Indicates the type of data that can be represented in the value of the data element
- **Maximum Occurrence** - Indicates any limit to the repeatability of the data element
- **Comment** - A remark concerning the application of the data element

Fortunately, six of the above ten attributes are common to all the Dublin Core elements. These are, with their respective values:

Version: 1.1
Registration Authority: Dublin Core Metadata Initiative
Language: en
Obligation: Optional
Datatype: Character String
Maximum Occurrence: Unlimited

The above attributes will not be repeated in the below definitions, however, they do represent part of the formal element definitions.

The definitions provided here include both the conceptual and representational form of the Dublin Core elements. The Definition attribute captures the semantic concept and the Datatype and Comment attributes capture the data representation.

Each Dublin Core definition refers to the resource being described. A resource is defined in [RFC2396] as "anything that has identity". For the purposes of Dublin Core metadata, a resource will typically be an information or service resource, but may be applied more broadly.

Element: Title
Name: Title
Identifier: Title
Definition: A name given to the resource.
Comment: Typically, a Title will be a name by which the resource is formally known.

Element: Creator

Name: Creator
Identifier: Creator
Definition: An entity primarily responsible for making the content of the resource.
Comment: Examples of a Creator include a person, an organisation, or a service. Typically, the name of a Creator should be used to indicate the entity.

Element: Subject
Name: Subject and Keywords
Identifier: Subject
Definition: The topic of the content of the resource.
Comment: Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.

Element: Description
Name: Description
Identifier: Description
Definition: An account of the content of the resource.
Comment: Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.

Element: Publisher
Name: Publisher
Identifier: Publisher
Definition: An entity responsible for making the resource available
Comment: Examples of a Publisher include a person, an organisation, or a service. Typically, the name of a Publisher should be used to indicate the entity.

Element: Contributor
Name: Contributor
Identifier: Contributor
Definition: An entity responsible for making contributions to the content of the resource.
Comment: Examples of a Contributor include a person, an organisation, or a service. Typically, the name of a Contributor should be used to indicate the entity.

Element: Date
Name: Date
Identifier: Date
Definition: A date associated with an event in the life cycle of the resource.
Comment: Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and follows the YYYY-MM-DD format.

Element: Type
Name: Resource Type
Identifier: Type
Definition: The nature or genre of the content of the resource.
Comment: Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the working draft list of Dublin Core Types [DCT1]). To describe the physical or digital manifestation of the resource, use the FORMAT element.

Element: Format
Name: Format
Identifier: Format
Definition: The physical or digital manifestation of the resource.
Comment: Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats).

Element: Identifier
Name: Resource Identifier
Identifier: Identifier
Definition: An unambiguous reference to the resource within a given context.
Comment: Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Example formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).

Element: Source
Name: Source
Identifier: Source
Definition: A Reference to a resource from which the present resource is derived.
Comment: The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

Element: Language
Name: Language
Identifier: Language
Definition: A language of the intellectual content of the resource.
Comment: Recommended best practice for the values of the Language element is defined by RFC 1766 [RFC1766] which includes a two-letter Language Code

(taken from the ISO 639 standard [ISO639]), followed optionally, by a two-letter Country Code (taken from the ISO 3166 standard [ISO3166]). For example, 'en' for English, 'fr' for French, or 'en-uk' for English used in the United Kingdom.

Element: Relation
Name: Relation
Identifier: Relation
Definition: A reference to a related resource.
Comment: Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

Element: Coverage
Name: Coverage
Identifier: Coverage
Definition: The extent or scope of the content of the resource.
Comment: Coverage will typically include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and that, where appropriate, named places or time periods be used in preference to numeric identifiers such as sets of coordinates or date ranges.

Element: Rights
Name: Rights Management
Identifier: Rights
Definition: Information about rights held in and over the resource.
Comment: Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource (“Dublin Core Metadata Initiative”, 1999).

Appendix B

Example of an OAI-PMH Metadata Record

```
<header>
   <identifier>oai:arXiv:cs/0112017</identifier>
   <datestamp>2002-02-28</datestamp>
   <setSpec>cs</setSpec>
   <setSpec>math</setSpec>
</header>
<metadata>
 <oai_dc:dc
    xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
    http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:title>Using Structural Metadata to Localize Experience of Digital
          Content</dc:title>
  <dc:creator>Dushay, Naomi</dc:creator>
  <dc:subject>Digital Libraries</dc:subject>
  <dc:description>With the increasing technical sophistication of both
   information consumers and providers, there is increasing demand for
   more meaningful experiences of digital information. We present a
   framework that separates digital object experience, or rendering,
   from digital object storage and manipulation, so the
   rendering can be tailored to particular communities of users.
  </dc:description>
  <dc:description>Comment: 23 pages including 2 appendices,
               8 figures</dc:description>
  <dc:date>2001-12-14</dc:date>
  <dc:type>e-print</dc:type>
  <dc:identifier>http://arXiv.org/abs/cs/0112017</dc:identifier>
 </oai_dc:dc>
</metadata>
<about>
 <provenance
    xmlns="http://www.openarchives.org/OAI/2.0/provenance/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/provenance/
    http://www.openarchives.org/OAI/2.0/provenance/oai_provenance.xsd">
   <originDescription>
     <baseURL>http://the.oa.org</baseURL>
     <identifier>oai:r2:klik001</identifier>
     <datestamp>2002-01-01</datestamp>
     <metadataPrefix>oai_dc</metadataPrefix>
    <harvestDate>2002-02-02T14:10:02Z</harvestDate>
   </originDescription>
  </provenance>
</about>  (Lagoze, Van de Sompel, Nelson & Warner, 2002)
```

Appendix C

Figure 1 - Number of DPs by Top-Level Domain

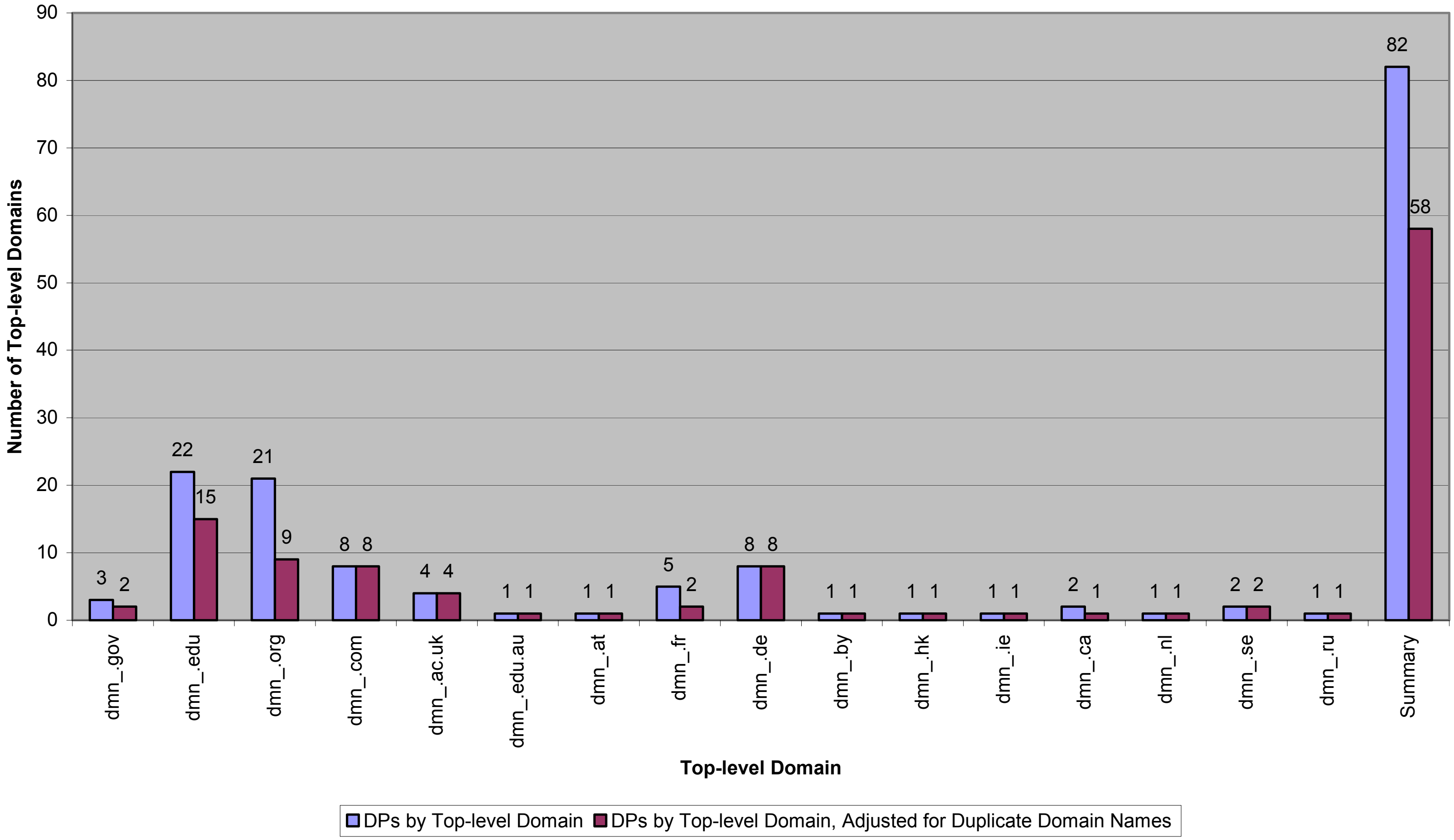


Figure 2 - % of Top-Level Domains Across 82 DPs

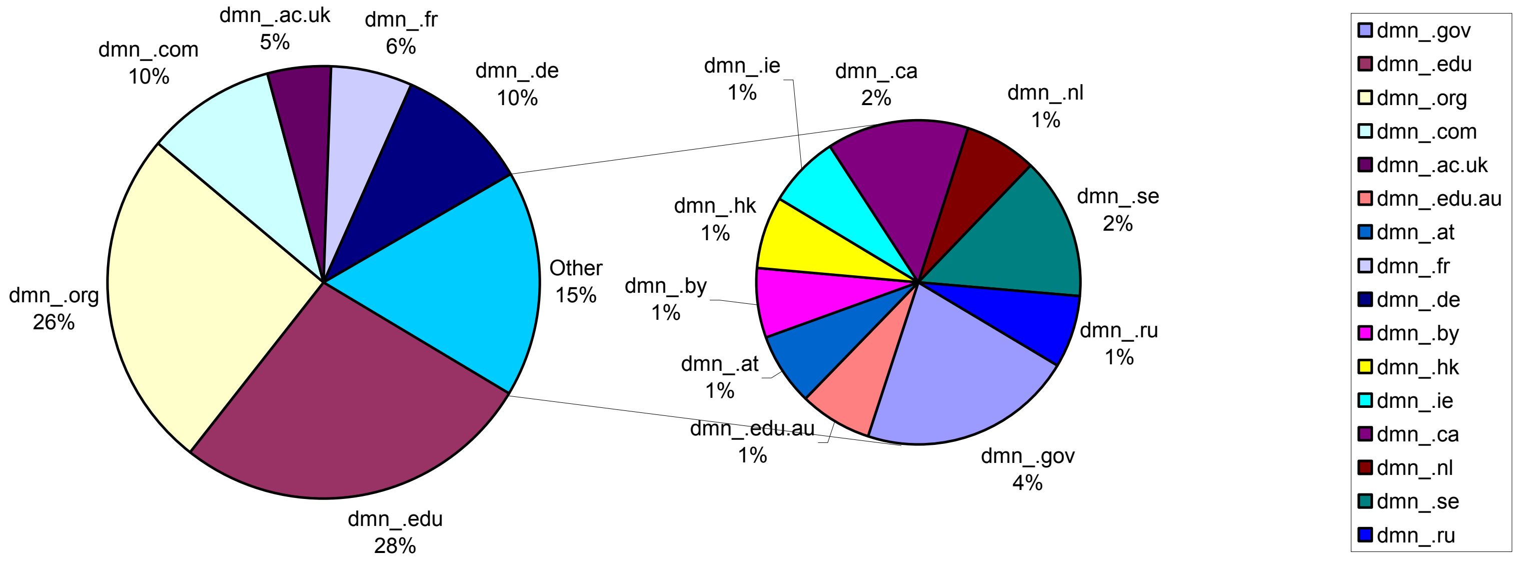


Figure 3 - % of Top-Level Domains Across 82 DPs (Adjusted for Duplicate Domain Names)

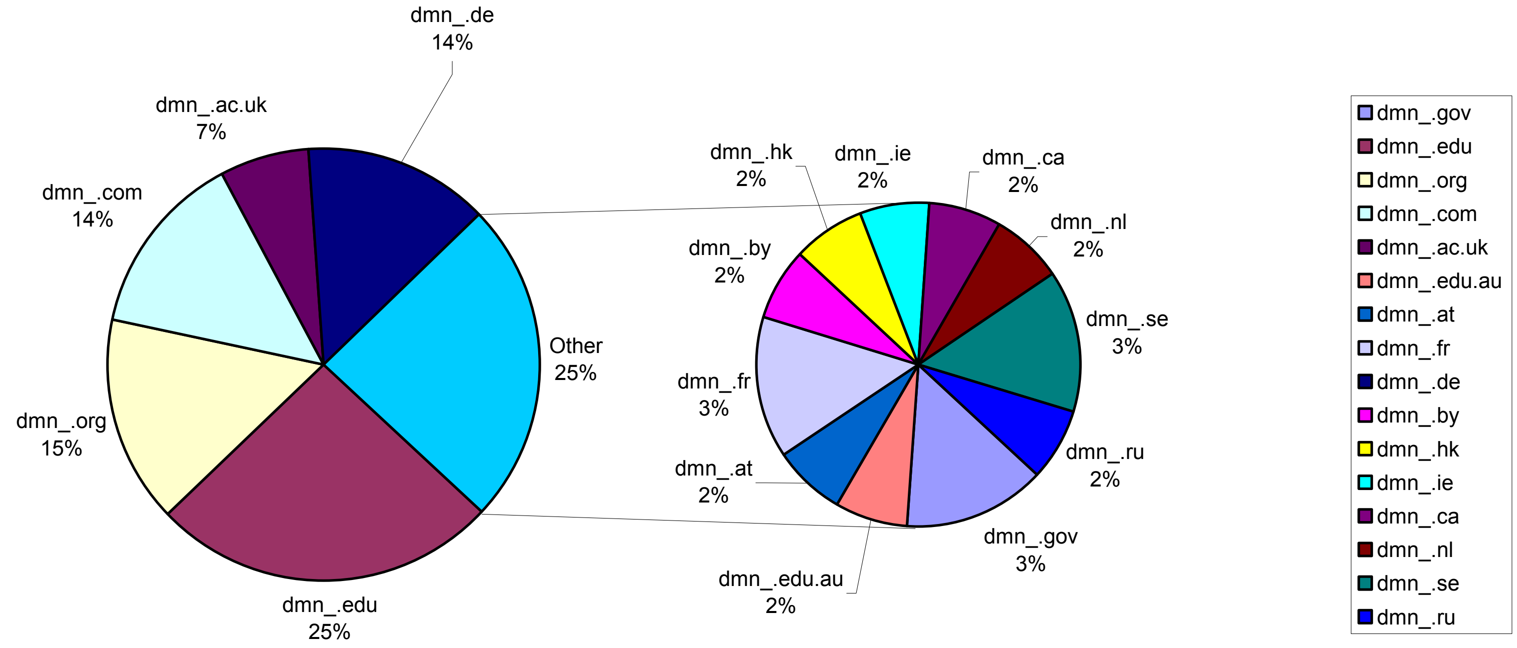


Table 2 – Number of DPs by Domain Name

Domain Name	Number of Repositories Contributed by Domain
*.language-archives.org	10
*.caltech.edu	4
*.cdlib.org	4
*.cnrs.fr	3
*.upenn.edu	2
*.in2p3.fr	2
*.nasa.gov	2
*.ubc.ca	2
*.uiuc.edu	2
*.vt.edu	2
*.ethnologue.com	1
*.uni-oldenburg.de	1
*.aim25.ac.uk	1
*.anu.edu.au	1
*.arXiv.org	1
*.berkeley.edu	1
*.biomedcentral.com	1
*.chemweb.com	1
*.conoze.com	1
*.cstc.org	1
*.davidrumsey.com	1
*.emich.edu (164.76.128.26)	1
*.eprints.org	1
*.gla.ac.uk	1
*.hku.hk	1
*.hofstra.edu	1
*.hray.com	1
*.hu-berlin.de	1
*.ibiblio.org	1
*.indiana.edu	1
*.infomotions.com	1
*.ioffe.ru	1
*.loc.gov	1
*.lsu.edu	1

Domain Name	Number of Repositories Contributed by Domain
*.mathpreprints.com	1
*.may.ie	1
*.mit.edu	1
*.nottingham.ac.uk	1
*.numismatics.org	1
*.open-video.org	1
*.paristech.org	1
*.riacs.edu	1
*.rug.nl	1
*.slu.se	1
*.slub-dresden.de	1
*.tu-chemnitz.de	1
*.tufts.edu	1
*.uaf.edu	1
*.ulst.ac.uk	1
*.umich.edu	1
*.umn.edu	1
*.unibel.by (195.50.11.247)	1
*.uni-bremen.de	1
*.uni-dortmund.de	1
*.uni-duisburg.de	1
*.uni-tuebingen.de	1
*.utk.edu	1
*.uu.se	1
*.wu-wien.ac.at	1
Sum:	82

Appendix D

Figure 4 - Summary of Records, All DPs and by Type

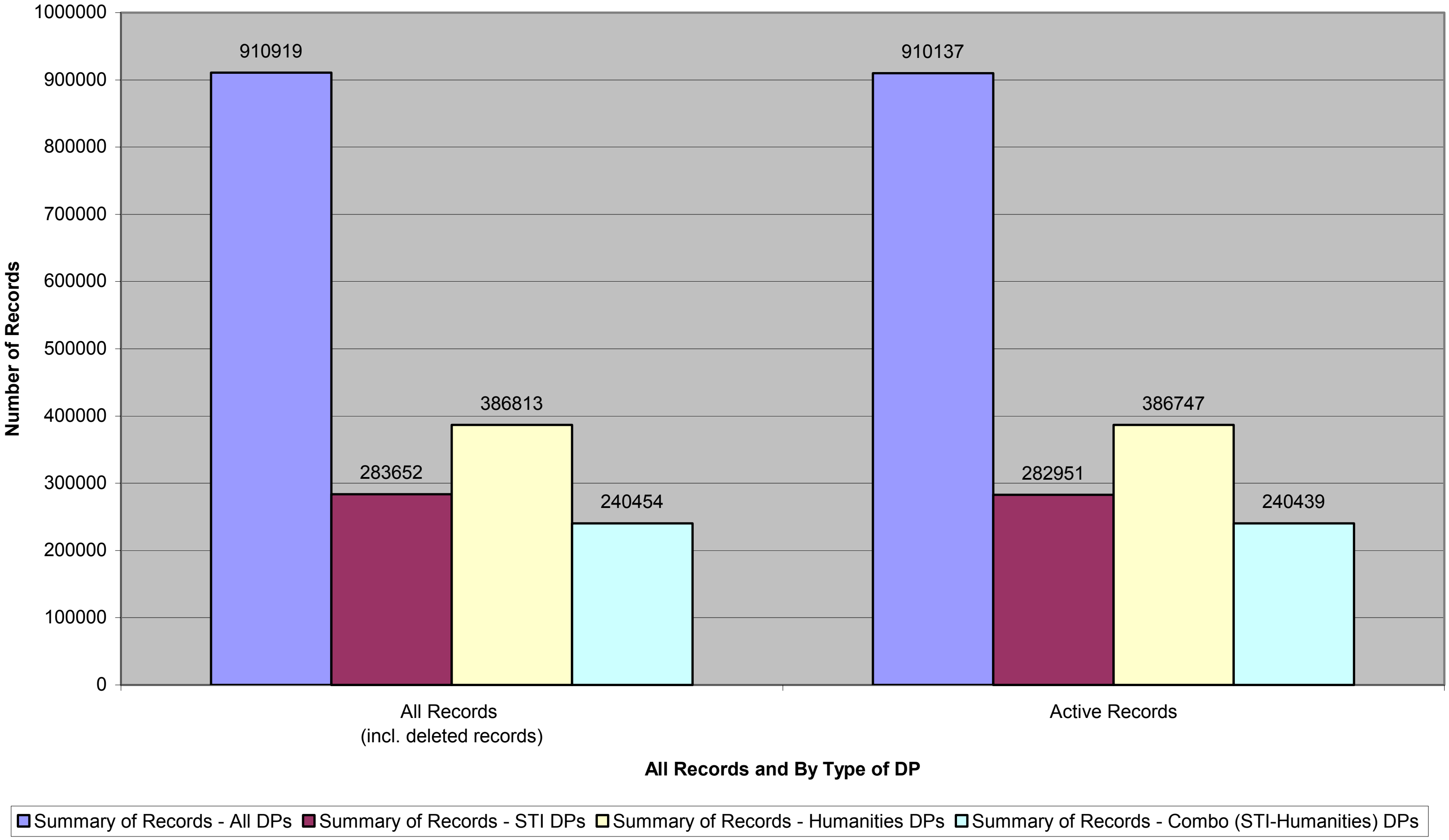


Figure 5 - Number of DPs per Type of Archive

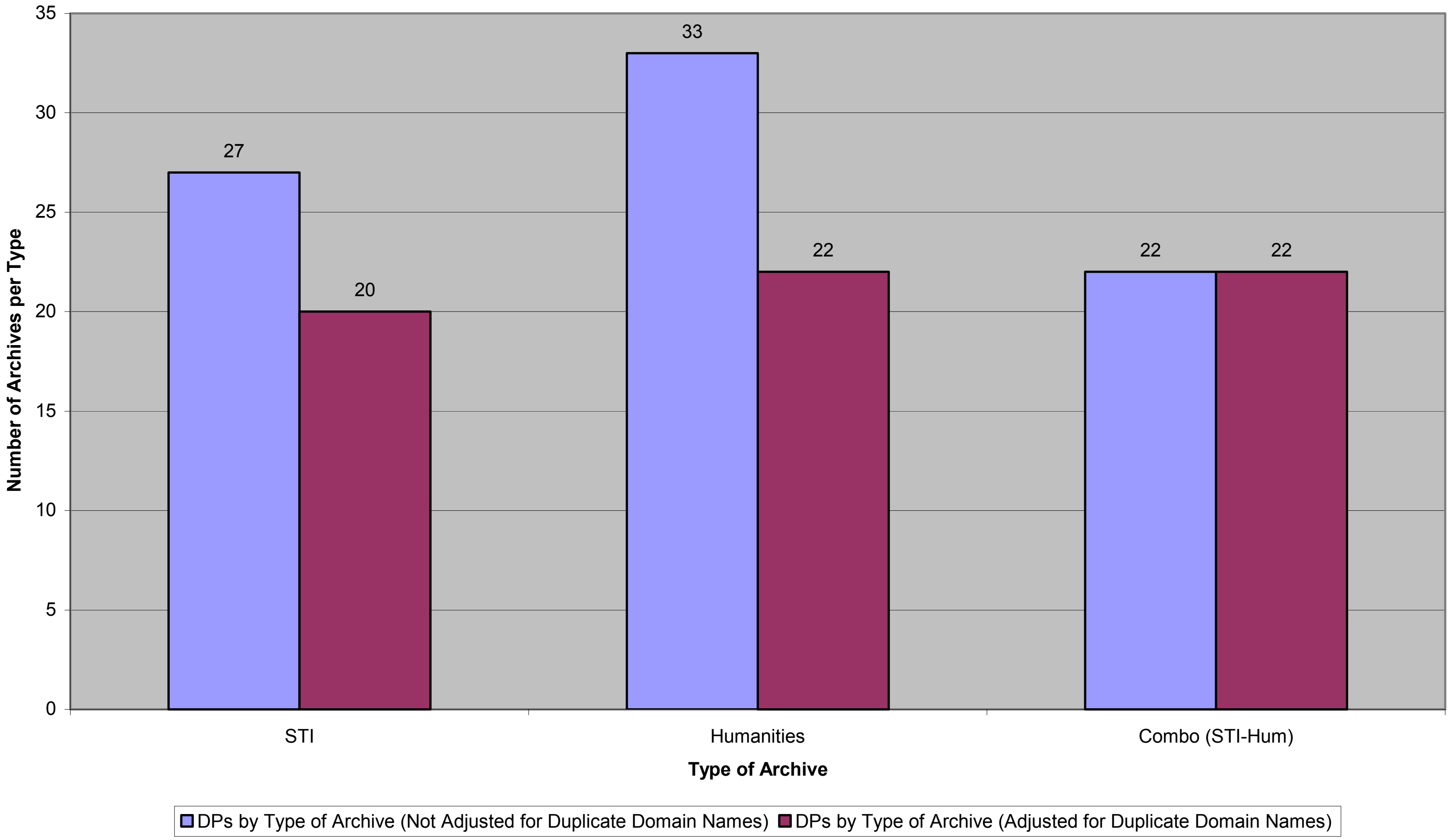


Figure 6 - % of DPs by Type of Archive, Not Adjusted for Duplicate Domain Names

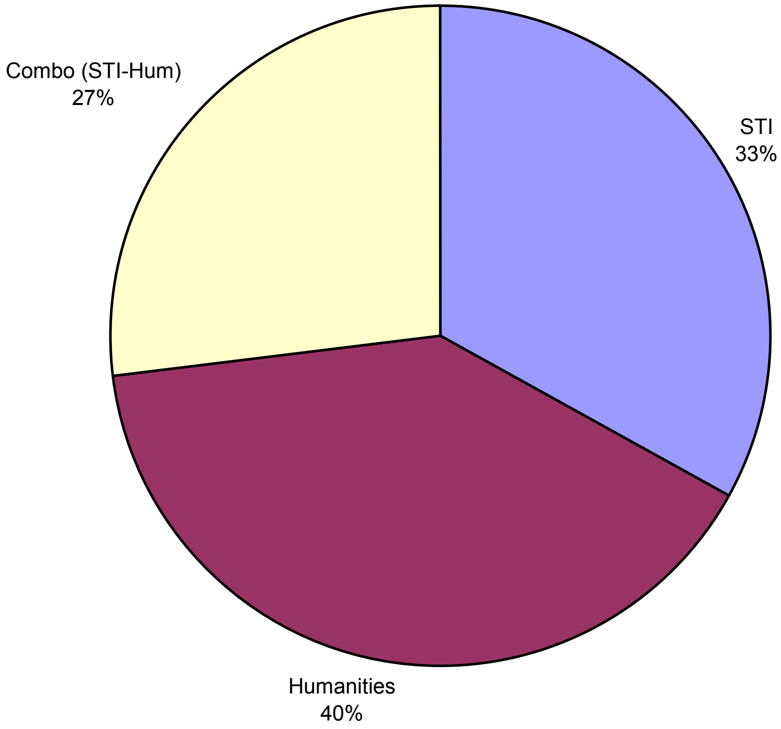


Figure 7 - % of DPs by Type of Archive, Adjusted for Duplicate Domain Names

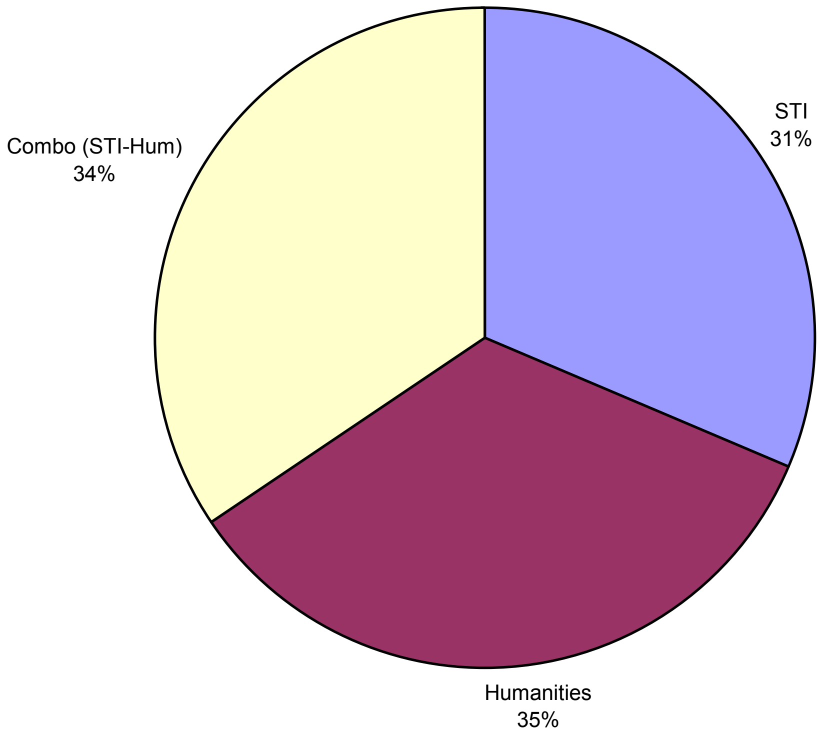


Figure 8 - % of Records Held by All DPs

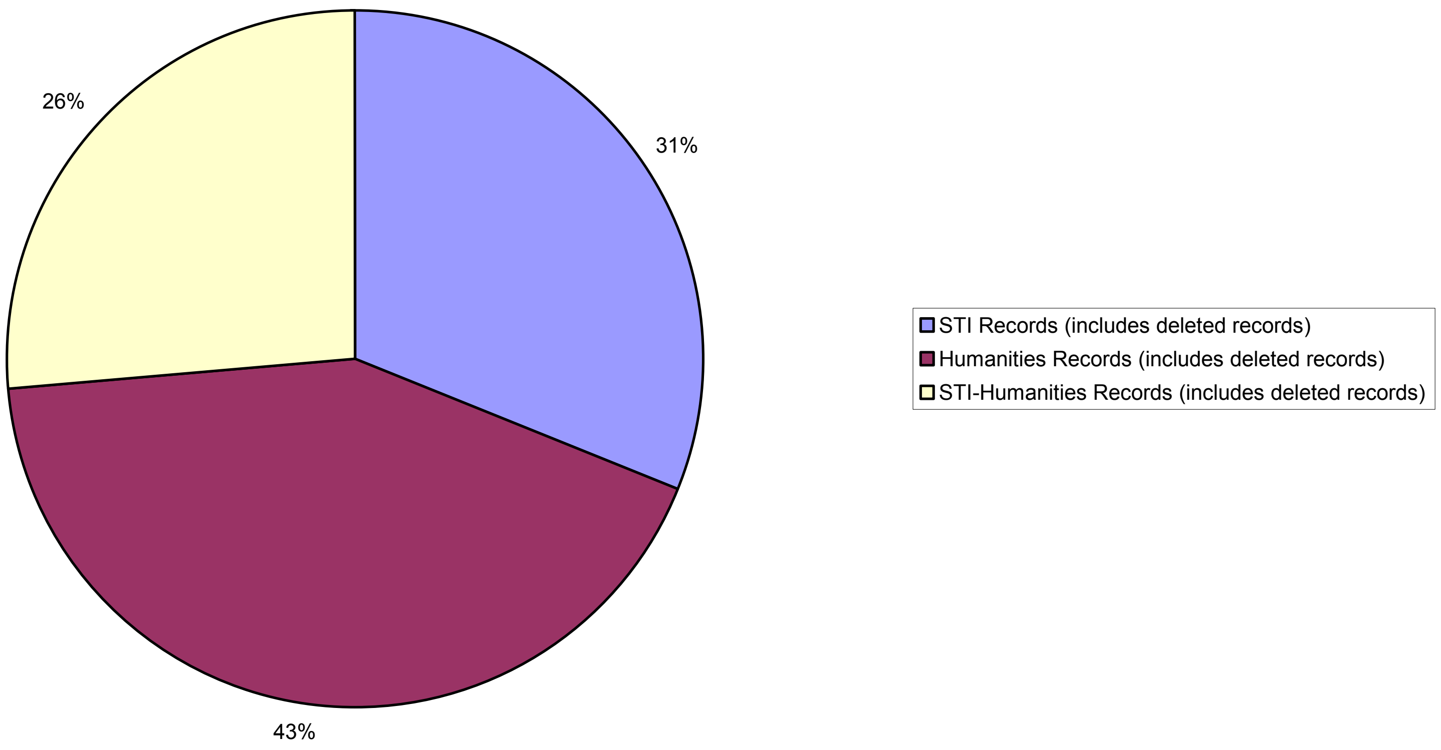


Figure 9 - % of Records Held by Each STI DP

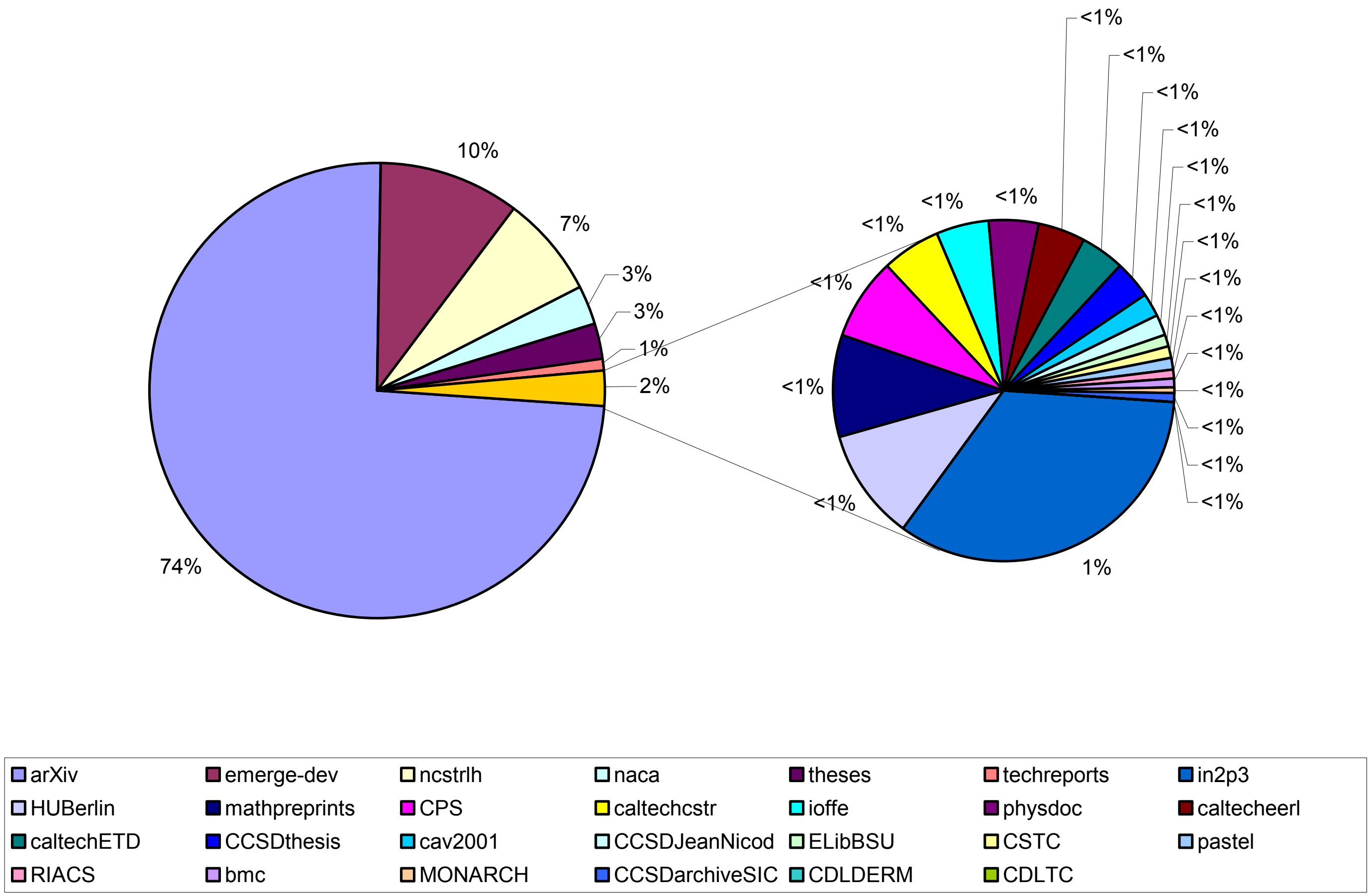


Figure 10 - % of Records Held by Each Humanities DP

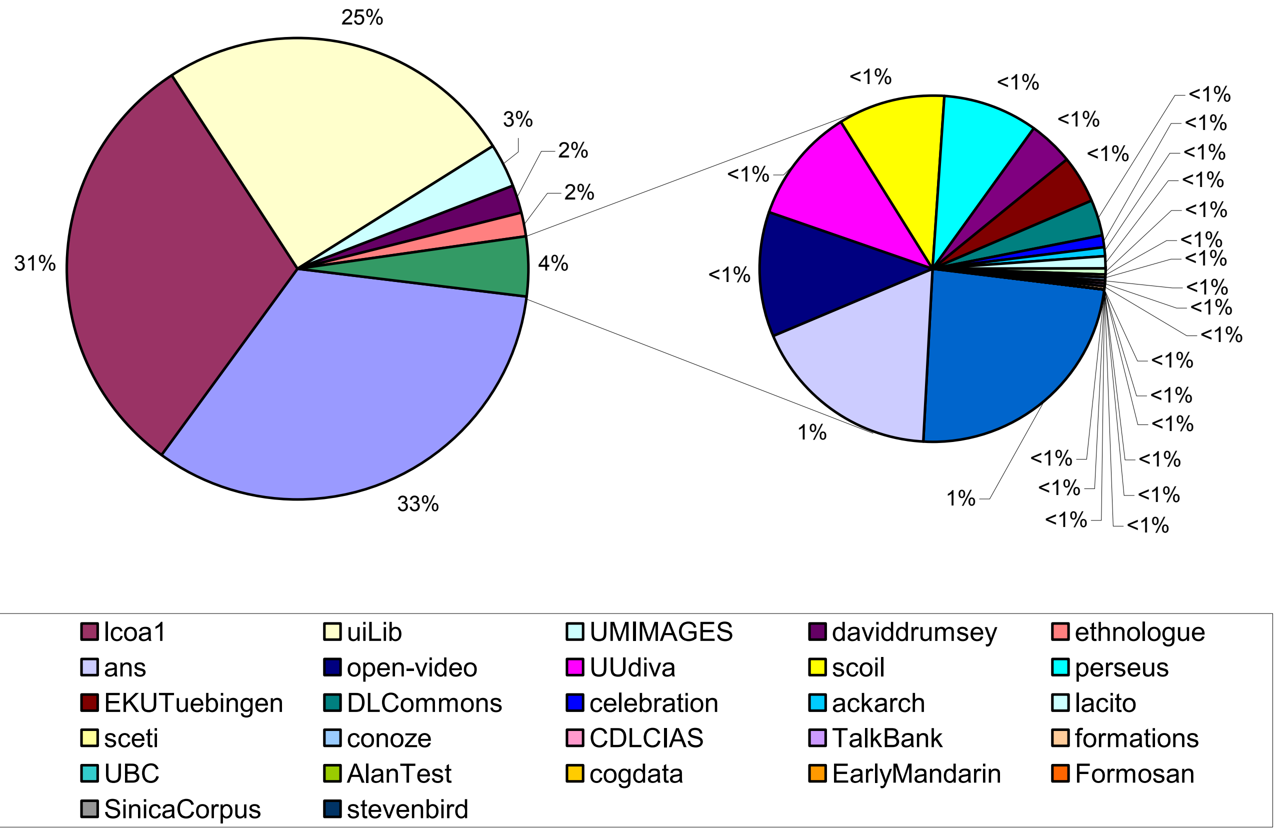
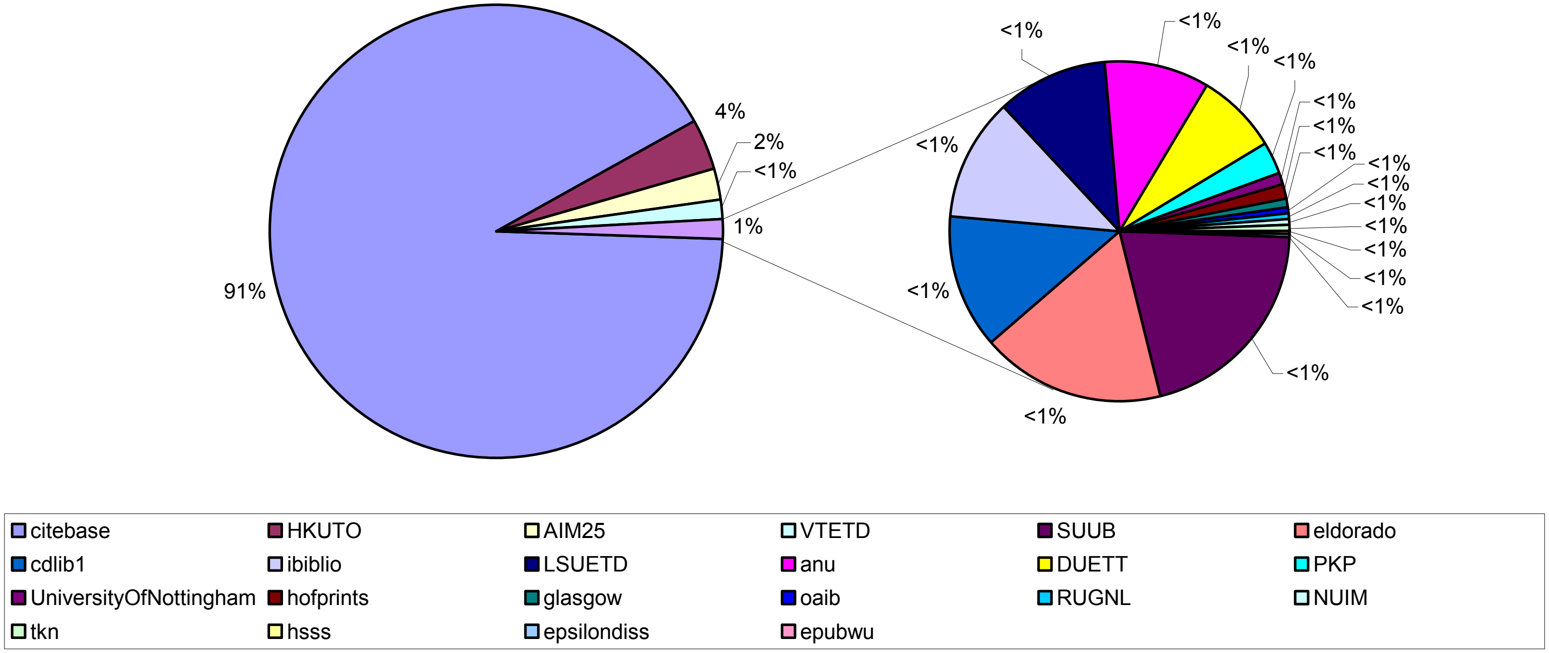


Figure 11 - % of Records Held by Each Combo (STI-Humanities) DP



Appendix E

Figure 12 - Total DP Records vs. the Number of DC Elements Used by All DPs

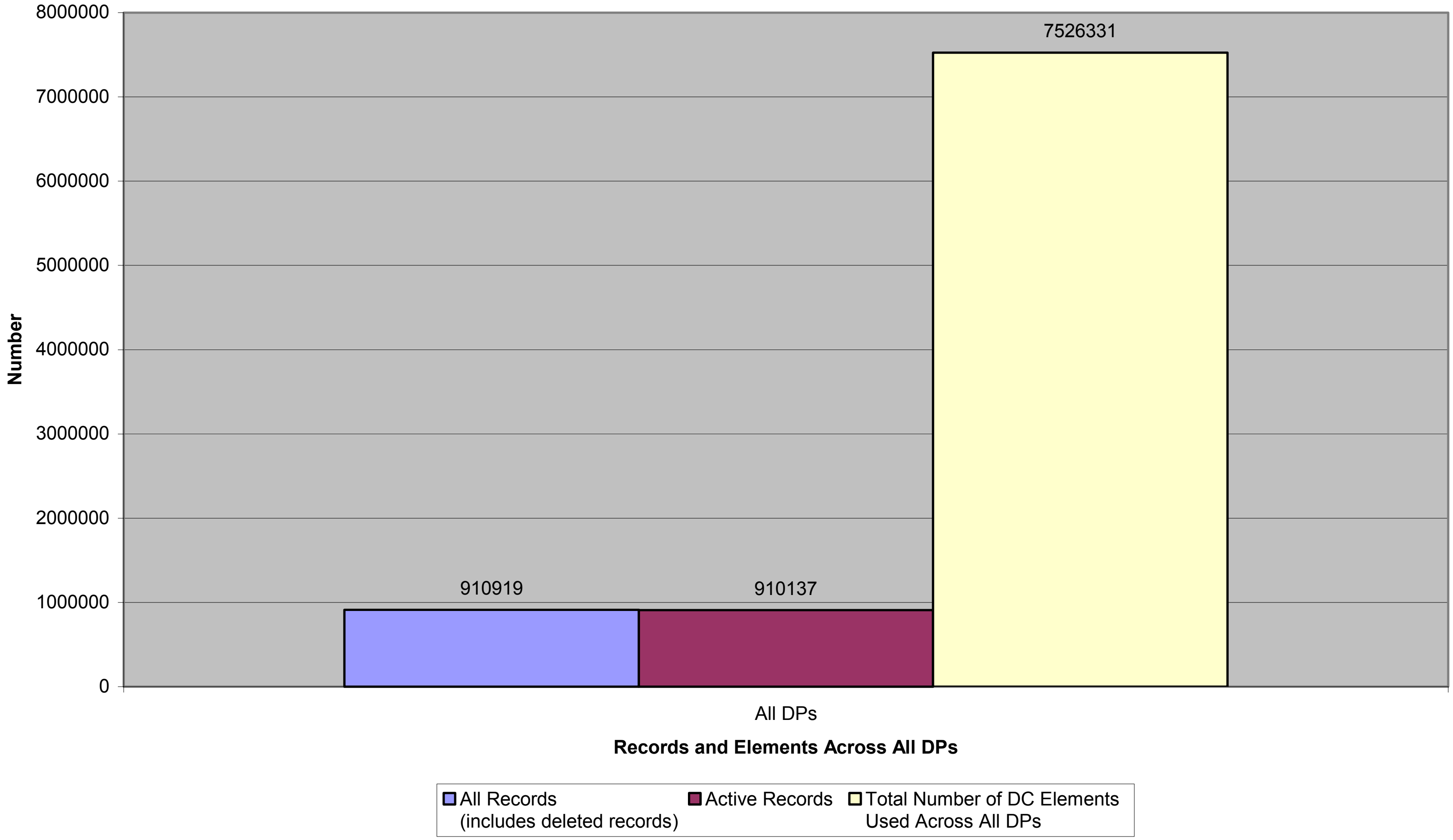


Table 3 - % of DC Elements by All DC Elements, All Records, and All DPs

Records (General Summary)			
DC Element	Number of Elements per Record	Each Element as a % of the Total Number of Elements Used (n/7,526,331)	Each Element as a % of the Total Number of Records Across All DPs (n/910,919)
creator	1,617,910	21.5	177.6
identifier	1,292,707	17.2	141.9
title	860,488	11.4	94.5
date	834,949	11.1	91.7
type	802,538	10.7	88.1
subject	495,414	6.6	54.4
description	463,833	6.2	50.9
rights	312,403	4.2	34.3
publisher	235,759	3.1	25.9
coverage	202,936	2.7	22.3
language	146,579	1.9	16.1
format	136,501	1.8	15.0
relation	47,748	0.6	5.2
contributor	39,743	0.5	4.3
source	36,823	0.5	4.0
Total:	7,526,331	100.0	826.2

DPs (Summary of Crosstabs Results)				
DC Element	Number of DPs That *Never* Used a Particular Element Out of 82 DPs		DPs That Used a Particular Element Out of 82 DPs	
	No.	%	No.	%
title	1	1.2	81	98.8
creator	4	4.9	78	95.1
date	6	7.3	76	92.7
identifier	7	8.5	75	91.5
type	10	12.2	72	87.8
subject	14	17.1	68	82.9
description	23	28.0	59	72.0
language	39	47.6	43	52.4
publisher	41	50.0	41	50.0
format	43	52.4	39	47.6
rights	46	56.1	36	43.9
contributor	50	61.0	32	39.0
source	52	63.4	30	36.6
coverage	66	80.5	16	19.5
relation	66	80.5	16	19.5

Appendix F

Figure 13 - Comparison of Total DC Elements Used, Overall and by Type

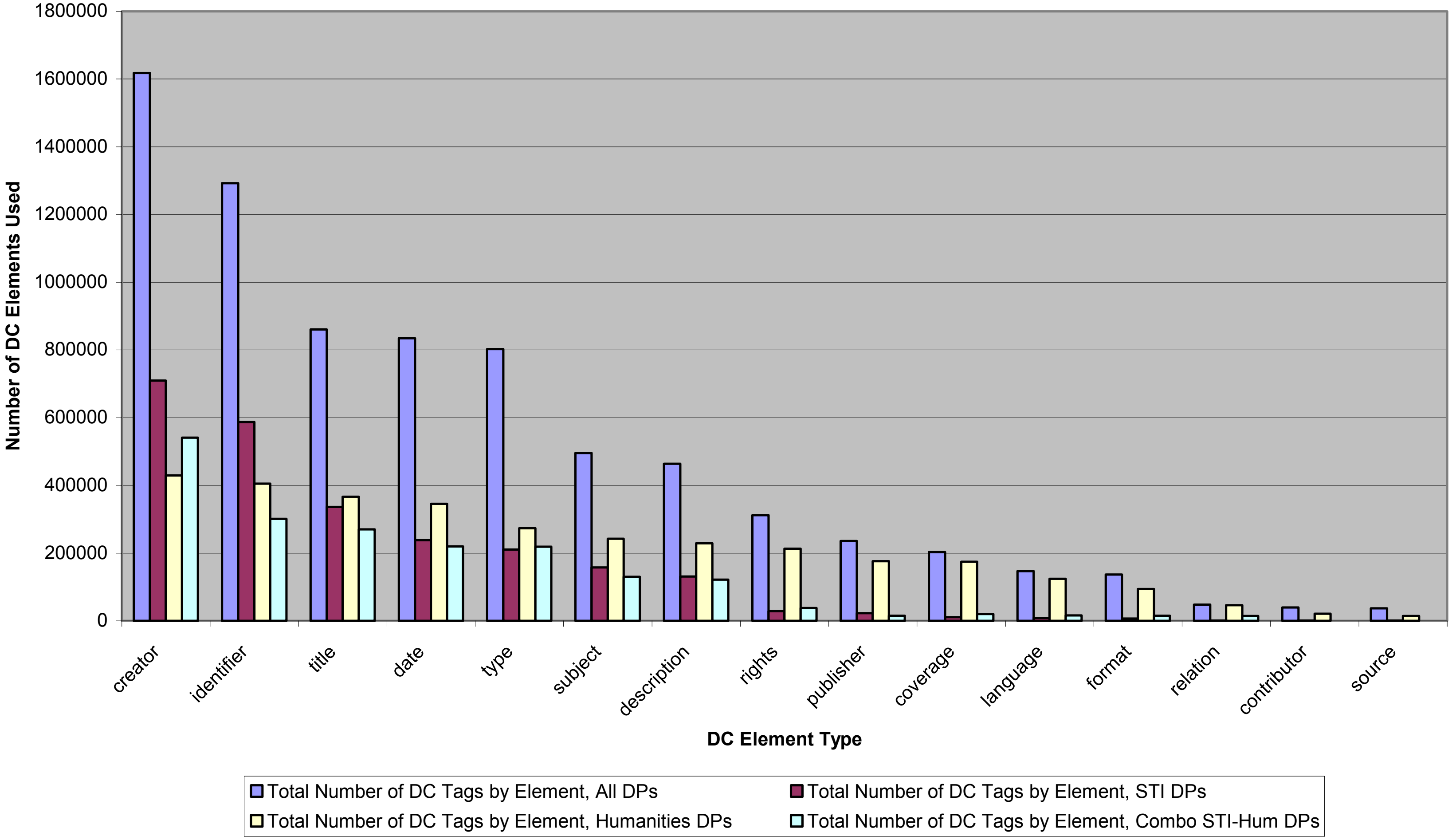


Figure 14 - DC Metadata Element Usage by % Across All 82 DPs

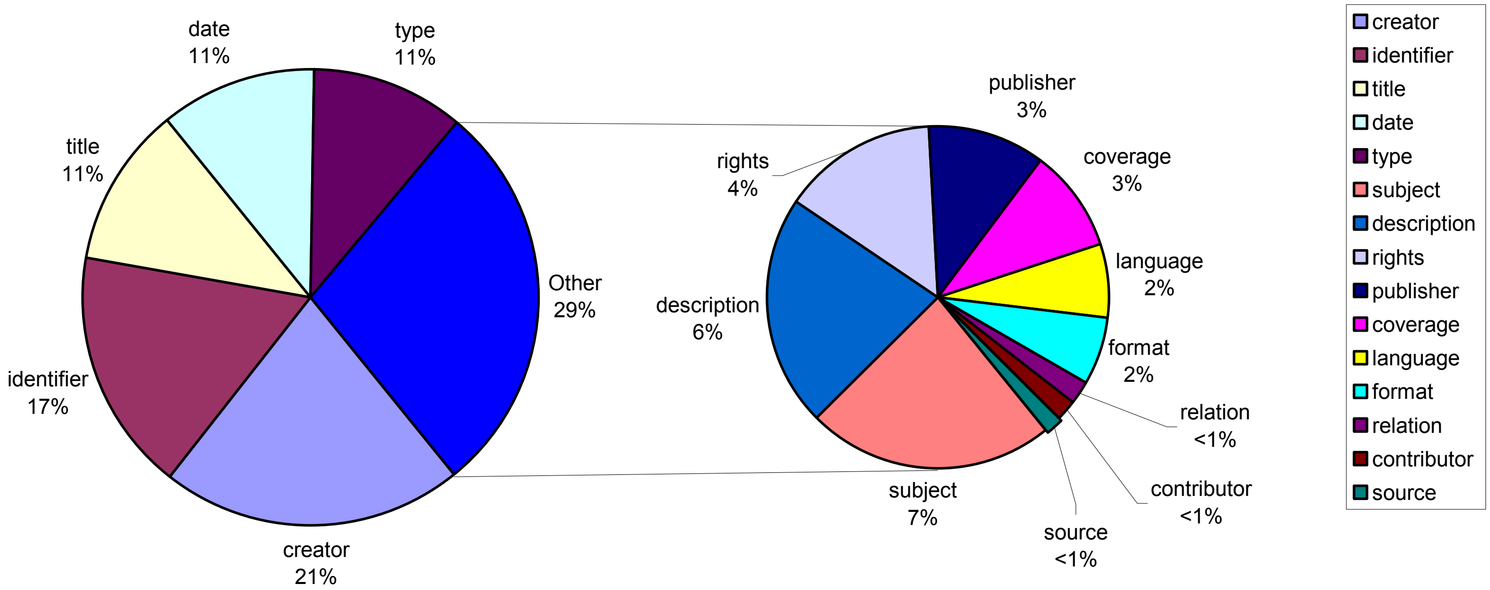


Figure 15 - DC Metadata Element Usage by % Across STI DPs

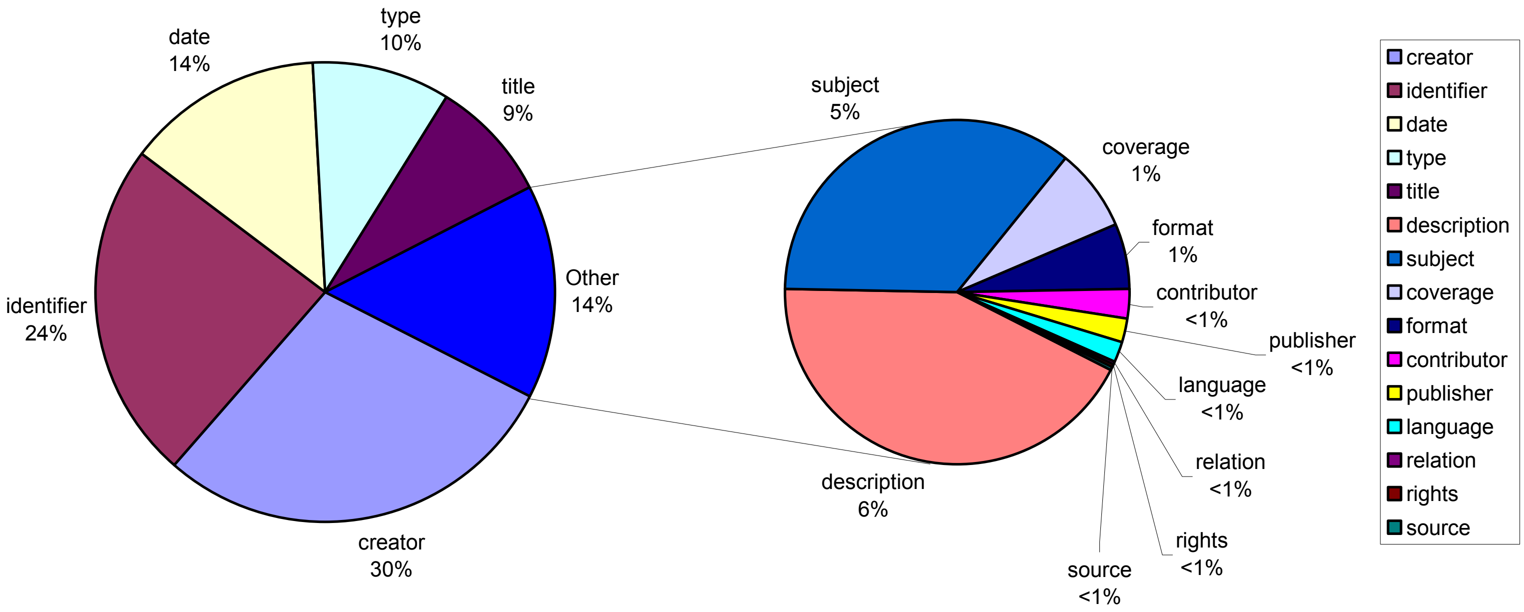


Figure 16 - DC Metadata Element Usage by % Across Humanities DPs

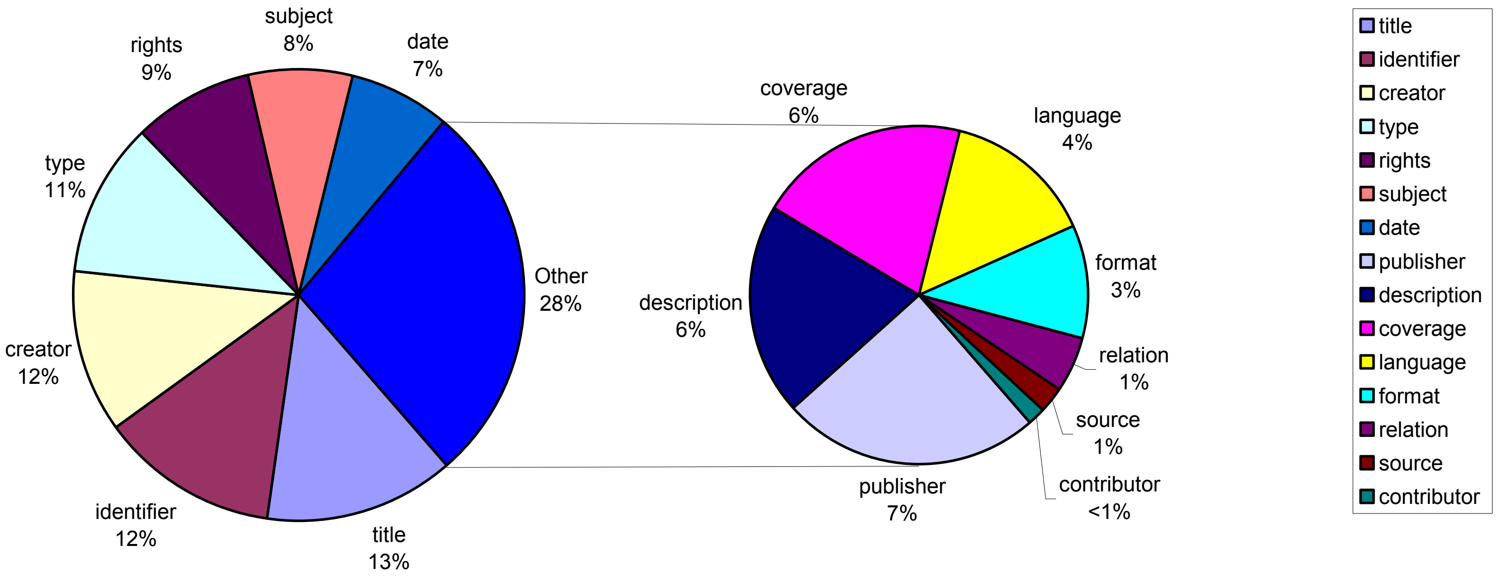
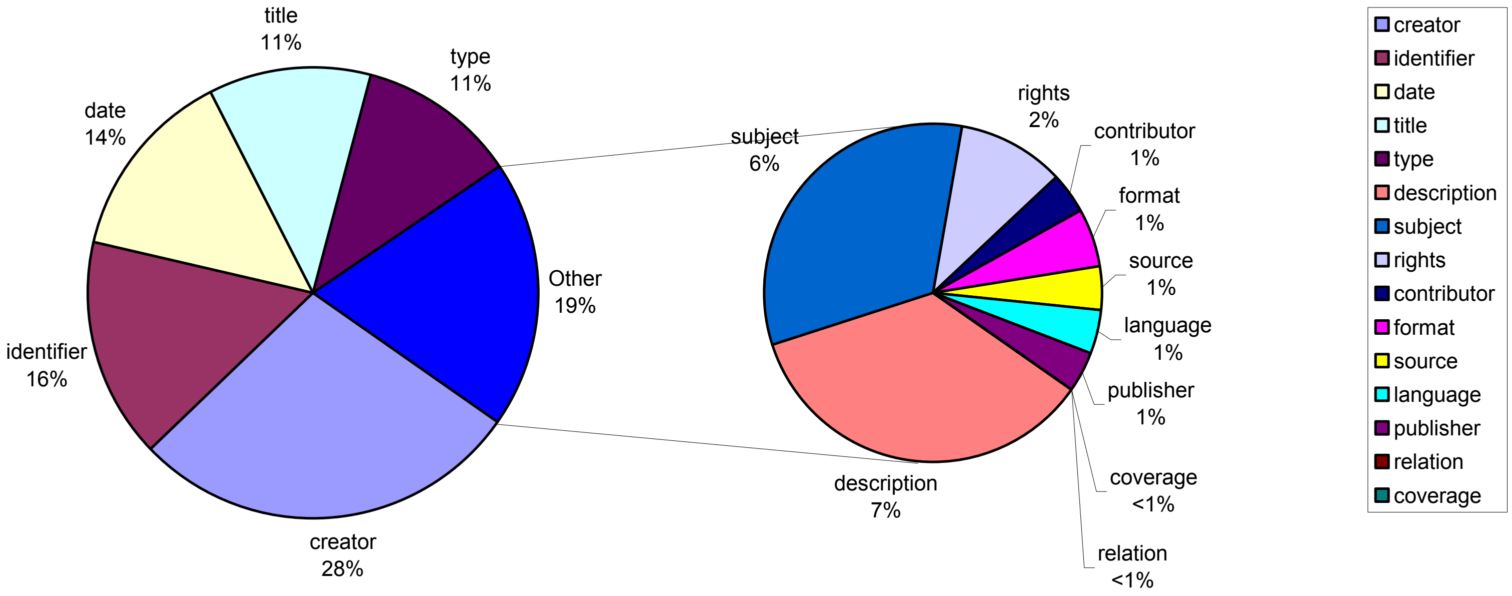


Figure 17 - DC Metadata Element Usage by % Across Combo (STI-Humanities) DPs


title
type
date
Other
identifier
creator
subject
rights
contributor
format
source
language
publisher
description
coverage
<1%
relation
<1%
creator
identifier
date
title
type
description
subject
rights
contributor
format
source
language
publisher
relation
coverage