Making Your Way through Grey: Metadata, MARC, and User Tools

Meagan Cooke and Sean S. Costigan

CIAO's Background

CIAO is the largest library of international affairs content on the web. Originally funded by the Mellon Foundation, CIAO became self-sustaining through library subscriptions after three years of operation. CIAO was built in a partnership with Columbia's libraries, the University Press, and its academic computing and information systems group (AcIS). Subject specialists, computer scientists and librarians all had a hand in its initial development. Today, such expertise is drawn on to further realize the service's goals of promoting a wide range of grey and published literature in international affairs. Currently over 200 institutions partner with CIAO, primarily contributing working papers, conference proceedings, reports, books, policy briefs and journals. CIAO boasts more than 800 subscribers, among them government agencies, militaries, academic institutions and businesses. In any given month, over 2000 pages of material from dozens of contributors will be posted on CIAO. Such a large and mature repository poses significant challenges with regard to data management, archiving and customization.

Many Organizations, Many Standards

At CIAO's inception in 1997 a variety of file formats were commonly in use. CIAO's production staff was likely to receive files from Word, WordPerfect, Quark, and a smattering of non-standard text editors. In keeping with our desire to make CIAO usable to as extensive an audience as possible, all files were converted to faster loading html. Initially, CIAO adhered to HTML 2.0 specifications and when additional HTML specifications came out, adjustments were made to new, but not existing, content.

Today the bulk of CIAO's contributors deliver content in PDF (Portable Document Format) or Microsoft Word. For some of CIAO's subscribers, particularly those overseas and from secure locations, low bandwidth continues to be an issue that we design around by producing HTML abstracts for the majority of PDFs. In addition, html abstracts afford us the opportunity to more comprehensively describe the content using our metadata. Where possible we add author and title in-

Address for correspondence: Mr. Sean S. Costigan, Senior Editor, Columbia University Press, 61 West 62nd Street, New York, NY 10023 Email : sc717@columbia.edu. Websites: www.ciaonet.org, www.earthscape.org.

formation to the PDFs, allowing our search engine to take advantage of that metadata as it indexes the site.

Metadata—A Primer

Several definitions of metadata exist. The W3C defines metadata as "machine understandable information for the web." For our purposes, metadata is simply "data about data"—particularly information like keywords, document type, title, abstract, location, ISBN, etc. In its paper *Understanding Metadata*,[1] the National Information Standards Organization details three main types of metadata:

- Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
- Structural metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters.
- Administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

To describe CIAO's content, our editorial team developed a metadata scheme that includes the following mixture of descriptive and administrative tags:

```
<meta name=”robots” content=”noarchive”>
<meta name=”ciao_title” content=””>
<meta name=”ciao_author” content=””>
<meta name=”ciao_type” content=””>
<meta name=”ciao_date” content=””>
<meta name=”ciao_subject” content=””>
<meta name=”ciao_subject” content=””>
<meta name=”ciao_subject” content=””>
<meta name=”ciao_subject” content=””>
<meta name=”ciao_institution” content=””>
<meta name=”ciao_language” content=””>
```

Several renditions of these tags have been employed over the life of the service, due in part to technical and editorial developments that could not be described with the old criteria. Additionally, several hands have worked on the site since its founding, producing variations in standards and use. A production management system, built by CIAO's editors and web developers, has assisted in the standardization of metadata.

Controlled Vocabulary

Taxonimist Amy Warner defines a controlled vocabulary (CV) as "organized lists of words and phrases, or notation systems, that are used to initially tag content, and then to find it through navigation or search."[2] Initially, nearly 100 subject tags (our controlled vocabulary) were chosen in consultation with CIAO’s editorial advisory board. Over time these subject tags were distilled to around 80 in number and some new tags were added.

CIAO’s subject tags are a mixture of regions, countries and themes. Four or more subject tags are chosen for each piece of content on CIAO, allowing users to find content in a variety of ways. Typically several region tags and thematic tags are used to describe documents. For CIAO, a complete tag might look like this:

```
<meta name=”robots” content=”noarchive”>
<meta name=”ciao_title” content=”Final Report of the Independent Panel to Review DoD Detention Operations”>
<meta name= “ciao_author” content =”Schlesinger, James R.”>
<meta name= “ciao_author” content =”Brown, Harold”>
<meta name= “ciao_author” content =”Fowler, Tillie K.”>
<meta name= “ciao_author” content =”Horner, Charles A.”>
<meta name= “ciao_author” content =”Blackwell, James A. Jr.”>
<meta name= “ciao_type” content =”wps”>
<meta name= “ciao_date” content =”200408">
<meta name= “ciao_subject” content =”Crime”>
<meta name= “ciao_subject” content =”International Law”>
<meta name= “ciao_subject” content =”United States”>
<meta name= “ciao_subject” content =”War>
<meta name= “ciao_subject” content =”Middle East”>
<meta name= “ciao_subject” content =”Arab Countries”>
<meta name= “ciao_institution” content=”U.S. Department of Defense”>
```

The Challenges of Inconsistencies

Insufficiencies have been addressed on two separate occasions through the introduction of new subject tags, though only after careful thought about the consequences. For a sobering analysis of metadata foibles, read Cory Doctorow’s paper *Metacrap*.[3] Since CIAO is built in a directory architecture with static html and PDFs, making new subject tags ripple through the site’s directories can pose significant challenges. For example, how does one ensure that new tags are properly deployed in older content? With a database as large as CIAO, going through manually is clearly not an effective option, although files were modified manually over the course of the summer of 2002.

Due to the volume of content and the size of our staff, asking a university programmer, often from outside the immediate group, to write a script generally helps us come closer to our goals. After running such scripts, we proceed to manual verification and modification of stray oddities. Initially all work is done on a test server, with the eyes of the editorial and production staff (as well as committed librarians) watching for inconsistencies. After achieving general agreement that all is well, we find a time to copy the changed files over to the production server. Since

CIAO has an international subscriber base, we pay careful attention to usage patterns around the world in order to minimize downtime.

Over the life of the service, we've made two additions to CIAO's administrative metadata. The first dealt with one of the ramifications of allowing Google to index the site. Shortly after allowing Google in, we realized that its robots had cached the site's pages. Adding a new administrative tag <meta name="robots" content="noarchive"> and requesting that Google crawl the site again restored our security. Additionally, taking advantage of Google's prominence has greatly increased the number of free trials worldwide and brought users from subscribing institutions to CIAO. Usage has doubled.

In consultation with our editorial advisors, CIAO is now striking agreements with international institutions to collect non-English language content. With an eye towards searching non-English content we have added a tag noting its language. The default language remains English, though we are aggregating content in French, Spanish, German, Romanian, Turkish, and are considering additional languages. All foreign language content includes abstracts in English.

Cataloging CIAO Content

In some disciplines the knowledge or information published in grey literature will never appear in published format[4] (Chilag 1982). The Internet has gone a long way towards making the reports, studies and proceedings that constitute grey literature easy to access. According to Luzi[5] (2000), grey literature databases were distributed commercially as early as the 1970s. To advertise these highly-valued, vetted resources we've worked with librarians and programmers in creating MARC records and disseminating citations. MARC is the acronym for MAchine-Readable Cataloging, an initiative that began over thirty years ago. Developed under the auspices of the Library of Congress, MARC formats are standards for "the representation and communication of bibliographic and related information in machine-readable form."[6]

When libraries subscribe to an electronic resource, the institutional libraries staff often has time to create one MARC record for the electronic resource, but not the materials it holds. Increasingly, academic libraries expect vendors to provide MARC records for materials within the resource. The benefit to libraries is that users can find publications across many resources through one search in the library OPAC (Online Public Access Catalog). This method is preferred to using popular search engines, such as Google, which index materials not vetted by information professionals. Users looking for a report will learn that the full text of that report is available electronically through their library's subscription to CIAO. Publishers benefit from the addition of many more access points to their resources and the likelihood that the files within those resources will be used increases.

In February 2003 we began working with the special materials cataloging staff at Columbia University libraries to catalog the published and grey literature in CIAO. As of November 2004, CIAO holds more than thirty full-text journals and 124 full text books. Because of the relatively manageable quantities of published literature, we decided that books and journals could be cataloged by library school students. Among the grey materials on CIAO, case studies were the only collection that could be cataloged manually. Library students cataloged CIAO case studies as part of their school practicum; the hours spent cataloging fulfilled their requirement for graduation.

The remaining grey literature collections in CIAO are vast; as of November 2004, the database holds more than 4,500 working papers and more than 5,000 policy briefs. Catalogers mapped descriptive metadata in CIAO abstract pages to MARC fields. Subjects in CIAO's pages were mapped to Library of Congress subject headings. With these values mapped to MARC fields and LoC standards, programmers are working to crosswalk CIAO metadata to XML and MARC.

We explored different models for distributing MARC records. CIAO case studies were uploaded to OCLC (Online Computer Library Center) and subscribing libraries were made aware of their availability through e-mail, allowing them to find these records through a title search in OCLC. As a distribution model, this was unsatisfactory for CIAO users and for CIAO staff. Catalogers at subscribing institutions found it time-consuming to search for the records and download them. Also, we did not have the ability to learn about this service's usage. As a result we chose instead to disseminate records from CIAO servers.

The MARC records for CIAO journals are now available and records for CIAO books will become available this spring. We expect the vast majority of these institutions to have OPACs, so usage in the majority of subscribing institutions should increase once MARC records for CIAO content are introduced. Sixty-five percent of CIAO subscribers are higher-education institutions. Instead of one access point to CIAO—the sole MARC record for the electronic resources—users will find more than 10,000.

Metadata Standards

Several standardization initiatives exist. For our purposes we have considered converting to Dublin Core. According to the Dublin Core Metadata Initiative's website, Dublin Core is "an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models." Dublin Core was intended to be a simple set of elements that could be used to describe web resources. Currently 15 elements make up the standard: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights. At this time CIAO continues to explore the value of converting its metadata to Dublin Core. Mapping our metadata remains the most likely outcome. Such mapping is likely to take place should we harvest the metadata for an OAI server.[7]

Rich Hyperlinking

It is certain that students comprise our largest user group. Specifically with students in mind, CIAO's editorial staff recently completed an atlas comprised of country data, maps and histories from the CIA, Department of State, Transitions Online and other sources. Database driven, the new atlas allows for data comparisons across a dozen variables. As an information resource, the atlas is rich enough to stand on its own, although we believe its true value will be realized when maps and country data are made available from within CIAO's content. The materials in the Atlas provide context for the technical grey literature on the site. To that end, we are currently introducing links to regional and country maps to all content in the working papers and books sections of the site. A link in the upper left corner of a paper or book about a country or region will open a map of the region. As with other changes on the site, scripts have been written to locate subject tags within the content and introduce the appropriate map links. We chose to add this functionality to working papers and books first because more often these materials given unique subject tags whereas materials appearing in serial form—including policy briefs and journals—are given tags according to the journal or policy brief subjects, not the article content. Without clean and standardized metadata, this important development would not have been possible.

E-mail this Citation

Once the metadata on CIAO was made consistent through editorial labor and programmer scripts, we began to create functionality based on the information. Of course, we looked forward to more and better search results but we also wanted users to be able to cite the content they found. CIAO includes a page devoted to guidelines for citing content found on the website. These guidelines were taken from the Columbia Guide to Online Style (1998). We went a step further, embedding an "e-mail this citation" link in all reports, conference proceedings and working papers on CIAO. The "email this citation" script examines the metadata from a given page and extracts the relevant information. It then puts this information together into a citation. On the user end, clicking on the "email this citation" link renders a form in a new window. Users who complete the form are able to send citations for CIAO content in emails to a specified address. Users can email citations that adhere to APA and Chicago styles. It is hoped creation of this tool will facilitate usage of papers and reports; and advertise their availability to researchers and scholars.

Search

Columbia University employs an Inktomi search software product called Ultraseek. Ultraseek was purchased by AcIS in June 1999 as a replacement for other search engines in use at the university. Prior to 2001, CIAO used a homebrew search engine that was fast but had one glaring weakness: it was unable to index PDFs. When more contributors began delivering content in PDF it became clear that our technology was no longer sustainable. CIAO's staff began searching for a replacement search engine and we realized that, as a member of the university community, CIAO could benefit from Columbia's site license. Since its initial deployment on CIAO we have completed three modifications of the search form, each time learning a little bit more about Ultraseeks's capabilities and shortcomings.

Ultraseek is programmed in Python, a little known language at Columbia, and so changes are made infrequently by AcIS programmers. Over the years we have made modifications to the search form based on requests from librarians and metadata cleanup initiatives. We have experimented with adjusting the weight of different search criteria, including title, body and subject metadata. Currently, criteria for searching include publication format, author, contributing institution, region, title, subject and date. At this time we cannot sort publications by date, but we are investigating other search technologies, including Lucene and Google, that can perform that function. The CIAO language metatag will be used when we have a larger body of foreign language content.

The Future Is Grey

In the next year, we will migrate to an XML-based site architecture. The promise of XML is that it will allow content to be easily modified while also separating content from its metadata scheme. Implementing site wide changes will be less labor intensive. We will also be able to break individual documents into their components, offering greater granularity in search results. The ability to search abstracts is an often-requested feature that will be included in the future XML version of CIAO. Prior to migrating to XML we are likely to harvest all CIAO's metadata and build it into an OAI server.[8] Such initiatives promise to further our goals of disseminating international affairs grey literature.

Notes

[1] http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

[2] http://www.lexonomy.com/publications/aTaxonomyPrimer.html

[3] http://www.well.com/~doctorow/metacrap.htm

[4] Chilag, J. *Non-conventional literature in agriculture – an Overview.* 1982. IAALD Quarterly Bulletin, Vol. 27, No 1.

[5] Luzi, Daniela. *Trends and Evolution in the Development of Grey Literature: A Review.* 2000. The International Journal on Grey Literature, Vol. 1, Issue 3. http://www.emeraldinsight.com/rpsv/cgi-bin/linker?reqidx=/cw/mcb/14666189/v1n3/s2/p106.idx&lkey=-1038598568&rkey=911673

[6] http://www.loc.gov/marc/

[7] "The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content." OAI FAQ (www.openarchives.org)

[8] http://www.openarchives.org/OAI/openarchivesprotocol.html

References

Dublin Core Metadata Initiative, www.dublincore.org

The Expanding Horizon of Grey Literature http://cf.hum.uva.nl/bai/home/jmackenzie/pubs/glpaper.htm

Grey Literature and Library and Information Studies: A Global Perspective http://www.emeraldinsight.com/rpsv/cgibin/linker?reqidx=/cw/mcb/14666189/v1n4/s3/p167.idx&lkey=-1810531140&rkey=785358

Grey Literature: Its History, Definition, Acquisition and Cataloguing http://www.moyak.com/researcher/resume/papers/var7mkmkw.html

Metadata Resources http://www.ukoln.ac.uk/metadata/resources

Open Archives Initiative, www.openarchives.org

Understanding Metadata http://www.niso.org/standards/resources/UnderstandingMetadata.pdf