

## FINAL DRAFT:

Cite as: Greenberg, J. (2004, *in press*). Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. *Journal of Internet Cataloging*, 6(4): 59-82.

Author: Jane Greenberg, janeg@ils.unc.edu

## **Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications.**

### **Abstract**

This research explores the capabilities of two Dublin Core automatic metadata generation applications, Klarity and DC.dot. The top level Web page for each resource, from a sample of 29 resources obtained from National Institute of Environmental Health Sciences (NIEHS), was submitted to both generators. Results indicate that extraction processing algorithms can contribute to useful automatic metadata generation. Results also indicate that harvesting metadata from META tags created by humans can have a positive impact on automatic metadata generation. The study identifies several ways in which automatic metadata generation applications can be improved and highlights several important areas of research. The conclusion is that integrating extraction of harvesting methods will be the best approach to creating optimal metadata, and more research is needed to identify when to apply which method.

### **Keywords**

Automatic Metadata Generation, Metadata Harvesting, Metadata Extraction, Metadata Applications

### **Introduction**

The need for standardized metadata supporting resource description and discovery is radically increasing as the World Wide Web (Web) becomes a major means for communicating and disseminating information. Evidence of this need is clear from the Open Archives Initiative (<http://www.openarchives.org/>) and D-Space (<http://www.dspace.org>), both of which have adopted the Dublin Core metadata

standard as a fundamental component. A major challenge for these projects and other digital initiatives is the growing number of resources requiring metadata. Library metadata practices, whereby professionals (catalogers and indexers) generate metadata, are prohibitive due to the limited availability of both qualified persons and financial resources. Resource authors, offering one alternative, can produce fairly good metadata for selected elements (Greenberg, et al. 2001). However, research has also found some authors view metadata creation as an additional task and they are reluctant to engage in this activity (Greenberg, et al. 2003a; Trigg, et al., 1999). A key reason for this circumstance is limited author knowledge as to the value of metadata for discovering their work and the absence of an organizational incentive. As the demand for standardized metadata increases, researchers need to identify metadata production methods that are more efficient and less costly than practices involving humans.

Automatic metadata generation can help address this need. Automatic metadata generation, based on knowledge about automatic indexing (Anderson & Perez-Carballo, 2001), is known to be more efficient, less costly, and more consistent than human-oriented processes. In fact, research indicates that automatic metadata generation can produce acceptable results for subject metadata (Liddy, et al., 2001). Results can, however, be problematic, particularly for subject metadata or other metadata requiring human intellectual discretion (Lancaster, 1998; Anderson & Perez-Carballo, 2001). Researchers have concluded that the most effective mean of metadata creation is to integrate both human and automatic methods (e.g., Schwartz, 2000; Craven, 2001). This conclusion makes sense when considering that many researchers agree the best retrieval results can be achieved by integrating natural language processing and controlled vocabulary searching (*see* Svenonius, E., 1986, for an excellent historical review of key studies in this area). Limited financial and human resources, the Web's tremendous growth, and clear support of automatic processes via the conclusions just shared, underscore the need to study and identify specifically how automatic metadata generation can complement or serve as an alternative to human-oriented metadata production methods.

The research presented in this paper addresses this need and explores the capabilities of two Dublin Core automatic metadata generation applications, *Klarity* and *DC.dot* (hereafter referred to by their name or generically as metadata applications). This analysis is part of a larger study conducted by the Metadata Generation Research (MGR) project (<http://ils.unc.edu/~janeg/mgr>), which is developing a model to facilitate the most efficient and effective means of metadata production by integrating automatic and human processes. The MGR project is based at the School of Information and Library Science, University of North Carolina at Chapel Hill. The research reported on in this study was conducted in collaboration with National Institute of Environmental Health Sciences (NIEHS) (<http://www.niehs.nih.gov>), an Institute of the National Institutes of Health located in the Research Triangle Park, North Carolina.

### **The Dublin Core Metadata Standard and Automatic Metadata Generators**

The Dublin Core metadata standard, version 1.1 in particular (Dublin Core..., 2003), has had a major impact on resource description in the context of the Web. Developed under the direction of the Dublin Core Metadata Initiative (DCMI), the Dublin Core is an international and interdisciplinary metadata standard that has been adopted by an array of communities wanting to facilitate resource discovery and build an interoperable information environment.

The Dublin Core has been formally endorsed as a standard by Comité Européen De Normalisation (CEN) as CEN Workshop Agreement (CWA) 13874 (<http://www.cenorm.be/cenorm/businessdomains/businessdomains/informationssystem/etystandardizationsystem/published+cwas/13874.pdf>), by the National Information Standards Organization (NISO) as NISO Z39.85-2001 (<http://www.niso.org/standards/resources/Z39-85.pdf>), and, most recently, by the International Standards Organization (ISO) as ISO 15836-2003 (<http://www.niso.org/international/SC4/n515.pdf>). Dublin Core literature (e.g., see bibliography maintained by the DCMI, Harper & Sharfe, 2003), the growing number of projects adopting and implementing the Dublin Core (see Dublin Core Projects:

<http://www.dublincore.org/projects/>), and its translation into over 20 languages (see Translations of DCMI Documents:

<http://www.dublincore.org/resources/translations/>) confirm the Dublin Core's adoption as a standard.

Standards activities are important because they provide guidance for the development of tools and improve processes towards an end-goal. In the metadata community, standardization of the Dublin Core, as noted here, has spurred development of a wide-variety of metadata generation applications for creating metadata. Among such applications are a series of tools called *metadata generators* that almost exclusively rely on automatic processes (Greenberg, 2003b). Generators, as automatic applications, are advantageous compared to manual processes when considering metadata creation time, consistency, and costs. These applications have the potential to greatly aid metadata initiatives and improve the Web's metadata infrastructure. Despite this fact, there is little research examining the performance of metadata generators. Research needs to investigate metadata generator performance, specifically the results stemming from the two primary algorithmic emphases, metadata extraction and harvesting, which is an underlying goal of this research.

## Metadata Extraction and Harvesting

Metadata extraction and harvesting are different automatic generation methods.<sup>1</sup> *Metadata extraction* occurs when an algorithm automatically extracts metadata from a resource's content displayed via a Web browser. (The research in this paper focuses on "textual content" for practical reasons.) At a rudimentary level, resource content is mined to produce structured ("labeled") metadata for *object representation*. Specific to Web resources, the structured metadata is extracted from the "body" section of a HyperText Markup Language (HTML) or Extensible Hypertext Markup Language (XHTML) document. (The acronym HTML will be used to refer to both markups

---

<sup>1</sup> Metadata generation literature also includes the term "derived metadata" in reference to both "extraction" and "harvesting" processes as described in this present article. The term "derived metadata" is not used in this article, due to its inconsistency use in the literature.

hereafter.) Automatic extraction may employ sophisticated automatic indexing (e.g., Salton, & McGill, 1983; Hlava, 2002) and classification algorithms (e.g., Shafer, 1997, *see also* Scorpion Project at: <http://orc.rsch.oclc.org:6109/>) to improve the metadata quality.

A common extraction example is the “Web resource extract” (similar to an abstract) that many commercial search engines dynamically produce in response to a search. A fine distinction can be made between “extracts” and “abstracts” in that extracts are composed of a series of sentences taken directly from a resource’s content according to a *programmatically order* (e.g., the first few sentences of a document or the first sentence of each paragraph), and abstracts are constructed in an *intellectually logical order* (e.g., overview statement, methods, results, conclusions) (Lancaster, 1998; Johnson, 1995). Regardless of this distinction, both examples demonstrate “extraction” in that they draw from the document’s content.

*Harvesting*, the other key automatic metadata generation method, occurs when metadata is automatically collected from META tags found in the “header” source code of an HTML resource or encoding form another resource format (e.g., Microsoft WORD documents). The “harvesting process” relies on the metadata produced by humans or by full or semi-automatic processes supported by software. For example, Web editing software (e.g., Macromedia’s Dreamweaver and Microsoft’s Frontpage) and selected document software (e.g., Microsoft WORD and Adobe) automatically produce metadata at the time a resource is created or updated for “format,” “date of creation,” “revision date,” without human intervention. Software can also support a semi-automatic approach to metadata creation by presenting a person (e.g., resource author or Web architect) with a “template” that guides the manual input for “keywords” and “description” metadata, and additional metadata. The software automatically converts the metadata to META tags (or another tagged form depending on the document format) and places them in the resource header. These methods provide metadata that will not only aid searching, but can be harvested by a generator to create a structured metadata record.

Metadata extraction and harvesting are central to metadata generator functionality, and yet there is little analysis of the effectiveness of these techniques for creating Dublin Core metadata. If metadata initiatives are to benefit from automatic capabilities, researchers must examine how these methods impact metadata quality and identify specifically how these applications should be employed.

### **Research Goals and Questions**

The research presented in this paper explored the capabilities of two Dublin Core automatic metadata generation applications, *Klarity* and *DC.dot*, each with different algorithmic emphases. Research examined the quality of the results and determined if there were specific metadata elements that might generally be good candidates for automatic metadata generation with either or both of the generators. A larger goal was to consider how automatic processes may assist future metadata initiatives. Questions guiding the study were:

1. Can the application produce good quality metadata for any of the Dublin Core elements? If so, which elements?
  - a. Does the amount of textual content impact metadata quality?
  - b. Does the availability of source code META tags impact metadata quality?
2. Is there an overall difference between generators emphasizing extraction or harvesting?
3. How can automatic metadata generation applications be improved?

### **Method and Procedures**

The investigation was an experiment testing the performance of *Klarity* and *DC.dot*. The research was conducted over a one year period; the data was gathered March through May 2002, and the data analysis was completed in February 2003. An analysis

of each generator was conducted prior to the experiment to determine the processing algorithms.<sup>2</sup>

### *Dublin Core applications*

*Klarity* and *DC.dot* were selected for this research because they were among the most sophisticated generators accessible for research at the time of the investigation. Sophistication was defined by the ability to produce multiple Dublin Core elements. These generators were also selected due to their different processing emphases: *Klarity* emphasizes the extraction method and *DC.dot* emphasizes the harvesting method, although both generators have elements of each method.

*Klarity* is a commercial service that was produced by the Australian company tSA, now named Intology (Intelligent Technology) (<http://www.intology.com.au/>). The *Klarity* archive is found at: <http://archive.klarity.com.au/>. Dublin Core metadata is automatically generated when an identifier (e.g., URL) is submitted, and the metadata is converted into HTML META tags or eXtensible Markup Language (XML) within the Resource Description Framework (RDF) (Brickley and Guha, 2004). *Klarity* automatically generates metadata for the following five elements: Identifier, title, concepts, keywords, description. "Identifier" metadata is copied from the Web browser's "address prompt," "title" metadata is harvested from the resource source code, and "keywords" and "description" metadata are extracted from resource text. *Klarity*'s "concepts" element is a unifying concept representing keywords and functions more like a classificatory node. An algorithm based on term frequencies is matched against an underlying vocabulary to create this element. *DC.dot* does not have a "concepts" elements, so this *Klarity* function was not considered in the analysis. *Klarity*'s editor feature, where a human can manually enter additional metadata by answering a series of questions, was not analyzed because it was beyond the scope of this study.

---

<sup>2</sup> Both generators have been revised and enhanced since the research reported on here was conducted.

DC-dot (<http://www.ukoln.ac.uk/metadata/dcdot/>) is a generator developed by UKOLN (UK Office for Library and Information Networking) based at the University of Bath. DC.dot is open source and can be redistributed or modified under the terms of the GNU General Public License as published by the Free Software Foundation. DC-dot produces Dublin Core metadata, and can format output according number of different metadata schemas (e.g., USMARC, SOIF, IAFA/ROADS, TEI headers, GILS, RDF, and IMS). Metadata creation with DC-dot is initiated by submitting a URL. DC.dot, copies resource “identifier” metadata from the Web browser’s “address prompt,” and harvests “title,” “keywords,” “description,” and “type” metadata from resource META tags. If source code metadata is absent (meaning META tag are absent), DC-dot will automatically generate “keywords” by analyzing anchors (hyperlinked concepts) and presentation encoding, such as bolding and font size, but will not produce “description” metadata. DC.dot also automatically generates “type,” “format” and “date” metadata, and can read source code programming that automatically tracks date. For example, “last modified” might be coded as: “Last Modified” + `lm_day+’ ’+monthName[lm_month-1]+’ ’+lm_year`” for last updated date. This example illustrates DC.dot’s ability to partially conform with the recommended Dublin Core Qualifiers (see, DCMI Metadata Terms, <http://www.dublincore.org/usage/documents/overview/>). This enhanced feature of DC.dot was not examined in this study because Klarity does not produce equivalent output that would allow for a comparison. DC-dot, like Klarity, has an edit feature that enables a human to modify or enhance automatically generated metadata. DC-dot’s editor feature was not analyzed because it was beyond the scope of this study.

### *Web resource sample*

The sample consisted of a corpus of 29 resources obtained from National Institute of Environmental Health Sciences (NIEHS) Web site. The sample consisted of Web resources for which authors (scientists) created metadata in a larger study. Each resource had been archived during data gathering for the larger study (December 2001

through February 2002) at a maximum of four levels, or less, depending on the depth of the resource. The following examples illustrate two levels of a resource: The Environmental Genome Project resides at: <http://www.niehs.nih.gov/envgenom/>, with more specific information on the Polymorphism and Genes at: <http://www.niehs.nih.gov/envgenom/polymorf.htm>. These two pages represent the top level and a secondary level of a single Web resource and were archived at: <http://152.2.81.69.888/evaluation/envgenom/> and <http://152.2.81.69.888/evaluation/envgenom/polymorf.htm> respectfully. Web resource genre, textual density, and the availability of source code META tags were analyzed for each resource. These results are presented below in the data analysis section, *Web resource characteristics*.

### ***Metadata generation and evaluation***

The top-level identifiers (URLs) for each of the 29 archived Web resources were submitted to both Klarity and DC.dot, and metadata was automatically generated. Both applications tested only read the top level of the resource (that is, the single Web page represented by the URL submitted). The generated metadata was parsed and placed in a form for evaluation (see Appendix A, Example 1, "Metadata Evaluation Form"). Each metadata record was evaluated by three metadata professionals (hereafter referred to as evaluators). The evaluators were given a copy of the NIEHS Application Profile (Harper, et al, 2002), which is a Dublin Core-based metadata schema. The overriding metric was that the metadata being evaluated should facilitate resource discovery for the "intelligent health consumer" and "NIEHS staff." The evaluators were asked to assess the quality of each individual occurrence of each metadata element discreetly and according to a three-tier scale:

- *Good*. Good metadata accurately represented the resource and would facilitate accurate resource discovery for "intelligent health consumers" and "NIEHS staff." A good metadata element did not require any revision.

- *Fair*. Fair metadata would be somewhat useful for resource discovery of the resource being represented. In this case, a revision(s) would generally improve the quality of the metadata element.
- *Reject*. Reject (poor quality) metadata was inaccurate. In this case, the metadata element required substantial revision to be useful for resource discovery.

The individual evaluations for each element were aggregated per element and then averaged per metadata record.

In addition to the element evaluations, the evaluators were asked to answer questions about each metadata record's overall subject specificity and exhaustivity. This aspect of metadata was studied, given the prominence of subject searching viewed in NIEHS Web logs and the general popularity of subject searching on the greater World Wide Web. A final set of questions asked the evaluators about generator functionality and overall accuracy of metadata.

## Results

Data analysis focused on the characteristics of the Web resource sample and the quality of the metadata generated via both Klarity and DC-dot.

### *Web resource characteristics*

Web resource genre, textual density, and the availability of source code META tags were analyzed for each resource. The genre results are summarized in Table 1.

**Table 1: Web Resource Genre**

Resource Genre	Frequency	Percent
Program / division	8	27.6%
Research group / laboratory	14	48.3%
Fact sheet (resources providing a synopsis or overview of a topic)	3	10.3%
Research study (resources summarizing an investigation)	1	3.4%
Personnel information (resources about NIEHS staff, such as curriculum vitae, publication lists, and biographical statements)	3	10.3%

The sample consisted of resources for which authors created metadata in a larger study, and is not statistically representative of NIEHS' complete Web resource population.

A textual density analysis was conducted by classing the top-level Web page of each resource, as having “ample,” “fair,” or “limited” textual content. “Ample” indicated that at least two or more pages (based on computer screen pages) of text were available for automatic processing, “fair” indicated that at least a paragraph of text was available, and “limited” indicated that almost no text was present – and what was available was generally in the form of a hot-list of links to other Web pages and resources. The classification provided a more in-depth view of the Web resource sample. Web resource textual density results are summarized in Table 2.

**Table 2: Web Resource Textual Density**

Degree of content	Frequency	Percent
Ample	10	34.5%
Fair	11	37.9%
Limited	8	27.6%

A cross-tabulation was conducted to determine whether any relationship could be found between resource “genre” and “textual density.” The majority of resources in the “research group/laboratory” and “program/ division” genres (17 of 22 Web resources, 77.3%), had fair to limited amounts of textual content, while the majority of resources falling into the “fact sheet,” “research study,” and “personnel information” genre of (71.4%, 5 of 7) of resources classed had ample text. These results make sense because “research group/laboratory” and “program/ division” serve as pointers to more textually dense documents, such as a “fact sheet,” “research study,” and “personnel information.” These results suggest a relationship between Web resource genre and textual density. A larger sample would strengthen this conclusion. Textual density was also considered in examining the effectiveness of the metadata generators, and is discussed in the section below, Metadata Quality – An Element Assessment.

The final Web resource characteristic analysis focused on the availability of source code META tags. The source code was examined for the top Web page for all 29 Web resources processed via the two applications. Source code data for “title,” “description,” “keyword,” and “type,” metadata was recorded for this study. Results

found that all 29 Web resources (100%) had an HTML “title” tag, eight resources had a “description” META tag, 13 had a “keyword” META tag, and three had information on resource “type” in the source code. Results of this source code analysis were considered when analyzing the metadata quality evaluations for both applications.

## 6.2 Metadata Quality – An Element Assessment

The quality for “title,” “description,” and “keyword” metadata was evaluated for each metadata record generated via Klarity and DC-dot, and “type” metadata was evaluated for DC-dot.

*Title metadata.* All 29 of the resources tested had metadata in an HTML title tag. Given that both tools harvest this metadata from source code, it was anticipated that the quality assessment for this element would be exactly the same. The results of the evaluation (Table 3) were almost exactly the same, with a slight discrepancy in the results for two resources. In one case, DC.dot extracted a full title, including a subtitle, whereas Klarity only extracted the first part of a title. The resource’s title was lengthy (over 100 characters), and Klarity’s harvesting algorithm was found to have a character limit when extracting metadata from the “title” tag. In this case, the DC.dot title was evaluated to be “good,” and the Klarity title was evaluated to be “fair.”

**Table 3: Title Metadata**

Evaluation	Klarity	DC-dot
Good	12 (41.4%)	13 (44.8%)
Fair	10 (34.5%)	8 (27.6%)
Reject	7 (24.1%)	8 (27.6%)

In the second case, Klarity and DC.dot harvested the exact same title. No immediate reason for this result was apparent, so the full metadata records generated via both applications were examined. The Klarity metadata record included the harvested “title,” an extracted “description,” and 12 extracted “keywords” – all of which were consistently evaluated as being “fair.” The DC.dot metadata record was more sparse than the Klarity-generated metadata record; it included the harvested “title,” and only 5 “keywords” – all of which were evaluated as a “reject.” It is possible that the evaluation context provided by the full metadata record generated via both

applications influenced the title evaluation. In other words, if DC-dot had generated a more complete metadata record, which included a “description” element and perhaps a greater number of keywords, the DC.dot title may have been evaluated as being “fair.”

A final analysis of “title” found that, although the highest percent were evaluated as “good” (41.1% for Klarity and 44.8% for DC-dot), the majority of the results fell into the “fair” and “reject” categories. This result may be attributed to the fact that the titles, as part of the HTML markup, were created by NIEHS technical staff, and that they were often abbreviated, and several appeared akin to file names, rather than a formal title one finds on a resource.

*Description metadata.* Eight Web resources in the sample had a “description” META tag. Klarity extracts description metadata from the resource’s textual content, and DC.dot harvests description metadata from the META tag when it is available. Examples of “description” metadata generated via Klarity’s extraction algorithm and DC-dot’s harvesting algorithm, for the same Web resource, are given in the Appendix. (Compare the Klarity record in Appendix A to the DC.dot record in Appendix B.) Results for this analysis are given in Table 4, and suggest that if META tags are available for harvesting, as demonstrated by DC-dot, description metadata has a better chance of being evaluated as “good” compared to the description metadata generated via Klarity’s extraction algorithm. It is, however, difficult to generalize findings related to harvesting for the description element, given the small sample of eight records description META tags.

**Table 4: Description Metadata Evaluation Results**

generator→	For Web resources with “description” META tags		For Web resources w/out “description” META tags	
	Klarity	DC-dot	Klarity	DC-dot
<b>Good</b>	0	3 (37.5%)	1 (4.8%)	NA
<b>Fair</b>	4 (50.0%)	3 (37.5%)	9(42.9%)	NA
<b>Reject</b>	4 (50.0%)	2 (25.0%)	10(47.6%)	NA
<b>total no. of resources evaluated:</b>	8 Web resources	8 Web resources	20 Web resources	NA

\*Evaluations for one Web resource were not sufficient for Klarity’s output, thus the results reported on here are limited to 28 of the sample of 29 Web resources.

The description analysis considered whether textual density had an impact on metadata quality. This examination was limited to Klarity, given that this application’s extraction algorithm relies on the resource’s textual content. It was hypothesized that a greater amount of text would result in better quality metadata. Fifty percent (50%, 7 of 14) of the “description” evaluations that were “rejected” came from resources with ample text, while only 21.4% (3 of 14) the “descriptions” evaluations that were considered “fair” or “good” had ample text. These results suggest that the greater amount of resource text available, the more likely the description element produced by Klarity would be “rejected,” and is contrary to what was initially hypothesized.

**Keyword metadata.** Klarity extracts keyword metadata from document text, and DC-dot harvests this metadata from keyword META tags. If there are no keyword META tags, DC.dot appears to use resource anchors (hyperlinked concepts) and presentation encoding, such as bolding and font size, to automatically generate “keywords.” Table 5 summarizes evaluation results for the keyword element. Given that Klarity’s underlying algorithm is the same, whether or not keyword META tags are present, Klarity’s results have been combined (Table 5, column 2). DC-dot’s results are parsed by the presence or absence of keyword META tags (Table 5, column 3 and 4 respectfully).

**Table 5: Keyword Metadata Evaluation Results**

generator & META tag availability→	Klarity: for resource with & w/out "keyword" META tags	DC-dot: for resources with "keyword" META tags	DC-dot: for resources w/out "keyword" META tags
Good	0	1 (7.7%)	0
Fair	12 (41.1%)	7 (53.8%)	3(18.8%)
Reject	17 (58.6%)	5 (38.5%)	13(81.3%)
total no. of resources evaluated:	29 Web resources	13 Web resources	16 Web resources

The results suggest that keyword metadata harvested from humanly-generated META tags via DC.dot has a slightly greater potential of being evaluated as “good” compared with keyword metadata extracted from a document’s text via Klarity’s algorithm. Even so, over one-third (5 of 13, or 38.5%) of the harvested keywords (evaluations averaged per metadata record) were actually “rejected.” Klarity, which extracts keywords from document text failed, however, to elicit any “good” evaluations for this element, although approximately 40% of “keyword” evaluations were evaluated as “fair,” indicating that the algorithm was able to produce results that were somewhat acceptable. Even so, the majority of keywords produced with Klarity were rejected. It seems that Klarity’s algorithm extracted better keyword results than DC-dot for resources that did not have keyword META tags (see Table 5, column 4 for DC-dot results). More research is needed in this area and it must consider any modifications made to both applications since this study was conducted.

Special attention was given to keyword metadata *specificity* and *exhaustivity*, given that keyword searching (of a subject or topical nature) is a primary way people search for information on the Web (Ahronheim, 2002). Specificity examined the subject depth represented by the keywords, and exhaustivity focused on the extent of topics that the keywords represented. Keywords extracted via Klarity’s algorithm were neither sufficiently specific nor exhaustive. However, keywords harvested by DC.dot satisfied the measure of specificity and exhaustivity in three of the records generated; and keywords, in an additional record, were found to be exhaustive, although not specific.

The data analysis had intended to consider if textual density had an impact on specificity or exhaustivity – a reasonable question to examine for Klarity, which relies on document content to extract keywords. Given that none of the results generated via Klarity satisfied the measure of specificity or exhaustivity, however, this analysis was not pursued. It was not reasonable to pursue this question with DC-dot because this generator relies on keyword META tags, and if they are not available, additional source code features, such as anchors and presentation formatting are analyzed to generate keywords.

**Type metadata.** The last metadata element evaluated in this study was type. Three resources had identified “type” metadata in the source code. Only DC-dot automatically generated metadata for this element, and all 29 metadata records were evaluated as being “good.”

### ***Generator Functionality and Accuracy***

Beyond the element analysis, generator functionality and metadata accuracy were evaluated. In terms of functionality, the metadata generators were evaluated on their ability to correctly sort and label the metadata being generated. For example, a contributor’s names should *not* be labeled as a “keyword,” unless the resource was *about* the contributor. Klarity’s overall functionality results were quite promising, with a large majority of the metadata records (27 of 29, 93.1%) having the generated metadata labeled correctly. DC.dot found that a little under half metadata records (13 of 28, 46.4%) examined placed metadata in the correct field, and a little over half (15 of 28, 53.6%) labeled the generated metadata incorrectly.

The accuracy question was presented to get an overall evaluation of the quality of the metadata generated via both applications. Accurate metadata would support accurate resource discovery. Although the DC applications studied in this investigation were functional and the individual element evaluations found these applications useful, their accuracy evaluations were not ideal, with 25 of the 29 (86.2%)

metadata records evaluated as “poor” for Klarity, and 22 of 28 (78.6%) of the metadata records were evaluated as “poor” for DC.dot.

## Discussion

The Dublin Core applications examined in this study stand as important contributions to automatic metadata generation. The metadata generated, while not always ideal, appears useful for resource discovery in many cases. The fact that both applications have edit features indicates that the automatically generated metadata does not necessarily need to be viewed as an end-product. However, this study viewed the generated metadata as an “end-product” in order to further our understanding of automatic processing and to identify how automatic metadata generation applications might be improved.

The metadata applications tested in this study only read the top level of the sampled Web resources. That is, extraction and harvesting were based on a single, top-level, Web page represented by the URL submitted, rather than the multiple levels that often comprise a Web resource. The focus on a single Web page appeared to be the state-of-the-art for Dublin Core-based applications available at the time the research was conducted. An application reading multiple Web pages (the different levels that comprise a single resource) would likely produce different and perhaps better results. A human-computer interaction (HCI) component or additional programming could support this sort of processing. In the HCI arena, a person could inform an application about the measure of *generator feed*; that is, how many levels of the resource to analyze for automatic metadata generation. Another option would be to program the extraction/harvesting algorithm to read a Web address and a selected number of extended levels. The overriding challenge here is in determining what actually comprises a Web resource – an issue that the metadata community has been grappling with since the advent of the Web. The recent *Functional Requirements for Bibliographic Records* (International Federation..., 1998) may offer promise in this area, and it is recommended that automatic metadata generators explore this model.

Related to understanding what a “Web resource is” are Web resource characteristics. This study examined Web resource “genre” and “textual density” and the relationship between these two characteristics. The results indicate that Web-resources classed as a “division/program” and “research group/laboratory,” which are top-level units in the NIEHS organizational hierarchy, are less textually dense than Web resources classed as “study information,” “fact-sheet,” and “personnel information,” which are from lower-level units in the NIEHS organizational hierarchy. These results make sense in that top-level organizational units are skeletal umbrellas that oversee a number of specific units or activities. Web-resources representing top-level organizational units are generally synoptic, containing summaries and lists of links to lower-level Web resources that are more textually dense (e.g., a fact-sheet about a project or an employee’s vitae). It’s likely that a “genre/ textual density” analysis of another organization’s Web resources would yield similar results.

The “genre/textual density” analysis connected to the element analysis, where it was found that “description” metadata was more frequently rejected for Web resources with greater textual density. On one hand, these results are specific to Klarity’s extraction method—and a generator with a different extraction algorithm may produce more positive results for textually dense resources. On the other hand, it is also possible that text is a better source for extraction when pulled from less textually dense resources. In other words, *less is more* (resources with less text contains more significant content), and textually dense Web resources may have a tendency to be too extensive for extraction. Research ought to explore how a textual density analyses (automatic counting of terms) and automatic genre identification, facilitated via electronic organizational charts, could be combined and aid in the selection of appropriate “extraction” methods. With more knowledge in this area, an algorithm similar to the one underlying Klarity might be automatically employed to Web resources representing lower level organizational units, while another algorithm may be automatically employed to resources found at different levels in the organizational hierarchy.

The “title” metadata examination in this study suggests that character limitations in applications can impact metadata quality. As presented in the results section above, DC-dot extracted a full title, including a subtitle, for one resource, where Klarity only extracted the first part of a title. In this case, the DC-dot title was evaluated as “good,” and the Klarity title was evaluated as “fair.” It was surprising that the applications harvested different titles from the HTML title tag. Effective applications, it seems, should not limit the amount of text harvested from certain HTML tags, particularly for “title.” A character limit may, however, be very useful when harvesting metadata from “description” and “keyword” or other META tags.

Central to this study was the “harvesting” and “extraction” comparison. The DC.dot / Klarity comparison suggests that “description” and “keyword” metadata have a slightly better chance of being evaluated as “good” when it is harvested from META tags, than when it is extracted via Klarity’s algorithm. The results demonstrate that humans are an important asset in metadata generation and suggest that humanly generated metadata should be harvested whenever it is available. A counter argument could be presented when considering metadata abuse, that is the excessive or unethical use of metadata for resource representation. Digital signatures, whereby an agency’s stamp indicates a degree of metadata quality assurance could help address this abuse problem. The support for human generated metadata does not extend to the use of HTML anchors and other presentation encoding for extracting keywords, as demonstrated by DC.dot, despite the fact that a human oversees this design aspect.

Although a portion of the harvested metadata examined in this immediate study was evaluated as good, it was not always optimal. A consideration here is that the metadata was created by NIEHS technical staff (e.g., Web architects and database and systems personnel), rather than people trained as metadata professionals. It would be useful to know if NIEHS technical staff involved in this activity understand how metadata, embedded in the header of an HTML resource, is used by search engines for information retrieval, and if they understand how an application might harvest this metadata for representation. The immediate results presented in this paper highlight a

need to improve human understanding of the value of META tags for information retrieval and harvesting.

The Klarity algorithm emphasizes extraction. Although Klarity failed to elicit any “good” evaluations for “keyword” metadata, approximately 40% of the keywords generated were evaluated as “fair.” Thus, Klarity's algorithm was able to produce somewhat acceptable keyword metadata. Klarity's results for “description” metadata were less promising. The difference found between the results for these two elements are clues that research on extraction methods should continue. Automatic indexing and classification, natural language processing, and machine-learning have an array of algorithms, general to domain-specific, that ought to be integrated into applications that generate that generate subject and other metadata elements, such as those examined in this study.

## **Conclusion**

This research indicates that generators, using both extraction and harvesting methods, have the potential to create useful metadata. Although the sample was small and the number of Dublin Core elements examined was limited, the results help identify how applications might be improved and highlight important areas of automatic metadata generation research.

Among some immediate suggestion for improving automatic metadata applications are the following:

- The level of Web resource analysis should extend beyond the top-level Web page. Human-computer interaction and computer programming options could be used to define what a “Web resource” is and support this type of analysis.
- Web resource genre and textual density should be used to determine the appropriate extraction algorithm. It seems that less textually dense resources are probably better candidates for extraction via an algorithm similar to that supported by Klarity.

- Character limitations, initially programmed into applications, should be eliminated for certain metadata elements, such as “title.”
- Automatic metadata applications should harvest metadata initially created by humans, even if the emphasis is on extraction.

Several areas of research were also highlighted in this study. Automatic metadata generation applications should explore how to incorporate FRBR and current thinking on what denotes a Web resource. As indicated above, HCI and programmatic research could aid developments here. Research needs to test the effectiveness of different algorithms for automatic metadata generation. The field of automatic indexing is rich, with many algorithms, ranging from simple/general domain to domain-specific. Exploring how to incorporate these algorithms and integrate them into applications could improve automatic processes for more intellectually demanding metadata, such as “keywords” and “description.”

Finally, research using automatic techniques needs to explore the optimal time for human intervention in the metadata creation sequence. The template features supported by both Klarity and DC.dot were not explored in this study, due to practical considerations. However, their existence indicates that human input is a likely step in creating optimal metadata. The results discussed above and the mention of this feature together invite questions about the timing and sequence of human involvement in an automatic metadata creation activity. Should the human start the metadata creation process, by perhaps assigning keywords, and then initiate an automatic sequence, or should the automatic processing take place first, and then allow for human input? Factors, such as metadata creator experience, Web resource genre, and Web resource textual density, may help to determine the appropriate heuristics. What is clear is that, as stated in the beginning of this paper, the best metadata generation option is to integrate both human and automatic processes. In conclusion, and specific to this study, is that integrating extraction of harvesting methods will be the best approach to

creating optimal metadata, and more research is needed to identify when to apply which method.

## References

Ahronheim, J. A. (2002). Introduction: High-level Subject Access Tools and Techniques in Internet Cataloging. *Journal of Internet Cataloging*, 5(4): 1-4.

Anderson J. & Perez-Carballo. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. part I: research, and the nature of human indexing. *Information Processing Management*, 7: 231-254.

Brickley, D. & Guha, R.V. (Eds.). (2004). RDF Vocabulary Description Language 1.0: RDF Schema. (W3C Recommendation 10 February 2004):  
<http://www.w3.org/TR/2004/REC-rdf-schema-20040210/>.

Craven, T. (2001). DESCRIPTION meta tags in public home and linked pages. LIBRES: library and information science research. 11 (2):  
<http://libres.curtin.edu.au/LIBRE11N2/index.htm>.

Dublin Core Metadata Element Set, Version 1.1: Reference Description. (2003):  
<http://dublincore.org/documents/2003/06/02/dces/>.

Greenberg, J., Crystal, A., Robertson, D., & Leadem, E. (2003a). Iterative Design of Metadata Creation Tools for Resource Authors. In Sutton, S., Greenberg, J., and Tennis, J. (Eds.). 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice—Metadata Research & Applications. September 28 - October 2, 2003, Seattle, Washington. New York: ERIC Clearinghouse on Information and Technology:  
[http://www.siderean.com/dc2003/202\\_Paper82-color-NEW.pdf](http://www.siderean.com/dc2003/202_Paper82-color-NEW.pdf).

Greenberg, J. (2003b). Metadata and the World Wide Web. In *Encyclopedia of Library and Information Science*, 2nd ed. vol. 3 (New York: Marcel Dekker, 2003): 1876-88

Greenberg, J., Pattuelli, M. C., Parsia, B., & W. D. Robertson. (2001). Author-generated Dublin Core metadata for Web resources: a baseline study in an organization. *Journal of Digital Information (JoDI)*, 2(2):

[http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/.](http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/)

Harper C.A. & Sharfe, E. (2003). DCMI Bibliography:

<http://dublincore.org/documents/2003/08/26/usageguide/bibliography.shtml>.

Harper, C.A., Greenberg, J., Robertson, D.W., and Leadem, E. (2002). Abstraction versus Implementation: Issues in Formalizing the NIEHS Application Profile. *DC-2002: Metadata for e-Communities: Supporting Diversity and Convergence. Proceedings for the International Conference on Dublin Core and Metadata for e-Communities, 2002, Florence, Italy. October 13-17. Firenze University Press, pp. 213-215. Also available at:*

<http://www.bncf.net/dc2002/program/ft/poster7>.

Hlava, M. (2002). Automatic indexing: A matter of degree. *Bulletin of ASIS&T*, 29(1):

<http://www.asis.org/Bulletin/Oct-02/hlava.html>.

International Federation of Library Associations and Institutions. (1998). Functional Requirements for Bibliographic Records. <http://www.ifla.org/VII/s13/frbr/frbr.pdf>

Johnson, F. (1995). Automatic abstracting research. *Library Review*, 44(8): 28-36.

Lancaster, F.W. (1998). *Indexing and abstracting in theory and practice*. Champaign, IL: GSLIS, University of Illinois.

Liddy, E. D., Sutton, S. A., Paik, W. Allen, E., Harwell, S. Monsour, M., Turner, A, & Liddy J. (2001). Breaking the metadata generation bottleneck: preliminary findings. *JCDL 2001*: 464.

Salton, G. & McGill, M.J. (1983). Text analysis and automatic indexing. In: *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, pp. 52-117.

Schwartz, C. (2002). *Sorting out the web: approaches to subject access*. Westport, Connecticut: Ablex publishing. Part of the *Contemporary Studies in Information Management, Policies, and Services* series by Hernon, P. (Ed.).

Shafer, K. (1997). Scorpion Helps Catalog the Web. *Bulletin of the American Society for Information Science*, 24(1): <http://www.asis.org/Bulletin/Oct-97/>.

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science* 37: 331-340.

Trigg, R. H., Blomberg, J., & Suchman, L. (1999). Moving document collections online: the evolution of a shared repository. Paper presented at the Sixth European Conference on Computer-Supported Cooperative Work, Copenhagen, Denmark, September 12 - 16, 1999, pp. 331-350.

### **Acknowledgements**

I would like to acknowledge Microsoft Research and OCLC, Online Computer Library Center for funding that made this research possible, and the following people for helping to run the study and conduct statistical analyses: Davenport Robertson (NIEHS), Ellen Leadem (NIEHS), Cathy Zimmer (Odem Institute for Research in Social Science), Abe Crystal (SILS/UNC), and Michelle Mascaro (SILS/UNC).

Appendix A, Example 1 part 1: Web Evaluation Form (Also, Klarity output that was evaluated for Web resource 29)

<b>TITLE</b>	Impact Of Environmental Exposures On Special Populations (Women, Children and Minorities)	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	Environment	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	pollution	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	environmental	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	exposures	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	health	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	chemicals	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	disease	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	affected	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	reduce	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	dose	<input type="checkbox"/> Good <input checked="" type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	effects	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>DESCRIPTION</b>	By understanding the combined effects of exposure, genetics and age or timing of exposure, researchers supported by NIEHS are working to reduce the burden of environmentally induced disease in populations most affected by environmental pollution. particular chemicals, high dose) or characteristics which make them more susceptible to the effects of pollutants (genetic susceptibility, metabolism, gender differences). NIEHS encourages research on the environmental determinants of human disease in populations which may be particularly susceptible to such exposures.	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject

## Appendix A, Example 1, part 2

1. In general, does this record's metadata appear in correct data fields? (For example, identifier [URL] metadata is not given in date-created field.)  Yes  No

2. Do you think the keywords given are specific enough?  Yes  No

3. Do you think the keywords sufficiently cover all the subjects on the webpage?  Yes  No

4. Based only on the metadata content provided (not the metadata that you think is missing), how good do you think the metadata given is? (Good metadata supports accurate resource discovery.)  Good  Fair  Poor

5. How complete is the metadata record? (Does it provide a good, comprehensive surrogate?)

Very complete  In-between  Very limited

## Appendix B: Example 1, NIEHS Web resource no. 29



The screenshot shows a web browser window with the address bar containing the URL: <http://www.niehs.nih.gov/dert/programs/special/specpops.htm>. The page header includes the NIEHS logo and the text "Division of Extramural Research and Training". A navigation menu contains links for Home, About DERT, Funding Opportunities, Research Programs, Training, Outreach and Educational Programs, and Environmental Health Info. The main content area is titled "Special Emphasis Areas" and "Impact Of Environmental Exposures On Special Populations (Women, Children and Minorities)". The text describes NIEHS's focus on research in vulnerable populations and lists three groups: Women, Children, and Minority Populations. Contact information for Gwen Collman is provided at the bottom.

## Appendix B, Example 2, Header Source Code for Web resource 29

```
<!doctype html public "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
<head>
<title>Impact Of Environmental Exposures On Special Populations
(Women, Children and Minorities) - DERT</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<meta name="description" content="Funding Opportunities - Division of Extramural Research
and Training">
<meta name="keywords" content="funding opportunities, funding">
<link href="../../dert.css" rel="stylesheet" type="text/css">
</head>
```

Appendix B, Example 3, DC-dot output evaluated for Web resource 29 (\*note, DC-dot output for “date,” “format,” and “identifier” were eliminated from the evaluation due to practical research constraints.)

<b>TITLE</b>	Impact Of Environmental Exposures On Special Populations (Women, Children and Minorities)	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	funding opportunities	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>SUBJECT</b>	funding	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>DESCRIPTION</b>	Funding Opportunities - Division of Extramural Research and Training	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject
<b>TYPE</b>	text/html	<input type="checkbox"/> Good <input type="checkbox"/> Fair <input type="checkbox"/> Reject

1. In general, does this record's metadata appear in correct data fields? (For example, identifier [URL] metadata is not given in date-created field.)  Yes  No

2. Do you think the keywords given are specific enough?  Yes  No

3. Do you think the keywords sufficiently cover all the subjects on the webpage?  Yes  No

4. Based only on the metadata content provided (not the metadata that you think is missing), how good do you think the metadata given is? (Good metadata supports accurate resource discovery.)  Good  Fair  Poor

5. How complete is the metadata record? (Does it provide a good, comprehensive surrogate?)

Very complete  In-between  Very limited