Developing an Automatic Metadata Harvesting and Generation System for a Continuing Education Repository: A Pilot Study

Jung-Ran Park[1], Akshay Sharma[1], Houda El Mimouni[1]
[1]Drexel University, College of Computing and Informatics

Abstract
The goal of this pilot study is to assess the effectiveness and reliability of an automated metadata generation and harvesting system developed for a project repository which hosts continuing education resources for cataloging and metadata professionals. Using a web crawler developed for the repository, 500 web resources are selected as seed pages for metadata extraction and generation. This paper summarizes the processes as well as the results of the study. The metadata harvesting system combined with powerful article analysis and data generation tools such as Adlegant's Article Anaylsis API produces significant improvement in metadata generation.
Keywords: Automated metadata generation; metadata harvesting system; web crawler; continuing education repository
doi: 10.9776/16498

Acknowledgements: This study is supported through an award from IMLS 21st Century Laura Bush program for the project entitled *Building a Workforce of Information Professionals for 21st Century Global Information Access* for a three year period (2014-2017).
Contact: Author will add e-mail address.

1 Introduction

Considering the overwhelming proliferation of digital resources managed by libraries and the relatively high cost of generating manual metadata, the utilization of (semi)automatic metadata generation is crucial for the development of sustainable digital repositories (Park & Brenza, 2015; Park & Lu, 2009; Greenberg et al., 2006). The goal of this pilot study is to assess the accuracy and reliability of an automated metadata generation and harvesting system developed for our IMLS funded project repository. Using a web crawler developed for the repository, 500 continuing education web resources are selected (Park et al, in press). This paper summarizes the processes and the results of the metadata extraction, harvesting and generation for the selected web resources.

2 Lack of Consistency

One of the main challenges for harvesting metadata is the lack of uniformity in terms of the format, type and extent of availability of metadata. The resources with HTML meta tags present further challenges due to the multiple sources of the metadata including Highwire Press tags, Facebook's OpenGraph (OG) tags, PRISM tags, Dublin Core tags, Twitter tags, uncommon proprietary tags (e.g. *SailThru*) as shown in Figure 1 below.

```
66
67                    <meta name="citation_publisher" content="IEEE">
68
69
70
71                    <meta name="citation_author" content="Camp, L. Jean">
72
73
74                  <meta name="citation_author_institution" content="Indiana Univ.,
75
76
77
78                    <meta name="citation_title" content="Digital identity">
79
80
81                    <meta name="citation_date" content="Fall 2004">
82
```

Highwire Press® Citation Meta Tags

```
1457 <meta property="og:type" content="website" />
1458 <meta property="og:title" content="Digital identity" />
1459 <meta property="og:description" content="The lack of conceptual clarity reflected by the overload
     management. Thus we clarifies the nature of identification in a digital networked world, as oppose
1460 <meta property="og:url" content="http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=133788
1461 <meta property="og:image" content="http://ieeexplore.ieee.org/assets/img/logo-ieee-200x200.png" />
1462 <meta property="og:site_name" content="IEEE Xplore" />
1463 <meta property="fb:app_id" content="179657148834307" />
```

Facebook's OpenGraph Meta Tags

```
<meta name="Description" content="IEEE Xplore. Delivering full text ac
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />

<title>IEEE Xplore Abstract
```

basic HTML meta tags

Figure 1. Source Code of an IEEE Webpage

For resources lacking metadata elements, we used third party tools and APIs to generate metadata and mapped them to Dublin Core metadata elements. The extraction of technical metadata is fairly routine inasmuch they comprise basic file details such as file name, size, MIME type and format of the resource; technical metadata elements are available through most file I/O APIs. File Size and MIME type fields were merged and stored in DC.format.

3 The Workflow for Metadata Extraction

The workflow for metadata extraction consists of the following steps:

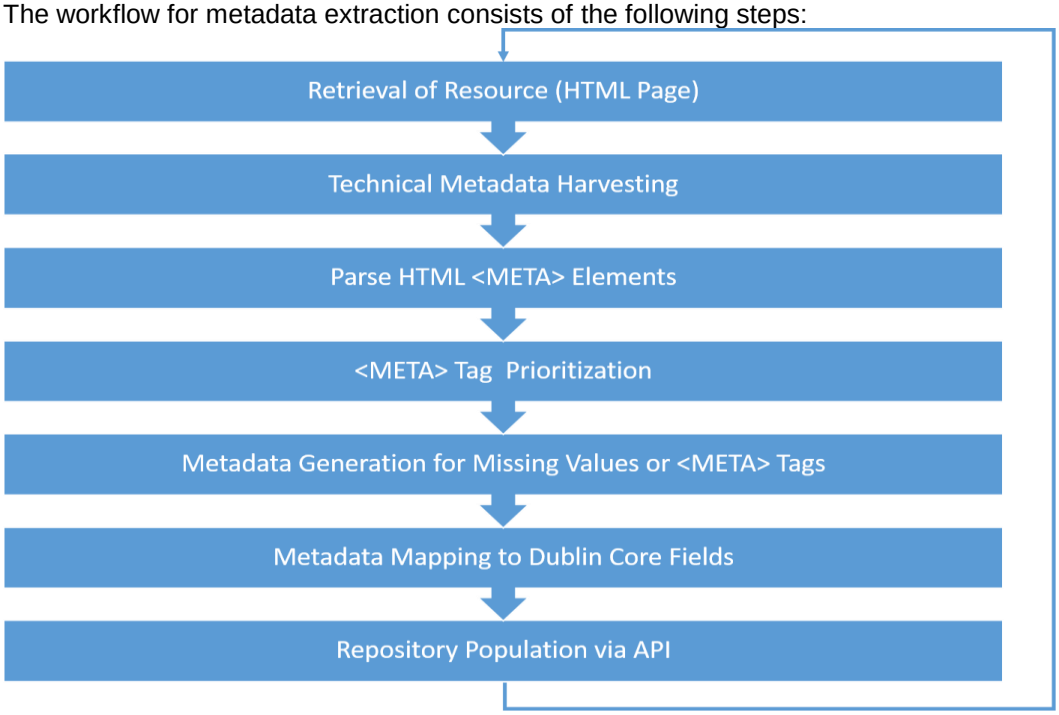


Figure 2. Metadata Harvesting and Generation System Workflow Diagram.

Prior to initializing the metadata harvesting and generation system, a custom crawler is used to collect continuing education web resources. The crawler requires an initial set of URLs referred to as "seed URLs". The resources present at the "seed URLs" are parsed for links which are then recursively crawled, in turn fetching more links. The crawler then generates CSV files listing the links of the relevant resources.

The metadata generation system reads the crawler-generated CSV file to produce metadata as follows:

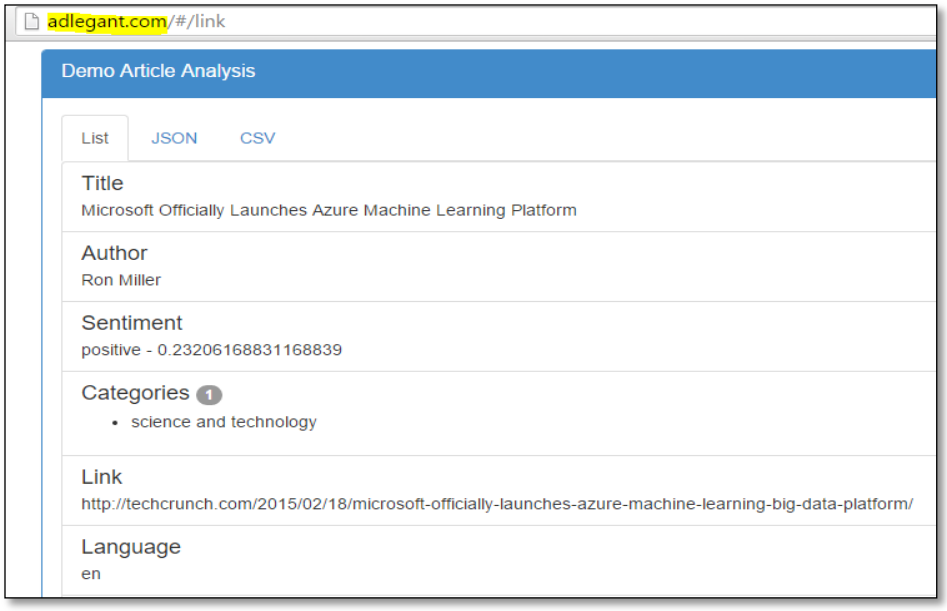
a) Retrieval: For every URL in the CSV file, the resource (file) is retrieved and stored by the system.
b) Technical Metadata Harvesting: Technical metadata is extracted from the file including MIME type, size and filename. Dates (created, modified, accessed, etc.) are not extracted as these are

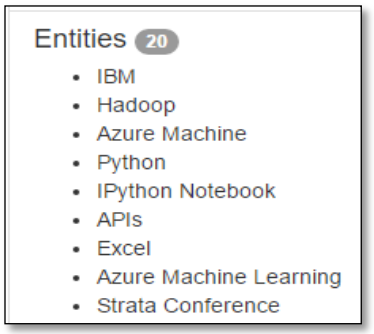
altered when the resource is retrieved from the URL. File creation date does not represent the date of publication or creation of the resource.

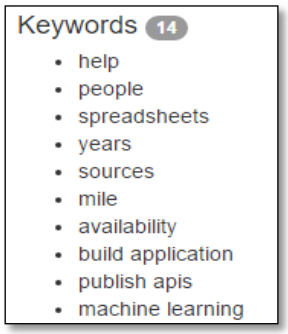
c) Meta Elements Processing: The HTML tags in the resource are parsed.

d) Prioritization: A priority order is followed to address conflicts and redundancy when two or more meta tags of different sources are found for the same field. Preference to certain meta tags over others is given based on the following order:
   - Basic HTML: find basic HTML tags such as *title, author* and *description* of the document.
   - Highwire Press®: find Highwire Press® Citation tags used in the web resources published by IEEE, ACM and other research community websites. Once found, overwrite any previously captured relevant values with the ones present in the Highwire Press® tags.
   - PRISM Tags: find PRISM tags. Once found, only store the values that are not previously captured.
   - Dublin Core (DC) Tags: find Dublin Core (DC) tags and prioritize them over others.
   - OpenGraph (OG) Tags: Lastly, the harvesting program searches for Facebook's OpenGraph tags.

e) Generation: For resources lacking data elements such as *title, author (creator), publisher, subject* or *description* the generator is called. The Adlegant's Article Analysis API, available free of cost, is used for the generation and population of keywords. Our choice of Adlegant's API was largely based upon the quality of classified textual data responses returned by the API, the speed of the service, the restriction-free license, and the ease of use of the API. The service being a REST API is platform-independent and requires minimal overhead to work with our existing metadata extraction infrastructure. The API simply fetches the resource from a URL provided as input and mines the resource to return structured data. The primary data elements returned by the API include *author, language* and the *subject* categories of the resource. A summary of the text and most relevant keywords ("tags") are also returned. The figure 3 below illustrates this:







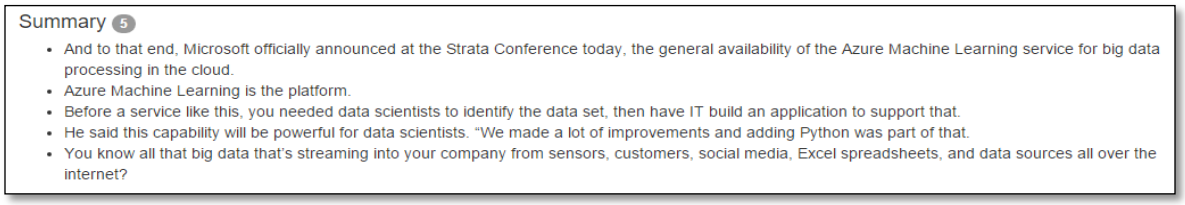


Figure 3. Screenshots Summarizing Different Types of Metadata Obtained from Adlegant's Article Analysis API.

f) Mapping: Before populating the repository with the resource URLs and their metadata, the extracted metadata are mapped to the corresponding Dublin Core fields.

g) Populate Repository: Populate the repository with the metadata, resource identifier (URL) via Omeka's API.

4 Analytics

4.1 Metadata Harvesting Success Rates

The success rates of the extraction (harvesting) of metadata, without making use of any 3rd party generation tools or techniques, are as follows:

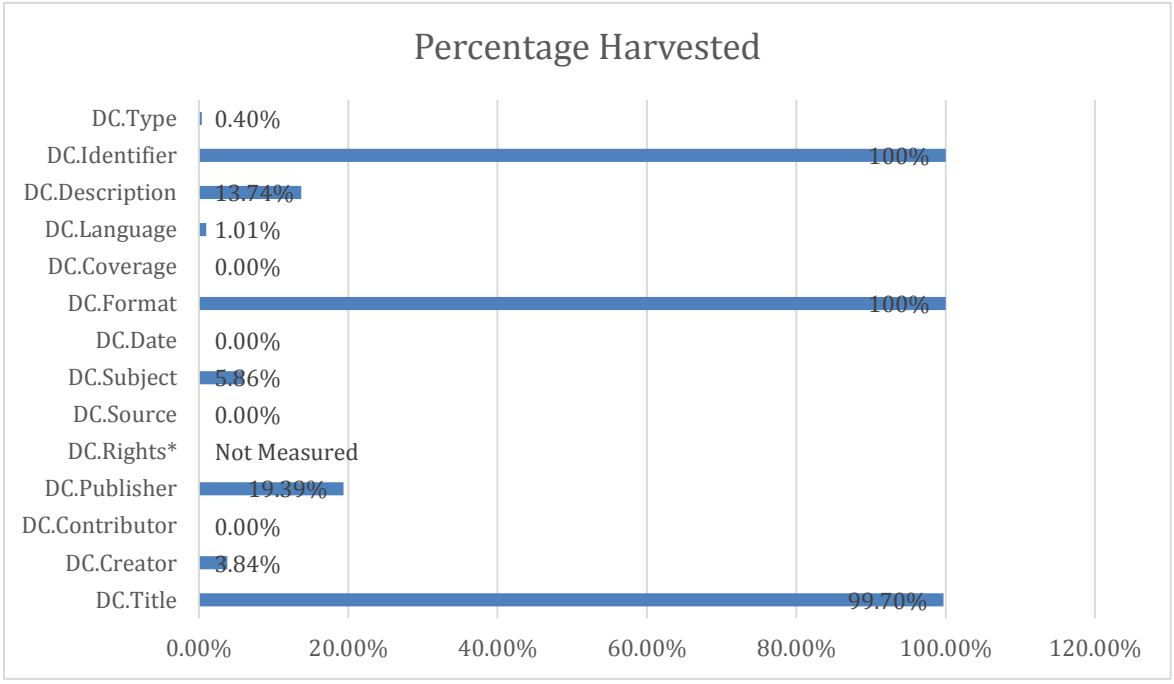


Figure 4. Metadata Extraction without Making Use of the Third Party Generation Tools

4.2 Metadata Harvesting and Generation Success Rates

The metadata harvesting system combined with powerful article analysis and data generation tools such as Adlegant's Article Analysis API produces significant improvement of the results as shown below:

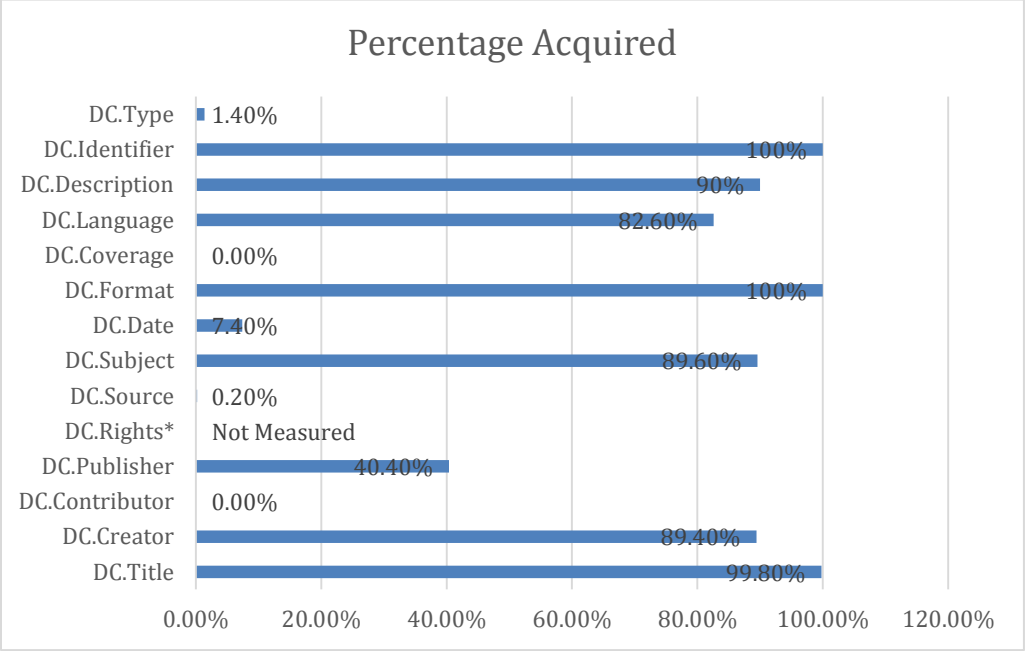


Figure 5. Metadata Extraction Making Use of Third Party Generation Tools

5 Conclusion and Further Improvements

In this pilot study, we examined the workflow processes for automatic metadata generation. The study presents challenges of automatic metadata generation owing to the lack of consistency and uniformity of metadata format and type and the lack of metadata in web resources. The metadata harvesting system combined with powerful article analysis and data generation tools such as Adlegant's Article Anaylsis API produces significant improvement in metadata generation. One of the areas for further improvement concerns speed optimizations. Quality evaluation of automatically generated metadata is also a critical next step for improving the metadata harvesting system.

6 References

Greenberg, J., Spurgin, K., & Crystal, A. (2006). Functionalities for automatic metadata generation applications: A survey of metadata experts' opinions. *International Journal of Metadata, Semantics and Ontologies*, 1(1), 3−20.

Park, J. R., Yang, C., El Mimouni, H., & Ping, Q. (in press, 2015). Developing an automatic crawling system for populating a digital repository of professional development resources: A pilot study. *Journal of Electronic Resources Librarianship*.

Park, J. R., & Brenza, A. (2015). Evaluation of (semi)automatic metadata generation tools: A survey of the current state of the art. *Information Technology and Libraries, 34 (3).* doi: 10.6017/ital.v34i3.5889

Park, J. R., & Lu, C. (2009). Application of semi-automatic metadata generation: Types, tools, and techniques**.** *Library & Information Science Research,* 31(4), 225–231.