A Metadata Generation System for Scanned Scientific Volumes

Xiaonan Lu[1], Brewster Kahle[2], James Z. Wang[1]*, and C. Lee Giles[1]
[1]The Pennsylvania State University, University Park, PA
[2]Internet Archive, San Francisco, CA
xlu@cse.psu.edu, brewster@archive.org, jwang@ist.psu.edu, giles@ist.psu.edu

ABSTRACT

Large scale digitization projects have been conducted at digital libraries to preserve cultural artifacts and to provide permanent access. The increasing amount of digitized resources, including scanned books and scientific publications, requires development of tools and methods that will efficiently analyze and manage large collections of digitized resources. In this work, we tackle the problem of extracting metadata from scanned volumes of journals. Our goal is to extract information describing internal structures and content of scanned volumes, which is necessary for providing effective content access functionalities to digital library users. We propose methods for automatically generating volume level, issue level, and article level metadata based on format and text features extracted from OCRed text. We show the performance of our system on scanned bound historical documents nearly two centuries old. We have developed the system and integrated it into an operational digital library, the Internet Archive, for real-world usage.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Collection, Systems issues*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Design, Experimentation

Keywords

Metadata Generation, Scanned Journal

*J. Wang is also affiliated with Carnegie Mellon University, Pittsburgh, Pennsylvania.

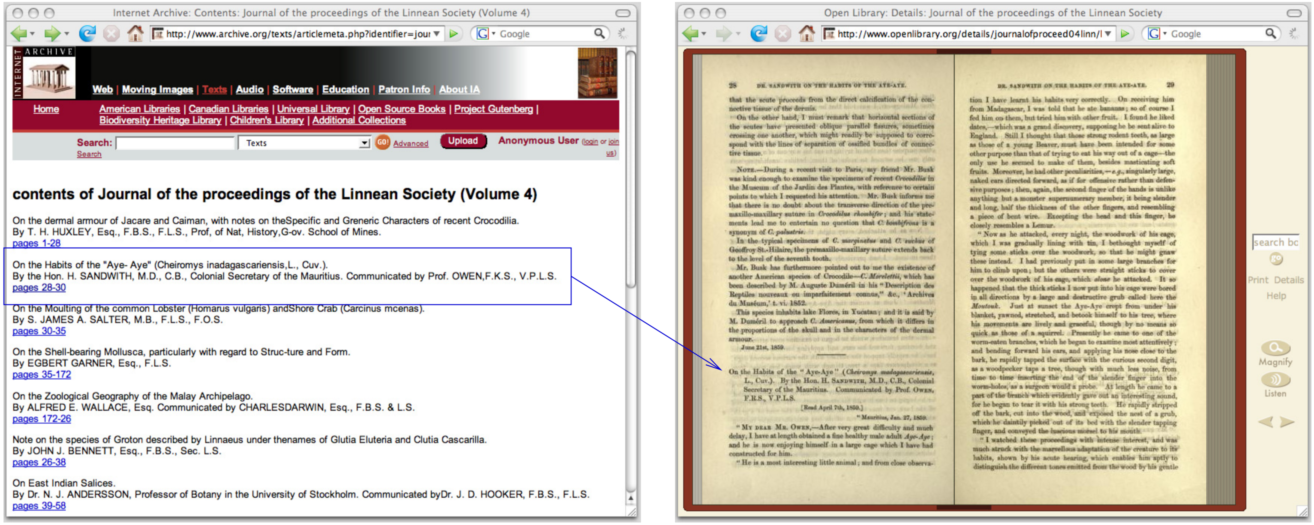
JCDL'08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.
Copyright 2008 ACM 978-1-59593-998-2/08/06 ...$5.00.

1. INTRODUCTION

Large scale digitization projects are underway at public and commercial digital libraries to preserve cultural artifacts in digital format and to provide easy web access. As one example, ten major natural history museum libraries, botanical libraries, and research institutions have joined to form the Biodiversity Heritage Library (BHL). The BHL partners will digitize the published literature of biodiversity held in their respective collections and provide basic and important content for immediate research and for multiple bioinformatics initiatives [1]. Another example in the public domain, the Universal Library Project (also called the Million Book Project) [8] has completed the scanning of one million books and has made the entire database accessible. In the commercial domain, the database underlying Google Book Search [5] continues to grow, with more than a hundred thousand titles added by publishers and authors and some 10,000 works in the public domain now indexed and included in search results. Managing this growing amount of digitized resources and providing efficient content access to them present challenging issues for digital libraries.

For many digitized works, such as scanned books, journals, and diaries, it is not enough to merely display the work [10] and provide sequential access to its content. Users often need to navigate through the work. For example, imagine a bound volume of published journal consisting of hundreds of digital files, each one the scan of a single page. A user may need to view the table of contents, and then switch to a particular article. Or a user may need to switch to the bibliography to read a citation as the article is being read and then come back to the initial location. It is necessary for digital libraries to provide these convenient content access functionalities for digitized resources, which further requires tools and methods that will efficiently analyze large collections of digitized resources.

In digital libraries, digitalized resources are often compound objects consisting of a large number of scanned images of pages, OCRed text, and viewable PDF files generated by automatic scanning and recognition processes. In contrast, there is usually only a very limited amount of manually-generated metadata available to describe the structure and content of digitized resources. Thus, digital libraries could effectively use automated tools to generate metadata for digitized resources, to describe the intellectual content of compound objects and to connect different components [24]. The extracted metadata will enable a broad range of content access functionalities, including


Internet Archive: Contents: Journal of the proceedings of the Linnean Society (Volume 4)
http://www.archive.org/texts/articlemeta.php?identifier=jour
Google
INTERNET ARCHIVE
Web | Moving Images | Texts | Audio | Software | Education | Patron Info | About IA
Home
American Libraries | Canadian Libraries | Universal Library | Open Source Books | Project Gutenberg |
Biodiversity Heritage Library | Children's Library | Additional Collections
Search:
Texts
GO!
Advanced Search
Upload
Anonymous User (login or join us)
Open Library: Details: Journal of the proceedings of the Linnean Society
http://www.openlibrary.org/details/journalofproceed04linn/
Google
search bc
Print
Details
Help
Magnify
Listen

contents of Journal of the proceedings of the Linnean Society (Volume 4)

On the dermal armour of Jacare and Caiman, with notes on theSpecific and Greneric Characters of recent Crocodilia.
By T. H. HUXLEY, Esq., F.B.S., F.L.S., Prof, of Nat, History,G-ov. School of Mines.
pages 1-28

On the Habits of the "Aye- Aye" (Cheiromys inadagascariensis,L., Cuv.).
By the Hon. H. SANDWITH, M.D., C.B., Colonial Secretary of the Mauritius. Communicated by Prof. OWEN,F.K.S., V.P.L.S.
pages 28-30

On the Moulting of the common Lobster (Homarus vulgaris) andShore Crab (Carcinus mcenas).
By S. JAMES A. SALTER, M.B., F.L.S., F.O.S.
pages 30-35

On the Shell-bearing Mollusca, particularly with regard to Struc-ture and Form.
By EGBERT GARNER, Esq., F.L.S.
pages 35-172

On the Zoological Geography of the Malay Archipelago.
By ALFRED E. WALLACE, Esq. Communicated by CHARLESDARWIN, Esq., F.B.S. & L.S.
pages 172-26

Note on the species of Groton described by Linnaeus under thenames of Glutia Eluteria and Clutia Cascarilla.
By JOHN J. BENNETT, Esq., F.B.S., Sec. L.S.
pages 26-38

On East Indian Salices.
By Dr. N. J. ANDERSSON, Professor of Botany in the University of Stockholm. Communicated byDr. J. D. HOOKER, F.B.S., F.L.S.
pages 39-58

Automatically–generated Table of Content

The Article Display Window

Figure 1: Article navigation services in test mode at the Internet Archive.

turning pages; navigating to a particular chapter, section, or page in a venue; navigating to a particular article in a bound volume of scholarly publications; or switching views among multiple formats including image and text. These user-friendly services will improve some of the needed accessibility of intellectual content.

In this paper, we present our work on automatic metadata generation for bound volumes of scientific publications. The main goal is to generate structural and descriptive metadata for digitized volumes and support various content access functionalities. Compared with related work on extracting metadata for a single article, our work attempts to solve specific problems associated with digitized volumes. A bound volume of journals contains rich content, including the hierarchy of issues, articles and various types of pages. Different from born-digital documents, the digitized documents do not have reliable formatting information, such as fonts for specific textual content. Based on our manual check, high variation of elements and styles are common in scanned historical journals. In addition, the accuracy of the OCR process is relatively low for historical volumes since they have old type and formatting styles. All these characteristics present challenging problems for automatic analysis of scanned volumes.

Our metadata generation system has been integrated into the Internet Archive [6], a large scale digital library digitizing and hosting historical scientific publications. The metadata generation system extracts metadata from volumes of printed journals which is used to support internal navigation functionalities. Specifically, the system takes OCRed text and related metadata of a volume of journal as input, identifies different types of content within the volume, generates multi-level metadata, and then outputs the metadata in XML form. The automatically-generated metadata consists of descriptive information about the collection of published articles within the volume as well as correspondence between scanned images of pages, pages within viewable PDF file, and pages within the original volume. As an example, Figure 1 shows services based on automatically-generated metadata for a volume of journal publications. The left side shows a screen shot of the navigation service listing descriptive metadata of identified articles contained within a scientific volume as well as links to corresponding articles. The right side shows the article display window after a user switches to a particular article within the volume.

The main contributions of our work include:

- A system which processes digitized scientific volumes and generates metadata describing their content, providing support for efficient access to intellectual content in digital libraries;
- A supervised learning based method for generating descriptive and structural metadata for digitized volumes based on estimated style, linguistic, and format features of content;
- Testing of the system on real-world scientific volumes collected from various domains.

The remainder of this paper is organized as follows. In section 2, we analyze and compare closely related prior work. Section 3 gives an overview of the metadata extraction process for scanned scientific volumes. We present techniques used for analyzing digitized resources and extracting text line features in Section 4. We present techniques for automatic generation of volume level, issue level, and article level metadata in Section 5. In Section 6, we describe the operational system, real-world data sets used for testing, and evaluations of the performance of our system. Finally, we draw conclusions and suggest future directions in section 7.

2. RELATED WORK

Our system aims at generating descriptive and structural metadata for scanned scientific volumes, which will be used by digital libraries to support efficient browsing and retrieval of information contained within scanned volumes. The problem and related design issues are closely related to prior work on descriptive and structural metadata generation, and on digital library systems.

2.1 Descriptive Metadata Generation

With the popularity of digital libraries and the establishment of large-scale underlying databases, automatic metadata generation systems have been developed to assist the management and retrieval of resources in various domains. In scientific literature digital libraries, techniques have been proposed to extract document metadata from research papers. Rule-based [14] and machine-learning based [15] methods have been used to identify metadata elements within a document using formatting and textual features. At the U.S. National Library of Medicine (NLM), a system [22] has been developed to generate descriptive metadata, including title, author, affiliation, and abstract, from scanned medical journals. In educational digital libraries, the Gateway for Education (GEM) metadata [4] as well as the Dublin Core metadata [3] have been extracted from educational materials. For general documents, especially born-digital documents which have been created by document processing software such as Microsoft Word, PowerPoint, and LaTeX, a supervised learning based method[16] has been proposed to identify title of the document using formatting and font features embedded within the document. CiteSeer [13], a digital library for computer and information science, also automatically extracts Dublin Core metadata.

In prior work on document metadata extraction, a document almost always represents one logical unit, such as a single research paper. As such, there is no need to identify a collection of units within a compound digital resource, which is the major goal of our work.

2.2 Structural Metadata Generation

Providing effective navigation and search tools for digital content is an advantage of digital libraries versus conventional libraries. For compound digital objects, including text, audio, and video resources, it is necessary to provide convenient random access to digital contents. In order to achieve this, digital libraries need structural metadata to describe content of digital resources and join together the components of a compound resource. In [12], Dushay introduced personalized content access, an application of structural metadata in digital libraries. In this work, digital libraries expose structural metadata in repositories so that third-party context brokers can utilize the metadata to localize the experience of a digital object.

In regards to digitized books, Cesarini et al. [11] proposed a method for classifying book pages into predefined classes including title page, index, and normal page. Book pages are classified to improve the metadata extraction and increase the accessibility of collections of scanned documents. In speech recognition applications, Liu et al. [21] proposed a structural metadata detection system to improve automatically-generated speech transcripts. The system detects various types of structural information, including sentence boundaries, filler words, and disfluencies, within speech transcripts using lexical, prosodic, and syntactic features.

In our work, a digitized volume corresponds to a collection of objects, including scanned images of pages, OCRed text, manually-generated metadata, among others. In order to support efficient content navigation and retrieval, digital libraries need both structural and descriptive metadata about the volume which describes the internal structure of the volume, correspondences between different objects, and individual articles contained within the volume.

2.3 Digital Library System

Digital library systems have been evolving in order to efficiently organize, store, and provide access to the rapidly increasing amount of digital information. The early work of Arms et al. [9] reports an experimental system developed by the National Digital Library Project (NDLP) at the Library of Congress. The work described how technical building blocks are used to organize collections of material and how these methods fit into a general distributed computing framework. Later, Besser [10] proposed a conceptual framework for interoperable digital library development and discussed how we might move from isolated digital collections to interoperable digital libraries which will enable searching across collections. As an example of the integration of digital libraries into the semantic web, Petinot et al. [23] presented a service oriented architecture of a scientific literature digital library. The architecture enables the integration of services provided by the digital library through an application programming interface.

3. SYSTEM OVERVIEW

The complete process of digitizing printed volumes, extracting metadata, indexing the content of digital resources, and providing web access is illustrated in Figure 2. At the beginning of the process, volumes of published journals are scanned and saved as page images, and are then converted to text information by OCR techniques. After the scanning and text recognition process, the metadata generation system generates metadata describing the internal structure of the scanned volume and published articles contained within the volume. Finally, generated metadata information and OCRed text are integrated to support navigation and retrieval of content within scanned volumes.

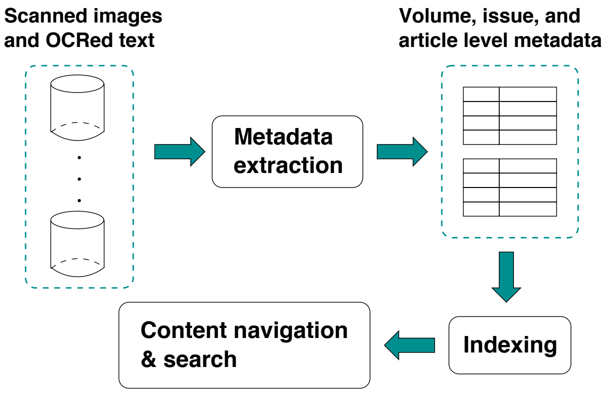

Scanned images and OCRed text
Volume, issue, and article level metadata
Metadata extraction
Indexing
Content navigation & search

Figure 2: Analysis of scanned scientific volumes.

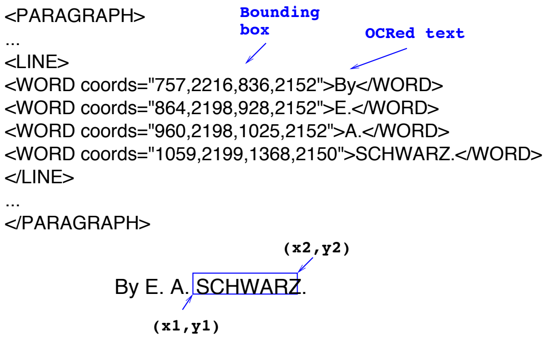
This section presents various digital resources of each scanned volume, selection of input for the metadata generation system, the method for automatic metadata generation, and the set of metadata elements generated by the system.

3.1 Digital Resources of Scanned Volumes

A digital library can scan volumes of journals and store them in four types of digital resources in a repository: scanned images of pages, OCRed text, PDF files, and manually-generated metadata. Specifically, scanned images of pages are stored in both the JPEG and the DjVu formats; OCRed text is stored in text and the DjVu XML format; PDF files are stored in color and binary PDF file format; while manually-generated metadata is presented in the XML format. The metadata extraction system needs to generate information which relates objects in the different sources.

We choose the DjVu XML [2] file as the main input of the metadata generation system for several reasons:

- The DjVu XML file contains full OCRed text.
- The DjVu XML file presents logical structures of the OCRed text. In each DjVu XML file, the OCRed text is organized in a page, paragraph, line, and word hierarchy. This logical structure information can be used to help the metadata extraction process. For instance, based on our observations, an article title almost always starts a new paragraph.
- The DjVu XML file retains the bounding box information of every single OCRed word, from which we can estimate format features. As an example, the average height of words within a line is a good indicator of the font size used for the line, which can be calculated from bounding boxes of all words within the text line. Figure 3 shows logical structure and bounding box information embedded within a DjVu XML document.



```
<PARAGRAPH>
...
<LINE>
<WORD coords="757,2216,836,2152">By</WORD>
<WORD coords="864,2198,928,2152">E.</WORD>
<WORD coords="960,2198,1025,2152">A.</WORD>
<WORD coords="1059,2199,1368,2150">SCHWARZ.</WORD>
</LINE>
...
</PARAGRAPH>
```

Bounding box
OCRed text
(x2,y2)
By E. A. SCHWARZ.
(x1,y1)

Figure 3: A portion of a DjVu XML file used to store information for the OCRed volume.

Besides the information about OCRed text, a DjVu XML file also contains references to corresponding scanned images of pages, which can be used to establish correspondences between OCRed text and scanned page images.

3.2 Metadata Generation

Based on manual check of a randomly selected real-world data set, each scanned volume often contains multiple issues and many articles (approximately 50-150 articles within a volume). In addition, there exists other information, such as notes and announcements. In this work, our efforts focus on identifying the internal structure of a scanned volume and describing articles within that volume.

There are two steps in the automatic metadata generation process: feature extraction and metadata labeling. The feature extraction step uses OCRed text and the bounding box information to calculate line features for every text line contained within a scanned volume. The metadata labeling process uses rule-based and machine-learning based methods to extract metadata about the issues and the articles contained within the scanned volumes. Specifically, rule-based pattern match is used to recognize and analyze volume and issue title pages, while a machine-learning based approach is used to detect article title blocks and to generate article metadata.

The metadata extraction system generates multi-level metadata information at the volume level, the issue level, and the article level. The volume level metadata describes the whole volume, such as the number of pages in the volume; the issue level metadata describes a single issue, such as the issue number; while the article level metadata describes a single article, such as the article title and the list of authors. Figure 4 shows the data flow inside the metadata generation system.

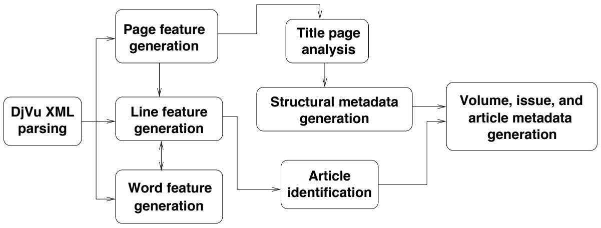


Figure 4: Data flow of the metadata generation system.

4. FEATURE EXTRACTION

For each scanned volume, the metadata generation system takes the DjVu XML file as input and parses the hierarchy of objects contained within the file. During the parsing of the XML file, the system calculates features for every word, line, paragraph, and page of the OCRed text.

We then combine page features and line features for volume level and issue level metadata generation. In particular, matching of specific patterns is used in the detection of the title page. For identifying articles and extracting descriptive metadata for articles, we choose the text line as the unit for feature extraction. We believe the text line is the appropriate unit for metadata labeling because the article title and the author information usually occupy one or a few consecutive lines. Additionally, text within the same line usually has the same style. Thus, line features are designed to estimate properties of OCRed text within a line, which can be calculated based on OCRed text and bounding box information in the DjVu XML file.

4.1 Style Features

Style features characterize the basic formatting style of a text line.

Capital Letter Mode

This feature represents the usage of capital characters in a text line. The feature takes one of the three possible values, $\{0, 1, 2\}$, corresponding to non-capital, first character capital, and full capital, for letters, respectively. The feature is designed based on the observation that article title blocks often contain capital characters.

The capital letter mode of a line is dependent upon the capital letter mode of all words within the line, while the capital letter mode of a word is obtained by analyzing its characters. Here, we use n_0, n_1, n_2 to represent the number of non-capital, first character capital, and full capital words respectively. Based on the observation that certain stop words, such as "a", "the", and "in", in a title line are usually non-capital while others words in the title are first character capital, we compile a list of common preposition and article words and record the number of their appearances in every line, represented by n_p. Thus, the capital letter mode of a line, represented by CM_l, is defined as the following:

$$CM_l = \begin{cases} 2 & \text{if } max\{n_2, n_1, n_0\} = n_2; \\ 1 & \text{if } max\{n_2, n_1, n_0 - n_p\} = n_1; \\ 0 & \text{else.} \end{cases}$$

Alignment

This feature approximates the horizontal alignment of a line. It is a numerical value defined as s_l/s_r where s_l is the distance in pixels from the left edge of a page to the left side of the first character in a line and s_r is the distance in pixels from the right side of the last character in a line to the right edge of a page.

4.2 Semantic and Linguistic Features

Semantic and linguistic features indicate certain text patterns in a text line.

Person Name

This is a binary feature indicating if person names are detected within a text line. We downloaded 16,000 surnames and 5,200 first names from the Internet and use common patterns of person names to detect occurrence of person name in a text line.

Special Words

This is a binary feature indicating if certain keywords appear in a text line, such as "By", which is often used before the name of the first author.

Word Count

This feature records the number of words within a text line.

4.3 Structure and Context Features

Structure and context features capture the relative position of a text line within its paragraph and page.

Paragraph Begin

This is a binary feature indicating if the text line starts a new paragraph.

Line ID

This feature represents the relative position of a line within all the text lines on the same page. The line ID starts from 1 for the first line and increases sequentially until the last line of the page.

Vertical Position

This feature represents the vertical position of a line within a page. It is calculated as the y/h_{page}, where y represents the average vertical position of a text line and h_{page} represents the height of a page.

Distance to Previous Line

This features measures the vertical distance in pixels from a text line to the previous text line.

Distance to Next Line

This features measures the vertical distance in pixels from a text line to the following text line.

4.4 Font Features

Font features are calculated to estimate the size of text within a line.

Word Height

This feature represents the average height of words within a line. It is calculated as:

$$\frac{1}{m} \sum_{i=1}^{m} \left(y_1^i - y_2^i \right)$$

where m represents the number of words in a line and y_1^i and y_2^i are bounding box coordinates of the ith word in the line. This feature is designed to approximate the height of the font used for a line.

Character Width

This feature represents the average width per character for a text line. It is calculated as:

$$\sum_{i=1}^{m} \left(x_2^i - x_1^i \right) \Bigg/ \sum_{i=1}^{m} n_i$$

where m represents the number of words in a line, x_1^i and x_2^i are bounding box coordinates of the ith word in the line, and n_i represents number of characters in the ith word. The feature is designed to approximate the width of the font used for a line.

Normalized Word Height

This feature represents the relative word height of a line compared with the average heights of all lines in the same page. If n represents the number of lines within a page, the normalized average word height of the lth line is calculated as:

$$\frac{\frac{1}{m_l} \sum_{j=1}^{m_l} \left(y_1^{l,j} - y_2^{l,j} \right)}{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{m_i} \sum_{j=1}^{m_i} \left(y_1^{i,j} - y_2^{i,j} \right) \right)}$$

where m_i represents the number of words in the ith line and $y_1^{i,j}$ and $y_2^{i,j}$ are bounding box coordinates of the jth word in the ith line. The feature indicates how the font size compares with the typical fonts in the page. If the font of a line is larger than normal, the line might contain metadata information.

Normalized Character Width

This feature represents the relative character width of a line compared with average character width of all lines in the same page. If n represents the number of lines within a page, the normalized average character width of the lth line is calculated as:

$$\frac{\sum_{j=1}^{m_l}\left(x_2^{l,j}-x_1^{l,j}\right)\Big/\sum_{j=1}^{m_l}k_{l,j}}{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m_i}\left(x_2^{i,j}-x_1^{i,j}\right)\Big/\sum_{j=1}^{m_i}k_{i,j}\right)}$$

where m_i represents the number of words in the ith line, $x_1^{i,j}$ and $x_2^{i,j}$ are bounding box coordinates of the jth word in the ith line, and $k_{i,j}$ represents the number of characters in the jth word in the ith line.

5. METADATA GENERATION

A scanned volume often has multiple issues of a periodical bound in a book form. In order to describe the content of a scanned volume, we define three levels of metadata: the volume level metadata, the issue level metadata, and the article level metadata. The volume level metadata describe the whole scanned volume, such as the title of journal and the number of issues. The issue level metadata describe a single issue, such as the issue number. The article level metadata describe a single article, such as the article title and the list of authors.

We use both rule-based pattern matching and machine learning-based approaches in the process of automatic generation of multi-level metadata. Specifically, we use certain heuristics to detect the volume and the issue title pages, the table of contents pages, and the lines indicating volume and issue information. In terms of the article level metadata, we use a supervised learning based method to train a model and to label the starting lines of articles using the model.

In this section, multi-level metadata elements and the generation techniques will be presented in detail. To facilitate presentation, in Tables 1, 2, and 3, “S”, “N” ,”T” in the Type column represent “String”, “Numerical”, and “Table”, respectively.

5.1 Volume Level Metadata

A set of volume level metadata elements, as shown in Table 1, has been defined to describe properties of the whole volume, such as issues in the volume, the mapping between the page number in the original publication and the index of the page in the digitized document. Here, the page number refers to numbers printed on the original document, while page index refers to the index of a page in the digitized document. For instance, if the viewable PDF file for a scanned volume has 485 pages, the range of page index is $\{1, 2, ..., 485\}$. Normally, the range of a page index is larger than that of the printed page numbers, because there are empty pages, title pages, and other pages contained within a scanned volume which may not have page numbers printed on the document. We need to keep track of both set of page numbers.

Volume level metadata elements which are not related to article information, i.e., excluding “number of articles” shown in Table 1, are generated by a rule-based approach. For instance, the volume and issue information are generated by the detection of title pages and special text lines within title pages. Specifically, title pages are detected by their page format features and special line patterns containing volume and/or issue numbers, etc. In order to tolerate OCR errors in special pattern matching, we use the Levenshtein Distance[20] metric to measure the degree of match between a string possibly corrupted by OCR errors and a target string. For instance, for a word “volume”, which is critical for title page detection, there are some observed variations caused by OCR errors, such as “voluime”, “volm”, etc. By setting the threshold Levenshtein Distance value, we can adjust the system to tolerate certain degree of OCR errors. Based on our manual check, we found that OCR errors appear quite often in title pages, partly due to special fonts used in title pages.

Table 1: Volume level metadata elements.

Name	Type	Description
title	S	journal title
volume	N	volume number
number of issues	N	number of issues within the volume
number of articles	N	number of articles within the volume
number of pages	N	number of scanned pages for display
page number	N	the maximum page number in the original printed journal
page mapping	T	the mapping of scanned pages to original printed pages

5.2 Issue Level Metadata

A set of metadata elements, as shown in Table 2, have been defined to describe a single issue within a scanned volume. Similar to the volume level metadata, the issue level metadata has been generated by title page detection. Metadata elements which depend on extracted articles, such as number of articles within an issue, are collected after the article metadata generation process.

Table 2: Issue level metadata elements.

Name	Type	Description
volume	N	volume number
issue	N	issue number
start	N	page index of the start page of the issue
end	N	page index of the end page of the issue
number of articles	N	number of articles within the issue

5.3 Article Level Metadata

Article level metadata elements are designed to describe articles contained within a scanned volume, such as article title and author information. Besides, they are also designed to link various digital resource corresponding to a volume. The current article level metadata element is listed in the following Table 3. Compared with article metadata for current research papers, there is no abstract or reference information, because almost none of the publications in our historical dataset have them.

Table 3: Article level metadata elements.

Name	Type	Description
Title	S	article title
Author	S	article author
Volume	N	volume number
Issue	N	issue number
Start Page	N	start page number
End Page	N	end page number
Start Page Index	N	start page index
Start Page Image	N	start page image

In Table 3, start page, start page index, and start page image triples define a mapping of three information sources. Start page refers to the start page number printed on an original printed volume, start page index refers to the index of the start page in the viewable PDF file, while start page image refers to the file name of the scanned image of the start page. The mapping of these resources will be used to support navigation and related functions on the web. For instance, the system can switch from the article list view to another view of a particular article which displays the start page image of the selected article.

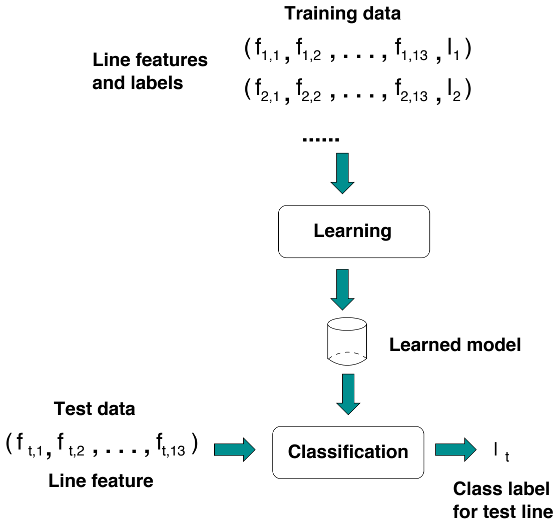

Training data
Line features and labels
$(f_{1,1}, f_{1,2}, \ldots, f_{1,13}, l_1)$
$(f_{2,1}, f_{2,2}, \ldots, f_{2,13}, l_2)$
......
Learning
Learned model
Test data
$(f_{t,1}, f_{t,2}, \ldots, f_{t,13})$
Line feature
Classification
l_t
Class label for test line

Figure 5: Supervised learning based article metadata generation.

We use a machine learning, in particular, a support vector machines (SVM) based method to generate article level metadata. The occurrence of an article, i.e., the start line of an article is detected by classification of text lines based on line features. After the start line of an article is detected, the article tile and the author information are generated by analyzing the article start line and limited following lines, since article title and author are nearly always the first part of an article. Other article level metadata elements, such as the volume and issue number, are generated based on volume and issue level metadata. For instance, the volume and issue of a page where an article starts is generated based on the volume and issue metadata which defines the start and end of an issue. Furthermore, the corresponding page number and page image for an article start page, which can be inferred from volume level metadata, will serve as the start page and start page image of the detected article.

Figure 5 illustrates the supervised-learning based article detection. In the learning and classification process, text line is the basic unit. Various types of features, as described in the previous section, have been extracted for every text line. Thus, every line in the scanned volume corresponds to one vector of feature values. In order to train the machine, we manually label every line in the training set, i.e., determine if it is the start of an article. The set of training data, including line features and labels, are fed into the learning module to create the learned model. Finally, the classifier takes feature vector of an unlabeled text line and the previously trained model as input and output the class label of the text line, i.e., tells if the text line starts an article.

6. EXPERIMENTAL RESULTS

Our metadata generation system is implemented using Java, Perl, and SVM Light [18]. The Java API for XML Processing (JAXP) provides the facilities for working with XML documents through the Document Object Model (DOM), which have been used to parse DjVu XML files of digitized volumes. The developed program has been integrated into the Internet Archive [6], in testing mode, to support the online navigation of scanned scientific volumes.

The metadata generation system parses the DjVu XML document, processes the hierarchical structure of the document, calculates various features, and generates the metadata for a digitized volume. The metadata output is stored in XML form, which facilitates the sharing of metadata across different systems.

To evaluate performance of the metadata generation system, we have randomly selected real-world volumes of journals hosted at the Internet Archive. The program is then tested on those volumes and the automatically-generated metadata values are compared with manually-generated ground truth metadata. Precision and recall are used to measure the performance of automatic metadata generation. Specifically, in our experiments, these two values are calculated as:

$$Recall = \frac{number\ of\ correctly\ extracted\ article\ metadata}{number\ of\ article\ within\ the\ volume}$$

$$Precision = \frac{\#\ of\ correctly\ extracted\ article\ metadata}{number\ of\ extracted\ article\ metadata}$$

6.1 Data Set

Testing of the metadata generation system has mainly focused on journals contributed by the Smithsonian Institute [7] based on the needs of the Biodiversity Heritage Library (BHL) project [1]. Those collections of journals have been digitized by the Internet Archive and hosted on the Internet Archive website [6] for open public access. For illustration purposes, we list sample journal titles and their publication time in Table 4.

Table 4: Sample journal in real-world data set.

Name	Century
Proceedings of the Entomological Society of Washington	19th
Proceedings of the Biological Society of Washington	19th
Journal of Natural Philosophy, Chemistry & the Arts	19th
Magazine of Natural History and Journal of Zoology, Botany, Mineralogy, Geology and Meteorology	18th - 19th

Most journals we have seen and manually checked were published nearly two centuries ago, and are represented in styles very different from that of contemporary journal publications. For instance, these articles seem to have no standard formatting: they may start from any place in a page; their titles may have a mixture of fully capitalized, first letter capitalized, and non-capitalized words; there may be no clear separation between the article title and the author names; or the article author name may not be available.

6.2 Volume and Issue Metadata Generation

Title page detection and page number recognition are two major tasks in the volume and issue level metadata generation. Specifically, the program identifies the volume number, the issue number, the beginning and the end of issue by detecting volume and issue title pages within the scanned volume. In terms of the mapping between page index, the index of a scanned page in the viewable PDF file, and page number, the number printed on the original volume, the program recognizes available page numbers on scanned pages by analyzing the OCRed text in particular areas of pages.

We use rule-based approach for title detection using page and line features calculated from OCRed text, bounding box information, and context analysis. Based on the observation, title pages have relatively fewer number of text lines and larger average distance between text lines, and they contain text lines indicating volume number (and issue number in issue title pages). Thus, we design several adjustable system parameters corresponding to these features and use them in finding special patterns in title pages. For instance, in order to tolerate OCR errors in volume and issue number line, we set the Levenshtein Distance[20] between an examined string and the target “volume”and “issue”keywords as a parameter and choose the optimal value based on experiments. As another example, in case the program can not recognize the volume and issue number due to OCR error, such as “IV” was OCRed as “it”, the program will use the previous or the following title page information, if available, to construct the current volume or issue metadata.

In the current system, the page number of a scanned page is recognized by analyzing the OCRed text. Based on the observation that page number is printed at certain areas of a page, such as the top and the bottom margin, and that it should be a number, specific pattern matching is performed in those restricted areas. Besides, certain heuristics are applied in order to tolerate OCR errors. For instance, page number “32” may be recognized as two separate numbers “3”, and “2”, or worse, page number “16” may be recognized as a number “1” and a letter “b”. In the first case, two neighboring numbers in the margin area will be treated as one number if the distance between them is below certain system parameter values. In the latter case, where there is no available information, the page number recognition for that particular page fails. As a result, the algorithm will try to fill in correct page number by checking page numbers of preceding and following scanned pages.

6.3 Article Metadata Generation

Three major steps are performed in article metadata generation: feature extraction, learning, and classification. The feature extraction step calculates style, linguistic, context, and font features for every text line within a scanned volume, as presented in the previous section. To facilitate presentation, the set of text line features is summarized in the following Table 5.

Table 5: Line features.

ID	Feature Name
1	Capital letter mode
2	Alignment
3	Person name
4	Special words
5	Word count
6	Paragraph begin
7	Line ID
8	Vertical position
9	Distance to the previous line
10	Distance to the following line
11	Word height
12	Character width
13	Normalized word height
14	Normalized character width

After text line features have been generated, we use SVM Light [18] for learning and classification tasks. We use two separate indicators, title begin and title end, for article metadata generation. A text line is title begin if it starts a new article, i.e., the first line of an article title. A text line is title end if it is the last line of the title block, i.e., the set of text lines containing title and author information. Thus, there are two separate classification tasks: title begin classification and title end classification. The SVM training tool takes line features and manually-generated ground truth class labels to train two separate models.

For any test scanned volume, the SVM Light Classifier takes the generated line features as well as previously trained models as input and predicts class labels for every text line. After two classification tasks, title begin and title end, certain heuristics has been used to filter out invalid cases, for example, cases where the distance between a pair of title begin and title lines exceeds a threshold value.

For title begin and title end classification, there is an imbalance issue [17] due to the nature of our problem. For the complete set of text lines contained within a scanned volume, there is a very small ratio of title lines, which is obviously true for almost all publications. For instance, in the Proceedings of the Entomological Society of Washington Vol II., there are 484 scanned pages, which contains thousands of text lines. On the other hand, there are 86 articles contained in the volume, which means 86 titles. Thus, the number of positive instances of title begin (the same number as title end) is much smaller than the number of negative instances, and the ratio of positive instances in the whole set is only 0.1% to 0.2%. It is known, and also verified through our experiments, that imbalanced training set causes low performance in the statistical classification process. In order to overcome this problem, we adopt the idea of shrinking the majority instances [19] in order to make the dataset balanced. Specifically, we randomly select negative instances to make the number of negative instances close to the number of positive instances. In this way, we construct a balanced data set. In order to evaluate the classification performance, we use six-fold cross-validation in the training and testing process.

6.4 Results

We randomly selected scanned volumes from the real-world data set, manually checked several thousands of pages, and generated ground-truth metadata. Specifically, we selected two volumes from each of the following: Proceedings of the Entomological Society of Washington, Journal of Natural Philosophy, Chemistry & the Arts, and Magazine of Natural History. For every volume, we compared the automatically-generated metadata with the ground-truth metadata.

Based on manual checking, the rule-based title page detection works well for these volumes. In our test volumes, volume and issue title pages have been successfully detected.

Our main efforts focused on evaluation of article level metadata extraction. We manually compared automatically-extracted article metadata with ground-truth data. In the automatically-generated article metadata output, if a set of metadata correctly identifies an article, it is a success case. After manual checking, precision and recall are used to measure the performance of the metadata generation system on different journals. The following Table 6 summarizes the results for three different journals. To facilitate presentation, we use J1, J2, J3 to represent Proceedings of the Entomological Society of Washington, Journal of Natural Philosophy, Chemistry & the Arts, and Magazine of Natural History and Journal of Zoology, Botany, Mineralogy, Geology and Meteorology, respectively. For every journal, the result is based on two scanned volumes.

Table 6: Performance of the metadata generation system as measured by precision and recall.

	J1	J2	J3
number of articles	146	203	180
number of extracted articles	141	201	162
number of correctly extracted articles	138	134	147
Precision	98%	67%	91%
Recall	94%	66%	82%

From the results presented in Table 6, we see that the performance of our metadata generation system varies for different journals. After analyzing failed cases in detail and comparing scanned volumes from different journals, we contribute this variance to various degree of format consistency in different journals since the method is very sensitive to those features.

In order to analyze the effects of person name features on the performance the metadata generation system, we conducted two types of classifications: without the person name feature vs. with the person name feature. Four scanned volumes, two volumes from J2 and J3 respectively, have been tested, and a comparison of performance is presented in Table 7.

Table 7: Impact of the person name feature to the system performance.

	J2-1 (19th)	**J2-2** (18th)	**J3-1** (19th)	**J3-2** (18th)
number of articles	74	129	73	107
w/o name feature				
extracted articles	47	128	56	87
correctly extracted	44	75	56	85
Precision	94%	59%	100%	98%
Recall	59%	58%	77%	79%
with name feature				
extracted articles	59	151	66	96
correctly extracted	52	82	57	90
Precision	88%	54%	86%	94%
Recall	70%	64%	78%	84%

Results in Table 7 show that adding the person name features increase the recall significantly at the cost of a reduction in precision. It suggests that we should use the person name to increase of chance of article metadata detection, while we need to design new features to filter out incorrectly-extracted cases.

7. CONCLUSIONS

We have presented an automatic metadata generation system for scanned volumes of journals. The system is designed to parse scanned volumes and generate structural and descriptive metadata. The volume level, issue level, and article level metadata have been generated using supervised-learning based and rule-based methods. The developed metadata generation system has been integrated into an operational digital library and has been tested on real-world historical journals with promising performance.

In the future, we plan to analyze the tables of content information for article metadata generation. By integrating this extra information and establishing correspondences between items in the tables of content and individual articles within the volume, we hope to enable advanced browsing based on tables of content. In addition, the system will utilize multiple sources of information for article metadata generation so that it can have more tolerance to OCR errors.

8. ACKNOWLEDGMENTS

This work was supported in part by the Smithsonian Institute, the Internet Archive, the US National Science Foundation under grant nos. 0535656, 0347148, 0454052, and 0202007, and Microsoft Research. We thank Tom Garnett of the Smithsonian Institute for helpful discussions.

9. REFERENCES

[1] Biodiversity Heritage Library. *http://www.biodiversitylibrary.org.*

[2] Djvu Zone. *http://www.djvuzone.org.*

[3] Dublin Core Metadata Initiative. *http://dublincore.org.*

[4] Gem. *http://www.thegateway.org.*

[5] Google Book Search. *http://books.google.com.*

[6] Internet Archive. *http://www.archive.org.*

[7] Smithsonian Institute. *http://www.si.edu.*

[8] The Universal Digital Library. In *http://www.ulib.org.*

[9] W. Y. Arms, C. Blanchi, and E. A. Overly. *An Architecture for Information in Digital Libraries.* 1997.

[10] H. Besser. *The Next Stage: Moving from Isolated Digtal Collections to Interoperable Digital Libraries. http://www.firstmonday.dk/issues/issue7_6/besser.* 2002.

[11] F. Cesarini, M. Lastri, S. Marinai, and G. Soda. Page classification for meta-data extraction from digital collections. In *Database and Expert Systems Applications*, pages 82–91, 2001.

[12] N. Dushay. Localizing experience of digital content via structural metadata. In *Proceedings of the ACM/IEEE joint conference on Digital libraries*, pages 244–252, New York, NY, USA, 2002. ACM.

[13] C. L. Giles, K. Bollacker, and S. Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the International Conference on Digital Libraries*, pages 89–98, 1998.

[14] G. Giuffrida, E. C. Shek, and J. Yang. Knowledge-based metadata extraction from postscript files. In *Proceedings of the International Conference on Digital Libraries*, pages 77–84, 2000.

[15] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 37–48, 2003.

[16] Y. Hu, H. Li, Y. Cao, D. Meyerzon, and Q. Zheng. Automatic extraction of titles from general documents using machine learning. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 145–154, 2005.

[17] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. In *Intelligent Data Analysis*, pages 429–449, 2002.

[18] T. Joachims. Making large-scale svm learning practical. In *Advances in Kernel Methods - Support Vector Learning*, 1999.

[19] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets:one-sided selection. In *International Conference on Machine Learning*, pages 179–186, 1997.

[20] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, pages 707–710, 1966.

[21] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper. Structural metadata research in the ears program. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 957–960, 2005.

[22] S. Mao, J. W. Kim, and G. R. Thoma. A dynamic feature generation system for automated metadata extraction in preservation of digital materials. In *Proceedings of the International Workshop on Document Image Analysis for Libraries*, pages 225–232, Washington, DC, USA, 2004. IEEE Computer Society.

[23] Y. Petinot, C. Giles, V. Bhatnagar, P. Teregowda, H. Han, and I. Councill. A service-oriented architecture for digital libraries. In *International Conference on Service Oriented Computing*, pages 263–268, 2004.

[24] R. Wendler. LDI Update: Metadata in the Library. *Library Notes*, no. 1286 (July/August): 4-5, 1999.