Metadata and Data Structures for the Historical Newspaper Digital Library

Robert B. Allen[1] and John Schalow[2]

[1]College of Library and Information Services
[2]University Library
University of Maryland
College Park, MD 20742
rba@glue.umd.edu and js368@umail.umd.edu

ABSTRACT

We examine metadata and data-structure issues for the Historical Newspaper Digital Library. This project proposes to digitize and then do OCR and linguisting processing on several years worth of historical newspapers. Newspapers are very complex information objects so developing a rich description of their content is challenging. In addition to frameworks for the logical structure and physical layout, we propose metadata relevant to the image processing and to the historians who will use this collection. Finally, we consider how the metadata infrastructure might be managed as it evolves with improved text processing capabilities and how an infrastructure might be developed to support a community of users.

Keywords: Digital Libraries, History, Metadata, Newspapers, OCR.

1. INTRODUCTION

1.1. The Historical Newspaper Digital Library

We have been engaged in a project to digitize and produce searchable full-text archives of historical newspapers [8]. Many of the historical newspapers in the United States have been microfilmed. We plan to digitize 15 years worth of that microfilm for an urban daily newspaper (about 200,000 pages) and apply advanced optical character recognition (OCR) and linguistic processing techniques to the collection. We will then index this material, build user interfaces for accessing it, and encourage historians to use the collection for research. As a trial, we obtained an original copy of *The New York Brooklyn Eagle* [10] for November 11, 1917 (Armistice Day). The top of the first page is shown in Figure 1. We have used this sample to conduct zone-segmentation studies and to show how the OCR'd text can be searched from a prototype user interface [8].

In addition to the city newspaper, we plan to process African-American newspapers from the Reconstruction Era and some non-English-language historical newspapers from European immigrant communities. There are many challenging research issues in this work ranging from multilingual OCR to story continuations across pages. We believe that there must be effective interfaces for access by professional and student historians as a well as the genealogists and the public at large.

Other projects, such as JSTOR (www.jstor.org), have used OCR for building collections of electronic text. However, newspapers have a particularly complex and variable logical structure and physical layout so they present a much greater challenge for automated processing. In the longer-term, we can envision a nation-wide digital library of historical newspapers as resource for both historians and for the public.

1.2. The Need for Metadata

Reliable metadata and well-managed data structures are essential to this project in many respects. The amount of material to process is very large and it spans several levels of detail. The metadata should facilitate the automated processing of the text. Metadata will guide the OCR, the linguistic processing, and the search. Ideally, the OCR and linguistic analyses would be flexible enough to extract all the necessary information from profiles so as to be able to process any newspaper simply by adjusting the metadata values.

The metadata can provide reference points for the performance of the programs. For instance, the metadata can provide ground-truth for zoning studies. In addition, the inferred metadata would be helpful for evaluation and quality control.

Metadata may also be used as a guide for end users in their interaction with the collection. End users can se-

THE BROOKLYN DAILY EAGLE

7 SECTIONS. NEW YORK CITY, SUNDAY, NOVEMBER 11, 1917. 68 PAGES. THREE CENTS.

BRITISH GUNS ON ITALIANS' FIRST LINE; TEUTONS TAKE ASIAGO, IN TRENTINO; U. S. TO IGNORE TALK OF ARMISTICE

Austro-Germans Begin Flanking Movement Against Italian Left Wing—Capture Asiago After Hard Fighting. Opposing Armies Face Each Other Along the Piave — Great Battle Near — Venice Is in Danger.

Washington Will Ignore Lenine's Armistice Plan

Bolsheviki Forms Cabinet, With Lenine as Premier—Russian Embassy at Washington Refuses Recognition to Maximaliat Regime. Revolutionists in Control of Moscow Government — Americans in Petrograd Safe, Francis Reports.

NOT A PUFF!

6 GRAFT ARRESTS IN FOOD SCANDAL AT THE NAVY YARD

Two Navy Men and Provision Dealer Among Those Locked Up.

U.S. LOSES THOUSANDS

HAIG IN NEW DASH GAINS HALF MILE AT PASSCHENDAELE

Sweeps Forward 800 Yards on Wide Front in Flanders.

REPULSED—BERLIN

BEGIN SECOND DRAFT CLASSIFYING DEC. 15; PRESIDENT ASKS AID

Process Will Take Two Months, Crowder Orders Legal Ad. vice Bureaus.

ACTUAL DRAFT IN FEBRUARY

LOW PRICES An Old Story Here

Figure 1: Scanned image of *The Brooklyn Daily Eagle* for November 11, 1917.

lect metadata values to facilitate browsing and restrict searches. Sophisticated interfaces will be needed to allow the users to navigate at several levels of granularity – within a single page and a single newspaper, but also across issues of the newspaper, and, eventually, across collections of newspapers.

1.3. Relevant Systems of Metadata

The News Industry Text Format (NITF, www.iptc.org/-iptc/xmlnitf.htm) DTD and XMLNews-Story (www.-xmlnews.org) will provide the basis of the coding. However, these are uned to the needs of newspaper production and distribution more than management of a historical collection. We stress themes of structural metadata [1] and show how the many levels of metadata need to interoperate.

There have been several efforts to describe systems for document structure in conjunction with OCR (e.g., [4] [6]). PinkPanther [11] is a tool for creating geometric metadata for scanned document images. However, the tags that are produced use an ad hoc system that is not compatible with other established standards.

1.4. Overview of the Current Effort

The distinction between "metadata" and "data" in data structures is difficult to apply rigidly. This is especially true for values determined at one level of processing which guides processing at other levels. For example, the identification of a region of the newspaper as a news story may determine which lexicon is applied in the detailed processing of the text from that news story. Each part of a page image should be assigned as a news object. Specifically, we require that every page be composed of "base objects."

However, there is a need for standard descriptions of both the logical structure and the physical layout of newspaper content. To the extent possible, we separate the logical from the physical (layout) structure. This is in the spirit of XML(SGML) and XSL, which produce DTDs and DSSSLs (style sheets) respectively.

This is an initial specification rather than a final report. Some of the records described here are essential to the project while others are lower priority and may not be implemented in the early stages.

2. ORIGINS OF THE METADATA VALUES

The metadata framework described here is complex and highly interwoven. Table 1 shows codes which are used to identify the origins of the metadata in the remainder of this paper. These notes are necessary to determine the reliability of the observations and the need for updating them during different types of revisions.

Code	Type	Description or Example
A	Authority	LC, OCLC
D	Data Entry	Human supervisor for processing
L	Linguistic	Inferred during linguistic processing
O	OCR	Inferred during OCR
P	Project	Policy set by project
U	End User	Individual user

Table 1: Codes for the origins of metadata.

In some cases, there may be several possible origins of a single metadata item. In the following tables, we apply the codes only to the most likely origin. A rough hierarchy will be developed for the trust placed on the metadata such that human judgments will be preferred over programmatically inferred judgments. As described be-

low, there may be cases in which the user can override the inferred value.

in addition, the values will be extensively cross-linked with precedence carefully specified so as not to cause cycles during revisions. There is a tension between the default values, which may be seen as the newspaper's style, and the need to be able to process unexpected content flexibly.

3. CATALOG RECORDS FOR INDIVIDUAL ISSUES OF A NEWSPAPER

The library community has a long tradition of creating bibliographic metadata in the form of the MARC (Machine-Readable Cataloging) standard. MARC is a system of storing bibliographic information in predefined tags and subfields, which then enables automated manipulation of this information to facilitate computerized applications.

Two significant applications of MARC for newspapers are CONSER and the United States Newspaper Program (USNP). The CONSER (Cooperative ONline SERials) Program has its origin in the early 1970s when it was formed by the library community and OCLC to create MARC records for serials. The USNP is a cooperative national effort among the states and federal government to locate, catalog, preserve, and make available on microfilm newspapers published in the United States from the eighteenth century to the present. The National Endowment for the Humanities (NEH) and the Library of Congress (LC) jointly administer the USNP. Bibliographic records entered by USNP participants are part of the CONSER database and OCLC publishes the United States Newspaper Program National Union List. The USNP National Union List contains over 138,000 bibliographic records for newspapers published from 1690 to the present. As an example of one state's implementation, the Maryland Newspaper Project was completed by the Maryland State Archives and the University of Maryland at College Park. Over 2,500 titles were located and given title-level cataloging, and over 1,333,000 pages were microfilmed.

Although MARC has, in the the past, been focused on describing only entire books and documents, it is a possible framework for more specific metadata. MARC can be used to catalog the overall title of the newspaper (a title-level record), an individual article (an analytic record) or a collection of related articles (a collection level record). The MARC holdings and locations codes enables one to store details about enumeration, chronology, or publication patterns. They also support storing reproduction information when the bibliographic record applies to the original. The holdings record can be stored independently or be attached to the MARC bibliographic record.

Several layers of metadata are required for this project. Some of these are already established. For instance, there are extensive guidelines for components of titles because of their participation in MARC records. For describing individual issues, the closest structure would be MARC subrecord number 773; however, it is not very specific. It does not provide guidelines for storing physical dimensions of objects; neither does it provide a method for representing a series of objects (i.e., linked lists).

Table 2 shows a possible record structure for copies of an individual newspaper. The International Standard Serial Number (ISSN) will be used as the key around which these other standards will be referenced. ISSN can be used to obtain information such as circulation and the publisher.

Type	Field	Origin
Serial Record		
	OCLC Control number	A
	ISSN	A
Serial Subrecord		
	Date	A
	Edition	D
	# Pages	D
	Rights	D
	Sections (e.g., Metro, Sports)	D
	Layout Style Pointer	D
	List (Article IDs)	L

Table 2: Descriptors for individual issues of a newspaper in the Historical Newspaper Project.

4. PAGE-ORIENTED ATTRIBUTES

The quality of the page images dramatically affects the quality of the OCR and linguistic processing. Therefore, it is necessary to have detailed metadata about those images. Indeed, we hope to develop the page-image collection into a testbed for other researchers, so documentation about them is critical. Moreover, the page images themselves are an important resource whose history needs to be described in detail. For instance, many scholars may wish simply to browse the page images.

Table 3 shows the proposed record structure for the images. These values may include the resolution at which the copy is made and information about any earlier processing. Table 4 describes other values recorded for each page beyond the properties of the image.

5. BASIC OBJECTS

Each page image from the historical newspaper collection will be divided into basic objects. There are two types of these basic objects: text objects and graphical objects. Following the spirit of SGML/XML, we will try to identify the logical versus layout aspects of these objects.

Record	Attribute
Original	
	Dimensions
	Paper Stock
	Printing Mechanism
	Type of intermediate reproduction (if any)
Intermediate Reproduction - Microfilm	
	Control Number
	Compliant with NEH Guidelines
	Rights on Microfilm
	Date Microfilmed
	Reduction Factor
	Microfilm Type
	Comments on quality of film
Intermediate Reproduction - Fax	
	Control Number
	Date
	Reduction Factor
Digital Image Properties	
	ID number
	Date Scanned
	Scanner Manufacturer
	Scanner Model
	Scanner Reduction Factor
	Date Scanned
	Resolution in X
	Resolution in Y
	Encoding (JPEG, TIFF)
	Compression (IV)
	Pixel Characteristics (color, grayscale, binary)
	Rights
	Regions of degraded print or damaged paper

Table 3: Page-image metadata.

Sub-Record Types	Attribute	Origin
Processing of Page		
	OCR version	D
	linguistic processing version	D
	metadata version	D
Layout		
	list of articles	L
	number of columns	D
	page headers	D

Table 4: Page-content record.

Attribute	Values	Origin
News Object ID		N
Coordinates	linked list of corners	L
Area	in column inches	L
Page(s)		U
Object Type		L

Table 5: Basic object record.

It is worth exploring what we judge as constituting basic objects. For instance, should two related articles that are spanned by a single headline (see top left of Figure 1) be considered one object or two? For now we have claimed that they are two disjoint objects sharing the same headline. Indeed, there will probably be exceptions to any rules which could be developed to describe patterns within the logical or the layout structures.

6. BASIC TEXT OBJECTS

6.1. Text Object Types

To the extent possible the news articles will be structured with the DTDs developed by the NITF. This standard was developed for news-wire stories so it will not always be easy to fit the 1917 news stories into that framework. However, it does have advantages in providing a standard structure for names such as those used in bylines.

There are many types and sub-types of text objects: notices, want ads, editorials, etc. Some types of material such as poetry and plays, often appear in newspapers. The IPTC Subject Reference System (http://-www.iptc.org/iptc/) (see Figure 6) will be used to code news objects.

6.2. Fine-Grained Text Objects

OCR deals primarily with text objects at a fine level of detail. In order to support the OCR, many low-level

Code	Description
01000000	Arts, Culture & Entertainment
02000000	Crime, Law & Justice
03000000	Disasters & Accidents
04000000	Economy, Business & Finance
05000000	Education
06000000	Environmental Issues
07000000	Health
08000000	Human Interest
09000000	Labour
10000000	Lifestyle & Leisure
11000000	Politics
12000000	Religion & Belief
13000000	Science & Technology
14000000	Social Issues
15000000	Sport
16000000	Unrest, Conflicts & War
17000000	Weather

Table 6: Top-level ITPC subject codes.

tables will be required. This level of characterization is similar to that found in (e.g., [4, 6]).

Attribute	Sample Value	Origin
font script	Roman	O
reading direction	left-right	A
character orientation	upright	A
font size	10 pt	O
language	English, German	O

Table 7: Description of text characters.

The OCR processing makes many inferences. Every character and every word decision has a confidence level. Without storing unrealistic amounts of detailed data, some of the most likely alternatives need to be maintained.

7. BASIC GRAPHICAL OBJECTS

Any object which is not a text object is classed as a graphical object. A graphical object may include text entities such as a caption and a credit. Moreover, a graphical object may be linked with a text object (as in a drawing which illustrates a news story), or it may be free-standing. We examine two main groups of graphical objects: tables and pictorial material.

Newspapers typically have many different types of tables such as financial tables and box scores. Automatic extraction of values from tables is known to be a difficult problem for OCR [2].

Pictorial materials may include weather maps, advertisements, logos, crosswords, and schematics. Some graphical materials follow templates. In particular, the name of the paper is generally presented in a banner. The CONSER format goes into considerable detail about terms for describing variations of the titles because they are used in the MARC records. Characteristics such as whether an image is grayscale or color and whether it is a line-drawing, an etching, or a photograph will be described by following the *Thesaurus of Graphical Materials* [9].

8. LAYOUT METADATA

Layout is particularly complex in newspapers. However, it is essential to model the layout. For instance, one might want to find out what articles appear adjacent to each other. The logical structure described in Table **??** provides rules for the structure of composite objects. For instance, one of the rules might be that the headline must be above the text of the story.

Layout is a major concern in the production of a newspaper. It is an essential element in editorial training (e.g., [3], [7]). The Society for Newspaper Design (http://-www.snd.org/) has been involved with creating a repository of information regarding layout information.

Given the division we have made of the newspaper page into basic objects, we distinguish between the layout of these basic objects on a page and the internal structure of each object. For instance, we are able to identify related basic objects based on having a similar layout repeated on one page or even across different issues (i.e., days) of the newspaper. Initially, we will use simple edges and adjacency for describing the layout of news objects on the entire page. However, there are many formalisms we can explore such as spatial grammars, bintrees, and quadtrees.

9. OTHER TYPES OF METADATA

Several other types of metadata in this project should be noted. For instance, there should be a count of the number of times each object is accessed. This may be considered as a type of "administrative metadata".

Some metadata would apply across an entire collection. [5] describes data models for supporting this task. This may include threads of news stories across days, or details about the processing of different material. For instance, we need records of which lexicons and sublexicons are used in different stages of processing.

We will also record annotations by readers. We expect several different classes of annotations, such as those which comment about contents and those which react to aspects of the system or the indexing. The latter type of comment be used for updates as described in Section 10.2.

10. MANAGEMENT OF METADATA

10.1. Validation of Metadata: Inter-Observer Reliability

Given that metadata is so integral to this project, it is essential to have estimates of the quality of the data and procedures for managing it. While some metadata

is unambiguous, other data and metadata may vary depending on human judgment.

If two persons are asked to mark metadata for a specific page, it is very likely that they will produce metadata that differ from each other. For example, one might mark one paragraph as a zone, while the other might mark an entire column of an article as a zone. Furthermore, if we ask a person to mark metadata on a page at two different times (e.g., a gap of a few months), it is very likely the metadata will differ. Thus, both inter- and intra- observer reliability can be calculated.

We seek to develop descriptions for metadata that are relatively unambiguous and will maximize inter-observer reliability. However, some subjectivity will be inevitable in many cases. We need a procedure which will minimize debate about the correctness of metadata, and evaluate the proficiency of human metadata inspectors.

10.2. Distributed Management of Metadata

There are bound to be errors in the dataset so we will implement a process by which users can log their changes. For example, the zone segmentation algorithm may mistakely merge two articles together. If one user discovers this mistake, he/she would want to log the mistake in a database, to be analyzed later.

We feel that the most workable solution will be a "cooperative" program in which groups of users with different interests and skills will log suggestions for making changes on specific aspects of the metadata or data. Periodically these log reports from users would be examined by a committee and, if appropriate, the archive would be updated. Different levels of authority could be assigned to different groups of users. The Cooperative Online Resource Catalog (CORC, http://www.oclc.org/-oclc/research/projects/corc/) project demonstrates that such an organization is possible.

10.3. Interfaces for Entering Metadata

There are two common tools for entering and updating metadata in the library community. The PRISM Passport system is a tool for creating and updating MARC records. SiteSearch is another tool for creating and updating OCLC records. Because our metadata will have many fields that are not in either MARC or OCLC records, we will create a new user interface for entering and updating metadata.

10.4. Naming and Organizing Metadata Resources

Because there will be such a large number of records, a unified approach to naming is required to keep them straight. We propose a hierarchical naming scheme consisting of:

collection.ISSN.IssueID.PageRecordIDs.BasicObjectID.

Both the PageRecordIDs and BasicObjectIDs will have version numbers. When updates are made at one level, changes will need to be propagated to all datasets lower in the hierarchy.

11. DISCUSSION

The Historical Newspaper Digital Library is still in its early stages. This specification outlines the issues relating to metadata. While our immediate goal has been to outline a framework for use in our own research project, developing a community consensus on this effort is desirable. Moreover, although we have focused on metadata for the Historical Newspaper Digital Library, many of the issues addressed here apply beyond newspaper to other printed materials. For instance, collections of journals, newsletters, and magazines will need similiar sets of metadata.

Acknowledgments

We thank Marietta Plank of the University Library for helpful discussions. We also thank Tapas Kanungo, our colleague in this project.

REFERENCES

1. Working Group 3: Structural and administrative metadata in page-image conversion projects: Discussion summary and recommendations. In *TEI and XML in Digital Libraries*. Washington, DC.

2. Alam, H., Chang, C. H., Shi, Z., and Tupaj, S. Extracting tables from printed documents. In *Symposium on Document Image Understanding and Technology* (1995), pp. 113–124.

3. Baskette, F. K., Sissors, J. Z., and Brooks, J. S. *The Art of Editing*. Allyn and Bacon, 1996.

4. Bunke, H., and Wang, P. S. P. *Handbook on Character Recognition and Document Image Analysis*. World Scientific, 1997.

5. Cabo, M. A. An approach to a digital library of newspapers. *Information Processing and Management 33*, 5 (1997), 645–661.

6. Document Processing Group. Page decomposition and related research at the University of Maryland. In *Symposium on Document Image Understanding and Technology* (1995), pp. 39–55.

7. Harrower, T. *The Newspaper Designer's Handbook*. McGraw Hill, 1997.

8. Kanungo, T., and Allen, R. B. Full-text access to historical newspapers. Tech. Rep. CS-TR-4014, Laboratory for Language and Media Processing, University of Maryland, Apr. 1999.

9. Library of Congress. *Thesaurus for Graphical Objects*. 1995.

10. SCHROTH, R. A. *The Eagle and Brooklyn: A Community Newspaper, 1841-1955.* Greenwood Press, Westport CT, 1974.

11. YANIKOGLU, B. A., AND VINCENT, L. Pink Panther: A complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognition 31* (September 1998), 1191–204.