DCPAPERS

Enhancing Discovery with AI: Volume Extraction and Summary Statements for Holdings Metadata

Myung-Ja K. Han[1,*], Owen Monroe[2,†]

[1] University of Illinois Urbana-Champaign, 1408 W, Gregory Dr. Urbana, Illinois, USA
[2] University of Illinois Urbana-Champaign, 614 E Daniel St, Champaign, Illinois, USA

Abstract

Serials volume information is essential for helping users and collection managers understand what volumes are available and to inform future collection strategies. However, due to historical practices of binding and recording summary statements varying by institution, inconsistent holdings metadata poses significant challenges in aggregated discovery environments. This research explores the use of Large Language Models (LLMs) to enhance holdings metadata through two approaches. The first approach employs a Python script that prompts Gemini AI to extract volume (year) information from title pages in digitized serial PDF files submitted by various institutions. The extracted data is used to generate accurate coverage ranges and identify missing volumes for entire digitized serial contents. The second approach trains a BERT model using labeled data from text files to detect title pages of annual reports and identify publication years present or missing from the digitized serial contents. Both approaches—using Gemini and BERT—have shown measurable success in extracting publication date information and generating summary notes that enhance holdings metadata that would support improved resource navigation and informs strategic collection decisions for digitized serials.

Keywords

AI in libraries, holdings metadata, summary statements, Gemini, BERT

1. Introduction

Mass digitization has enabled libraries and other cultural heritage institutions to identify collection gaps, support shared collection development, and highlight areas for future digitization that would enhance user discovery of resources. However, serials [1] present unique challenges for collection gap analysis due to varying practices in serials management—particularly in binding and inconsistencies in holdings metadata across institutions.

Unlike monographs, which typically have one bibliographic record describing a single item, serials bibliographic metadata represents the entire publication issued under a given title. Item-specific information—such as volume-level details and overall coverage—is recorded separately in holdings metadata, often summarized through statements that indicate the volumes held and any gaps or missing issues. While two international standards exist for recording holdings summary statements [2,3], in practice, implementations frequently vary due to local policies. These policies are influenced by factors such as the library system in use, the availability of serials cataloging expertise, and the institution's strategic priorities. As a result, inconsistent holdings metadata presents significant challenges in aggregated environments like HathiTrust

*Corresponding author.
†These authors contributed equally.
mhan3@illinois.edu (M. Han); omonroe2@illinois.edu (O. Monroe)
0000-0001-5891-6466 (M. Han); 0009-0003-5099-3397 (O. Monroe)

https://doi.org/10.23106/dcmi.952591060

[4], where users rely on volume-level information submitted by individual institutions. These records typically lack a comprehensive summary statement for the entire serial title, making it difficult to identify missing volumes or issues. Consequently, both users and institutions face barriers in understanding the full extent of digitized coverage for a given serial.

While Artificial Intelligence (AI) and related technologies are increasingly tested and applied in cultural heritage institutions for metadata creation and enhancement, most efforts to date have focused on bibliographic metadata, not holdings metadata. This paper explores the potential of AI—specifically Google's Gemini model [5] and BERT model [6]—to train and extract serials volume information from the full text of digitized serial content and generate summary statements that include information about missing volumes. This approach aims to improve user understanding of digitized serial coverage and support libraries in making informed decisions about collection development and future digitization priorities.

2. Methodology

2.1. Data Selection

The test sample was randomly selected from HathiTrust. Using the search term "Annual report" in the HathiTrust search interface, a serial title with digitized content contributed by multiple institutions was chosen. This particular title had 17 digitized bound volumes, which were downloaded in both .txt (11.8 MB total) and .pdf (385.1 MB total) formats. The corresponding MARC metadata was also retrieved using the HathiTrust Bibliographic API service [7]. For this serial title, four institutions contributed digitized volumes, each applying different practices for recording volume information. These variations in descriptive practices are summarized in Table 1 below.

Table 1
Practices of Recording Serials Volume Information

Institution	No. of volumes contributed	Practice of recording volume information	Example
1	3	Four-digit year for first volume, two-digit year for last volume	(1903-09)
2	6	Four-digit year for both first and last volume	(1903-1912)
3	4	Inclusive coverage years	(1900/02-1903/04)
4	4	Four-digit year for both first and last volumes	(1904-1908)

2.2. Gemini Tests

Two Python scripts were developed for this testing, as illustrated in Figure 1. The first script was designed to process all 17 downloaded files, identify title pages and extract volume information using a trained Gemini 1.5-pro-002 AI model. It works by extracting text from each identified title pages by applying a structured Gemini prompt to do the tasks. Based on the model's response, the script generates structured output summarizing the volume information for each file. These summaries are saved in a designated output folder for further analysis.

https://doi.org/10.23106/dcmi.952591060

The second script aggregates the extracted volume information produced by the first python script, generating a summary CSV file that includes a per-file breakdown and the complete volume range. During this process, any missing volumes are also identified. Using the full range and the missing volume data, the script generated a summary statement, which was then added to an existing MARC.XML record following the cataloging standard. This enriched MARC metadata holdings information provides clear, consolidated digitized volume information for a serial title.

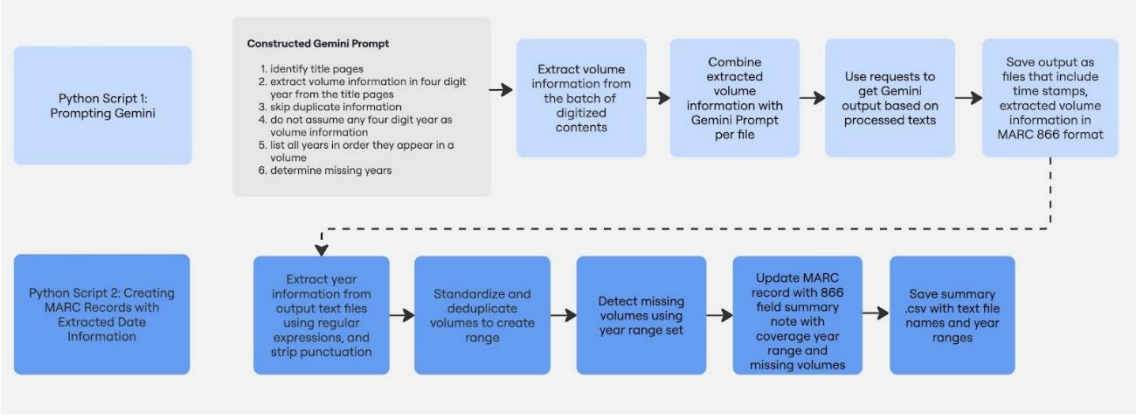

Figure 1: The workflows demonstrate how two Python scripts use the Gemini AI model to identify and extract volume information, then update MARC metadata with accurate holdings data.

2.3. BERT Tests

A labeled data set and python script were also developed to train a BERT base model for detecting title pages within annual report volumes as illustrated in Figure 2. BERT was chosen as a Large Language Model that is easy to run and can work with smaller amounts of data, and is able to perform classification tasks with complicated text structures through its bidirectional context learning [8].

The labeled data set contained texts from a random selection of historical digitized annual reports. Title pages of annual reports, which contain date information and appear throughout the volumes, are labeled as 1, and other text pages from the volumes, such as report information, organization member lists, schedules, and financial statements are labeled 0. 100 title pages were selected manually and 80 non-title pages with varying length of texts were selected manually, attempting to include the different kinds of information and texts on non-title pages.

A crossvalidation test was conducted using the labeled data, with splitting 80% of the texts into training data and 20% into test data. The training data was used to train a BERT base uncased model, which generated predictions for which elements of texts in the test data were title pages. The predictions were evaluated against the ground truth of the test data.

Following the training and evaluation, the BERT model was applied to five randomly chosen historical volumes of annual reports. These volumes were split into individual pages using the OCR page delimiter through regular expressions. The model was applied to these pages to

https://doi.org/10.23106/dcmi.952591060

determine which were title pages. The script produced a .csv file for each volume, including texts with coressponding page number, probability of the text being a title page, and a resulting predicted label of 1 or 0. Detected title pages can then be used to determine publication date information, year ranges, and missing volumes, improving MARC metadata, and splitting volumes into their constitutive reports.

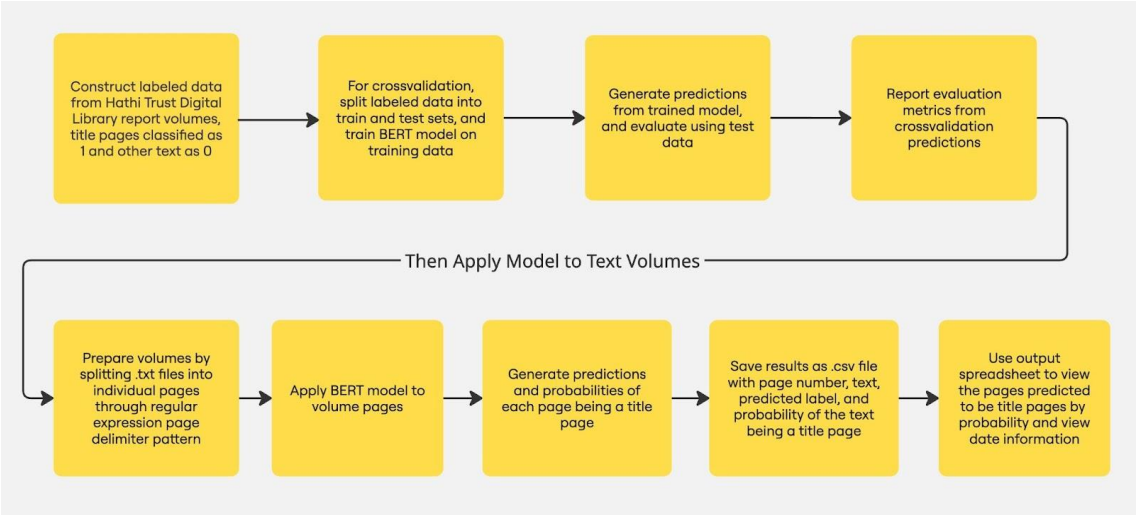

Figure 2: Flow chart of using labeled data and BERT model to detect annual report title pages within digitized volumes.

3. Results

3.1. Gemini

The results of title page identification, volume information extraction, summary statement generation, and metadata enhancement using AI technologies were encouraging. The Gemini API performed reliably with the designed prompt and effectively supported batch processing of digitized serial content.

The first Python script produced a structured list of volume information for each digitized item. The second script consolidated these outputs into a single summary statement, identifying duplicate and missing volumes. The final summary expressed the full digitized range as "first volume – last volume," with any gaps clearly noted. This summary was programmatically integrated into the existing MARC.XML metadata, as illustrated in Figure 3. The enhanced metadata enables more accurate and accessible representation of digitized serial holdings.

```
<datafield tag="866" ind1="4" ind2="1">
            <subfield code="a">1898-1922</subfield>
            <subfield code="z">Missing volume: 1900</subfield>
</datafield>
```

Figure 3: Updated MARC.XML record with accurate digitized serials volume coverage years and missing volume information.

https://doi.org/10.23106/dcmi.952591060

However, the process also raised important questions that merit further discussion—particularly in the areas of volume disambiguation, edge cases in OCR or AI misidentification, and long-term maintenance of AI-enhanced metadata.

3.2. BERT Model Evaluation

Through the initial cross-validation work, the BERT model was evaluated for its title page predictions against "ground truth" manually labeled test data. The model had an accuracy score of .8889, and a 0.9000 score for precision and recall. The model had strong accuracy and predictive ability. Moving forward, the model can be tested and continually evaluated on a larger data set.

The BERT model was used to detect title pages from five new annual reports contributed from different institutions within the HathiTrust Digital Library. These volumes are from the Ontario Historical Society digitized by Indiana University (ID: inu.30000117897003), the Board of Trade of the City of LaCrosse, WI digitized by the New York Public Library (ID: nyp.33433020893230), The Archaeological Institute of America digitized by Princeton University (ID: njp.32101067642437), The Michigan Historical Commission digitized by the New York Public Library (ID: nyp.33433091140131), and the American Seaman's Friend Society digitized by Harvard University (ID: hvd.32044100865286).

page_number	text	predicted_label	title_prob
392	TWENTY-FIFTH ANNUAL REPORT OF THE AMERICAN SEAMEN'S FRIEND SOCIETY. PRESENTED MAY 9, 1853. NEW-YORK: AMERICAN RAILROAD JOURNAL BOOK AND JOB PRINTING OFFICE. 1853.	1	0.9887802600860600
54	EIGHTEENTH ANNUAL REPORT OF THE AMERICAN SEAMEN'S FRIEND SOCIETY. 1846. NEW YORK: JENNINGS & DANIELS, PRINTERS, 12 NASSAU STREET.	1	0.9886954426765440
317	1802, Feb. 17. Girt of the Society	1	0.9886314272880550

Table 2: Example of detected title pages and a false positive from the American Seamen's Friend Society

The BERT model produced probabilities for identifying title pages from the complete digitized serial volumes. For each volume, it identified the title pages through the probability of prediction accuracy. The model was able to correctly label all title pages within these volumes. Pages with a determined title page probability of above 0.9850 captured almost all title pages, with a few false positives included for the American Seamen's Friend Society volume. An example of correct labels and a false positive is presented in Table 2. Duplicate title pages are detected as well. This establishes a strong confidence threshold for detecting title pages using this model. Detected title pages reveal further information about the annual report years within

https://doi.org/10.23106/dcmi.952591060

the volumes. The model detections confirmed the publication year range for volumes from three titles, the Archaeological Institute, LaCrosse Board of Trade, and Ontario Historical Society. The model reveals that the volume of the Michigan Historical Commission digitized by the New York Public Library only contains one report, May 28 to December 31, 1913, published in January of 1914. However, this volume was labeled as 'v.3-4 1913-1916.'

The model also reveals that the American Seaman's Friend Society volume from Harvard University contains reports from the American Bethel Society from 1860, Western Seamen's Friend Society in 1860 and 1861, and Boston Seaman's Society in 1860. Specific date information for these additional reports included in the volume are not mentioned in the metadata. The metadata includes these titles in the 580 field. The BERT model is successful in detecting title pages within historical annual reports, and should be improved through more data, evaluation, and application to other serial publications. The predictions of the model can improve metadata, confirming, correcting, and adding to information on included publication years and specific reports within historical volumes.

4. Discussion and Next Steps

The current approaches of using Gemini and BERT models and selected digitized content for the research were tailored for annual reports—among the simplest forms of serial publications. These reports are typically issued once per year, with the publication year commonly serving as the volume designation. While the script using a Gemini prompt includes logic to handle noise and duplication (e.g., "Some PDFs may contain duplicate title pages for the same year. Only include each year once"), the core prompt was optimized for a specific use case, with instructions such as: "Identify all title pages that follow the structure of an 'Annual Report' or similar government publication," and "From each matching title page, extract the year (as a four-digit number)." The BERT model learns to recognize title pages from multiple irregular examples, but only draws from historical annual report volumes. Further coding to handle outputs could also work to remove duplicates and noise to find the correct date information.

To adapt the approaches for broader batch processing of various serial types, the prompt and the process must be scalable and flexible. Two possible directions for enhancing scalability are:

- Incorporating Bibliographic Metadata for Publication Patterns - information already available in serial bibliographic records can guide and refine the extraction process. Specifically, control field 008 position 18 indicates publication frequency and position 19 indicates regularity, and datafield 310 provides the current publication frequency. These MARC fields in bibliographic metadata can help define expectations for how volumes are structured and labeled, enabling more intelligent filtering and pattern recognition across different serial types.
- Training the Model on Title Page Structures - a second path is training the AI models to recognize a variety of title page structures beyond annual reports. By building a training dataset of labeled title pages—representing diverse serial formats, government documents, journals, and other periodic publications—the model can generalize more effectively to new inputs. This would support broader applicability for title page detection and volume extraction in batch workflows.

https://doi.org/10.23106/dcmi.952591060

The current method demonstrates that AI can be effectively applied to assist in enhancing holdings metadata, beyond bibliographic metadata creation, that have not been addressed in recent years. At the same time, it reveals new opportunities for scaling these methods across various types of serial publications. Achieving this may require metadata-driven strategies and enhancements to AI models that account for the complexity and variability inherent in digitized serials.

Acknowledgements

This project was supported by the Berthold Family Research Fund for Information Access and Discovery.

References

[1] CONSER Cataloging Manual Module 2: What is a Serial?, 2025. URL: https://www.loc.gov/aba/pcc/conser/CCM/Module2.pdf.

[2] ANSI/NISO Z39.71-2006 (R2011) Holdings Statements for Bibliographic Items, 2011. URL: https://www.niso.org/publications/z3971-2006-r2011.

[3] ISO 10324:1997: Information and documentation — Holdings statements — Summary level, 1999.

[4] HathiTrust Digital Library. https://www.hathitrust.org/.

[5] Google Gemini 2.0 Flash. https://gemini.google.com/app.

[6] Devlin, Jacob, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. arXiv: https://arxiv.org/abs/1810.04805.

[7] HathiTrust Bibliographic API. https://www.hathitrust.org/member-libraries/resources-for-librarians/data-resources/bibliographic-api/.

[8] Koroteev, M. V. BERT: A Review of Applications in Natural Language Processing and Understanding. 2021. arXiv: https://arxiv.org/abs/2103.11943.

https://doi.org/10.23106/dcmi.952591060