Information Research – Vol. 31 No. iConf (2026)

‘Appears to be about’: an evaluation of AI-generated metadata quality for community archives

Nikki Wise, Katrina Fenlon, Diana Marsh, Amanda Sorensen, Ugoma Smoke, Candy Navarrete, and Lucy Havens

DOI: https://doi.org/10.47989/ir31iConf64144

Abstract

Introduction. We report on an evaluation of the quality of metadata generated by a general purpose chatbot using items from a community organisation archive.

Method. We developed an evaluation framework adapting quality dimensions from prior work and applied it to analyse a sample of 140 Dublin Core metadata records created by ChatGPT 4o from primary sources drawn from a community organisation collection, based on informal prompts.

Analysis. Using independent qualitative coders and a peer review process, we assessed accuracy, conformance, consistency, completeness, objectiveness, transparency, bias, engagement, meaning and context, understandability, and provenance.

Results. We found approximately 70% of elements to be accurate. Most records were substantially complete and objective but often vague. Records exhibited significant inconsistencies in how ChatGPT completed fields, conformed to the Dublin Core schema, and interpreted primary sources.

Conclusion. General purpose AI chatbots have the capacity to provide substantial ‘rough draft’ descriptive records for community collections, even with minimal prompting. These records require significant human intervention to ensure quality in terms of completeness, conformance to schema, accuracy, and meaningfulness to users. We offer insights for organisations and communities working with AI chatbots for description, along with implications for broader archival practice.

Introduction

Vast troves of cultural knowledge are held in analogue primary source collections in community and institutional archives. Previous research has shown that 70-80% of institutional collections remain unprocessed, left largely as they arrived at a given archive, and are not discoverable on the web (Prud'homme & Compton, 2020; Bucciferro, 2008; Greene and Meissner, 2005; Panitch, 2000; Marsh, 2019). By some estimates, 98-99% of collections remain undigitised (Weissner 2024). The labor required for digitisation and description present an overwhelming obstacle to making most collections accessible online. This challenge is most acute for small institutions and community archives. Generative artificial intelligence (AI) offers an opportunity to make analogue collections more discoverable through augmented approaches to metadata creation, to meet longstanding, widespread community needs.

This paper undertakes a qualitative evaluation of metadata produced by ChatGPT 4o (OpenAI, 2022) for a community organisation's archive. Small institutions and community-based repositories steward invaluable community histories and knowledge, often relying on small staff and volunteers, borrowed space, and minimal resources. These groups often conduct their work without access to significant computing resources or the rapidly evolving expertise required to train AI for special purposes or deploy it programmatically. This study therefore explores the use of an openly and freely accessible AI chatbot, in line with the resources accessible to many community archives and under-resourced cultural institutions. The goal of this study is to determine whether a general-purpose AI chatbot may represent a useful addition to the workflow of a community archive, given the quality of metadata it produces with informal, inexpert prompting.

In prior work, we asked ChatGPT to generate metadata for a collection of 140 digitised objects—mainly PDF and JPEG files representing a wide range of paper documents, photos, and other artifacts—from the National 4-H Council History Preservation Team (4-H History Preservation Program, n.d.). This collection, gathered and maintained by a group of volunteers, reflects the 120-year history of 4-H (4-H, n.d.), the largest youth development organisation in the United States. In this paper we conduct a qualitative evaluation of the content of the resulting records. The evaluation relies on a qualitative coding framework adapted from several prior metadata quality frameworks and builds upon the convergent bodies of work on evaluating human-generated metadata and machine learning and AI outputs (Brzustowicz, 2023; Taniguchi, 2024; Zavalin & Zavalina, 2025).

This study is one phase of a larger project, '*linking anthropology's data and archives*' (NSF Cultural Anthropology, #2314762), exploring sustainable infrastructures for making data from anthropological, Native, and Indigenous collections more accessible to communities of origin. The collection studied in this paper, however, is not a Native or Indigenous collection, but rather the digital archive of a community organisation. We started with a community archive to mirror the resource conditions confronting many anthropological archives, but to avoid experimenting with culturally sensitive materials until we have an established sense of pitfalls. Our ongoing research continues this study in partnership with anthropological collections of Native and Indigenous materials and their community representatives.

Prior work

AI in cultural heritage

In light of the enormous backlogs confronting description and processing in most cultural institutions (Prud'homme & Compton, 2020), work on using AI tools for collections processing and description is a burgeoning research area. AI tools have been applied to generate image captions and alt-text (Berger, 2024; Männistö et al, 2022; Marinescu et al., 2020); metadata, subject terms, and entity recognition (Brador, 2024; Carter et al., 2022; Magnus et al., 2025; Sun et al., 2025; Suominen, 2019); digital facsimiles and structured transcriptions (Nockels et al., 2022; Pepper et al., 2024; Hosseini et al., 2022); and new ancillary research datasets (Yu et al., 2024). Dedicated tools are proliferating both within extant professional applications and as new, bespoke platforms such as JSTOR's Seeklight (JSTOR, n.d.; Society for American Archivists, n.d.), CatalogerGPT (Taniguchi, 2024), and Alma's AI Metadata Assistant (Ex Libris, n.d.).

Recent work has explored the application of AI to reparative processes for human-generated metadata. In a user study of Anthropic's *Claude* model, Roke (2025) found the AI capable of flagging bias in existing records and supporting updates to original language. This study found that human users preferred machine- and hybrid-generated metadata over human-generated content, citing its increased context, specificity, and comprehensiveness (Roke, 2025). An expanded study evaluated the interpretive abilities of three further large language models (LLMs) to assist human judgment processes as part of reparative description workflows (Osti & Roke, 2024). Despite limited training and even without access to original primary sources, LLMs showed capacity to identify gaps, harmful language, and latent context in original records (Osti & Roke, 2024). Their findings, joining a growing consensus in the cultural heritage community, suggest promise for hybrid human-AI workflows in metadata creation and remediation.

Significant ethical concerns attend the use of AI in cultural heritage. The environmental impact of LLMs is potentially devastating (Crawford, 2024; Kneese, 2024; Kneese & Young, 2024; Rotman, 2025). Environmental costs disproportionately affect marginalised communities (Kneese, 2023; Kneese & Young, 2024; Temple, 2025). Calls for increased transparency and governance around AI energy use and research on using smaller models with less environmental impact, improving training and inference efficiency, and integrating renewable energy sources is growing, but remains too nascent to alleviate environmental concerns around AI use. In addition, there are ethical concerns around how data used in training are and were collected, including through cultural theft (Mollema, 2024; Sundararasan, 2024; Walter and Russo Carroll, 2020) what information LLMs make visible (Widder & Kneese, 2025;); and the precarious labor conditions of those who train AI (Jaillant & Aske, 2024; Kneese, 2023; Meaker, 2023; Wen, 2014). The adoption of AI for cultural heritage metadata is not without potential harm, though research and practice in this area (including this study) represent an infinitesimal sliver of AI's overall impact as AI integrations are increasingly ubiquitous and LLMs now field billions of queries daily. Given their demonstrable potential to improve description and therefore access (Fenlon et al., 2025; Jaillant et al., 2025; Osti & Roke, 2024; Ray et al., 2025; Roke, 2025), AI may offer a singular opportunity to provide meaningful access to collections for many under-resourced communities.

Metadata quality assessment

We build upon a large body of prior work on evaluating metadata quality, generally focused on human-generated metadata. Metadata quality assessment addresses both semantic and syntactic aspects (Bruce & Hillmann, 2004, p. 9), though in this paper we are focused on semantic evaluation. In practice, assessing quality can be challenging due to resource constraints, the diverse

backgrounds of those implementing metadata, and the variety of materials that metadata describe (Bruce & Hillmann, 2004). Stvilia and Gasser (2008) found that repositories and end-users could have different models of metadata quality, so a single quality assessment may not reflect quality for different groups and uses of metadata.

This study adapts several prior frameworks for metadata assessment, including the ALA Metadata Assessment Framework (Metadata Schema Assessment Framework, n.d.), Data Europa's Metadata Quality Assessment (*Metadata Quality*, n.d.), the DLF Metadata Assessment Framework and Guidance (*Metadata Assessment Framework and Guidance*, n.d.), and Europeana Report and Recommendations from the Task Force on Metadata Quality (*Report and Recommendations from the Task Force on Metadata Quality*, 2013). These frameworks have built on one another to encompass a wide range of '*dimensions*' for assessment, including relevance, accuracy, timeliness, accessibility, interpretability, coherence, completeness, conformance to expectations, provenance, logical consistency, timeliness, accessibility, and more (Bruce & Hillmann, 2004). Margaritopoulos et al. (2012) draw on how courts of law define the quality of a witness statement, in terms of '*the truth, the whole truth, and nothing but the truth*' – corresponding to correctness, completeness, and relevance. In our methods section, we explain how we identified and adapted dimensions from various frameworks. These dimensions are core to most evaluations of metadata quality in practice (Park & Tosaka, 2010). Further work has identified quantitative metrics and logic-based techniques for computational analysis of quality (Margaritopoulos et al., 2012; Gavrilis et al., 2015; Stvilia & Gasser, 2008), but we took a primarily qualitative approach in this study, in light of the expected idiosyncrasies introduced by AI-generated metadata that would be difficult to capture and assess quantitatively.

AI-generated metadata quality

Work on assessing AI-generated metadata quality emerges from a long tradition of evaluating machine learning and AI systems. Many decades of work on '*benchmarking*', or comparatively evaluating the outputs of algorithms against agreed-upon protocols or metrics, are now being adapted to evaluate the reasoning and accuracy of LLMs. However, these remain too general to assess context-specific tasks such as metadata generation in archival contexts (Hutchinson et al., 2022; Kugler 2025, p. 16; Raji et al. 2021). Our study aims to meet a need for context-specific, human-centered approaches to AI evaluation (Havens et al., 2025).

Because LLMs are designed to provide outputs based on linguistic likelihood rather than truthfulness or accuracy (Fisher 2024, Hicks et al. 2024) and because their training data are known to be problematic in terms of representation and bias (Birhane et al. 2023; Ciecko, 2020; Foka et al., 2025; Hoffman et al., 2024), questions around metadata evaluation are particularly acute for cultural institutions as trusted stewards of cultural knowledge. Many metadata quality assessments are geared toward assessing the quality of human-generated metadata, but automated approaches to creating and evaluating metadata are not new (Greenberg, 2005; Zavalin & Zavalina, 2025). Much of this literature suggests a growing consensus that LLMs can improve processes in terms of efficiency and comprehensiveness of description, but require significant human oversight (Brzustowicz, 2023; Chow et al., 2024; Taniguchi, 2024; Zavalin & Zavalina, 2025).

Zavalin and Zavalina (2025) assessed the quality of metadata output across four standards (Dublin Core, MODS, MARC, and BIBFRAME) and three generative AI tools (ChatGPT, Gemini, and Gemini Advanced). Assessing quality based on accuracy and completeness, they found that overall, AI generated metadata 'does not meet basic functional expectations' for metadata, though ChatGPT had better completeness outcomes than Gemini and Gemini Advanced. Taniguchi (2024) compared

ChatGPT-generated MARC 21 records to human-generated records for the same items finding that ChatGPT had promise and could assist human cataloguers but that it struggled with '*complex bibliographic patterns and nuanced cataloguing rules*' (p. 544). Breeding (2023) tested ChatGPT for creating a MARC record for one book and found that although title and author were correct, no other elements were reliably correct (p. 18). Brzustowicz (2023) used ChatGPT to generate MARC records in RDA and Dublin Core finding that ChatGPT can generate accurate records conforming to multiple metadata standards, for items in various formats and languages, and found that ChatGPT could accurately and efficiently extract essential metadata elements (p. 4). However, Brzustowicz recogises that even with this accuracy, any records generated by an LLM need review by human cataloguers to mitigate bias of the training data and mitigate the risks of copyright infringement, authorship misattribution, and disclosure of sensitive information (2023, p. 6).

Taniguchi (2024) asked ChatGPT to create MARC records for 105 different information sources, including books, maps, recordings, and sheet music, that already had MARC records (pulling from *Maxwell's Handbook for RDA* by Robert L. Maxwell). Although these items already had MARC records, Taniguchi used a '*zero-shot*' approach with ChatGPT asking that it create a MARC record based only on the information source they scanned, not providing the existing MARC record. Brzustowicz (2023) assessed AI-generated metadata quality for six different items, providing the AI with only the source's title. An example prompt from Brzustowicz (2023) is '*Can you generate a MARC record using RDA for Mood Rings*' 2013 single '*Pathos Y Lagrimas.*' Five of six of these items had an existing MARC record. The sixth item was 'specifically chosen to test ChatGPT's ability to generate an original record' (Brzustowicz 2023, p. 2). This sixth item most closely parallels our collection of unprocessed items and our project goal in understanding ChatGPT's ability to create original metadata records.

Our study builds on prior work by evaluating a larger collection of materials than was assessed in most other studies and focusing on a community archives context with minimal prompting.

Methods

In an earlier phase of this study, we used ChatGPT to generate item-level Dublin Core records (as well as transcripts and other data outputs) for 140 items from the 4-H collection, using a series of informal and dialogic prompts. Our sample size was constrained by the size of our research team and restrictions set by the free version of ChatGPT 4o, which limited users to 20 file uploads. The methods used to generate metadata are described fully in Fenlon et al. (2025). Our research team comprised non-expert AI users with varying degrees of expertise in relation to metadata schemas and technologies. We did not deploy systematised prompts because the goal of this work was not to evaluate AI effectiveness under ideal conditions or with optimised prompting, but instead in the context of intuitive use—how people without AI training or expertise in community contexts interact with the chatbot intuitively. Metadata records were added to a custom-built Airtable cloud-based relational database to support qualitative analysis by a team of coders.

We developed an evaluation framework by adapting dimensions of quality defined in many prior frameworks for evaluating the quality of metadata (*Metadata Assessment Framework and Guidance*, n.d.; *Metadata Quality*, n.d.; *Report and Recommendations from the Task Force on Metadata Quality*, 2013; Huang et al., 2016), metadata schemas (*Metadata Schema Assessment Framework*, n.d.), linked data (Zaveri et al., 2015), and data produced by AI (Dilmegani, 2025; *Implementing Australia's AI Ethics Principles in Government*, 2024; Schwabe et al., 2024). We also used existing scoping reviews on metadata quality frameworks to understand patterns and commonalities across such

frameworks (Kumar et al., 2025). The research team gathered and conducted a systematic comparison of prior frameworks and areas in which they overlap. Based on this analysis, and after merging overlapping dimensions, we identified a set of dimensions for analysis given in Table 1. We define selected dimensions in relevant sections of '*Findings*', below. Our codebook included each dimension with its original definition as given in prior frameworks (see Table 1), quantitative indicators of certain dimensions (Table 2), and an adapted set of prompts for our coders to guide consistent interpretation.

Accuracy • DLF Metadata Assessment Framework and Guidance (n.d.) • DLF Inclusive Metadata Toolkit (Huang et al., 2016) • Data Europa (Metadata Quality, n.d.)	**Objectiveness** Király, P. (2018). Metadata quality in cultural heritage institutions. Workshop on FAIR Principles for Digital Research Data Management.
Consistent Representation/Consistency • DLF Metadata Assessment Framework and Guidance (n.d.) • DLF Inclusive Metadata Toolkit (Huang et al., 2016)	**Provenance** • DLF Metadata Assessment Framework and Guidance (n.d.)
Completeness • DLF Metadata Assessment Framework and Guidance (n.d.) • Király, P. (2018). Metadata quality in cultural heritage institutions. Workshop on FAIR Principles for Digital Research Data Management.	**Transparency, Bias, and Engagement** • ALA Core Metadata Standards Committee Metadata Schema Assessment Framework (2024)
Meaning and Context • Europeana Report and Recommendations from the Task Force on Metadata Quality (2013)	**Understandability** DLF Metadata Assessment Framework and Guidance (n.d.) DLF Inclusive Metadata Toolkit (Huang et al., 2016)
	Conformance (Novel contribution)

Table 1. Dimensions of quality and frameworks from which they were adapted.

Using this codebook, each of the 140 records was assessed by one of five independent coders. Early in the coding process, an independent second coder conducted peer review of 40 records (approximately one third of the sample). Following the peer review exercise, the coding team discussed their observations and came to a shared understanding of the meaning and application of each code before completing coding. In parallel to this manual, qualitative analysis of metadata contents, we undertook a separate computational analysis of completeness and conformance to syntax standards. However, the evaluation reported in this paper does not take syntactical aspects of completeness and conformance into account.

Findings

In part because our team took an intentionally informal approach to prompting, the records demonstrate variety in their syntax and representation. Our dataset includes 95 records expressed

as Dublin Core XML, 37 expressed as simple, plain text lists of elements and values, and 8 expressed in JSON. Most of the plain text records ChatGPT created would readily convert to structured representations, while several (7) lacked punctuation or other syntactical elements that would support automatic conversion into, for example, JSON, XML, or CSV. Sometimes ChatGPT provided the Dublin Core record in an RDF wrapper. We could discern no pattern to when ChatGPT represented records in different ways. This evaluation does not consider aspects of syntax but focuses on the values or contents of Dublin Core metadata elements. Below we report selected results of our assessment, especially in relation to completeness, accuracy, conformance to the Dublin Core schema, objectiveness, transparency, meaning and contextuality, and provenance.

Quantitative indicators of completeness and accuracy

Our definition of '*completeness*' was based on the Digital Library Federation (Huang et al. 2016) and Király (2018): 'The element, property, and/or attribute is present. Number of metadata elements filled out.' We refined this for qualitative assessment by prompting coders with questions like:

> To what extent or how is the record complete (or incomplete), both in the sense of the extent or number of elements it uses to describe a primary source, and how fully or completely it describes each element? How fully described is it?

Our interpretation aligns with definitions of completeness and accuracy used in prior work (Aljalahmah & Zavalina, 2024; Zavalin & Zavalina, 2023; Zavalina & Burke, 2021). In addition to analysing completeness qualitatively, we identified four indicators of completeness that are countable:

- The total number of elements used in each record;
- The number of unique or distinct elements used in each record;
- The total number of elements given as '*unknown*', '*unspecified*', 'N/A', or left blank;
- The number of '*unknowns*' that a human rater judged to be correct or accurate to the primary source (as opposed to 'unknowns' that indicated a mistake).

Our definition of '*accuracy*' was based on the definition from Király (2018): '*data correspond to the resource that is being described.*' We refined this for qualitative assessment by prompting coders with: '*To what extent and how are the metadata in the Dublin Core record accurate (or not) to the primary source being described?*' As a quantitative indicator of accuracy, we counted the number of elements judged by our coders to be accurate to the primary source.

Quantitative indicators	Results (n=140 records)
Indicators of completeness	
Average number of elements in each record	14 per record
Average number of 'unknown' elements	1.49 per record
Mean correct 'unknown' elements	81.02%
Median correct 'unknown' elements	100%
Indicators of accuracy	
Mean accuracy of elements	69.55%
Median accuracy of elements	71.42%

Table 2. Quantitative indicators of completeness and accuracy.

In most records, ChatGPT used all or most of the 15 original Dublin Core elements (originally defined in the '*elements*' namespace). In just a few instances, ChatGPT drew on terms unique to the extended '*terms*' namespace, such as *audience* or *extent*. (It more frequently used the '*terms*' prefix while relying on the core set of 15 terms that exist in both namespaces.) The average number of elements for records that had one or more '*unknown*' elements was 14.4 while the average number of elements for records with no unknowns was slightly lower at 13.75. This may be because ChatGPT sometimes omitted elements it did not know. A little over half of the records (75 out of 140) contained one or more '*unknown*' elements. In 52 of those, ChatGPT correctly identified most or all unknown elements for an item. The mean and median accuracy of each record was around 70%, so a 15-element record had 10 or 11 accurate element values. Records with no unknown values had the same average accuracy as records that had one or more unknowns (69.6%).

Completeness

In addition to the quantitative indicators of completeness, we assessed completeness qualitatively. Most records were relatively complete in the strict sense of having a reasonable number of elements with values. Sometimes, ChatGPT provided overly complete metadata records despite a lack of information. One primary source, a document without any context, contained only the words, '*4-H / For Youth / For America*'. ChatGPT created a Dublin Core record of 15 elements for this item, with four unknowns, and only one of those unknowns was judged to be incorrect. However, many elements had vague or less than helpful values.

The elements *contributor*, *creator*, *rights*, *date*, and *publisher* were most frequently cited as unknown. The *rights* element was most likely to be correctly marked as unknown because ChatGPT could not derive such meta or contextual information directly from primary sources, and we did not include that information in our prompts.

Rather than mark an element as unknown, ChatGPT would sometimes create template or placeholder values with guidance for a human curator, or it would omit elements entirely. Sixty-four items (45%) had fewer than 15 elements, suggesting that ChatGPT left out unknowns. The metadata record created for Item #110 contained multiple placeholder elements such as '*Date:*

[Insert estimated date, if available, e.g., 1950s or 1960s]' and '*Relation: [Optional: Could link to related items, e.g., other 4-H promotional materials]*'. However, ChatGPT would inexplicably omit elements when the values were evident to a human curator or were inferable, judging by ChatGPT's previous inferences. In multiple cases, this incompleteness seemed to be caused by ChatGPT's unpredictable determination about how and when to use what it termed 'OCR' to recognise and interpret text content. For example, one 56-page text document (item #44) had the potential for a robust record but included a paltry 2-sentence description. In another case, ChatGPT seemed to use OCR to correctly summarise a long text document (item #55) in the *description* field, but then did not leverage that information to complete other possible fields in the same record.

Accuracy

As noted above, the records achieved an average 70% accuracy rating, in terms of the number of elements that were technically correct according to human judgment. However, this rating belies many cases in which element values were so vague that they verged on being unhelpful. For example, the *creator* listed for item #96 '*4-H Subcommittee Meeting Minutes*' was '*4-H Subcommittee*' but the item names a more specific individual creator.

Our qualitative accuracy observations revealed that ChatGPT often extracted or inferred information from the item's filename in addition to the content of the primary source. Many filenames in this collection included descriptive or title information, or indicators of the date of digitisation. There was no discernible pattern to when ChatGPT elected to use information from filename versus the content of the file. Sometimes ChatGPT used information from the filename when equivalent information was not discernible from the item's content; sometimes it used filename information instead of information available in the content; and sometimes it ignored filename information even when it might have added critical context to an item.

For example, ChatGPT often pulled *date* values from filenames—usually when dates were not provided in the item. ChatGPT correctly interpreted 'fa24' in a filename as referring to the fall semester of 2024, for example, adding that value as the *date*, but its presumptions were often incorrect. The date of digitisation was rarely the best date to include for an item and was sometimes used in place of a readily discernible date of creation.

In one notable case, ChatGPT ignored key format information provided in a filename. '*Tape.pdf*' constituted a scanned image of a VHS tape. Not recognising the format clue in the filename, and not recognising the image as a scan of a three-dimensional artifact, ChatGPT made a series of incorrect descriptive statements, including failing to recognise the name of a historical A/V production company written on the tape's cover, 'The Production Center at Arthur Young,' and interpreting 'Arthur Young' as the human creator of a brochure.

ChatGPT encountered difficulty distinguishing named entities into *creator*, *contributor*, and *publisher* roles. As many items did not make this information explicit, this confusion is not surprising. However, in some cases, ChatGPT included all or many names appearing in a document as *contributors* when they were simply named entities discussed in an item's content.

Meta- and collection-level elements such as *rights*, *source*, and *relation* were unsurprisingly most difficult for ChatGPT to accurately provide, given that we provided items one by one and offered no high-level context for the collection in our prompts. Rather than mark these elements as unknown or omit them, ChatGPT frequently conjured believable but inaccurate values. This is a known tendency of LLMs. For example, ChatGPT described multiple plausible but imaginary collections of materials to contextualise items in the *relation* field, such as '*part of the cooperative*

extensions system reports' (Item #111) or '*National 4-H Conference records and publications*' (Item #138).

Many other inaccuracies were unpredictable. For example, in one item, ChatGPT was able to recognise the *creator* and *publisher* within the content's handwritten script but missed the nearby *date* information. In one case, ChatGPT created a metadata record based on a transcription it had also created. In this case the transcription was completely inaccurate, including incorrect or made-up headings and sections of text, contact information, and more. The metadata record was accurate to the transcription, but the transcription was a fake. Sometimes, ChatGPT simply failed to follow instructions, and provided an explanation of its behavior in place of the requested metadata. For example, a subject value in one record was 'Extracted text from the first page of the PDF' (item #67).

Conformance

While prior evaluation frameworks have included dimensions of '*conformance*,' these generally refer to how well metadata conforms to user expectations or content standards (such as current subject authorities). Because we are evaluating AI metadata, we adapted the '*conformance*' factor to consider to what extent the metadata records conform to the requirements of the Dublin Core schema or adhere to its recommended practices. In a few records, elements were invented or drawn from other schemas. One record included a *location* element (item #104). Another drew terms from a Schema.org schema, such as the *about* property to represent subjects (*Schema.org*, n.d.).

Almost all records fell short of Dublin Core best practice recommendations in two main ways. First, all records but one failed to leverage the repeatability of elements, instead lumping multiple values into one field (such as multiple terms listed within a single *subject* element). For example, this value in one *coverage* field conflated spatial and temporal aspects of coverage into one long statement: '*Historical events from 1923 to 1927, with reference to activities in multiple U.S. states and Canadian provinces*'. Such conflation might be acceptable in metadata that were generated as informal, plain text lists of element/value pairs, but this also occurred in records expressed in XML and JSON in almost every case.

Second, records rarely adhered to Dublin Core's recommended practices for representing values, such as the recommendation to use URIs in place of string literals, to draw on controlled vocabularies for types, subjects, and format, or to format dates in accordance with international standards. The only apparent authorities deployed in a few scattered records were language codes ('en', occasionally), the DCMI Type Vocabulary (*DCMI Metadata Terms*, 2020), and IANA Media Types (MIME Types) (*Media Types*, 2025), which are explicitly referenced by and hyperlinked within the Dublin Core documentation.

Most records included *subject* terms that resembled those a human curator might use. In some cases, terms were too vague to be helpful. In some cases, the *subject* field included an overly long description, verging on a full sentence, e.g., 'Draft document with comments related to donors, organisational fundraising, and historical clarifications in the context of National 4-H Council and other affiliated entities.'

Sometimes ChatGPT seemed to get into a '*rut*' of creating elements according to a syntactical pattern, record after record, within a chat session. For example, after successfully creating a record that used an accurate IANA Media type for its *format* element (e.g., '*application/pdf*'), ChatGPT went on to mimic the forward-slash syntax of IANA Media Types but using terms that are

not part of that vocabulary—e.g., '*text/text document*' and '*image/scanned document*'—in multiple subsequent records.

A few records violated Dublin Core's one-to-one principle, '*that conceptually distinct entities, such as a painting and a digital image of the painting, should be described by conceptually distinct descriptions*' (*One-to-One Principle*, 2011). For example, in one record, some fields accurately described the primary source as a conference booklet created by '*IFYE World Conference Organising Committee*' (*Item #61*), *while other fields made reference to the ChatGPT-generated transcription of the PDF booklet*: '*<dc:description>Transcription of the IFYE World Conference 2003 booklet detailing events, programs, and cultural exchanges.</dc:description>*' and '*<dc:format>PDF Transcription</dc:format>*'. The booklet and its text transcription are conceptually distinct entities that should not be described by one record.

Objectiveness and transparency

The evaluation framework judged '*objectiveness*' by whether '*values describe the resource in an unbiased way*' (Király, 2018). We adapted this definition using the following prompts for coding: 'To what extent or how is the record objective in its descriptions? Does the record make assumptions that are or are not warranted based on the primary source, or which suggest value judgments? Similarly, but separately, we evaluated '*transparency, bias, and engagement*' (adapting multiple definitions, but especially drawing on the Metadata Schema Assessment Framework) as *the* '*extent to which metadata schema acknowledges and documents possible bias*'. We coded this factor using the following prompts: *'To what extent or how does the record acknowledge, document, make explicit or make transparent its own possible biases or assumptions, or aspects in which the tool has played an interpretive role?*'

Coders found the ChatGPT-generated records demonstrated objectivity, mainly because the LLM often created metadata using excerpts from text transcriptions in this document-heavy collection. In a few records, the LLM added value terms, suggesting that a document contained '*important information*,' for example, without context. ChatGPT sometimes added hedging language to indicate uncertainty or presumptions, such *as* '*appears to be about*,' 'likely,' and '*potentially*'. Occasionally ChatGPT attributed the source of a presumption, for example by including the following parenthetical in a *subject* field: '*United States (assumed based on IRS reference)*'. In this way, the ChatGPT occasionally made its decisions or uncertainty explicit. More often, however, uncertainty, presumptions, and outright inventions were provided without notice.

Meaning and context

We combined multiple categories from prior frameworks in relation to whether metadata make items meaningful to users or provide adequate contextual information for use (*Report and Recommendations from the Task Force on Metadata Quality*, 2013) under the umbrella of 'Meaning and Context'. For this factor, we judged to what extent and how the records are made meaningful to audiences or provide additional contextual information to enrich their meaning and usefulness.

Most records included sufficient information to be meaningful to potential users but rarely included additional context, with a few exceptions. In one record ChatGPT correctly inferred that a single reference to '*Mr. Pressler*' in a scanned image, specifically a JPG representation of a page from the *Congressional Record*, referred to '*Senator Larry Pressler*,' and included his full name in the *creator* and *description* fields. Occasionally, ChatGPT also correctly interpreted acronyms in primary sources and spelled them out in the metadata, improving accessibility. For example, ChatGPT correctly inferred that '*UMCP*' and '*CES*' referred to the University of Maryland College

Park Cooperative Extension Service in a document explaining tenure and promotion guidance for affiliate faculty and included the full terms as well as acronyms in the metadata record. In many other cases, however, ChatGPT ignored contextual clues, leading to inaccuracies in the metadata based on poorly founded inferences—such as taking a recent date of digitisation from the filename and using it as date of creation, even for objects that a human curator would readily recognise as historical.

Provenance

To assess '*provenance*' we considered whether records provided any information about their source (*Metadata Assessment Framework and Guidance*, n.d.). The majority of records did not provide any provenance information. Sometimes, a metadata record would include provenance information for an item but not its metadata, e.g., '*original physical document scanned into digital format*' (Item #101). Sometimes ChatGPT included information about from where, within a primary source document, it obtained a metadata value, e.g., the *identifier* for Item #21 was given as '*file name*: *B-4.jpg*.' On rare occasions, ChatGPT would acknowledge its own role in creating the metadata. One item's *contributor* was given as, 'OCR processing by ChatGPT' (item 128).

Discussion

The accuracy of '*unknown*' elements was surprisingly high (between 80-100% correct) despite the known tendency of LLMs to guess when uncertain (Kalai et al., 2025; Xiong et al., 2024). When ChatGPT stated that it did not have sufficient information to complete a record, it was almost always correct—so the AI was capable of identifying '*true negatives*.' However, ChatGPT was not reliably expressing uncertainty, leading to a relatively high degree of 'false positives,' or inaccurate elements.

We did not take a systematic approach to prompting in this study, trying to mimic the practices of a nonexpert user interacting with a chatbot in an informal setting. In addition, our necessarily small sample limits the generalisability of our findings, as does the ongoing evolution of AI tools. Nevertheless, our findings shed light on elements of prompts that would make chatbot-generated metadata more useful. For example, in the future we would explicitly ask the LLM to follow Dublin Core recommendations for element values and suggest specific openly accessible authorities for various elements as Zavalin & Zavalina (2025) did. We would provide contextual information about the collection, including source and rights information, and ask ChatGPT to provide more contextual information about named entities in records. We could be more explicit about how ChatGPT handles and expresses uncertainty. We could have asked for records to be expressed in a particular expression language.

OpenAI and other AI providers and researchers are developing increasingly detailed guidance on effective, task-specific prompting (see Long et al., 2025; Marvin et al., 2024). Understanding and adapting more sophisticated prompting protocols could become a technical competency for future archivists and cultural heritage professionals. More likely, prompts will become enmeshed in custom tools developed to support professional practice, which are trained on relevant datasets and designed to gather contextual information without necessitating chatbot dialog. For many communities and institutions, however, such tools may be out of reach. For those working with openly accessible models using chatbot dialog, accessible, general prompting guidance and instruction could go a long way toward improving metadata accuracy.

Conclusion

Our findings resonate with prior work suggesting that foundational, general-purpose AI chatbots can provide substantial '*rough draft*' descriptive metadata records, even with minimal prompting, achieving a basic accuracy rate of around 70%. These records need significant human intervention to ensure metadata quality–in terms of completeness, conformance to schema, accuracy, and meaningfulness—and to prevent unethical disclosures or retention (e.g., of personally identifiable information or culturally sensitive knowledge). Setting aside the more troublesome questions about the environmental sustainability and community ramifications of using AI at scale for cultural collections, from a pragmatic perspective AI would make a useful addition to archival processing workflows, but it needs item-to-item oversight.

Our study addresses the contradictions embedded in this year's conference theme of '*digital enlightenment.*' The same tools and technologies that offer potential for democratising knowledge and access have been built on systems that reify societal stratification and disenfranchisement (Chun, 2004; Wen, 2014; Ziegler 2020). We remain aware of the hazards of technosolutionism and the reality that AI tools have major impacts on small communities and on the environment (Crawford, 2024; Kneese, 2024; Kneese and Young, 2024; Rotman, 2025). Training sets have also shown core biases with major implications for community representation, especially for marginalised or lesser represented communities (Jaillant and Aske, 2024; Widder and Kneese, 2025). Studies of output quality like this one is essential groundwork for enabling communities to weigh the risks and benefits of using AI for their own collections, but more research is needed to engage higher-order questions around cultural authority and sovereignty, private and sensitive knowledge, and other ethical dimensions of the use of AI.

Explainability—the capacity for an AI to explain its own decisions—is a critical component of trustworthy AI (e.g., Steyvers et al., 2025). Transparency, explanation, and communication have also been posited as key aspects of ethical, professional archival practice, especially in relation to data derived from community materials (Ziegler, 2020). Given the extremely unpredictable quality of metadata generated by ChatGPT, AI tools are not currently well suited to engender trust among users, particularly in vulnerable community contexts. As we look to the future of these tools and their development, cultural heritage organisations and professionals can depart from 'traditional' tech development cultures and data brokerism to embrace empathy and transparency in description, collaboration and openness to outside expertise, and active feedback channels to allow their community-informed cocreation (Ziegler, 2020, lines 32-41).

The next stages of our work will facilitate and study conversations about these processes and tools with both partner collections and communities as well as organisations designing AI tools and services for cultural heritage. As Roke (2025) and others have argued, we will need to embrace human-AI collaboration, in which we collaborate not only with community constituencies and repositories, but with the tools and tool creators–unless, looking across the evidence about community and environmental impacts, and as a form of refusal, we decide that the tools and their benefits are not worth the potential costs.

Acknowledgements

We thank the National Science Foundation Cultural Anthropology program for their generous support of this research (NSF CA-SR award #2314762), including for the Career/Life Balance Supplement that enabled project continuity through the co-PI's family leave and supported postdoctoral assistant Lucy Havens. We thank additional members of our data curation team, including Joseph Sioui and Desmond Mantle. We acknowledge the profound generosity of the

National 4-H Council History Preservation Team, especially Thomas Tate and Gwen El Sawi, for sharing this collection as well as their invaluable expertise and time with us. We are enormously grateful to our Council for the Preservation of Anthropological Records (CoPAR) Working Group, who have workshopped this project and its methods with us at key stages, and to CoPAR's Advisory Board, whose members provided important feedback as we drafted our grant proposal. At UMD, we acknowledge the work of Samantha Lee, Maura Matvey, Polly O'Rourke, and Susan Winter for their grant-writing, development, and early project support. We also thank the anonymous reviewers whose insightful comments shaped our final paper.

About the authors

Nikki Wise is a PhD student in the College of Information at the University of Maryland, College Park, USA. She earned her master's degree in Library and Information Science from the University of Maryland with a specialisation in archives and digital curation and a focus on digital archaeology. Her doctoral research examines information infrastructures and information collection and standardisation in the context of human trafficking, sexual assault, and domestic violence. She can be reached at nwise@umd.edu.

Katrina Fenlon is an Assistant Professor in the College of Information at the University of Maryland, College Park, USA. She earned her PhD and master's degrees in Library and Information Sciences at the University of Illinois at Urbana-Champaign School of Information Sciences. Her research interests include sustainability and preservation for digital humanities and scholarly communication, metadata, and data curation. She can be contacted at kfenlon@umd.edu.

Diana Marsh is an Assistant Professor in the College of Information at the University of Maryland, College Park, USA. She earned her PhD in Anthropology (Museum Anthropology) at the University of British Columbia. Her work draws on qualitative and ethnographic methods to better understand the discovery and use of archival collections. Her work focuses on improving discovery of Native and Indigenous collections held in colonial repositories. She can be contacted at dmarsh@umd.edu.

Amanda Sorensen is a PhD Candidate in the College of Information at the University of Maryland, College Park, USA. Her doctoral research examines museum collection management systems and their information infrastructures through time. Previously, Amanda has worked for the Smithsonian Institution, the School for Advanced Research, and the Field Museum. She holds an MA in anthropology with a focus in museum studies from the University of British Columbia. She can be contacted at asorens1@umd.edu.

Ugoma Smoke is a Research Assistant in native and indigenous archives and linked data in the College of Information at the University of Maryland, College Park, USA. She earned her master's degree in Library and Information Science at Kent State University's iSchool, Kent, Ohio, USA, with specialisations in archives and special collections, museum Studies, and digital preservation. Her research interests include community archives, data curation, Indigenous and African American heritage preservation, and metadata governance. She can be contacted at usmoke@umd.edu.

Candy Navarrete is a Repatriation Coordinator and Native Archives Specialist at the Autry Museum of the American West in Los Angeles, CA and Research Assistant for the Linking Anthropology's Data and Archives (LADA) project at the University of Maryland, College Park, USA. They received their MLIS from University of California, Los Angeles and their research interests include

Indigenous Archives, Archival Repatriation, and Indigenous Data Sovereignty. They can be contacted at candymnavarrte@gmail.com.

Lucy Havens is a Postdoctoral Research Fellow in human-data interaction at The Roux Institute at Northeastern University in Portland, Maine. She received her PhD from the University of Edinburgh, and her research interests include natural language processing approaches to analysing cultural heritage metadata and responsible AI, with a focus on AI literacy and the evaluation of AI systems. She can be contacted at hello@lucyhavens.com.

References

4-H. (n.d.). National 4-H Council. Retrieved September 11, 2025, from https://4-h.org/ (Archive Link)

4-H History Preservation Program. (n.d.). National 4-History Preservation Program. Retrieved August 7, 2025, from https://4-hhistorypreservation.com/ (Archive Link)

Aljalahmah, S., & Zavalina, O. L. (2024). Student-Created Dublin Core Metadata Representing Arabic Language eBooks: Comparison of Individual and Group Work Outcomes. Journal of Education for Library and Information Science, 65(3), 325–344. https://doi.org/10.3138/jelis-2023-0016

Berger, T. (2024, October 23). Can You Try Again?: Using Large Language Models to Generate Alt Text for Online Image Collections. The Virtual 2024 DLF Forum. https://osf.io/pc4rx/

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, A. (2023) Science in the Age of Large Language Models. Nature Reviews Physics, 5(5), 277-280. https://doi.org/10.1038/s42254-023-00581-4

Brador, I. (2024, November 19). Could Artificial Intelligence Help Catalog Thousands of Digital Library

Breeding, M. (2023). AI: Potential Benefits and Concerns for Libraries. Computers in Libraries, 43(4), 17–19.

Bruce, T. R., & Hillmann, D. I. (2004). The Continuum of metadata quality: Defining, expressing, exploiting. ALA Editions.

Brzustowicz, R. (2023). From ChatGPT to CatGPT: The Implications of Artificial Intelligence on Library Cataloguing. Information Technology and Libraries, 42(3). https://doi.org/10.5860/ital.v42i3.16295

Bucciferro, A. (2008). Attacking the Backlog: NARA Archivists Mobilise to Make Unprocessed Records Available to the Public. Prologue Magazine, 40(2), 46-51.

Carter, K. S., Gondek, A., Underwood, W., Randby, T., & Marciano, R. (2022). Using AI and ML to optimise information discovery in under-utilised, Holocaust-related records. AI & SOCIETY, 37(3), 837–858. https://doi.org/10.1007/s00146-021-01368-w

Chow, E. H. C., Kao, T. J., & Li, X. (2024). An Experiment with the Use of ChatGPT for LCSH Subject Assignment on Electronic Theses and Dissertations. Cataloguing & Classification Quarterly, 62(5), 574–588. https://doi.org/10.1080/01639374.2024.2394516

Chun, W. H. K. (2004). On Software, or the Persistence of Visual Knowledge. Grey Room, 18, 26–51.

Ciecko, B. (2020). AI sees what? The good, the bad, and the ugly of machine vision for museum collections. The Museum Review, 5(1). https://static1.squarespace.com/static/578a4d33e4fcb586152bc72d/t/5ea76766c971ba41c7ed4403/1588029296143/TMR_vol5no1_Ceicko.pdf

Crawford, K. (2024). Generative AI's environmental costs are soaring - and mostly secret. Nature, 626, 693. https://doi.org/10.1038/d41586-024-00478-x

DCMI Metadata Terms. (2020). DCMI. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (Archive Link)

Dilmegani, C. (2025, July 9). Data Quality in AI: Challenges, Importance & Best Practices. AIMultiple. https://research.aimultiple.com/data-quality-ai/ (Archive Link)

Ex Libris. (n.d.). The AI metadata assistant in the metadata editor. Ex Libris Knowledge Center. https://knowledge.exlibrisgroup.com/Alma/Product_Documentation/010Alma_Online_Help_(English)/Metadata_Management/005Introduction_to_Metadata_Management/The_AI_Metadata_Assistant_in_the_Metadata_Editor# (Archive Link)

Fenlon, K., Havens, L., Marsh, D. E., Wise, N., Smoke, U., Navarrete, C., Sioui, J., Mantle, D., & Sorensen, A. (2025). Linked data workflows for community collections: Experiments with open access AI. 88th Annual Meeting for the Association of Information Science and Technology Conference Proceedings, 62. https://doi.org/10.1002/pra2.1246

Fisher, S. A. (2024). Large language models and their big bullshit potential. Ethics and Information Technology, 26(4), 67. https://doi.org/10.1007/s10676-024-09802-5

Foka, A., Griffin, G., Ortiz Pablo, D., Rajkowska, P., & Badri, S. (2025). Tracing the bias loop: AI, cultural heritage, and bias-mitigating in practice. AI & Society. https://www.doi.org/10.1007/s00146-025-02349-z

Gavrilis, D., Makri, D.-N., Papachristopoulos, L., Angelis, S., Kravvaritis, K., Papatheodorou, C., & Constantopoulos, P. (2015). Measuring Quality in Metadata Repositories. In S. Kapidakis, C. Mazurek, & M. Werla (Eds.), Research and Advanced Technology for Digital Libraries (pp. 56–67). Springer International Publishing. https://doi.org/10.1007/978-3-319-24592-8_5

Greenberg, J. (2003). Metadata Generation: Processes, People and Tools. Bulletin of the American Society for Information Science and Technology, 29(2), 16–19. https://doi.org/10.1002/bult.269

Greene, M. A., & Meissner, D. (2005). More Product, Less Process: Revamping Traditional Archival Processing. The American Archivist, 68(2), 208-263. (Archive Link)

Havens, L., Bach, B., Terras, M., & Alex, B. (2025). Investigating the Capabilities and Limitations of Machine Learning for Identifying Bias in English Language Data with Information and

Heritage Professionals. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–22. https://doi.org/10.1145/3706598.3713217

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. Ethics and Information Technology, 26(2), 38. https://doi.org/10.1007/s10676-024-09775-5

Hosseini, K., Wilson, D. C., Beelen, K., & McDonough, K. (2022). MapReader: a computer vision pipeline for the semantic exploration of maps at scale. Paper presented at the Proceedings of the 6th ACM SIGSPATIAL International Workshop on Geospatial Humanities.

Huang, J., Provo, A. A., McKeehan, M., & Wittmann, R. (2016). Inclusive Metadata Toolkit. Digital Library Federation. https://osf.io/2nmpc/

Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K., & Prabhakaran, V. (2022). Evaluation Gaps in Machine Learning Practice. 2022 ACM Conference on Fairness Accountability and Transparency, 1859–1876. https://doi.org/10.1145/3531146.3533233

Implementing Australia's AI Ethics Principles in Government. (2024). Australian Government Department of Finance. https://www.finance.gov.au/government/public-data/data-and-digital-ministers-meeting/national-framework-assurance-artificial-intelligence-government/implementing-australias-ai-ethics-principles-government (Archive Link)

Jaillant, L., & Aske, K. (2024). AI and medical images: Addressing ethical challenges to provide responsible access to historical medical illustrations. Digital Humanities Quarterly, 18(3). https://dhq.digitalhumanities.org/vol/18/2/000755/000755.html (Archive Link)

Jaillant, L., Mitchell, O., Ewoh-Opu, E., & Urbaneja, M.H. (2025). How can we improve the diversity of archival collections with AI? Opportunities, risks, and solutions. AI & Society, 40, 4447–4459. https://doi.org/10.1007/s00146-025-02222-z

JSTOR (n.d.) JSTOR Digital Stewardship Services. https://about.jstor.org/get-jstor/digital-stewardship/ (Archive Link)

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why Language Models Hallucinate (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2509.04664

Király, P. (2018, November 28). Metadata quality in cultural heritage institutions. Workshop on FAIR Principles for Digital Research Data Management.

Kneese, T. (2023, August 2). Climate Justice & Labor Rights. http://dx.doi.org/10.2139/ssrn.4533853

Kneese, T. (2024, February 12). Measuring AI's environmental impact requires empirical research and standards. TechPolicy.Press. https://www.techpolicy.press/measuring-ais-environmental-impacts-requires-empirical-research-and-standards/ (Archive Link)

Kneese, T., & Young, M. (2024). Carbon emissions in the tailpipe of generative AI. Harvard Data Science Review. https://doi.org/10.1162/99608f92.fbdf6128

Kugler, L. (2025). How Do You Measure AI? Communications of the ACM, 68(4), 15–17. https://doi.org/10.1145/3708972

Long, D. X., Dinh, D., Nguyen, N.-H., Kawaguchi, K., Chen, N. F., Joty, S., & Kan, M.-Y. (2025). What Makes a Good Natural Language Prompt? Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 5835–5873. https://doi.org/10.18653/v1/2025.acl-long.292

Magnus, B., Priem, M., Vanderperren, N., Berghe, P. V., Keer, E. V., & Vissers, R. (2024). Metadata creation and enrichment using artificial intelligence at meemoo. Journal of Digital Media Management, 13(2), 110–123. https://doi.org/10.69554/NGFF5280

Männistö, A., Seker, M., Iosifidis, A., & Raitoharju, J. (2022). Automatic Image Content Extraction: Operationalising Machine Learning in Humanistic Photographic Studies of Large Visual Archives. arXiv. https://doi.org/10.48550/ARXIV.2204.02149

Margaritopoulos, M., Margaritopoulos, T., Mavridis, I., & Manitsaris, A. (2012). Quantifying and measuring metadata completeness. Journal of the American Society for Information Science and Technology, 63(4), 724–737. https://doi.org/10.1002/asi.21706

Marinescu, M.-C., Reshetnikov, A., & López, J. M. (2020). Improving object detection in paintings based on time contexts. 2020 International Conference on Data Mining Workshops (ICDMW), 926–932. https://doi.org/10.1109/ICDMW51313.2020.00133

Marsh, D. E. (2019). Research-Driven Approaches to Improving Archival Discovery. IASSIST Quarterly 43(2), 1–9. https://doi.org/https://doi.org/10.29173/iq955.

Marvin, G., Hellen, N., Jjingo, D., Nakatumba-Nabende, J. (2024). Prompt Engineering in Large Language Models. In Jacob, I.J., Piramuthu, S., Falkowski-Gilski, P. (eds), Data Intelligence and Cognitive Informatics. ICDICI 2023. Algorithms for Intelligent Systems, pp. 387-402. Springer. https://doi.org/10.1007/978-981-99-7962-2_30

Meaker, M. (2023, September 11). These prisoners are training AI. Wired. https://www.wired.com/story/prisoners-training-ai-finland/(Archive Link)

Media Types. (2025, September 2). Internet Assigned Numbers Authority. https://www.iana.org/assignments/media-types/media-types.xhtml (Archive Link)

Metadata Assessment Framework and Guidance. (n.d.). DLF Metadata Assessment Working Group. https://dlfmetadataassessment.github.io/projects/framework/ (Archive Link)

Metadata Quality. (n.d.). Data Europa. https://data.europa.eu/mqa/methodology?locale=en (Archive Link)

Metadata Schema Assessment Framework. (2024). ALA Core Metadata Standards Committee. https://hdl.handle.net/11213/22781

Mollema, W.J.T. (2024). 'AI colonialism' is a conceptual metaphor. [Masters thesis, Utrecht University]. Utrecht University Student Theses Repository. https://studenttheses.uu.nl/handle/20.500.12932/47214?show=full

Nockels, J., Gooding, P., Ames, S., & Terras, M. (2022). Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. Archival Science, 22(3), 367-392.

One-to-One Principle. (2011, May 1). DCMI. https://www.dublincore.org/resources/glossary/one-to-one_principle/ (Archive Link)

OpenAI. (2022, November 30). Introducing ChatGPT. OpenAI. https://openai.com/index/chatgpt/ (Archive Link)

Osti, G., & Roke E. R. (2024). Collaborating for Change? Assessing Metadata Inclusivity in Digital Collections with Large Language Models (LLMs). 2024 IEEE International Conference on Big Data (BigData), 2479-2488. Washington, DC. https://doi.org/10.1109/BigData62323.2024.10825858.

Panitch, J. M. (2001). Special Collections in ARL Libraries: Results of the 1998 survey sponsored by the ARL Research Collections Committee. Association of Research Libraries.

Pepper, J., Jones, E., Zhao, X., Furst, J., Langlois, K., Uribe-Romo, F., Breen, D., & Greenberg, J. (2024). AI-Ready Data: Knowledge Extraction from Archival Lab Notebooks. 2024 IEEE International Conference on Big Data (BigData), 2489–2495. https://doi.org/10.1109/BigData62323.2024.10825206

Prud'homme, P. A., & Compton, J. (2020). A Research Study of Inventory Practices in Archives in the United States: Scalability and Process. Society of American Archivists Research Forum, 1-12. https://www2.archivists.org/sites/all/files/Inventory%20Practices%20in%20Archives%20FINAL.pdf

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (n.d.). AI and the Everything in the Whole Wide World Benchmark.

Ray, A., Tirrell, J., & Sayers, A. (2025). From Assimilation to Autonomy: Rethinking Data Sovereignty in the Age of Large Language Models. Technical Communication Quarterly, 34(3), 353–372. https://doi.org/10.1080/10572252.2025.2490503

Report and Recommendations from the Task Force on Metadata Quality. (2013). https://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf?__cf_chl_tk=m0ZH3G7aVKKnPtAIdNE5EeLIzumIEVgs4GjY3VBNzJE-1745323769-1.0.1.1-h7uHrd2oK5A1wWZq6XY9ZpsXDGcLfW5bO.tAMDpDHxo#page=3.34

Roke, E. (2025). Metadata Remediation through AI Collaboration. Paper presented at the SAA Research Forum, Online. https://www2.archivists.org/sites/all/files/2.1.4-Roke.pdf

Rotman, D. (2025, May 20). AI could keep us dependent on natural gas for decades to come. MIT Technology Review, Climate Change and Energy Series.

Schema.org. (n.d.). Schema.org. https://schema.org/ (Archive Link)

Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schaeffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. npj Digital Medicine, 7(1), 203. doi:10.1038/s41746-024-01196-4

Society for American Archivists. (n.d.). C.F.W. Coker award: JSTOR Seeklight. https://www2.archivists.org/recipients/2025/cfw-coker-award-jstor-seeklight (Archive Link)

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L.W., Smyth, P. (2025). What large language models know and what people think they know. Nature Machine Intelligence 7, 221–231. https://doi.org/10.1038/s42256-024-00976-7

Stvilia, B., & Gasser, L. (2008). Value-based metadata quality assessment. Library & Information Science Research, 30(1), 67–74. https://doi.org/10.1016/j.lisr.2007.06.006

Sun, Z., Yan, Y., & Zeng, Y. (2025). How to get enriched metadata? A multi-model model fusion strategy for automatic metadata enhancement in GLAM art collections. 88th Annual Meeting for the Association of Information Science and Technology Conference Proceedings, 62.

Sundararasan, T. (2024). Data sovereignty: Indigenous ownership in the age of AI. In Artificial Intelligence in Education Editors (pp. 151-166). Mithra Publication Tamil Nadu. https://doi.org/10.1037/1528-3542.4.3.507

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. LIBER Quarterly: The Journal of the Association of European Research Libraries, 29(1), 1–25. https://doi.org/10.18352/lq.10285

Taniguchi, S. (2024). Creating and Evaluating MARC 21 Bibliographic Records Using ChatGPT. Cataloging & Classification Quarterly, 62(5), 527–546. https://doi.org/10.1080/01639374.2024.2394513

Temple, J. (2025, May 20). The data centre boom in the desert. MIT Technology Review, Climate Change and Energy Series.

Walter, M., & Russo Carroll, S. (2020). Indigenous Data Sovereignty, governance, and the link to Indigenous policy. In Indigenous Data Sovereignty and Policy, pp. 1–20. Routledge. https://library.oapen.org/handle/20.500.12657/42782

Weissner, M. (2024). Ready, set, scan: National Archives to digitise 500M records by 2026. Federal Times. https://www.federaltimes.com/it-networks/ai/2024/04/18/ready-set-scan-national-archives-to-digitise-500m-records-by-2026/ (Archive Link)

Wen, S. (2014, November 11). The Ladies Vanish. The New Inquiry, Essays, and Reviews. https://thenewinquiry.com/the-ladies-vanish/ (Archive Link)

Widder, D.G., & Kneese, T. (2025). Salvage anthropology and low-resource NLP: what computer science should learn from the social sciences. Interactions, 32(2), 46–49. https://doi.org/10.1145/3714996.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024, March 17). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. ICLR 2024. https://doi.org/10.48550/arXiv.2306.13063

Yu, L., Charlton, A., Terras, M., & Filgueira, R. (2024). Advancing frances: New Heritage Textual Ontology, Enhanced Knowledge Graphs, and Refined Search Capabilities. 2024 IEEE 20th International Conference on E-Science (e-Science), 1–10. https://doi.org/10.1109/e-Science62913.2024.10678663

Zavalin, V., & Zavalina, O. L. (2023). Exploration of Accuracy, Completeness and Consistency in Metadata for Physical Objects in Museum Collections. In I. Sserwanga, A. Goulding, H.

Moulaison-Sandy, J. T. Du, A. L. Soares, V. Hessami, & R. D. Frank (Eds.), Information for a Better World: Normality, Virtuality, Physicality, Inclusivity (Vol. 13972, pp. 83–90). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28032-0_7

Zavalina, O. L., & Burke, M. (2021). Assessing Skill Building in Metadata Instruction: Quality Evaluation of Dublin Core Metadata Records Created by Graduate Students. Journal of Education for Library and Information Science, 62(4), 423–442. https://doi.org/10.3138/jelis.62-4-2020-0083

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. Semantic Web, 7(1), 63–93. https://doi.org/10.3233/SW-150175

Ziegler, S. L. (2020). Open data in cultural heritage institutions: Can we be better than data brokers? Digital Humanities Quarterly, 14(2). https://dhq.digitalhumanities.org/vol/14/2/000462/000462.html