Exploring Computational Descriptions for Metadata Creation for E-Books at the Library of Congress, United States of America

Anna-Lea Schumann, Justus Kühler, and Ines Mergel

Abstract The Library of Congress' (LoC) project "Exploring Computational Description" (ECD) is investigating the use of machine learning (ML) to create metadata for e-books that have not yet been catalogued. In-house the LC Labs carried out this initiative with the U.S. Programs, Law, and Literature Division and an external vendor. An initial budget of $250,000 from the National Digital Trust Fund was allocated for this experimental AI endeavour, prompted by a massive backlog of e-books. During the first project phase, five ML models were evaluated, and in the second project phase, human-in-the-loop prototypes that offer machine-generated terms to librarians were introduced. The integration of AI at the LoC has the potential to enhance cataloguing efficiency by automating repetitive tasks, thereby allowing librarians to focus more on intellectual tasks. At the same time, the project faced several challenges, including ensuring the reliability of AI-generated records, copyright concerns, and managing potentially harmful language in older texts used for training the models. Improving the accuracy of these models remains essential and depends on access to extensive digital data. However, human expertise remains crucial for ensuring high quality, and librarians need to develop a foundational understanding of ML to leverage these technologies effectively. The aim of the project is to develop innovative approaches that contribute to improving library practices.

A.-L. Schumann · J. Kühler
Department of Politics and Public Administration, University of Konstanz, Konstanz, Germany
e-mail: anna-lea.schumann@uni-konstanz.de; Justus.kuehler@uni-konstanz.de

I. Mergel (✉)
Department of Politics and Public Administration, University of Konstanz, Konstanz, Germany

Fraunhofer FOKUS, Fraunhofer Institute for Open Communication Systems, Berlin, Germany

School of Management, Public Management, University of Vaasa, Vaasa, Finland
e-mail: ines.mergel@uni-konstanz.de

I. Mergel, C. Schmidt (eds.), *AI Innovations in Public Services*,
https://doi.org/10.1007/978-3-032-01344-6_13

Keywords Library of Congress · Machine Learning · Metadata Creation · Cataloguing Automation · Human-in-the-loop · Digital Transformation · AI Ethics · Co-Production

1 Introduction

The Library of Congress (LoC), established in 1800 in Washington, D.C., is the de facto National Library of the United States. Initially housing a collection of legal texts in the Capitol Building, the LoC faced devastation during the War of 1812 when the British set the building ablaze, destroying its holdings. Following this, Thomas Jefferson sold his extensive and diverse personal collection of books and maps to Congress, laying the foundation for the library's broader scope. The LoC expanded rapidly over the years, aided by milestones such as the 1870 US Copyright Deposit Act, which significantly broadened its acquisitions. To accommodate its growing collection, iconic buildings such as the Thomas Jefferson Building (1897), the John Adams Building (1938), and the James Madison Memorial Building (1981) were constructed [1]. Beyond its physical collections, the LoC encompasses specialised facilities like the National Audio-Visual Conservation Center and initiatives such as the National Library Services for Blind and Disabled, the US Copyright Office, and the Congressional Research Service [1]. The LoC has more than 3000 permanent employees [1].

In recent years, the LoC has emphasised digital transformation and innovation. The Digital Strategy Directorate spearheads efforts in leveraging emerging technologies, collaborating with internal and external partners, and fostering modernisation through its strategic plan, "A Library for All: The FY2024–2028 Strategic Plan for the Library of Congress" [2]. Central to these initiatives is the Digital Innovation Division (LC Labs), launched in 2016, which drives experimentation and research. Since 2018, LC Labs has been instrumental in developing and accessing artificial intelligence (AI) tools for data management, accessibility, and analysis, reflecting the LoC's commitment to embracing technological advancements and addressing their implications [3].

This chapter explores how the LoC utilises machine learning (ML) for cataloguing. We focus especially on the "Exploring Computational Description" (ECD) project and examine how the library managed the project's first two phases, ECD1 and ECD2. First, we outline the necessity for an AI-based cataloguing solution and provide an overview of various AI-driven projects at the LoC. Next, we delve into the "Exploring Computational Description" project. We highlight the different actors involved in the project and the skills and competencies required for librarians. Additionally, we discuss the challenges and the value that has been (co-)created through the project execution. Finally, we conclude by sharing the lessons learned.

1.1 Need Behind the Implementation

Libraries are essential and trustable institutions for citizens and the public: "*The Library of Congress and libraries in general for the most part are still seen as trustworthy.*" Therefore, as one interviewee stated, it is important to "*just get our stuff into people's lives more easily, get ourselves out of the way, and make things more exciting. But then also this alternative model of attention to care for information integrity or truth.*"

To achieve LoC's vision of connecting all Americans to the library, LoC must explore new approaches that could significantly change its practices [2, 4]. The LC Labs are responsible for trying new methods, approaches, and technologies with external partners like universities and research initiatives [4]. In the past several years, the LC labs have also started experimenting with ML and AI. Moreover, one interviewee stated: "*We've already had the idea of automation for a very long time. Machine learning is the next level of that. If you have all these routine tasks or you have this very predictive work and you have all this data to help you identify patterns, why not take advantage of it?*"

For example, the LoC has a backlog of e-books that it receives, which has come to the attention of higher library management. As a result, the Principal Deputy of the Librarian of Congress, Mark Sweeney, has earmarked $250,000 from the National Digital Trust Fund, which is separate from the money given by Congress. This fund is used to contract out experimental AI work to a vendor.

1.2 AI Solutions in the LoC

The LoC is actively exploring the potential of ML and AI through various use cases and experimental projects. Key use cases for AI include creating machine-readable text from digitised documents using optical character recognition (OCR), developing standardised catalogue records from e-books and other digital materials, extracting historical copyright data, parsing legislative data, and experimenting with ML in the National Library Service for the Blind and Print Disabled [3].

The library has conducted several experiments to date, showcasing innovative applications of AI and ML. These include the *Speech-to-Text Viewer*, collaborative explorations with the Project Aida team, and initiatives under *Experimental Access* and *Humans-in-the-Loop* [3]. Other notable projects are the *Newspaper Navigator*, *Citizen DJ*, and *America's Public Bible*: *Machine-Learning Detection of Biblical Quotations Across LoC Collections via Cloud Computing* [3]. Additionally, efforts have focused on enhancing access and discovery of documentary images and using neural networks to broaden the reach of metadata in the project *Situating Ourselves in Cultural Heritage* [3]. These initiatives demonstrate the library's commitment to leveraging AI and ML technologies to expand access to cultural heritage and

improve data usability. However, this illustration focuses on the project called *Exploring Computational Description (ECD)*.

2 Description of the Project

In *Exploring Computational Description (ECD)*, the LC Labs and the U.S. Programs, Law, and Literature Division in Acquisitions and Bibliographic Access Directorate are currently working on a project to determine the feasibility of using AI and ML to create descriptive metadata for e-books that have not yet been catalogued. The project consists of three project phases: ECD1, ECD2, and ECD3 [5]. It was initiated in December 2021 and sponsored by the Principal Deputy of the Librarian of Congress and co-led by the U.S. Programs, Law, and Literature Division Chief and the Sr. Innovation Specialist in LC Labs [6]. It began in the Summer 2022 and involves collaboration with an external vendor, Digirati. The overall research question in the first project phase (ECD1) was: "*What are examples, benefits, risks, costs, and quality benchmarks of automated methods for creating workflows to generate cataloguing metadata for large sets of Library of Congress digital materials? And, what technologies and workflow models are most promising to support metadata creation and assist with cataloguing workflows? What similar activities are being employed by other organizations?*" [6].

Digirati tested five different models for ML and used two workflows for assisted cataloguing, creating a subject assistant suggestion prototype. In September 2023, the project team concluded the first part of the project, called "Toward Piloting Computational Description," and the team has moved on to the second part, ECD2 (2023–2024), which focuses on diverse human-in-the-loop prototypes that suggest machine-generated terms to librarians who then review them. Finally, in the last phase of the project, ECD3, the team uses the BIBFRAME system instead of the MARC system to test the most effective human-in-the-loop models and prototypes identified during the first two phases [5]. Since the interviews were conducted at the beginning of 2024, this case study focuses on the ECD1 and ECD2 projects.

Each project phase follows a framework consisting of three stages: (1) understand, (2) experiment, and (3) implement. The understand stage was described by one interviewee as follows: "*… understand what are you doing, who are the people, what is the data, what are the models, really understand and understand your own appetite for risk, understand how important is quality or how close to perfection does this need to be.*" Moreover, "*There are use cases where it's okay if there's a lot of error, where we were never going to process the stuff anyway and maybe it doesn't matter. And then, there are use cases where we're providing the only authoritative source on this. The trustworthiness of our whole institution relies on us getting this right…*" The expermentation stage "*… is the right place to do an experiment, … and that experiment is all about gathering data.*" One interviewee stated the benefits of experimentation as follows: "*… we absolutely get information about the models that's wonderful, but we also learn a lot through really looking at results what is*

good enough or questions that are much easier answered when you're actually looking at the machine-generated stuff than when you're imaging." Finally, the implementation stage "*… rests on policies, principles, a roadmap, understanding staff*". All three phases are designed to support the evaluation of the following elements of ML: (1) data, (2) models, and (3) people. Data is crucial in ML, serving as input, output, and training material for models. The data includes collections, historic copyright, and legislative data, often in unique digital formats [7]. These real-world datasets can confuse AI models due to their messiness, imbalance, and historical content [7]. Models refer to the technologies and tools supporting ML algorithms' training, processing, and prediction. ML programs learn patterns from data without explicit instructions [7]. The effectiveness of a model depends on training, data processing, and delivery methods. In ML, people are essential to creating and using data and models. They design and program AI tools, and their expertise shapes potential ML use cases [7]. People are responsible for the quality of AI systems and must decide how to implement AI responsibly.

Additionally, a set of worksheets and questionnaires, freely available at GitHub (https://github.com/LibraryOfCongress/labs-ai-framework), have been created to help staff and stakeholders identify priorities for future AI enhancements and services [7].

In the project's initial phase, ECD1, the project team and the vendor undertook tests on five models and methods to detect or generate comprehensive bibliographic records from e-books. They used around 23,000 existing machine-readable cataloguing (MARC) records and e-books to train the models. Digirati was granted access to these 23,000 e-books, which were primarily in English, along with their corresponding MARCXML records sourced from various collections: open-access e-books, legal reports, e-deposit registration e-books, and cataloguing-in-publication e-books [8]. The primary objective was to generate token and text-classified records, a goal that was ultimately achieved [8]. Token classification is generally defined by identifying group tokens in records and assigning them to specific classes or categories [8]. In contrast, text classification involves identifying the sentiment, subject, or topic of the e-book [8]. However, the data provided was insufficient for the machine to accurately predict all the categories: "*… we gave them access to almost 23,000 e-books and the associated MARCXML to use as training data, which was not nearly enough training data.*" Nevertheless, for token classification methods, such as identifying title, the accuracy of the generated information ranged from 80% to 90% [8]. Regarding text classification, particularly for subject headings, the models encountered more challenges in predicting the correct information [8].

In the second project phase, ECD2, the project team concentrated on gaining a deeper understanding of cataloguers' needs and priorities of the LoC, with an emphasis on developing testable human-in-the-loop prototypes for catalogue assistance workflows [9]. This phase created three initial prototypes and one "clickable" prototype. Like ECD1, the team used e-books and MARCXML records from four collections: open-access e-books, legal reports, e-deposit registration e-books, and cataloguing-in-publication e-books [9]. In total, nearly 120,000 e-books were leveraged to train the models. The primary goal was to generate fully valid MARC

records, replete with subfield information, and extract data from entire e-books [9]. Cataloguers reviewed the subject predictions produced by the models, indicating whether the suggestions were accurate or incorrect. The findings revealed that it is feasible to produce MARC records using ML methods, achieving an accuracy of 80% for the majority of fields and subfields [5]. However, accuracy dropped to approximately 50% when assessing subject fields [5].

2.1 Actors Involved

In the ECD, various internal and external actors come together in a process known as co-production. Co-production entails five interdependent phases: co-commissioning, co-design, co-implementation, co-delivery, and co-assessment [10]. Each phase engages various stakeholders, underscoring the collaborative essence of the initiative. In the following, we present the key actors in the ECD1 and ED2.

The ECD is led by the Senior Innovation Specialist in the LC Labs and the Chief of the U.S. Programs, Law, and Literature Division in the Acquisitions and Bibliographic Access Directorate.

The LC Labs were created to plan for the library's digital future and explore ways to utilise digitisation and digital collections to benefit the public. LC Labs act as an experimental space to test various projects and are included in the Digital Strategy Directorate. Since its inception, LC Labs has collaborated on many projects using AI and ML. The LC Labs consists of a small team of five full-time employees who recently welcomed their first programmer. One interviewee stated: "*There's no technical team in LC Labs, and none of the developers, designer, coders of the library work on our stuff. Everything LC Labs has done has been through partnerships in contracts, a lot of contracts. A lot of what the team has done over the years has led a program of exploration, experimentation—I think of it as broadening the horizons at the library or lowering barriers to innovation…*."

Part of the Acquisitions and Bibliographic Access Directorate is the U.S. Programs, Law, and Literature Division. This division, for example, catalogues copyright materials, supports ISSN infrastructure, or catalogues legal materials [11].

The U.S. Programs, Law, and Literature Division and the LC Labs collaborate. This relationship is described as follows: "*…collaboration has been great… it's been fantastic.*" About 12 cataloguers were also involved in testing, along with various other IT staff who helped move data and support the experiment.

In addition, both divisions work with the vendor Digirati in the ECD project. The vendor manages the development of the ML application. This collaboration is complemented by Digital Innovation Indefinite Delivery Indefinite Quantity (IDIQ), which is managed by LC Labs. This contract is valid for 5 years, during which vendors can submit task orders. Digirati was one of the vendors that successfully bid on a task order under this contract vehicle.

External vendors and digital service providers are contracted to contribute technical expertise and resources to the project. Research institutes and academic partners provide domain-specific knowledge and collaborate on various aspects. Users and patrons, such as cataloguers, participate through user testing and feedback sessions to ensure the project meets their needs.

Moreover, the LC Labs' planning framework helps connect staff with patrons and the library community: "*We're doing both, we are connecting with people, learning from them about what they care about and also, we are doing experiments with people, with our materials. We are exploring, we are not waiting until we get it because we don't know what it should be; how would we know? I think the fact that the framework is very vigorous and rigorous and it came out of actually engaging with people on these questions. It also came out of working with staff, it also came out of working with the library community and within the federal community.*"

Table 1 gives a brief overview of the co-production process based on Mergel et al. (2025) in the ECD 1 and ECD2 at the LoC:

2.2 Skills and Competencies Required for Librarians

Librarians need, besides cataloguing skills, necessary skills and competencies to effectively and responsibly utilise AI- and ML-based solutions. Therefore, it always requires a librarian with the necessary skills to comprehend and apply cataloguing principles effectively: "*You need to know how to catalogue and what to expect from*

Table 1 Co-production process in ECD1 and ECD2 at LoC

Co-commissioning	*Prospective co-production phase*	ECD was sponsored by the Principal Deputy of the Librarian of Congress and co-led by the U.S. Programs, Law, and Literature Division Chief and the Sr. Innovation Specialist in LC Labs.
Co-design	*Concurrent co-production phases*	Together with external vendor Digirati, the responsible project them tested different ML models (ECD1) and human-in-the-loop prototypes for cataloguing e-books (ECD2).
Co-implementation		These ML models were fed with data from e-books and checked for accuracy.
Co-delivery		The machine-generated records were reviewed and verified by professional cataloguers for their correctness.
Co-assessment	*Retrospective co-production phase*	Together with external vendors, the responsible project team noted that the initial set of 23,000 e-books (ECD1) was insufficient to create a dataset adequate for machine learning. They subsequently expanded the dataset by adding approximately 100,000 more e-books (ECD2).

what goes in a bibliographic record, for one thing, how to identify valid or invalid information, make decisions on name authorities, for instance, or subjects." This would involve identifying errors and generating new subject headings when required. This is not something that can be expected from a machine. Bias is another factor that needs to be taken into consideration. If cataloguers rely solely on machines to identify and apply subject-specific subjects to a particular resource, they may suggest biased options. Without someone with the necessary understanding and skills to identify and rectify this, the bias may perpetuate: "*There needs to be somebody going: 'Oh, wait. I don't think the machine did a good job here. I think we need to backtrack, start over.*'" In addition, librarians need at least a basic understanding of what ML and AI mean and how they work: "*I wish I had had a better understanding of what even machine learning meant, quite frankly, because it was sort of a duh moment when I said, 'What are you testing?*'" Furthermore, "*So understanding the basic concept was important, and I wish I understood that.*" Conversations with the vendors and workshops help to gain an understanding of the phenomena. In addition, it could be helpful if AI and ML are already part of the library science curriculum: "*So, having some sort of grounding basic and education in what that really is probably would be useful, to have a better understanding of what machine learning really means, how it can be used.*" Also, library staff should develop evaluation and assessment skills: "*So, it is going to always require careful curation, careful evaluation.*"

2.2.1 Organisational Level

The project has the support of the higher management of the LoC, as stated: "*In fact, this project was highlighted by the Librarian of Congress and the Chief Information Officer at a hearing before Congress… because Congress is, of course, interested in what are you doing with artificial intelligence, how do it's safe, how do we know that you can trust it, is it going to do horrible things, what about the ethical concerns…*" Furthermore, the tasks are: "*So, this current task order is going to be having librarians helping the machine make decisions, not expecting the machine to make decisions, but also using large language models to do some more of the work and see how that goes as opposed to say, natural language processing.*"

2.3 Challenges

During the project's first two phases, challenges were also encountered that should not be ignored. One challenge the LoC faced was deciding how transparent it should be with its patrons, mostly other libraries that use its catalogued records. Specifically, there is a question of whether patrons should be made aware if a record was created using ML and whether the record should include information about the confidence

level of the ML-produced record: "*It's something we have to think about, what do we want to say, do we need to say something in our record that says 'this record was created with machine learning', do we need to say something about its confidence level?*"

A second challenge that needs to be addressed is the issue of copyright for the submitted e-books. One of the interviewees stated, "*If you're using rights-restricted material, where there's a copyright on content, can you use machine learning to feed the content into this black box of a machine and then what will it do with the content once the modelling is done? Will it delete it? Will it store it? Will it use it for some other purpose that you didn't realize?*"

Thirdly, the LoC is currently facing the challenge of determining the accuracy of ML-generated output. They are still figuring out what level of accuracy is considered acceptable: "*I think something we're still grappling with is accuracy and still trying to figure out what's good enough, what's close, what's relevant, is relevant the same as good enough or does it have to be a 100% one-to-one match.*" Moreover, "*… it's not close enough to acceptable levels to allow it to go on without human interventions and the human in the loop has been this mantra…*" To ensure the accuracy of the output, a cataloguer is required to identify possible mistakes and correctly input the information into the record.

Fourthly, identifying, for example, the title and author of an e-book through AI is a complex task due to the varying design of title pages. Each e-book has a unique typography, layout, and font size, making it difficult for ML to differentiate between, for example, the title and author. Moreover, predicting the content of an e-book requires the machine to scan more than just the initial 50 pages to identify the most frequently used terms in the book. However, identifying words and phrases that are not part of the controlled vocabulary remains challenging, and it is uncertain whether the machine can identify them successfully. In addition, the predicted words and terms generated by the machine are not always helpful nor provide any information about the e-book's content: "*Some of them are not useful, like 'it', 'one', 'she', 'this', not useful.*" As they are only machine-generated output, they cannot be corrected.

Fifthly, there are challenges when it comes to handling the content of books. One question that arises is how to deal with harmful and disrespectful language in these books, as well as how to handle cultural heritage. This can be particularly problematic with old books that contain such language and when these are used to train ML models. Therefore, it is important to consider how best to handle this issue because: "*If you give the machine, if you feed the machine harmful language, it will predict harmful language. Garbage in, garbage out. That is an ethical concern for sure.*"

In addition, training AI requires large volumes of data, including born-digital materials, which can be challenging: "*I think one of the most complicated things it is getting the training data right. So, definitely more is better.*"

Finally, there is a risk of ineffective solutions connected to the fear of investing resources into solutions that may not work or could worsen existing issues: "*The biggest risk is that we're going to pay for something that doesn't work, that we're*

going to implement something that makes things worse. … and lose our credibility with the public and undermine people's faith in reality and truth."

3 Results

3.1 Organisational Level

The LC Labs promote the importance of experimenting with AI and ML at an organisational level. They are doing public affairs and presentations within the LoC: "*… we've done, as you can imagine, a lot of presentations internally.*" In addition, there are Congressional Hearings where the LoC must report their progress to Congress. For example, they provide updates on their experiments in the ECD project. One interviewee stated: "*So when we have these two Congressional hearings, a lot of what they wanted to hear about at the hearings were actually how are you using AI in copyright or Congressional Research Services. Members of Congress wanted to know are you doing to get our work done faster with AI.*" Furthermore, an AI working within the LoC provides input on policy and planning aspects of AI implementation.

However, LC Labs lacks the resources to educate more staff about the ML and AI technologies in the LoC: "*So, the idea of how do we tell more people that we can connect with them, we can't. So, we do the best we can. When we're invited to an internal meeting, we always go; we present.*"

In addition, an interviewee stated that AI is not changing the mission and the library's role: "*It's not like the machine is ever going to take over any concerns people have. Like, 'What are we going to need cataloguers for?' Using AI in a way is even more incumbent on expertise of cataloguers to be able to identify what's right, what's good enough.*"

3.2 Value Created and Co-created

Implementing and using ML and AI can create value for library staff and patrons within the LoC. The use of AI has the potential to enhance cataloguing output in numerous ways. For instance, AI can provide extractive or abstractive summaries, which can assist cataloguers in understanding the essence of a book without having to go through the entire book. Additionally, AI can take care of routine tasks, such as identifying information on the title page and freeing up librarians to focus on intellectual tasks. This can lead to more efficient use of time and resources. However, many library staff members are concerned that AI is replacing their jobs, but this is not necessarily the case. In the short term, AI creates more work because someone has to ensure it is functioning correctly, train and revise it, and continually monitor

it. This leaves library staff more time to focus on intellectual tasks such as reviewing the model or correcting records: "*But I don't think that it's ever going to replace cataloguing or cataloguers or librarians. It'll augment, and for a while, it might even make it more work because it's more work to do some of these projects than it is to just catalogue the thing. But once you can make some headway with it, then maybe that project can go more smoothly and then you can focus attention on other things to study.*"

Successfully using AI-based technologies in the LoC also helps to create added value for other libraries in the country. After completing the records, the LoC sends them to the Online Computer Library Center (OCLC). Once there, the records are accessible not only in the LoC catalogue but also to everyone else. This means that other libraries can download the metadata and use it in their local library catalogue: "*Once they're complete, they go to OCLC, which means the world can access them. But they're also available in our catalogue that anyone could go to our catalogue and you put it into your record and your local catalogue.*" However, staff capacity and time thwart the project implementation: "*I'm not sure if we'll ever get to the point where we can use that metadata in any meaningful way or it's going to sit there, sort of this training data we had from our first task order and never actually use it. … So, the cataloguing part won't even go live until probably spring of 2025.*"

Moreover, adopting AI in libraries is seen as a means of creating public value for society. However, one interviewee stated: "*The part about connecting with the public, for me, the degree to which our materials are—the materials that we steward are not meaningfully available to people in the ways that they would care to find them or have them is heart-breaking. We just don't have the resources to help make our materials part of people's lives because the paradigm that we're operating—every library is widely under resourced. We all are barely holding on to the processing that we can do with decreasing resources and increasing materials coming in and increasingly complex digital materials*…."

3.3 Lesson Learned

After the project's first two phases, some lessons were also learned. Especially after ECD1, the project team recognised that ML applications need more data to improve accuracy: "*… definitely would have given them as much of our content as possible*". In ECD2, they have increased the number of records to more than 120,000 e-books, hoping this will result in a more precise model. One interviewee stated: "*We only gave them 23,000 e-books. We could have given them 100,000. Now we're giving them 100,000, and hopefully, they'll have more success with the models they're using in this task order.*" This massive amount of digital content is necessary to describe the content of the e-books more accurately.

A second lesson learned is that there is a better understanding of what ML means necessary: "*I didn't understand and I don't think any of us really understood what*

a cataloguer-assisted workflow, what that meant, what does cataloguer-assisted workflow mean. It couldn't even—the concept didn't even make sense to me." However, upskilling helped to gain basic knowledge about the technology and process: "*And then we had a conversation, workshop conversation with different cataloguers, with people from the company, to try to talk about what would make your life easier or how do you find cataloguing e-books, what's difficult about that, what would make it easier.*"

4 Conclusion

In this chapter, we explored how the LoC utilised ML to catalogue e-books that have not yet been catalogued. We focused on the first two phases of the "Exploring Computational Description" (ECD) project. The project is run by the LC Labs and the U.S. Programs, Law, and Literature Division in cooperation with the external vendor Digirati. Notably, due to a backlog of e-books, the Principal Deputy of the Librarian of Congress has allocated $250,000 from the National Digital Trust Fund to contract experimental AI work.

In the first project phase, ECD1, five different models for ML were tested, and the vendor used two workflows for assisted cataloguing, creating a subject assistant suggestion prototype. The second project phase, ECD2, focused on diverse human-in-the-loop prototypes that suggest machine-generated terms to librarians. However, there were some challenges during the first two project phases, such as the trustable confidence level of AI-generated records, dealing with copyright content, or handling harmful or disrespectful language in older books, especially those used for training ML models. Moreover, the models' accuracy still needs improvement, and therefore, the ML models need large amounts of digital data.

Implementing and using AI in the LoC has the potential to enhance cataloguing output in numerous ways. Initial results show that AI can take over repetitive tasks, such as recording title information, which gives library staff more time for more intellectually demanding activities. Nevertheless, human expertise remains crucial for recognising errors and ensuring quality. However, introducing AI in libraries has also created new requirements: Library staff need to acquire basic understanding of ML to use and develop the technologies effectively.

Acknowledgements The work for this chapter was supported by the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement ID 101061516.

Appendix

Overview of the ECD project at the Library of Congress

Case and project name			
Exploring Computational Description (ECD) at the Library of Congress (LoC)			
Country	**Number of employees**	**Type of AI solution**	**Year and maturity level**
USA	More than 3000 permanent employees	Using ML to create metadata for e-books that have not yet been catalogued	ECD 1: 2022–2023 ECD 2: 2023–2024 ECD 3: start 2024
Project description			
The LC Labs and the U.S. Programs, Law, and Literature Division are currently working on a project called Exploring Computational Description to determine the feasibility of using ML to create metadata for e-books that have not yet been catalogued. The project began in September 2022 (ECD1) and involves collaboration with an external vendor, Digirati. Digirati tested five different models for ML and used two workflows for assisted cataloguing, creating a subject assistant suggestion prototype. In September 2023, the first part of the project was concluded, and the team moved on to the second part (ECD2), which focused on diverse human-in-the-loop prototypes that suggest machine-generated terms to librarians			
Need(s) behind implementation	**Actors involved**	**Challenges**	
In order to fulfil LoC's goal of connecting all Americans to the library, LoC needs to investigate new approaches that could bring significant changes to its practices. LC Labs are tasked with experimenting with new methods, approaches, and technologies, and collaborating with external stakeholders such as universities and research initiatives. Additionally, there is support from the library management and funding from the National Digital Trust Fund to experiment with AI and ML	The LC Labs serve as an experimental space for testing different projects and are part of the Digital Strategy Directorate. The U.S. Programs, Law, and Literature Division primarily catalogues books and e-books and is part of the Acquisitions and Bibliographic Access Directorate. These two departments collaborate and work with external vendors such as Digirati or with research institutions and academic partners. Furthermore, the Planning Framework, developed by the LC Labs, helps to connect library staff with patrons and the library community	1. Should patrons, mostly other libraries, be informed if a record was generated using AI and whether the record should contain details about the confidence level of the AI-generated record? 2. Can we use machine learning and AI to process content that is copyrighted? What will the machine do with the content after the modelling is complete? 3. What level of accuracy is considered acceptable for an AI-generated output? To ensure accuracy, a cataloguer is still required to identify possible mistakes and input the information correctly into the record. This is closely related to creating quality standards and policies specifically for AI 4. AI struggles to identify the title and author of an e-book due to the varying design elements and typography. Additionally, predicting e-book content and identifying non-controlled vocabulary words remain challenging for AI. Additionally, the predicted words and terms may not always be useful or informative and cannot be corrected 5. How should we deal with harmful and disrespectful language in books and cultural heritage? This can be particularly problematic with old books that contain such language and when these are used to train ML models	

(continued)

Case and project name		
Exploring Computational Description (ECD) at the Library of Congress (LoC)		
Results		
Organisational level	**Value created and co-created**	**Lesson learned**
AI can provide extractive or abstractive summaries, which can assist cataloguers in understanding the essence of a book without having to go through the entire book. Additionally, AI can take care of routine tasks, such as identifying information on the title page and freeing up librarians to focus on intellectual tasks. This can lead to more efficient use of time and resources. Many library staff members are concerned that AI is replacing their jobs. However, this is not necessarily the case. In the short term, AI creates more work because someone has to ensure it is functioning correctly, train and revise it, and continually monitor it. This leaves library staff more time to focus on intellectual tasks such as reviewing the model or correcting records	LC Labs conducts public affairs has to report to Congress about its progress in experimenting with and implementing AI. However, LC Labs lacks the resources to educate more staff about ML and AI technologies in the LoC. The expertise of the cataloguers is even more valuable in identifying wrongly predicted content and errors. It always requires librarians with the necessary skills to comprehend and apply cataloguing principles effectively. Nevertheless, library staff need at least a basic understanding of what ML and AI mean and how they work	Create a guideline on how to deal with harmful and disrespectful language use Start with a large-enough dataset so that the AI tool can learn and provide more accurate results over time

References

1. *General Information* (n.d), https://www.loc.gov/about/general-information/. Accessed 9 Feb 2025
2. *A Library for All* (n.d.), https://www.loc.gov/strategic-plan/. Accessed 9 Feb 2025
3. *AI at LC* (n.d.), https://labs.loc.gov/work/experiments/machine-learning/. Accessed 9 Feb 2025
4. Library of Congress Blog, *Why Experiment: Machine Learning at the Library of Congress* (Library of Congress Blogs, 2023)
5. C. Saccucci, A. Potter, *Exploring Computational Description: Experiment Results*. (2024)
6. *Experiment: Exploring Computational Description* (n.d.), https://labs.loc.gov/static/labs/work/experiments/Documents/ECD1_Executive_Summary.pdf. Accessed 10 Feb 2025
7. Library of Congress Blog, *Introducing the LC Labs Artificial Intelligence Planning Framework* (Library of Congress Blog, 2023)
8. Library of Congress Labs, *Exploring Computational Description: Final Presentation* (Library of Congress Blog, 2023)
9. Library of Congress Labs, *Toward Piloting Computational Description: Final Presentation* (Library of Congress Blog, n.d.)
10. I. Mergel, N. Edelmann, N. Haug, Co-production phases in the development and implementation of digital public services. Perspect. Public Manag. Govern. **8**(1), 1–13 (2025)
11. *About the Organization* (n.d.), https://www.loc.gov/aba/about/. Accessed 9 Feb 2025

Anna-Lea Schumann is currently pursuing a master's degree in Politics and Public Administration at the University of Konstanz. Her research interests focus on applying modern working practices within public administration and their implications for public value creation. In her bachelor's thesis, she explored the use and implementation of design thinking in libraries. She has gained initial practical experience through various internships in public administration, providing her with valuable insights into the operational methods and processes of the public sector.

Justus Kühler is studying for a bachelor's degree in political science and public administration at the University of Konstanz. He is particularly interested in the latest developments in public administration that reflect the trend towards democratisation. He is particularly interested in the application of new forms of participation at the local level. He has already gained initial experience with municipal administration through his involvement in local politics.

Ines Mergel is a University Professor of Public Administration in the Department of Politics and Public Administration at the University of Konstanz and a Fellow of the National Academy of Public Administration (Class of 2018). After holding positions at the Harvard Kennedy School of Government as a Doctoral Fellow and Postdoctoral Fellow (2002–2008), Professor Mergel was awarded tenure at the Maxwell School of Citizenship and Public Affairs (Syracuse University, NY) in 2014 (2008–2014) and then served as Associate Professor with tenure (2014–2016). Professor Mergel is a member of the supervisory board of the e-Governance Academy (Estonia). Other board memberships can be found here. Professor Mergel is a founding member of the international initiative Teaching Public Service in the Digital Age, which aims to integrate digital skills into the teaching and training of public managers. As part of this initiative, Professor Mergel was appointed a Schmidt Futures Innovation Fellow in 2022.