LIBRARY OF CONGRESS
LC LABS

Through experimentation, research, collaboration, and reflection, LC Labs works to realize the Library's vision that "all Americans are connected to the Library of Congress" by enabling the Library's Digital Strategy.

Since 2017, LC Labs has been exploring how the Library of Congress could use machine learning and artificial intelligence technologies to make digital collections more accessible and easier to use. The promise of these technologies is immense. The ability to recognize patterns and irregularities in data or make predictions across vast and heterogeneous collections will allow the Library to surface new types of collection metadata that could transform how people use and understand the historic record. However, the excitement for the potential of these technologies must be balanced against the effects of human bias, error and potential harms to the organization, users, staff and those depicted in Library collections.

Humans in the Loop is the latest LC Labs experiment to test approaches for scaling metadata creation. The experiment combined essential subject matter expertise and human quality assessment with machine learning methods to extract accurate information from a set of digitized historic telephone directories. The experiment gauged user attitudes about this work and yielded recommendations for how to blend machine and human expertise in ways that are ethical, useful, and engaging while mitigating potential harms and enhancing collection discovery.

The LC Labs team has built expertise in machine learning by hosting the Machine Learning & Libraries Summit, sharing the speech-to-text viewer experiment; collaborating on the experiment and white paper report, Digital Libraries, Intelligent Data Analytics, and Augmented Description; sponsoring Machine Learning + Libraries: A Report on the State of the Field; and investigating how to enable successful computational use of collections in the cloud in the Andrew W. Mellon Foundation-supported Computing Cultural Heritage in the Cloud initiative. LC Labs has been the home for Innovator in Residence experiments Citizen DJ and Newspaper Navigator, which were developed using machine learning technologies. The LC Labs team has championed the practice of crowdsourcing at the Library by hosting Beyond Words; designing and incubating By the People; and co-leading the Collective Wisdom Project to author the Collective Wisdom Handbook and research agenda-setting white paper (forthcoming January 2022).

LC Labs shares the outcomes of its experiments for free and for the benefit of the broader community. Learn more about LC Labs and our work at https://labs.loc.gov and stay up-to-date on our progress and opportunities by subscribing to the monthly LC Labs Letter.

Release date: 30 November 2021

Humans-in-the-Loop

RECOMMENDATIONS REPORT

Final: November 29, 2021

PREPARED BY

Shawn Averkamp, AVP, shawn@weareavp.com
Kerri Willette, AVP, kerri@weareavp.com
Amy Rudersdorf, AVP, amy@weareavp.com
Meghan Ferriter, Library of Congress, mefe@loc.gov

253 36th Street, Suite C302 | Brooklyn, NY 11232
634 W. Main Street. Suite 202 | Madison, WI 53703
+1.917.475.9630 | weareavp.com

CONTENTS

INTRODUCTION

THE HUMANS-IN-THE-LOOP INITIATIVE

A "human-in-the-loop"[1] process in machine learning (ML) is one in which humans and algorithms work together to solve problems more efficiently and accurately than each could on their own. Algorithms attempt to solve problems at a much greater scale than humanly possible while humans offer feedback to algorithms in the form of model examples of successful answers or corrections of mistakes, iteratively improving the accuracy of the algorithms' predictions.

Human-in-the-loop processes offer a great opportunity for cultural heritage organizations to expand discovery and use of the content of digitized collections through mechanisms such as text or audio transcription, extraction of structured data, image classification, geocoding of locations, or georectification of maps. By enlisting the help of the public through crowdsourcing endeavors, cultural heritage organizations can manually generate highly accurate structured data in volumes necessary to train ML algorithms to generate similar data on a larger scale. Establishing a feedback loop where humans continually train and validate the work of algorithms can help to refine and improve the output over time, resulting in trustworthy structured data.

The Humans-in-the-Loop (HITL) initiative builds on the foundation of the Library of Congress' (LC) success with crowdsourcing and machine learning initiatives, and seeks to address the challenges inherent in each approach. In addition to the challenges mentioned above, previous LC experiments have uncovered examples where crowdsourcing efforts may not be enough to efficiently approach data enrichment tasks at scale, while ML models usually need significant amounts of training data to achieve acceptable measures of accuracy.

BACKGROUND

In recent years, the Digital Strategy Directorate and its Digital Innovation Lab (LC Labs) have undertaken a range of data transformation and crowdsourcing experiments that aim to better support emerging research methods and further maximize the use of digital collections at the Library of Congress.

Current LC Labs experiments are in dialogue with the Library of Congress Digital Scholarship Working Group report published in March 2020[2], which recommends making a greater portion of the Library's digital and digitized collections "online, ready for computation, and ready for users."[3] Library initiatives such as Flickr Commons[4], Beyond Words[5], and the Library's By the

[1] While the HITL initiative is called "Humans-in-the-Loop," the term in its singular form ("human-in-the-loop") will be used throughout this document as this is the term of art used by the AI community.

[2] "Digital Scholarship Working Group Report: Published!," Library of Congress, 2020, accessed July 1, 2021, https://blogs.loc.gov/thesignal/2020/04/digital-scholarship-working-group-report-published/.

[3] "Digital Scholarship at the Library of Congress," Library of Congress, 2020, p.14, accessed July 1, 2021, https://labs.loc.gov/static/labs/work/reports/DHWorkingGroupPaper-v1.0.pdf.

[4] "The Commons," Flickr, accessed July 1, 2021, https://www.flickr.com/commons.

[5] "Beyond Words," Library of Congress, accessed July 1, 2021, https://labs.loc.gov/work/experiments/beyond-words/.

People[6] program have pushed to “throw open the treasure chest,”[7] engaging many users in the creation of new knowledge. Ongoing Library efforts have also informed and deeply impacted broader cultural heritage approaches to "collections as data"[8] and machine learning. LC’s work and leadership in these areas have clearly demonstrated that a wealth of more user-friendly ML tools in the hands of a ready and engaged public can be utilized to rapidly increase the breadth of machine-actionable, discoverable, and accessible collections data for many kinds of re-use.

In recent years, LC Labs have sustained exploration of machine learning in cultural heritage for tasks such as pre-processing, segmentation, classification, clustering, transcription, and extraction. In 2019, the team partnered with the Project AIDA researchers[9] on a series of demonstration experiments applying machine learning to Library of Congress collections in different ways. The experiment results and Library-specific recommendations are contained in their *Digital Libraries, Intelligent Data Analytics, and Augmented Description*[10] report and GitHub code repository[11]. In September 2019, LC Labs hosted the Machine Learning + Libraries Summit[12], convening over 75 cultural heritage practitioners and machine learning experts. The event coincided with the announcement of Ben Lee as one of the 2020 Innovators in Residence[13]. His Newspaper Navigator experiment was released in 2020 and used a ML algorithm trained by crowdsourced data to identify, segment, and search all of the visual content in the *Chronicling America* database[14] of historic newspapers. 2020 Innovator in Residence Brian Foo used machine learning to extract, classify, and package[15] samples of music from Library of Congress collections in order to enable creative use and help people make hip hop. Finally, LC Labs commissioned Professor Ryan Cordell[16] to conduct a comprehensive survey of the state of the field regarding machine learning and libraries. In his final report[17], Cordell built

[6] “Be a Virtual Volunteer,” Library of Congress, accessed July 1, 2021, https://crowd.loc.gov/.

[7] “Digital Strategy for the Library of Congress,” Library of Congress, accessed July 1, 2021, https://loc.gov/digital-strategy.

[8] “Always Already Computational,” Collections as Data, accessed July 1, 2021, https://collectionsasdata.github.io/.

[9] “Summer of Machine Learning Collaboration with the University of Nebraska-Lincoln,” Library of Congress, 2019, accessed July 1, 2021, https://blogs.loc.gov/thesignal/2019/09/summer-of-machine-learning-collaboration-with-the-university-of-nebraska-lincoln/.

[10] Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack, “Digital Libraries, Intelligent Data Analytics, and Augmented Description,” January 10, 2020, accessed July 1, 2021, Redirecting you to https://labs.loc.gov/static/labs/work/experiments/final-report-revised_june-2020.pdf.

[11] “Exploring-ML-with-Project-Aida,” Github, accessed July 1, 2021, https://github.com/LibraryOfCongress/Exploring-ML-with-Project-Aida.

[12]LC Labs, Digital Strategy Directorate, “Machine Learning + Libraries Summit,” 2020, accessed July 1, 2021, https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf.

[13] “Introducing Ben and Brian, the Library’s new Innovators in Residence!,” Library of Congress, 2019, accessed July 1, 2021, https://blogs.loc.gov/thesignal/2019/11/introducing-ben-and-brian-the-librarys-new-innovators-in-residence/.

[14] “Chronicling America: Historic American Newspapers,” Library of Congress, accessed July 1, 2021, https://chroniclingamerica.loc.gov/.

[15] “citizen-dj,” Github, accessed July 1, 2021, https://github.com/LibraryOfCongress/citizen-dj.

[16] “Machine Learning + Libraries: A Report on the State of the Field,” Library of Congress, accessed July 1, 2021, https://blogs.loc.gov/thesignal/2020/07/machine-learning-libraries-a-report-on-the-state-of-the-field/.

[17] Ryan Cordell, “Machine Learning + Libraries: A Report on the State of the Field,” 2020, accessed July 1, 2021, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf.

on some of the Aida team's recommendations and laid out steps for cultivating responsible ML in libraries.

The Library's work in this sphere has played a key role in growing awareness in galleries, libraries, archives, and museum (GLAM) communities around the challenges involved in realizing the potential of crowdsourcing and ML approaches. For example, further examination is needed on the types of bias that may occur in applying ML methods to metadata enrichment activities. This careful consideration extends to developing ethical approaches for selecting appropriate algorithms and human review processes to mitigate those risks. ML and crowdsourcing outputs must meet quality standards (and those standards must be defined, understanding that they may vary for different uses of the data). Additionally, the provenance and possible inaccuracies of machine-generated or crowd-generated data must be communicated to users who are accustomed to trusting the Library's catalog, resources, and services as authoritative sources.

INITIATIVE GOALS

HITL aims to model, test, and evaluate various relationships and interactions between crowdsourcing and ML methods in ways that will expand the Library's existing efforts to ethically enhance usability, discovery, and user engagement around digital collections.

The HITL initiative encompasses ML and crowdsourcing prototypes, proof-of-concept experiments, reports, and accompanying recommendations that deepen LC Labs' exploration of the opportunities and challenges that come from operationalizing emerging technologies at scale.

Based on the experiment RFP and conversations with LC Labs in September 2020, AVP set an objective to help LC Labs develop a framework for designing "human-in-the-loop" data enrichment activities that are **engaging, ethical,** and **useful.**

- **Engaging** — users are inspired to participate in the initiative and learn something from it.
- **Ethical** — users understand the purpose of the initiative and provenance of the data outputs; user privacy and social impact are considered in its design; peers/colleagues and downstream users of resulting data and code also understand the initiative design and boundaries of the data.
- **Useful** — data outputs are useful to future collection description and research tasks; users feel their time spent was worthwhile.

This experiment aimed to further enact three core goals of the Library's Digital Strategy[18]:

[18] "Digital Strategy for the Library of Congress," Library of Congress, accessed July 1, 2021, https://loc.gov/digital-strategy.

Throw open the treasure chest

The core team for the initiative looked to uncover and document approaches to improve the utility and accessibility of content that may be difficult to use as data, even though it has been digitized.

Connect

The core team worked directly with crowdsourcing and collection users, as well as numerous library staff, to learn more about the ethical considerations of applying ML processes to specific content and collections. We are sharing findings, code, and processes in an effort to inform and improve practice in the broader cultural heritage community.

Invest in our future

The core team evaluated the feasibility of technical, design, communication, and data generation approaches. Our work highlights challenges and opportunities that emerge along the way.

HITL initiative deliverables included the design of an experimental prototype that serves as proof of concept for two human-in-the-loop workflows — one in which humans create training data for ML processes, and another where humans correct the output of ML processes. As part of the design process, the prototype was presented and tested directly with users to help evaluate how crowdsourcing volunteers might feel about participating in tasks that interact with ML processes.

The HITL initiative also included the design of an experimental interface for presenting the resulting data outputs of the prototype human-in-the-loop pipeline to:

- identify possibilities for user discovery of and interaction with large volumes of data,
- explore integrations with source digital collections, and
- investigate the challenges in presenting data from ML and crowdsourcing sources alongside librarian-created metadata.

Wireframes for the presentation interface were also tested with users to gather researcher perspectives on how the data was generated and produced, as well as how successfully the mockup conveys data provenance.

Through this work, the core team built upon existing library-based crowdsourcing approaches while enhancing and further implementing ML processes through the Library's "existing commitments to responsibility and care."[19] As LC explored the intersections of machine learning and crowdsourcing data enhancement processes through human-in-the-loop workflows, the team sought to bring to the forefront ethical approaches such as proactive communication and

[19] Ryan Cordell, “Machine Learning + Libraries: A Report on the State of the Field,” 2020, p.7, accessed July 1, 2021, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf.

consent-seeking with crowdsourcing contributors, while harnessing the promise of machine learning to aid discovery across the Library's collections.

APPROACH

In support of these initiative goals and deliverables, AVP applied a user-centered design methodology approach. Intended to serve as an integral element of the design work itself, this research methodology puts people front and center so prototypes and recommendations for data management and delivery solutions are informed by user research. To achieve this outcome, the core team used a structured approach to human-in-the-loop processes that maps to traditional design-thinking stages.

In keeping with the design methodology approach, the core team defined and iterated upon the research goals and methods outlined below as each phase of the initiative unfolded.

Research Questions

(1) How can crowdsourcing and machine learning be used together in engaging, ethical, and useful ways by LC Labs to improve access to LC collections and content through data enrichment activities and programs?
(2) Can a high-level, repeatable framework be designed that enables LC Labs to reuse ML and crowdsourcing methodologies in ways that engage users in useful and ethical ways?
(3) What design patterns might best and most responsibly be employed to present human- and/or machine-generated metadata into digital collections discovery interfaces?

Hypotheses

- Machine learning and crowdsourcing can be used together in engaging, ethical, and useful ways in data enrichment activities.
- There are design approaches that can responsibly and clearly present human- and machine-generated metadata in digital collections discovery interfaces.
- It is possible to create a repeatable approach/methodology for selecting collections, data pipelines, and tasks for ML and crowdsourcing workflows.
- It is possible to create a repeatable approach to identifying and engaging users.

Methods

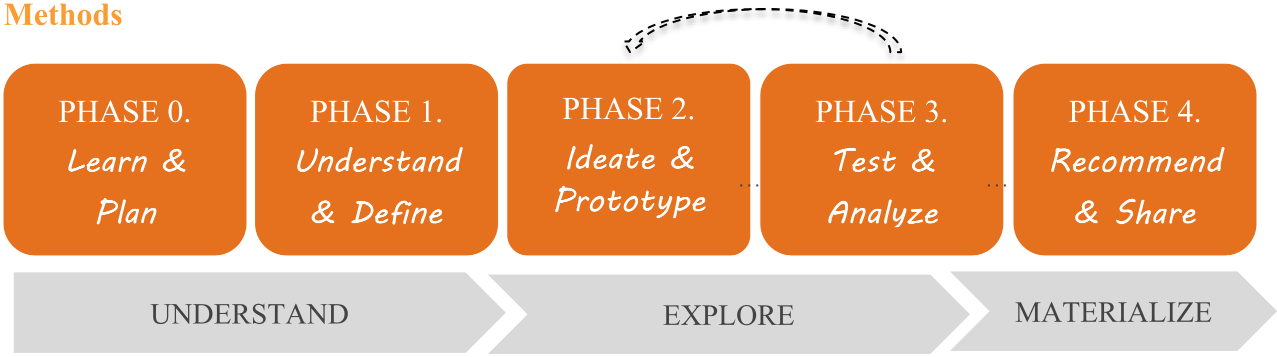


Image 1. Flowchart of the design methodology approach applied across the 5 HITL initiative stages.

Phase 0. Learn and Plan

Objective: Reach a shared understanding of HITL initiative needs and goals, and create a realistic plan based on budget and expectations.

Methods:

- Research review. Review relevant past and present initiatives; absorb LC Labs mission, work, and outputs to date; identify target users, use cases, goals, and requirements for LC collections data enrichment and presentation interface; and confirm and/or refine initiative goals and outcomes based on discovery results.

- **Phase 1. Understand and Define**

  Objective: Select collection(s) for enhancement and draft initial workflow development and testing plans.

  Methods:

  - Collection selection workshop. With the LC core team and collection managers, review experiment ideas and possibilities for selected collections, review the benefits and value as well as risks and ethical implications of candidate collections, and draft initial user stories for finalist collections.
  - Post-workshop survey. Evaluate participant feedback to the collection selection approaches and topics addressed through the workshop, and further refine criteria and processes for collection selection in the future.
  - Interviews with core team and stakeholders. Seek to fill in gaps about the feasibility of finalist candidate collections, uncover lessons learned from other ML initiatives at LC, review current user research techniques and methodologies employed by LC, identify potential LC user test participant groups and recruitment approaches that AVP may wish to access during this collaboration.
  - Technical review. Assess the technical feasibility of working with finalist collections.

- **Phase 2a. Ideate and Prototype: Data generation workflows**

  Objective: Brainstorm and build crowdsourcing or ML workflows that have the potential to meet users' needs based on insights from earlier phases.

  Methods:

  - Prototype design and development. Design data model for structured data to generate from collection content, identify and implement relevant ML processes for generating data from content, ideate on potential user journeys through data enhancement to develop user flows for crowdsourcing tasks, select and customize appropriate open-source code bases on which to develop ML processes and crowdsourcing workflows, and design and develop a database for

  - tracking workflow processes and data outputs for both ML and crowdsourcing pipelines.
  - Documentation report. Outline process, accuracy measures, and user-testing plan for crowdsourcing prototypes and machine learning pipeline.

- **Phase 3a. Test and Analyze: Data generation workflows**

  Objective: Observe users' experiences of experimental workflows and assess accuracy of results.

  Methods:

  - User testing interviews. Gather participant perspectives on the use of machine learning in libraries, gather participant perspectives on crowdsourcing activities that contribute to ML processes, and evaluate user understanding of the intended tasks and pipelines specific to the HITL initiative.
  - Usability survey. Evaluate LC core team users' experiences of experimental workflows, gather direct user feedback on prototype usability, evaluate levels of effort required to perform crowdsourcing tasks, and identify pain points in the workflows.

- **Phase 2b. Ideate and Prototype: Presentation interface**

  Objective: Brainstorm, build, and test ideas for integrating data into an experimental presentation interface that has the potential to meet users needs based on insights from earlier phases.

  Methods:

  - Interface design workshop. With the LC team, develop user personas representing potential interface users and their research needs, brainstorm and prioritize necessary interface functionalities to meet user persona needs, imagine possible interface features incorporating prioritized functionalities.
  - Wireframe development. Design an experience that engages front-end users, clearly communicate data provenance, and help mitigate previously identified risks.

- **Phase 3b. Test and Analyze: Presentation interface**

  Objective: Gather participant perspectives and evaluate proposed interface design of the presentation interface.

  Methods:

  - User testing interviews. Evaluate effectiveness of the mockup in conveying metadata provenance and collection coverage to researchers, gather participant perspectives on data provenance and attitudes toward how the presented data was generated or produced, and assess attempts made through the interface to mitigate previously identified risks to users.

- **Phase 4. Recommend and Share**

  Objective: Demonstrate methods and outcomes of integrating human- and machine-metadata-generation processes and recommend future methods for human-in-the-loop approaches.

  Methods:

  - Presentation. Share findings with LC core team and other LC stakeholders on HITL initiative outcomes and recommendations.
  - Final analysis and recommendations report. Deliver final analysis report on outcomes and recommendations that include results from user testing and feedback.

In working through the above phases, the core team identified numerous challenges that will likely arise for any organization attempting to develop an engaging, ethical, and useful human-in-the-loop initiative Methods employed throughout the course of the HITL initiative offered insights into how an organization might elucidate more specific goals with regards to engagement, ethics, and usefulness, and introduce feedback mechanisms to help measure and achieve success in reaching those goals.

These lessons are generalized here to frame the human-in-the-loop approach as a distinct set of phases, challenges, and goals, and offers guidance on staffing, resourcing, and assessing progress. The resulting framework was designed for use by the Library of Congress, but ideally it should benefit any cultural heritage organization with similar values and goals.

HITL INITIATIVE TEAM

AVP worked collaboratively with LC Labs and staff from the Digital Strategy Directorate; Digital Services Directorate; Science, Technology & Business Division; and User Experience Design, to design and develop this initiative.

AVP

Shawn Averkamp, HITL Initiative Lead, Subject Matter Expert
Kerri Willette, User Testing Lead, Subject Matter Expert
Amy Rudersdorf, HITL Initiative Manager, Subject Matter Expert
Dan Fischer, Machine Learning Expert, Software Engineer
Casey Arendt, UI/UX Designer
Wes Doyle, Software Engineer

Library of Congress

Meghan Ferriter, HITL Initiative Lead & Senior Innovation Specialist, LC Labs, Digital Strategy Directorate

Lauren Algee, Community Manager, By the People; Senior Innovation Specialist, Digital Content Management section, Digital Services Directorate
Natalie Burclaff, Business Reference and Research Specialist in the Science, Technology & Business Division
Eileen Jakeway Manchester, Innovation Specialist, LC Labs, Digital Strategy Directorate
Jaime Mears, Senior Innovation Specialist, LC Labs, Digital Strategy Directorate
Trevor Owens, Chief, Digital Content Management section, Digital Services Directorate
Abbey Potter, Senior Innovation Specialist, LC Labs, Digital Strategy Directorate
Leah Weinryb Grohsgal, Program Advisor to the Director of Digital Strategy, Digital Strategy Directorate
Jamie Bresner, Section Chief, User Experience Design, IT Design & Development Directorate
Amanda Perez, Art Director/Senior Designer, User Experience Design, IT Design & Development Directorate
Wendy Stengel, IT Section Chief, User Experience Design, IT Design & Development Directorate

PROCESS

FRAMEWORK

The unique challenges of machine learning as a technical solution require libraries to consider resources and staffing throughout the lifecycle of an initiative, beyond mere design and initial implementation. As machines learn from human training, outputs may improve but may also change or evolve in unexpected ways to negatively impact engagement, ethics, or usefulness. Slight variations in collection content or digital quality may be imperceptible to humans, but present new challenges to trained algorithms, so constant iteration is necessary in monitoring accuracy and bias of results and adjusting models or parameters or even ML approaches to course-correct.

User-centered design offers tools and principles to engage humans in the design and success of an initiative, also with a focus on human feedback and iteration, or “Humans-in-the-Loop.” We offer a framework for development of human-in-the-loop approaches in cultural heritage using the lens of iterative, user-centered design to help guide not just algorithm design but all phases in the lifecycle of an initiative, from collection selection to data integration and discovery.

The following framework outlines four stages of human-in-the-loop initiative development — **collection selection, design, implementation, and presentation/sharing** — addressing the challenges and goals of each stage in relation to engagement, ethics, and usefulness, the humans to be involved in each phase, and tools for incorporating feedback into the identification and mitigation of risk or potential harm to humans.

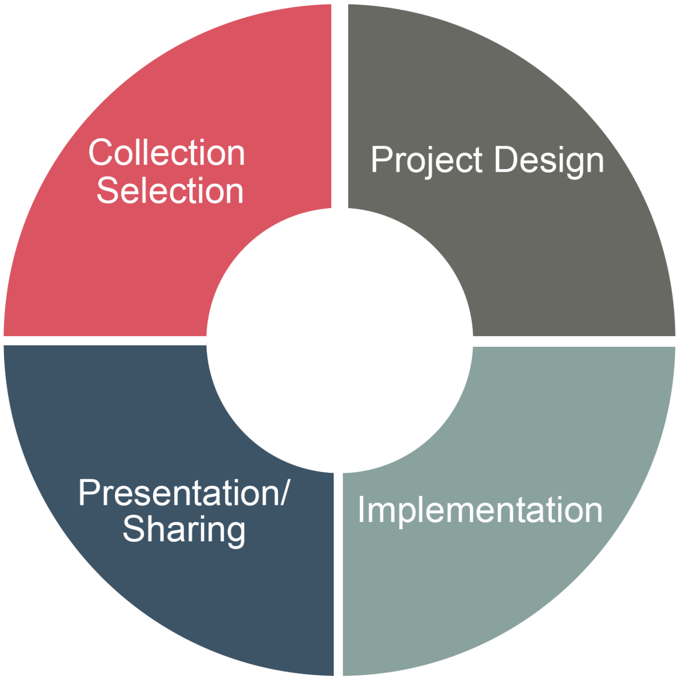


Image 2. Chart depicting four stages of the human-in-the-loop framework.

The following sections of this report will summarize the process and findings of the HITL experiment as undertaken by the team within the four stages in the human-in-the-loop framework defined above. Each section will detail the framework stage's objectives, goals, challenges, human requirements and impact ("humans"), and feedback mechanisms, incorporating lessons learned and recommendations for future endeavors. (See Appendix A for the full framework without initiative details.[20])

[20] "A Humans-in-the-Loop Framework." Library of Congress, accessed November 24, 2021, https://github.com/LibraryOfCongress/hitl/blob/main/hitl-recommendations-report-appendices/Appendix%20A_%20Framework.pdf.

STAGE 1: COLLECTION SELECTION

Objectives

Defined objectives for the collection selection stage included:

- Select collection(s) that will inform initiative design
- Identify data outputs

.Goals

Throughout each of the HITL initiative stages we defined goals in terms of the broader central concepts: ***engaging***, ***ethical***, and ***useful***. At the collection selection stage, the core team focused on achieving the following goals with a focus on creating a replicable selection process for future human-in-the-loop activities.

Engaging	Ethical	Useful
● Selecting collection content that is interesting to crowdsourcing users ● Crowdsourcing volunteers feel a personal connection, a sense of ownership to the selected content	● Exposing data from the selected collection respects the privacy of collection subjects or creators ● Potential risks to users and collection creators/ subjects can be identified and mitigated	● Replicable collection selection processes for human-in-the-loop approaches are modeled ● Data generated from the selected collection are useful to library users ● Data generated from selected collection improves discoverability of that content ● Collection content is free of permissions restrictions to enable broad use

.Challenges

During collection selection, the team needed to define who should be consulted in the collection selection process, trying to strike a balance between including a diversity of perspectives, while continuing to meet initiative deadlines. The AVP technical team also faced challenges around defining feasible ML processes for each proposed experiment, while identifying complementary crowdsourcing tasks that would be interesting and engaging for volunteers.

To address these challenges we asked:

How do we…

- Engage a broad diversity of perspectives in selection without bogging down the process?
- Select a collection large yet homogeneous enough to benefit from a ML approach?
- Select a collection that can attract and sustain the interest of crowdsourcing users?
- Find ML methods advanced enough to generate the desired data from the collection?

Humans

In order to achieve the goals and address the challenges of the Collection Selection phase:

We involve…

- **Community managers**, who understand what tasks are engaging to volunteers
- **Reference specialists**, who understand what data researchers are searching for and why
- **Collection curators**, who understand the content within collections and can speak to potential risk
- **Digital collection specialists**, who understand how to work with digitized collection objects and metadata
- **Machine-learning experts**, who understand what data generation tasks are possible to do with algorithms
- **Program specialists**, who understand how to connect humans across organizational divisions in support of the initiative

Feedback Mechanisms

In order to achieve the goals and address the challenges of this phase:

We learn from…

- **Cross-functional brainstorming workshops** to bring diverse collaboration to idea generation
- **Risk/benefit analysis** to help to identify risks and mitigation strategies early in the collaboration
- **User stories** to show how and what collection content will be useful as data

Process

This first stage of the HITL initiative centered around selecting a LC collection and goal that would serve as the focus of the experiment, and to identify data outputs that could be generated for proposed candidates. The team focused attention on defining replicable processes for collection selection that could be applied for identifying future human-in-the-loop activities.

During an initial discovery and research review process in November and December 2020, AVP worked closely with the LC core team and other LC employees and stakeholders to better understand LC's current experimental digital transformation efforts and human-in-the-loop opportunities.

Specific research conducted in support of collection selection included:

- Collection Selection Workshop with LC staff and stakeholders (November 6, 2020)
- Post-Workshop Survey with workshop participants (November 9-11, 2020)
- Staff and stakeholder interviews (November-December 2020)

These activities were designed to not only select a collection for the HITL experiment, but to inform the creation of a replicable, documented selection approach in support of the framework described above so that future LC staff seeking to evaluate collections' suitability for computational approaches will be able to apply the recommended approaches. Through the process of workshopping, interviewing, and brainstorming with stakeholders, the core team outlined and defined potential end-users, datasets, tasks, risks, and other pertinent background information for eight finalist candidate collections for HITL, found in Appendix H.

Narrowing the Candidate Pool

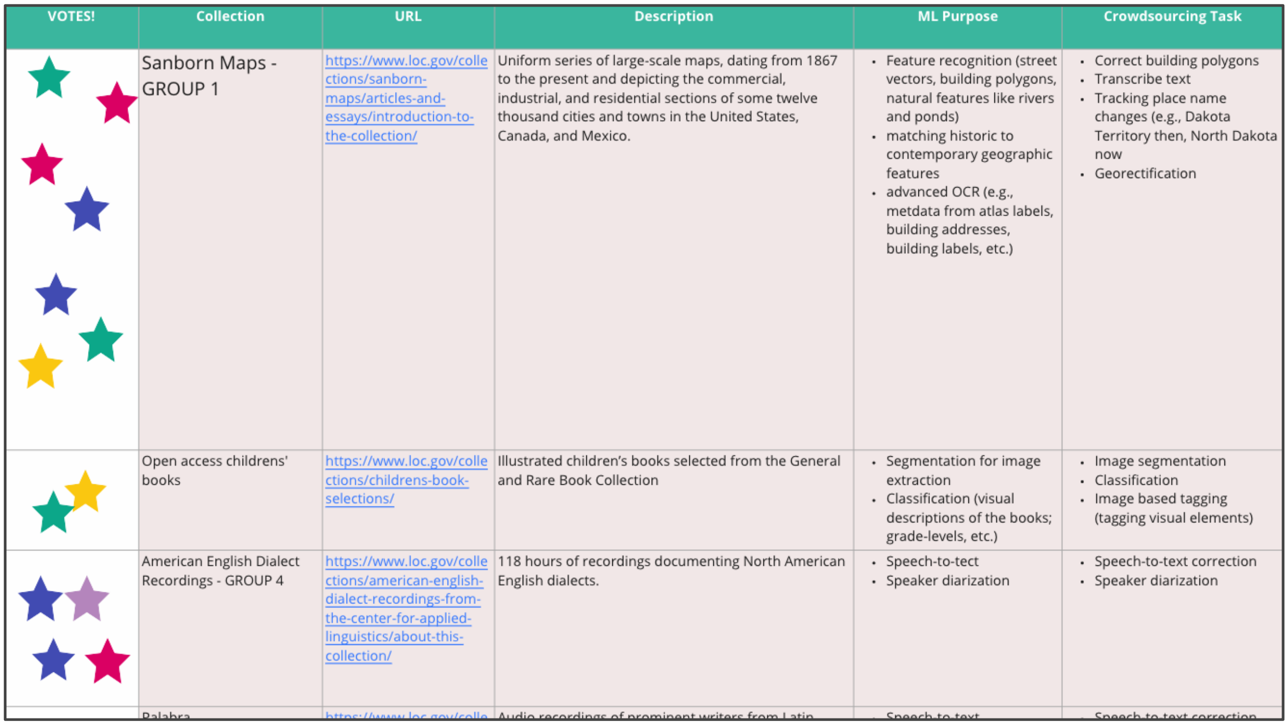
To develop a pool of viable collection candidates, AVP asked the LC team to nominate candidate digital collections, identify data-enhancement goals, and outline potential ML and crowdsourcing tasks that would support each. Defined types of ML tasks included content segmentation, OCR of text materials, entity recognition, speech to text transcription, and others. Identified crowdsourcing tasks included various possibilities for each collection such as classification, speech-to-text correction, and image-based tagging.

In November 2020, AVP assembled a group of representative LC Labs staff and stakeholders for a workshop to narrow the identified candidates down to four finalists. During the workshop, attendees discussed the eight candidates with a "champion" for each collection providing thorough background information and collection context to the larger group. Once all of the candidates were reviewed and discussed, the full team voted to select four finalist collections, which included Sanborn Map Collection[21], U.S. Telephone Directory Collection[22], By the

[21] "Introduction to the Collection," Library of Congress, accessed July 1, 2021, https://www.loc.gov/collections/sanborn-maps/articles-and-essays/introduction-to-the-collection/.

[22] "U.S. Telephone Directory Collection," Library of Congress, accessed July 1, 2021, https://www.loc.gov/collections/united-states-telephone-directory-collection/about-this-collection/.

People[23], and American English Dialect Recordings[24]. Once the four finalists were identified, workshop attendees engaged in brainstorming activities as illustrated in the image below[25] to uncover the potential benefits and values that human-in-the-loop approaches might offer human volunteers and end users. The exercises helped uncover potential risks and biases that might be introduced to human volunteers, end users, and collection creators or subjects through the specific tasks proposed for each goal. Finally, attendees discussed potential avenues to mitigate the risks or biases related to proposed tasks for each candidate.



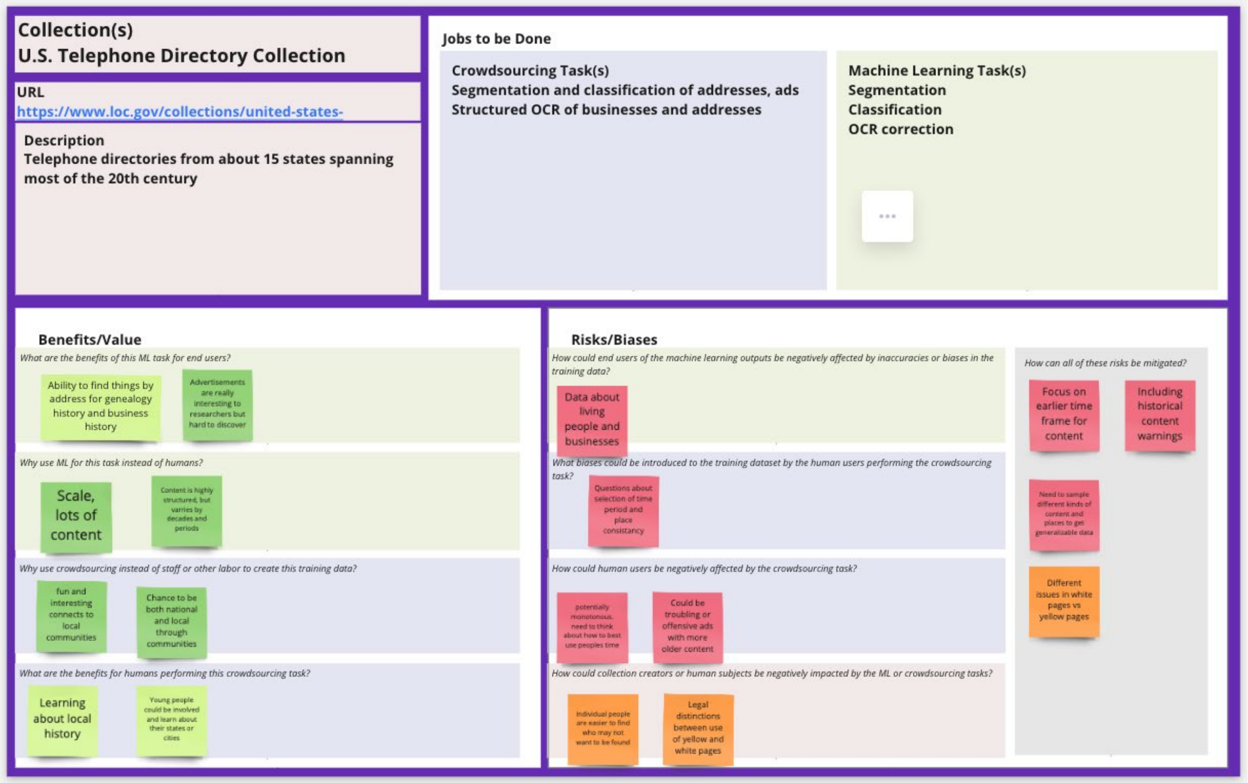
VOTES!	Collection	URL	Description	ML Purpose	Crowdsourcing Task
★★★★★★★	Sanborn Maps - GROUP 1	https://www.loc.gov/collections/sanborn-maps/articles-and-essays/introduction-to-the-collection/	Uniform series of large-scale maps, dating from 1867 to the present and depicting the commercial, industrial, and residential sections of some twelve thousand cities and towns in the United States, Canada, and Mexico.	• Feature recognition (street vectors, building polygons, natural features like rivers and ponds) • matching historic to contemporary geographic features • advanced OCR (e.g., metdata from atlas labels, building addresses, building labels, etc.)	• Correct building polygons • Transcribe text • Tracking place name changes (e.g., Dakota Territory then, North Dakota now • Georectification
★★	Open access childrens' books	https://www.loc.gov/collections/childrens-book-selections/	Illustrated children's books selected from the General and Rare Book Collection	• Segmentation for image extraction • Classification (visual descriptions of the books; grade-levels, etc.)	• Image segmentation • Classification • Image based tagging (tagging visual elements)
★★★★	American English Dialect Recordings - GROUP 4	https://www.loc.gov/collections/american-english-dialect-recordings-from-the-center-for-applied-linguistics/about-this-collection/	118 hours of recordings documenting North American English dialects.	• Speech-to-tect • Speaker diarization	• Speech-to-text correction • Speaker diarization
	Palabra	https://www.loc.gov/colle	Audio recordings of prominent writers from Latin	• Speech-to-text	• Speech-to-text correction

Image 3. Screenshot of Miro board used to facilitate brainstorming and voting on experiment finalists. See Appendix B for a larger view of this graphic.

[23] "About by the People," Library of Congress, accessed July 1, 2021, https://crowd.loc.gov/about/.

[24] "American English Dialect Recordings: The Center for Applied Linguistics Collection," Library of Congress, accessed July 1, 2021, https://www.loc.gov/collections/american-english-dialect-recordings-from-the-center-for-applied-linguistics/about-this-collection/.

[25] "HITL Collection Selection workshop," facilitated virtually November 6, 2020, accessed July 1, 2021, https://miro.com/app/board/o9J_kgLUJ7g=/.



Collection(s)
U.S. Telephone Directory Collection

URL
https://www.loc.gov/collections/united-states-

Description
Telephone directories from about 15 states spanning most of the 20th century

Jobs to be Done

Crowdsourcing Task(s)
Segmentation and classification of addresses, ads
Structured OCR of businesses and addresses

Machine Learning Task(s)
Segmentation
Classification
OCR correction

Benefits/Value

What are the benefits of this ML task for end users?

- Ability to find things by address for genealogy history and business history
- Advertisements are really interesting to researchers but hard to discover

Why use ML for this task instead of humans?

- Scale, lots of content
- Content is highly structured, but varries by decades and periods

Why use crowdsourcing instead of staff or other labor to create this training data?

- fun and interesting connects to local communities
- Chance to be both national and local through communities

What are the benefits for humans performing this crowdsourcing task?

- Learning about local history
- Young people could be involved and learn about their states or cities

Risks/Biases

How could end users of the machine learning outputs be negatively affected by inaccuracies or biases in the training data?

- Data about living people and businesses

What biases could be introduced to the training dataset by the human users performing the crowdsourcing task?

- Questions about selection of time period and place consistancy

How could human users be negatively affected by the crowdsourcing task?

- potentially monotonous. need to think about how to best use peoples time
- Could be troubling or offensive ads with more older content

How could collection creators or human subjects be negatively impacted by the ML or crowdsourcing tasks?

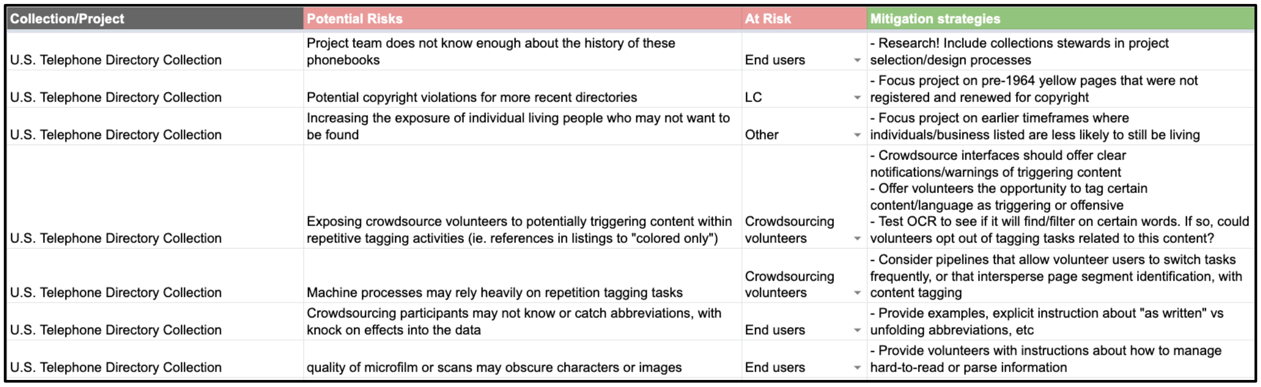
- Individual people are easier to find who may not want to be found
- Legal distinctions between use of yellow and white pages

How can all of these risks be mitigated?

- Focus on earlier time frame for content
- Including historical content warnings
- Need to sample different kinds of content and places to get generalizable data
- Different issues in white pages vs yellow pages

Image 4. Screenshot of Miro board depicting crowdsourcing and ML tasks, benefits/value, and risks/bias for the U.S. Telephone Directory Collection. See Appendix B for a larger view of this graphic.

Risk identification and mitigation were identified by LC Labs as foundational components of this and other human-in-the-loop approaches and experiments. Identifying potential risks and biases inherent in source data, as well as those that may be introduced through human or machine-driven data generation processes, began here at the collection-selection stage, but were revisited and updated at multiple stages in the lifecycle of the initiative. Identified risks and potential mitigation strategies were added to a tab in the candidate collection spreadsheet (Appendix H) for ongoing tracking and further refinement.[26]



Collection/Project	Potential Risks	At Risk	Mitigation strategies
U.S. Telephone Directory Collection	Project team does not know enough about the history of these phonebooks	End users	- Research! Include collections stewards in project selection/design processes
U.S. Telephone Directory Collection	Potential copyright violations for more recent directories	LC	- Focus project on pre-1964 yellow pages that were not registered and renewed for copyright
U.S. Telephone Directory Collection	Increasing the exposure of individual living people who may not want to be found	Other	- Focus project on earlier timeframes where individuals/business listed are less likely to still be living
U.S. Telephone Directory Collection	Exposing crowdsource volunteers to potentially triggering content within repetitive tagging activities (ie. references in listings to "colored only")	Crowdsourcing volunteers	- Crowdsource interfaces should offer clear notifications/warnings of triggering content - Offer volunteers the opportunity to tag certain content/language as triggering or offensive - Test OCR to see if it will find/filter on certain words. If so, could volunteers opt out of tagging tasks related to this content?
U.S. Telephone Directory Collection	Machine processes may rely heavily on repetition tagging tasks	Crowdsourcing volunteers	- Consider pipelines that allow volunteer users to switch tasks frequently, or that intersperse page segment identification, with content tagging
U.S. Telephone Directory Collection	Crowdsourcing participants may not know or catch abbreviations, with knock on effects into the data	End users	- Provide examples, explicit instruction about "as written" vs unfolding abbreviations, etc
U.S. Telephone Directory Collection	quality of microfilm or scans may obscure characters or images	End users	- Provide volunteers with instructions about how to manage hard-to-read or parse information

[26] "Appendix H – Collection Candidate Evaluation." Library of Congress, accessed November 24, 2021, https://github.com/LibraryOfCongress/hitl/blob/main/hitl-recommendations-report-appendices/Appendix%20H_%20CollectionProjectCandidateEvaluation.xlsx.

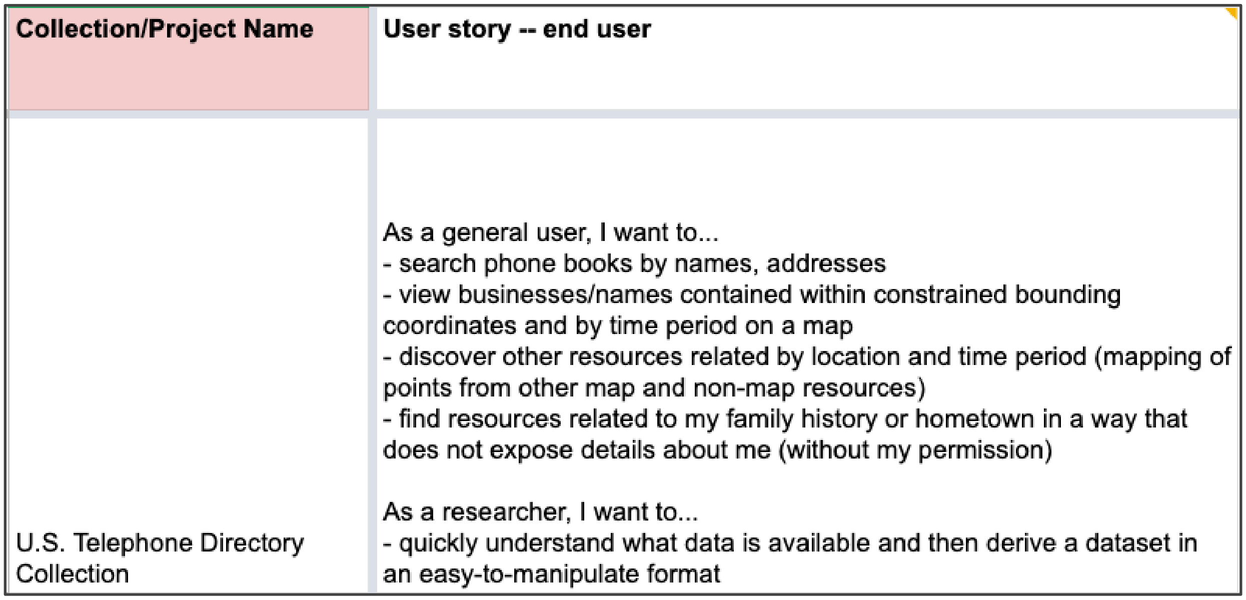
Image 5. Screenshot of Collection Candidates spreadsheet representing potential risks and mitigation strategies associated with the U.S. Telephone Directory Collection. See Appendix B below for a larger view of this graphic.

The candidate collections sheet referenced above contains additional compiled findings from the workshop and a post-workshop survey[27] captured feedback from participants on specific aspects of the workshop approach. The survey results indicated a positive response to the workshop approach generally. Participants expressed a desire for the current experiment to model tools that would support tracking candidate criteria, risks, and biases in the selection of collections and material to be included in future human-in-the-loop approaches.

Technical Review

With the candidate pool narrowed down to four potential experiments, each with clearly identified tasks, benefits, and risks, the team began investigating the technical feasibility of the finalist goals.

During this phase, the team drafted specific end-user stories to help clarify how end users might use and apply the data derived from each of the HITL experiment candidates. User stories identified specific user types for each experiment (e.g., general collection end users, crowdsourcing volunteers, and researchers), and then specified particular activities each user type might like to perform with support of the derived data, as shown in the illustrations below:

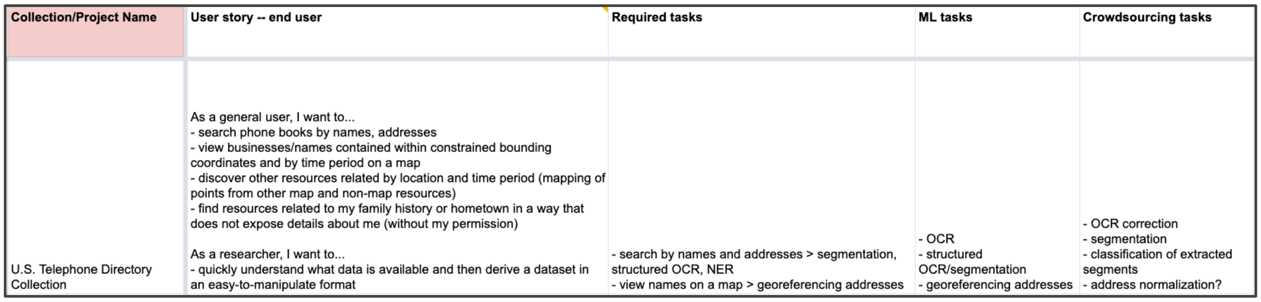


Collection/Project Name	**User story -- end user**
U.S. Telephone Directory Collection	As a general user, I want to... - search phone books by names, addresses - view businesses/names contained within constrained bounding coordinates and by time period on a map - discover other resources related by location and time period (mapping of points from other map and non-map resources) - find resources related to my family history or hometown in a way that does not expose details about me (without my permission) As a researcher, I want to... - quickly understand what data is available and then derive a dataset in an easy-to-manipulate format

Image 6. Screenshot of Collection Candidates spreadsheet representing potential user stories associated with the U.S. Telephone Directory Collection. See Appendix B below for a larger view of this graphic.

As user stories emerged, the core team further refined the human-in-the-loop tasks that would be required to inform data creation in support of identified end-user needs.

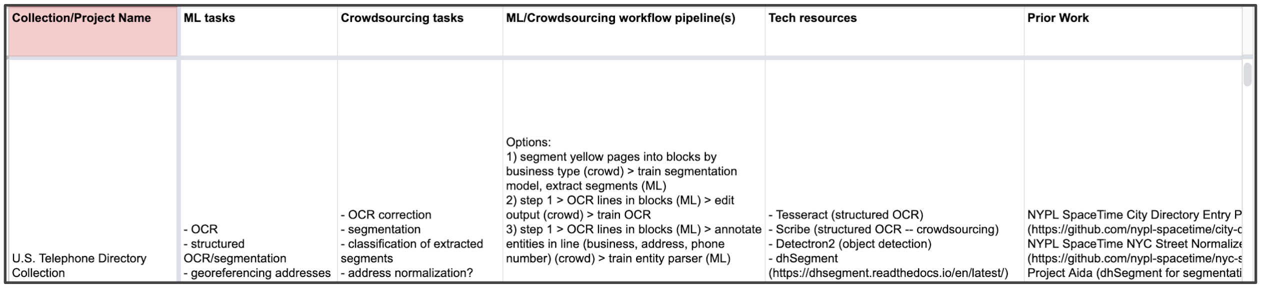
[27] "HITL Collection Selection: Post-Workshop Survey," Google form, accessed July 1, 2021, https://forms.gle/cVfaFVR7LYmiJ9MA8.



Collection/Project Name	User story -- end user	Required tasks	ML tasks	Crowdsourcing tasks
U.S. Telephone Directory Collection	As a general user, I want to... - search phone books by names, addresses - view businesses/names contained within constrained bounding coordinates and by time period on a map - discover other resources related by location and time period (mapping of points from other map and non-map resources) - find resources related to my family history or hometown in a way that does not expose details about me (without my permission) As a researcher, I want to... - quickly understand what data is available and then derive a dataset in an easy-to-manipulate format	- search by names and addresses > segmentation, structured OCR, NER - view names on a map > georeferencing addresses	- OCR - structured OCR/segmentation - georeferencing addresses	- OCR correction - segmentation - classification of extracted segments - address normalization?

Image 7. Screenshot of Collection Candidates spreadsheet representing potential user stories and necessary crowdsourcing and ML tasks to fulfill the user stories associated with the U.S. Telephone Directory Collection. See Appendix B below for a larger view of this graphic.

Once specific ML and crowdsourcing tasks were more clearly defined, AVP identified existing open-source technical resources and code bases from other ML and crowdsourcing initiatives that might potentially support the identified tasks and pipelines necessary for each candidate experiment. For example, the NYPL Space/Time Directory[28] offered insight into and open-source code for parsing and structuring entities found in the text of city directory business listings, which could easily translate to the task of structuring entities found in the text of LC's collection of telephone directories.



Collection/Project Name	ML tasks	Crowdsourcing tasks	ML/Crowdsourcing workflow pipeline(s)	Tech resources	Prior Work
U.S. Telephone Directory Collection	- OCR - structured OCR/segmentation - georeferencing addresses	- OCR correction - segmentation - classification of extracted segments - address normalization?	Options: 1) segment yellow pages into blocks by business type (crowd) > train segmentation model, extract segments (ML) 2) step 1 > OCR lines in blocks (ML) > edit output (crowd) > train OCR 3) step 1 > OCR lines in blocks (ML) > annotate entities in line (business, address, phone number) (crowd) > train entity parser (ML)	- Tesseract (structured OCR) - Scribe (structured OCR -- crowdsourcing) - Detectron2 (object detection) - dhSegment (https://dhsegment.readthedocs.io/en/latest/)	NYPL SpaceTime City Directory Entry P (https://github.com/nypl-spacetime/city-c NYPL SpaceTime NYC Street Normalize (https://github.com/nypl-spacetime/nyc-s Project Aida (dhSegment for segmentati

Image 8. Expanded view of screenshot of Collection Candidates spreadsheet, including a column on the far right representing prior digital initiatives and source code. See Appendix B below for a larger view of this graphic.

Through this process, AVP was able to identify existing technical resources, documentation, and code bases and to assess the technical feasibility of working with each of the finalist candidates within the given timeframe and specific technical constraints of the HITL collaboration. Specific considerations involved in choosing the technical resources and codebases for the final candidate are detailed in the *Design* section below.

Selected Collection and Goals

Based on findings from the activities above, AVP selected the **U.S. Telephone Directory Collection**,[29] scanned pages of telephone directory white and yellow pages from about 15

[28] "NYC Space/Time Directory," New York Public Library, accessed July 1, 2021, http://spacetime.nypl.org/.

[29] "U.S. Telephone Directory Collection," Library of Congress, accessed July 1, 2021, https://www.loc.gov/collections/united-states-telephone-directory-collection/about-this-collection/.

states spanning most of the 20th century, for enhancement. Initial discovery and user stories identified for this collection indicated a strong need to generate data that facilitates search across directories by business names, addresses, locations, and time periods in order to support a wide range of historical research centered on local communities, industries, and family histories. However, the core team also identified legal and ethical risks of working with white pages directories, which were excluded from final consideration and yellow pages business directories were drawn into focus.

In addition to the potential to reach a wide range of researcher needs, this collection was chosen for its homogeneity in subject matter, layouts, fonts, and structure across most digitized items, which makes it a prime candidate for applying machine learning and other batch processes. The task of transforming digitized images into structured business listing data is complex enough that multiple types of machine processes could be tested, but not so complex that it could not be completed (for a small subset of pages) within the very limited prototype timeframe. The team also considered available open-source crowdsourcing platforms and determined that the anticipated crowdsourcing workflows could be performed using Scribe[30] or PyBossa[31] with minimal customization.

With the end goal of extracting structured data on businesses listed in Yellow Page directories (e.g., business names, types, addresses, phone numbers), this collection became the basis for the HITL workflows detailed in the *Design* section below.

Recommendations

Based on experience from this experiment, the team recommends repurposing some of the specific approaches and tools applied to this HITL experiment, including:

- **Consult and include relevant collection experts**
  To identify collection candidates, the HITL team interviewed collection curators and reference teams who support the candidate collection. These interviews helped the team surface and document risks and user stories based on actual research scenarios and use cases. Collection experts were also included in the collection selection workshop, and were consulted at various stages throughout the lifecycle of the initiative.

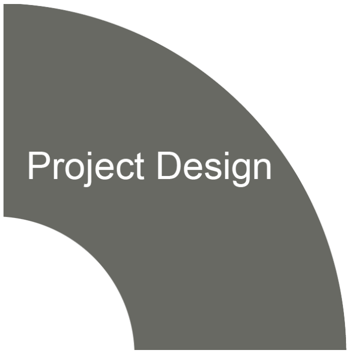
- **Build and maintain an ongoing list of human-in-the-loop Collection Candidates**
  The core team consistently relied on a specific tool throughout the HITL initiative: the collections candidate worksheet (Appendix H). AVP recommends maintaining a similar worksheet where human-in-the-loop candidate collections can be tracked, documented, and fully explored as opportunities to develop new human-in-the-loop approaches emerge.

[30] "scribeAPI," Github, accessed July 1, 2021. https://github.com/LibraryOfCongress/scribeAPI.

[31] PyBossa, accessed July 1, 2021, https://pybossa.com/.

- **Surface and continually reassess potential benefits, risks, and biases for each candidate collection**
  As candidate collections are considered for human-in-the-loop experiments, it will be necessary to identify and document potential benefits, risks, and biases associated with a given initiative. In particular, risks and biases should be identified during the collection selection stage so that mitigation strategies can be revisited and considered at multiple points throughout the lifecycle of the initiative.

STAGE 2: DESIGN



Objectives

Defined objectives for the Design stage included:

- Model collection content as structured data
- Design ML pipeline
- Design crowdsourcing tasks
- Define quality control (QC) measures
- Define volunteer outreach plan

Goals

Design stage goals for the HITL initiative centered around defining crowdsourcing tasks that engage volunteers in a variety of ways. The HITL initiative design aimed to be mindful of how defined tasks might expose volunteers to sensitive or offensive collection content, and transparent about how volunteer tasks would interact with ML processes.

Engaging	Ethical	Useful
● Defined crowdsourcing tasks are enjoyable to volunteers ● Volunteers get deep exposure to collection content ● Volunteers understand the value of their contributions to the greater good	● Volunteers understand how their contributions will be used especially in relation to ML processes ● Risks to volunteers from potentially offensive content are identified and mitigated ● Potential unintended consequences of ML processes are identified and documented	● Data is modeled in a way that it can be shaped for various kinds of use

Challenges

Challenges in this phase including addressing uncertainty around the potential of ML processes for data generation from the selected collection content.

To address these challenges we asked:

How do we…

- Support a wide range of structural complexities in collection content as they surface?
- Generate enough training data to test and select initial ML processes?

- Design crowdsourcing tasks that support the ML pipeline and are also interesting to volunteers?
- Identify a “good enough” threshold for ML accuracy?
- Ensure crowdsourcing data is accurate enough to use as ground truth?

Humans

In order to achieve the goals and address the challenges of the Design phase:

We involve…

- **Collection curators**, who understand the collection content to be modeled
- **Reference specialists**, who understand how researchers could use collection content as data
- **Metadata and digital collection specialists**, who understand how to model content as data
- **Machine learning experts**, who understand ML processes for extracting data
- **Community managers**, who understand what tasks are engaging to volunteers
- **Library staff**, who can create training data for initial ML explorations
- **Volunteers**, who can offer early feedback on possible crowdsourcing tasks

Feedback Mechanisms

In order to achieve the goals and address the challenges of this phase:

We learn from…

- **User interviews** with volunteers to help identify engaging types of tasks
- **Ground-truth accuracy testing** to help measure the fit of an ML process for a data generation task

Process

After selecting the collection, defining user stories and high-level data enrichment goals, and identifying potential ML and crowdsourcing tasks, the core team began designing the HITL initiative experiments. The design process focused on ideation and iteration to develop concrete strategies for selecting and applying the most effective ML processes for extracting data and complementary crowdsourcing tasks that would be engaging and ethical for volunteers. The core team kept user needs and potential risks that were defined for the collection in the foreground as they began developing ML and crowdsourcing pipelines for the next stage of implementation.

Data Modeling the Yellow Pages

In order to extract structured data from the Yellow Pages images with ML processes, the core team first needed to define what data would be important to users. Data modeling is a method for understanding a domain at a high level, scoping that domain to a set of user needs or goals. With those needs and goals in mind, the major entities, attributes of entities, and relationships between entities may be defined to frame and guide development of applications and processes that support those user needs.

Some of the user stories created in the collection selection process spoke directly to the types of structured data that would need to be extracted or enhanced, such as business names, addresses, types, and current geocoordinates:

- “As a family genealogist, I want to search phone books by individual names, business names, and addresses, so that I can identify where my ancestors lived or worked and move on to additional research about those people.”
- “As a small business owner, I want to search the directory for historical listings and ads at the address of my current business, so that I can locate previous businesses that were in this location.”
- “As a historian, I want to group listings from a directory by ‘type’ or ‘industry’ of business (automobile, television, telephone, etc.), so that I can understand the representation of various industries and compare them over time.”

Other user stories required a bit more imagination around what data would support their research questions, what structure it should take, and how that data could interact with or be integrated with data from other resources:

- “As a feminist literary historian, I want to find the names of female-owned businesses in the yellow pages, so that I may analyze trends in female business ownership over time.”
- “As a historian, I want to compare data from the yellow page listings to data from a particular city to the *Motorist Green-Book* data sets from NYPL, so that I can identify and dig more deeply into businesses in a specific city that were open to African American travelers in the 1940s.”
- “As a special collections staff person, I want to identify Boston businesses owned by or serving under-represented communities during a specific timeframe, so that I can work with local community organizations/historians to uncover and share historical records from Boston's underrepresented communities.”

In addition to the user stories crafted by the larger team, a conversation with Natalie Burclaff (Business Reference and Research Specialist in the Science, Technology & Business Division), illuminated areas of the data model. Specifically, Natalie reflected on supporting the research uses of the physical Yellow Pages collection, challenges and idiosyncrasies of the content, and function of the collection within the context of related library resources. These details suggested points where the team might need to address ambiguity, incompleteness, and bias and also enable points of connection to other potential collections-as-data in the future.

While the opportunities for potential data enrichments based on these hypothetical user needs heavily outweighed the available resources for this experiment, thinking big about possible collection enhancement helped the team design a flexible and extensible data model that would support additional future enrichment of the collection, as resources become available or as new ML approaches mature. AVP crafted an initial set of business rules, a human-readable narrative describing the entities, attributes, relationships, and constraints of the domain, and shared it with the LC collaborators for feedback and iteration.

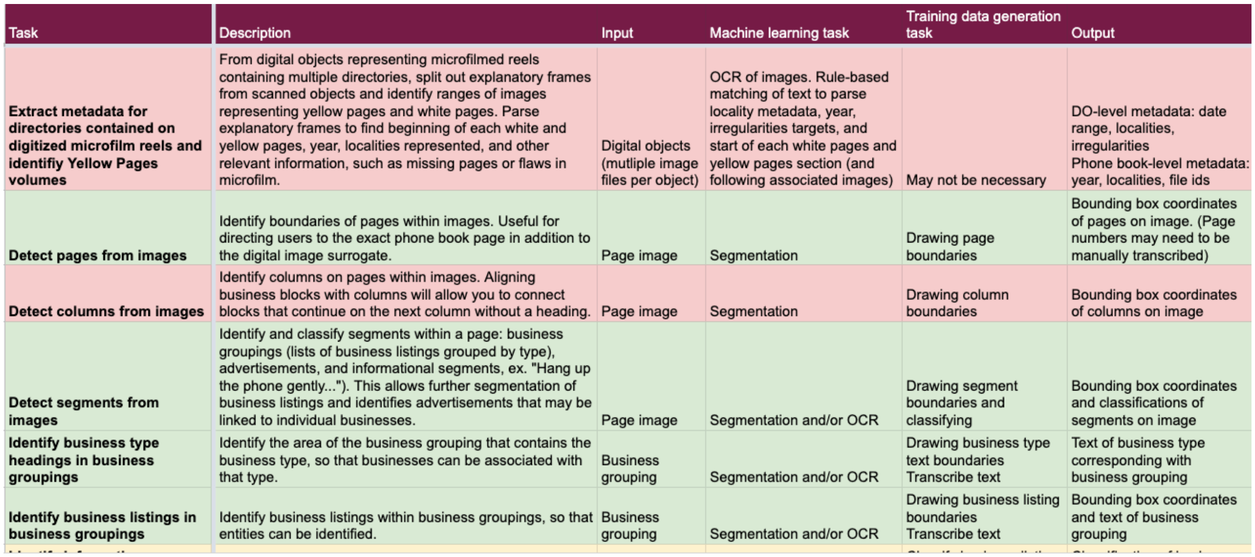
The full set of business rules and a data dictionary can be found in Appendix C below, but this excerpt shows how narrative business rules can establish a common vocabulary for use when developing a human-in-the-loop approach and help promote a shared understanding of scope and objectives:

> The Library of Congress U.S. Telephone Directory **Collection** contains **Phone Books** digitized from microfilm that may contain either **White Pages** (individual listings) and/or **Yellow Pages** (business listings). Digitized microfilm **Images** represent 1-2 phone book **Pages**, microfilm technical targets, or frames of explanatory material (such as indicators for where the White Pages or Yellow Pages sections start, where material is missing, or metadata for the original object).
>
> Pages of the Yellow Pages are usually divided into 2-4 **Columns**, though sometimes advertisements may span several columns, often creating shorter columns. Columns usually include **Groupings** of **Business Listings**, **Advertisements**, or **Tips/Information** about using the phone book. **Business groupings** are organized by business or service type displayed in a larger font above the listings. Business listings may be **Informational Listings** that give additional information about the business, sometimes in the form of an advertisement, with graphical elements. Information listings are usually set apart from standard listings with boxes or bounded by horizontal lines...

From the business rules, the core team further defined the entities (in bold, above), attributes of those entities, and relationships between entities in a data dictionary. In addition to providing a semantic definition to entities and attributes, the data dictionary specifies the data types of each attribute to support both the design of systems for storing the data, and data structures, such as JSON or CSV, for sharing the data output with other applications or human users.

Task Definition

Once the target data was defined, AVP sketched out at a high level the tasks necessary to extract the data from images through ML and supporting crowdsourcing tasks. The excerpt below shows the possible tasks envisioned in starting with digitized microfilmed images from the U.S. Telephone Directory Collection in LC Digital Collections to generating structured business listing data, all the way to possible future enhancement endeavors like geocoding business listing addresses and linking businesses across years of directories. (Full task list is available in Appendix H in the “YP Tasks” tab.)



Task	Description	Input	Machine learning task	Training data generation task	Output
Extract metadata for directories contained on digitized microfilm reels and identifiy Yellow Pages volumes	From digital objects representing microfilmed reels containing multiple directories, split out explanatory frames from scanned objects and identify ranges of images representing yellow pages and white pages. Parse explanatory frames to find beginning of each white and yellow pages, year, localities represented, and other relevant information, such as missing pages or flaws in microfilm.	Digital objects (mutliple image files per object)	OCR of images. Rule-based matching of text to parse locality metadata, year, irregularities targets, and start of each white pages and yellow pages section (and following associated images)	May not be necessary	DO-level metadata: date range, localities, irregularities Phone book-level metadata: year, localities, file ids
Detect pages from images	Identify boundaries of pages within images. Useful for directing users to the exact phone book page in addition to the digital image surrogate.	Page image	Segmentation	Drawing page boundaries	Bounding box coordinates of pages on image. (Page numbers may need to be manually transcribed)
Detect columns from images	Identify columns on pages within images. Aligning business blocks with columns will allow you to connect blocks that continue on the next column without a heading.	Page image	Segmentation	Drawing column boundaries	Bounding box coordinates of columns on image
Detect segments from images	Identify and classify segments within a page: business groupings (lists of business listings grouped by type), advertisements, and informational segments, ex. "Hang up the phone gently..."). This allows further segmentation of business listings and identifies advertisements that may be linked to individual businesses.	Page image	Segmentation and/or OCR	Drawing segment boundaries and classifying	Bounding box coordinates and classifications of segments on image
Identify business type headings in business groupings	Identify the area of the business grouping that contains the business type, so that businesses can be associated with that type.	Business grouping	Segmentation and/or OCR	Drawing business type text boundaries Transcribe text	Text of business type corresponding with business grouping
Identify business listings in business groupings	Identify business listings within business groupings, so that entities can be identified.	Business grouping	Segmentation and/or OCR	Drawing business listing boundaries Transcribe text	Bounding box coordinates and text of business grouping

Image 9. Excerpt of Yellow Pages task definitions. See Appendix B below for a larger view of this graphic.

This exercise helped the core team to:

- understand the steps necessary in generating structured data as defined by the data model,
- identify where ML processes or human effort would be needed,
- consider what ML approaches might be successful, and
- determine where crowdsourcing tasks would be desirable to train ML methods or to validate their accuracy.

Breaking tasks out in this fashion also helped the team get a sense of the size, complexity, and necessity of each task, so they could scope the experiment appropriately to achieve a minimum viable product (MVP) within resource constraints. (In the above image, green rows indicate tasks within the scope of this experiment, and red denotes tasks that were out of scope.) For this experiment, the team decided that the goal of splitting images into pages, detecting business groupings in pages, detecting business types and listings within business groupings, and identifying entities within business listings, in order to create structured business listing data, would be achievable within the timeframe of the initiative.

Machine-Learning Pipeline Design

Once tasks were defined and scoped, AVP team members began to test ML approaches to each task, keeping in mind there may be more than one possible pipeline solution. The choice of which possible pipeline may depend on a number of factors, such as available ML expertise on the team or the ability to assess potential success with a limited amount of training data.

While some ML tasks, such as OCR (optical character recognition) and some object detection methods, could be tested using built-in models, others would require training data specific to the problem domain of the Yellow Pages. For instance, to test the ability of an object-detection algorithm to segment Yellow Pages images into advertisements, business groupings (groupings of business listings organized by business type), and telephone tips, the algorithm would need examples of each of these segment types to learn how to recognize them. Supplying this training data proved challenging as the core team would need to find a way to efficiently download images from LC servers, select a tool to quickly mark up and save image segments for training, and then spend time creating enough of this training data to assess the viability of the ML method. Without an existing instance of a crowdsourcing tool to do this work, the team spun up an instance of the lightweight open-source image annotation tool, VoTT,[32] to generate training images. Using a local client application, AVP team members marked up sample images downloaded from LC into a shared VoTT instance installed on a cloud server.

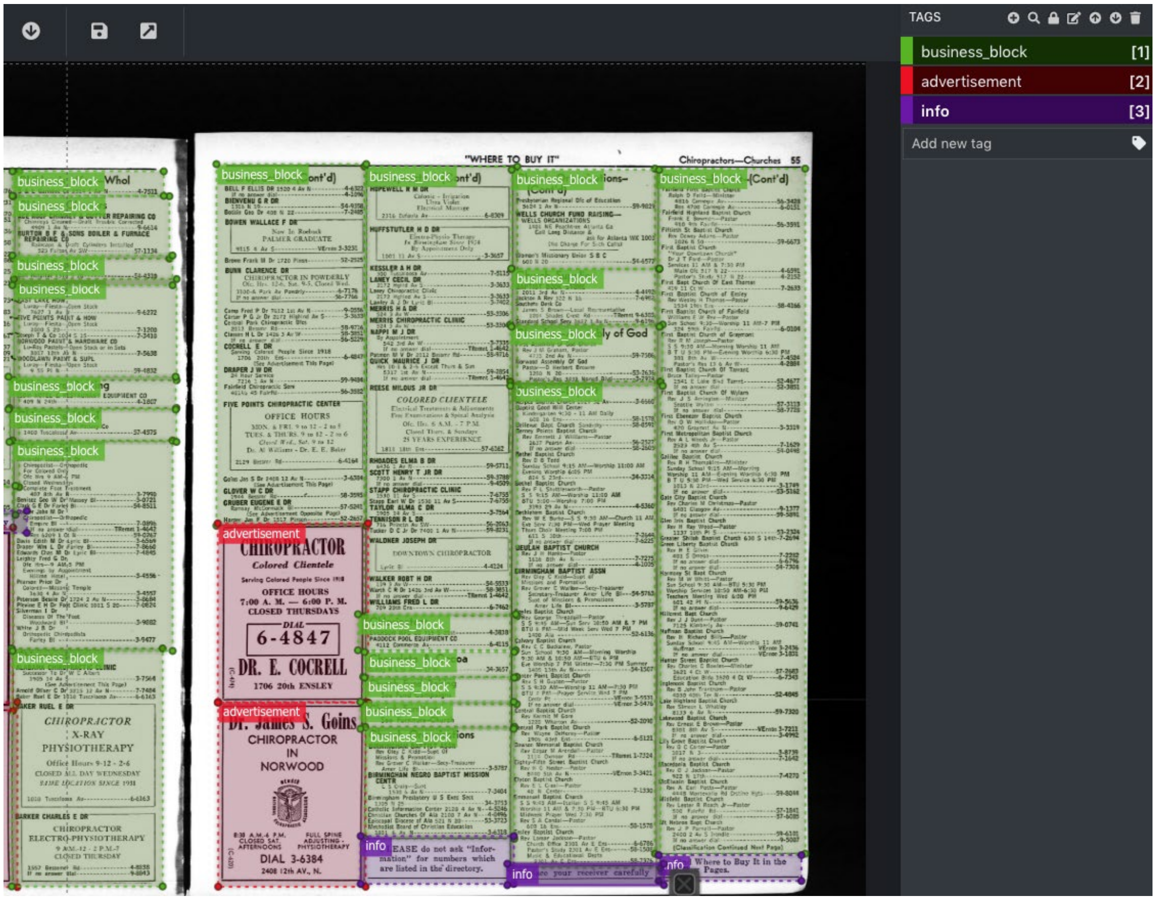


Image 10. Screenshot of marked up segments in the VoTT tool to be used for training an object detection model. See Appendix B below for a larger view of this graphic.

Though training a model from marked up segments from 36 two-up page images required a significant investment in staff time, it was clearly not nearly enough to produce remotely acceptable accuracy levels. While the process showed some promise for identifying advertisements, which were usually bounded by borders, it appeared that distinguishing distinct groupings of business listings from each other would be difficult, so AVP decided to explore alternate methods of segmentation.

[32] VoTT (Visual Object Tagging Tool), accessed July 1, 2021, https://github.com/microsoft/VoTT.

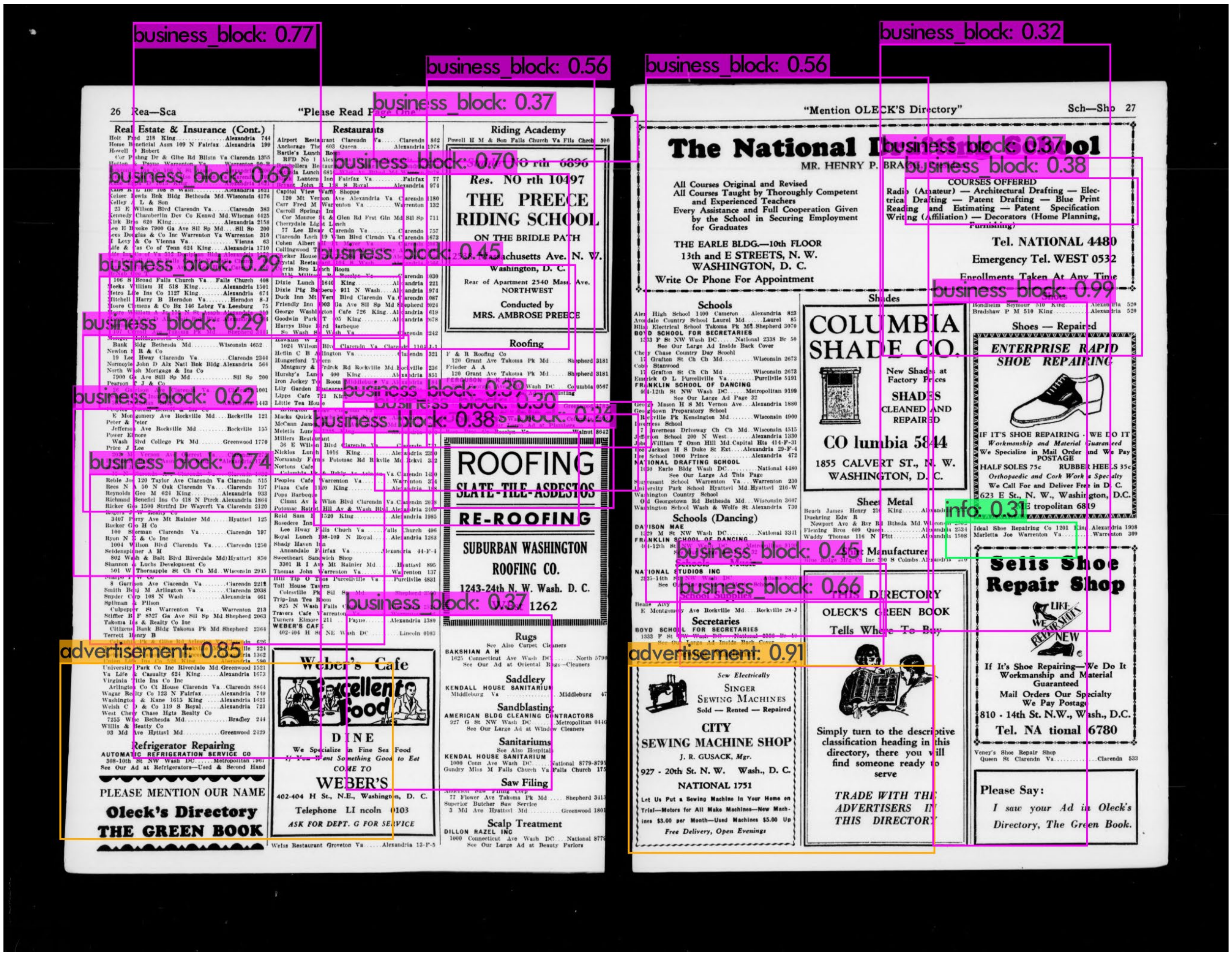


Image 11. Results of segmentation, with confidence levels, of the YOLO object detection system used with the open-source Darknet neural network framework[33], using a model trained on segments from 36 page images. See Appendix B for a larger view of this graphic.

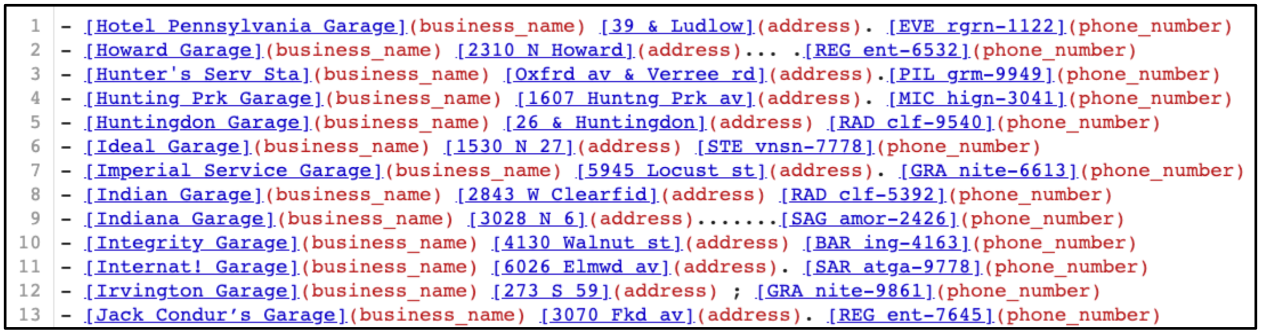
Similarly, in order to test the viability of a conditional random field (CRF)[34] algorithm[35] for identifying entities within a business listing, such as business name, address, and phone number, AVP had to manually create training data in Markdown format, as well as generate samples of text from OCR. A CRF is a type of natural language processing (NLP) method that identifies entities within text based on patterns it learns from model training data. AVP team members attempted first to train the CRF on 67 marked-up business listings without any adjustment to the default configurations. Based on a spotcheck of an initial limited attempt run over 100 OCRed business listings, as well as the success of the CRF experiment run by cultural

[33] Joseph Chet Redmon, "YOLO: Real-Time Object Detection," accessed July 1, 2021, https://pjreddie.com/darknet/yolo/.

[34] "Conditional Random Field," Wikipedia, accessed July 1, 2021, https://en.wikipedia.org/wiki/Conditional_random_field.

[35] The team used spacy-crfsuite (https://github.com/talmago/spacy_crfsuite), a wrapper around the scikit-learn Python library, sklearn-crfsuite (https://sklearn-crfsuite.readthedocs.io/), designed for use in SpaCy NLP pipelines.

heritage developer Bert Spaan and the (now-defunct) NYPL Labs on similar business directory text[36], the team decided to select this method as part of the pipeline.



```
- [Hotel Pennsylvania Garage](business_name) [39 & Ludlow](address). [EVE rgrn-1122](phone_number)
- [Howard Garage](business_name) [2310 N Howard](address)... .[REG ent-6532](phone_number)
- [Hunter's Serv Sta](business_name) [Oxfrd av & Verree rd](address).[PIL grm-9949](phone_number)
- [Hunting Prk Garage](business_name) [1607 Huntng Prk av](address). [MIC hign-3041](phone_number)
- [Huntingdon Garage](business_name) [26 & Huntingdon](address) [RAD clf-9540](phone_number)
- [Ideal Garage](business_name) [1530 N 27](address) [STE vnsn-7778](phone_number)
- [Imperial Service Garage](business_name) [5945 Locust st](address). [GRA nite-6613](phone_number)
- [Indian Garage](business_name) [2843 W Clearfid](address) [RAD clf-5392](phone_number)
- [Indiana Garage](business_name) [3028 N 6](address).......[SAG amor-2426](phone_number)
- [Integrity Garage](business_name) [4130 Walnut st](address) [BAR ing-4163](phone_number)
- [Internat! Garage](business_name) [6026 Elmwd av](address). [SAR atga-9778](phone_number)
- [Irvington Garage](business_name) [273 S 59](address) ; [GRA nite-9861](phone_number)
- [Jack Condur’s Garage](business_name) [3070 Fkd av](address). [REG ent-7645](phone_number)
```

Image 12. Sample of training data in Markdown format used to train a CRF to identify entities, such as business names, addresses, and phone numbers in Yellow Pages business listings. See Appendix B for a larger view of this graphic.

Though a goal for the HITL initiative was to use crowdsourced data to train ML processes, generating enough training data to test the viability of ML processes before the implementation of the experiment and collection of crowdsourced data was a major challenge. While the CRF approach to entity recognition appeared to be promising, it was hard to determine if that success would fall short of acceptable thresholds for accuracy once the experiment workflows integrated a wider variety of content. Likewise, without more of an investment in creating training images for the object detection of page segments (or, quite possibly, greater expertise in object detection of textual layouts), it would be difficult to know if accuracy would also scale to make the approach worth including in the pipeline.

In the interest of generating relatively accurate output in the early parts of the pipeline to produce somewhat accurate results at the end, AVP team members explored other options for segmenting page images that relied more on built-in models and rule-based algorithms than custom-trained models. After some experimentation, AVP settled on a combination of object contour detection and OCR to segment advertisements from business groupings and telephone tips and to isolate individual business grouping headings and business listings. Altogether with the CRF process, the final pipeline was designed as follows. (Fuller descriptions of these processes can be found in the HITL initiative GitHub code repository and in Appendix F.[37]):

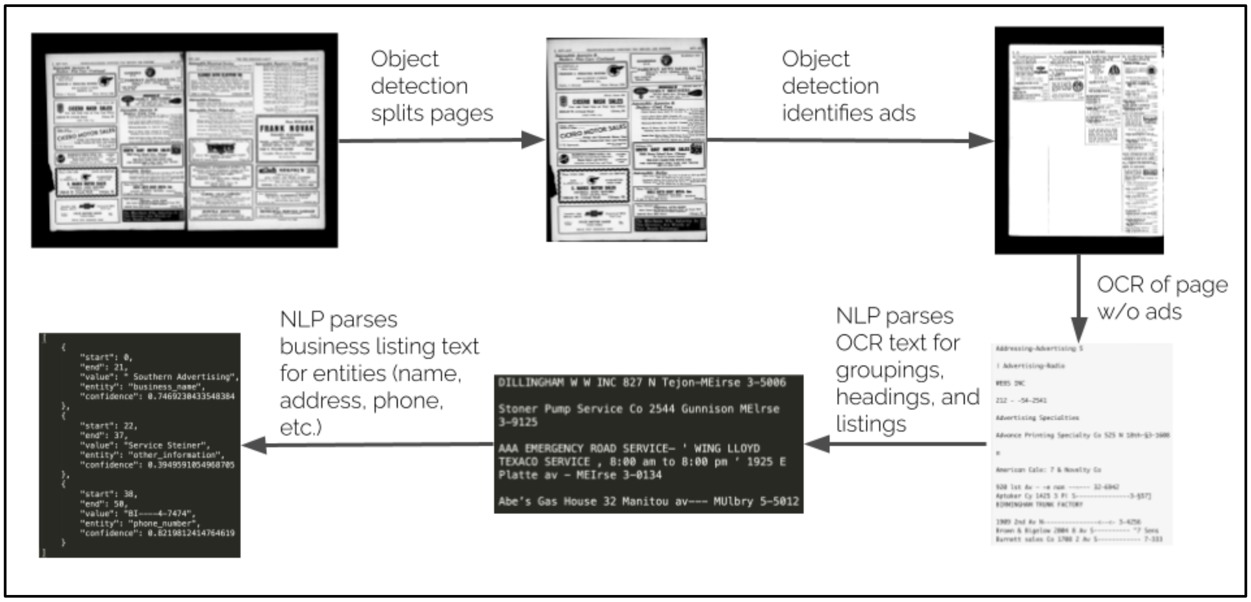
1) **Split two-up page images into individual pages** using contour detection with OpenCV,[38] an open-source computer vision programming library. Contour detection is an object-detection algorithm that looks for edges of objects to identify them.
2) **Find all advertisements in a page** using contour detection. (Advertisements in the Yellow Pages are typically bounded in boxes and within a certain size range relative to the page.)

[36] “NYPL-Spacetime City Directory Entry Parser,” Github, accessed July 1, 2021, https://github.com/nypl-spacetime/city-directory-entry-parser.

[37] “HITL.” Library of Congress, accessed November 24, 2021, https://github.com/LibraryOfCongress/hitl.

[38] OpenCV, accessed July 1, 2021, https://opencv.org/.

3) **Identify business types** by running Tesseract OCR[39], isolating larger-than-average OCR characters, filtering invalid text patterns. The business-type text is extracted by overlapping the OCRed "large" characters coordinates with the full page OCR coordinates
4) **Find business listings** using regular expressions to identify phone numbers. The full listing is built by reading the OCR text between the business type and a phone number, or the text between a phone number and a phone number.
5) **Identify business groupings** by aggregating the listing and type coordinates. Listings are assigned to business types based on horizontal overlap and vertical proximity.
6) **Identify entities in a business listing** using a conditional random fields (CRF) algorithm trained on human-generated examples.


Object detection splits pages
Object detection identifies ads
OCR of page w/o ads
NLP parses OCR text for groupings, headings, and listings
NLP parses business listing text for entities (name, address, phone, etc.)

```
Addressing-Advertising 5

| Advertising-Radio

WEBS INC

212 - -54-2541

Advertising Specialties

Advance Printing Specialty Co 525 N 18th-$3-1606

#

American Cale: 7 & Novelty Co

920 1st Av - -e nan ------- 32-6842
Aptoker Cy 1425 3 Pl S----------------3-$57]
BIRMINGHAM TRUNK FACTORY

1909 2nd Av N------------------<--<- 3-4256
Brown & Bigelow 2004 8 Av S----------- "7 Sens
Burrett sales Co 1708 2 Av S------------- 7-333
```

```
DILLINGHAM W W INC 827 N Tejon-MEirse 3-5006

Stoner Pump Service Co 2544 Gunnison MElrse
3-9125

AAA EMERGENCY ROAD SERVICE- ' WING LLOYD
TEXACO SERVICE , 8:00 am to 8:00 pm ' 1925 E
Platte av - MEIrse 3-0134

Abe's Gas House 32 Manitou av--- MUlbry 5-5012
```

```
[
    {
        "start": 0,
        "end": 21,
        "value": " Southern Advertising",
        "entity": "business_name",
        "confidence": 0.7469230433548384
    },
    {
        "start": 22,
        "end": 37,
        "value": "Service Steiner",
        "entity": "other_information",
        "confidence": 0.3949591054968705
    },
    {
        "start": 38,
        "end": 50,
        "value": "BI-----4-7474",
        "entity": "phone_number",
        "confidence": 0.8219812414764619
    }
]
```

Image 13. Illustration of the machine learning pipeline. See Appendix B below for a larger view of this graphic.

Crowdsourcing Pipeline Design

The design goals for the crowdsourcing pipeline were twofold: 1) design tasks that provide an enjoyable and safe experience for volunteers and 2) design tasks that generate the necessary data to train and validate accuracy of the ML processes.

In using contour detection, OCR, and OCR output parsing for the first parts of the machine learning pipeline, AVP only required training data for one ML process, the CRF natural language processing algorithm that would identify entities in business listings. Though the crowdsourcing path to get to this structured data for use in training would require several steps, the outputs from each could be used to validate earlier steps of the machine learning pipeline along the

[39] "Tesseract OCR," Github, accessed July 1, 2021, https://github.com/tesseract-ocr/tesseract

way, providing input to machine learning experts on how to adjust parameters or OCR parsing scripts to achieve greater accuracy.

In an interview about the By The People crowdsourcing initiative with Lauren Algee, Trevor Owens, and Carlyn Osborn (Digital Collection Specialist/Community Manager) of the Digital Content Management Section, Elaine Kamlley, Head of Product in the Design & Development Directorate, and Meghan Ferriter of LC Labs, AVP learned more about volunteer user engagement, as well as challenges with and goals for enhancing user experience on the crowdsourcing platform. While LC staff were able to share only anecdotal data on user experience from emails and forum messages, the conversation helped the AVP team members to better understand the motivations of volunteers and types of tasks some of them found engaging. Fortunately the tasks that would be necessary to support ML validation and training for the Yellow Pages met some of the recommendations for volunteer engagement, including providing a variety of types of tasks for users with different interests, and several degrees of difficulty requiring different levels of attention or ability.

With these considerations in mind, the team decided upon five crowdsourcing tasks:

1) **Identify segments (business groupings, advertisements, and telephone tips) on a page.** On being shown an individual page image, volunteers draw bounding boxes around and classify segments. This data will serve as ground truth to verify accuracy of ML processes detecting these segments from contour detection and OCR parsing.
2) **Identify business types and business listings.** On being shown a business groupings, volunteers draw bounding boxes around and classify business types and business listings. This data will serve as ground truth to verify accuracy of ML processes detecting business types and business listings from OCR parsing.
3) **Transcribe business types.** On being shown a business type, transcribe the text of the type and any “see” references to other types. This data can be used to verify accuracy of ML recognition of business-type text from OCR and OCR parsing.
4) **Identify business listing entities.** On being shown a business listing, volunteers draw bounding boxes around and classify the entities: business name, address, phone number, other information, “see advertisement” references, and graphics. This data will be used along with the data in the next task to train and test the CRF method for identifying entities.
5) **Transcribe business listing entities.** On being shown entities from a business listing, volunteers transcribe the text of the entity. This data will be used along with the data in the previous task to train and test the CRF process for identifying entities.

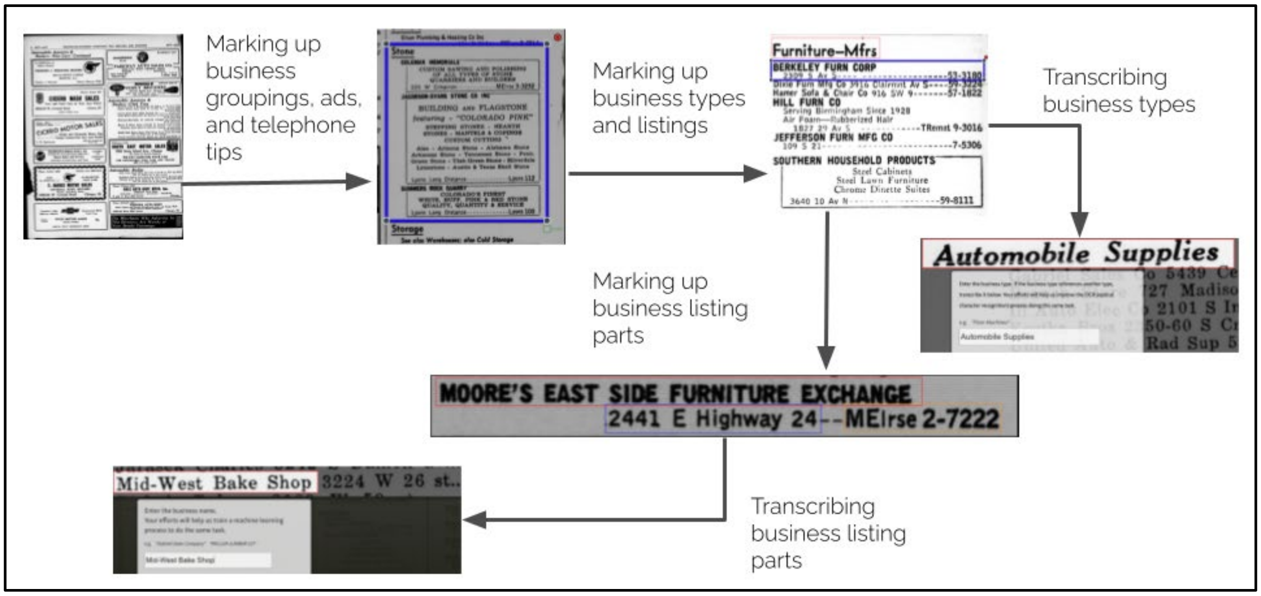

Marking up business groupings, ads, and telephone tips
Marking up business types and listings
Furniture–Mfrs
Transcribing business types
Automobile Supplies
Marking up business listing parts
MOORE'S EAST SIDE FURNITURE EXCHANGE
2441 E Highway 24--MElrse 2-7222
Mid-West Bake Shop 3224 W 26 st..
Transcribing business listing parts

Image 14. Illustration of the crowdsourcing pipeline. See Appendix B for a larger view of this graphic

The team considered the functionality of these tasks in selecting an appropriate open-source crowdsourcing platform to use for the prototype and also looked at the list of risks and mitigation strategies created during the collection selection stage (and updated after stakeholder conversations) for additional requirements to ensure an engaging and ethical crowdsourcing experience. For example, the core team was already aware that as historical documents, the Yellow Pages contained racially and culturally insensitive graphics, language, outdated terminology, and other potentially triggering content. To mitigate the risks of exposing volunteers to this content, strategies included adding content warnings in the footer or navigation sidebar of the crowdsourcing site or including an option to skip and/or report offensive content when encountered in the application. After a review of several platforms, the team chose the open-source Scribe platform,[40] originally developed by NYPL Labs and Zooniverse, for the prototype because of its support of marking and transcription tasks and customizable crowdsourcing workflow design, tutorials, and static web pages.

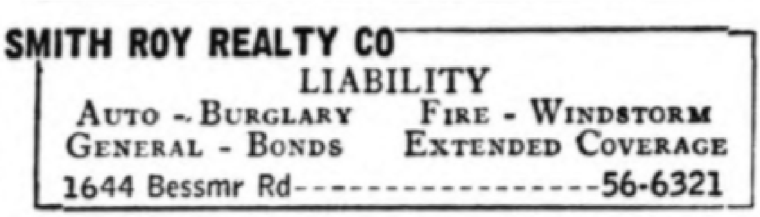
Recommendations

Based on experience from this experiment, the team recommends early investment in exploration and ideation, as well as continued discovery of user needs to set up technical design for a successful implementation.

- **Prototype tasks to uncover complexities early on in the design process.**
  During the implementation stage, the core team had more opportunity to get familiar with

[40] "Scribe Project," Github, accessed July 1, 2021, https://scribeproject.github.io/.

the wide variety of Yellow Pages content in the U.S. Telephone Directory Collection, which introduced new complexities in the content not previously noticed in the Design phase. For instance, while many business listings contain a simple set of information — one business name, one address, one phone number, some additional information — on closer inspection, some listings may include multiple addresses and phone numbers, or the listing may be for a product, with sublistings for authorized dealers.



SMITH ROY REALTY CO

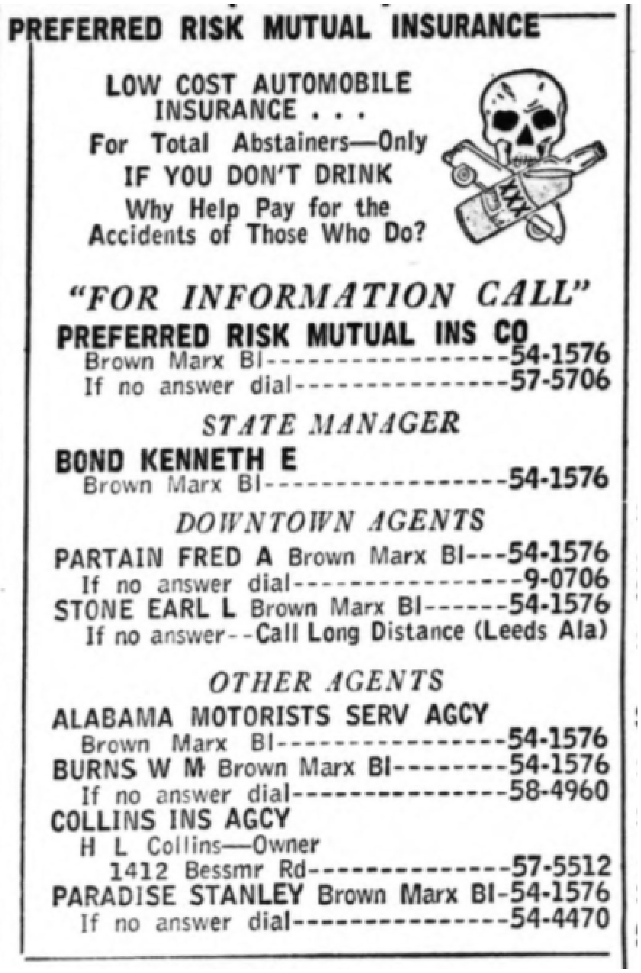
LIABILITY

AUTO – BURGLARY   FIRE - WINDSTORM

GENERAL - BONDS   EXTENDED COVERAGE

1644 Bessmr Rd----------------56-6321

Image 15. A “simple” business listing.



PREFERRED RISK MUTUAL INSURANCE

LOW COST AUTOMOBILE
INSURANCE . . .
For Total Abstainers—Only
IF YOU DON'T DRINK
Why Help Pay for the
Accidents of Those Who Do?

“FOR INFORMATION CALL”

PREFERRED RISK MUTUAL INS CO
Brown Marx Bl----------------54-1576
If no answer dial---------------57-5706

STATE MANAGER

BOND KENNETH E
Brown Marx Bl----------------54-1576

DOWNTOWN AGENTS

PARTAIN FRED A Brown Marx Bl---54-1576
If no answer dial---------------9-0706
STONE EARL L Brown Marx Bl------54-1576
If no answer--Call Long Distance (Leeds Ala)

OTHER AGENTS

ALABAMA MOTORISTS SERV AGCY
Brown Marx Bl----------------54-1576
BURNS W M Brown Marx Bl--------54-1576
If no answer dial---------------58-4960
COLLINS INS AGCY
H L Collins—Owner
1412 Bessmr Rd--------------57-5512
PARADISE STANLEY Brown Marx Bl-54-1576
If no answer dial---------------54-4470

Image 16. A complex business listing, including sublistings and multiple phone numbers.

This discovery not only complicated the data model created at the outset of the design stage, but it also posed challenges for the design of crowdsourcing tasks and ML processes. Adding further complexity to the crowdsourcing task could increase the difficulty level and thus user frustration, and it could also reduce the accuracy or consistency of the crowdsourced data with correct answers being much more open to interpretation. Complexities or variations in content can also make it difficult to control accuracy in the ML processes as well. In the case of the Yellow Pages, additional rules may need to be added to the OCR parsing process, so that the business listings and sublistings are identified and grouped in a way that aligns with the defined data model.

Early prototyping of tasks and testing of a wide variety of collection content by a wide range of users can help uncover these complexities and identify additional risks during the data modeling and task-design stage. Whether these discoveries help improve the user experience and resulting data or inform decisions on moving forward or not, this early investment will likely save time and effort that can be spent later on implementation.

- **Invest in tools that support early prototyping.**
  Uncovering complexities, generating training data for assessing potential machine-learning processes, and testing engagement of crowdsourcing tasks require that team members have easy access to tools that allow them to explore collection content, user experience, and training of ML processes. While the AVP team was able to generate training data and experiment with marking tasks with the VoTT tool, this tool did require installation on a server and download of a client to local computers, requirements that may be out of reach for some organizations with strict limitations on local software installation.

  Access to web-based tools for a wide range of staff users to test crowdsourcing tasks and quickly generate training data for early ML assessment is imperative to a Design phase that aims to maximize usefulness of data and minimize risk to users. While many open-source tools exist for training data annotation, access to these tools will usually require IT assistance in local or cloud server installation. A collaborative effort in the GLAM community towards shared rapid prototyping tools could help forward the state of human-in-the-loop approaches (or even crowdsourcing initiatives without a ML component), benefitting individual organizations in early experimentation and definition and the community as a whole by providing a platform for learning from others' experience.

- **Begin or plan ground-truth testing of machine-learning accuracy.**
  The limited timeframe of the experiment prevented the team from fully integrating ground-truth testing into the HITL initiative design and implementation to assess accuracy of ML options. However, considering how to incorporate these feedback mechanisms into the pipeline will help to provide a foundation for continuous iteration and improvement cycles during the implementation phase. During early experimentation stages with potential ML methods, creating ground-truth data can be a time-intensive task that may not yet be worthwhile when a subjective assessment or spot-checking of results can often reveal whether a method shows promise for further development. However, at this stage, teams should begin thinking about what measures they could

use to test accuracy, how they will create ground-truth data (as this will inform the design of crowdsourcing tasks), and how they will even define quality for their use cases.

For the Yellow Pages, the team hypothesized that they could test ML outputs on a page-by-page basis, testing segment-identification processes (machine learning steps 2 and 4) by comparing image coordinates between machine-generated data and human-generated ground truth using Intersection over Union (IoU), an evaluation metric for object detection that looks at the overlap of two boxes. Accuracy would need to be fairly high at this stage in the pipeline, as the outputs from these steps are passed as inputs to the next. Testing for business type and business listing could use similar IoU measures, and testing for OCR quality and entity recognition would need further consideration.

- **Conduct early-stage interviews or user tests with potential crowdsourcing volunteers.**
  While the team conducted interviews with LC collection and crowdsourcing experts, time and budget for the initiative did not allow for early testing with potential volunteers. In the early prototyping stage, task design would benefit from conversations with potential volunteers to gauge volunteer interest in engagement with specific collection content, as well as any concerns they may have around proposed collection content or ML methods.

- **Design a community outreach plan to begin developing support and engagement from potential volunteers.**
  Sustained engagement of crowdsourcing volunteers will be key to the success of any human-in-the-loop endeavor. Early planning and outreach around volunteer community development will help build the foundation for sustained engagement, and will provide an opportunity to plan and organize gathering volunteer feedback around early-design prototypes and tasks.

  An outreach plan for an initiative should include identification of long-term volunteer engagement goals and strategies, and define necessary staffing to support an active volunteer community. Other potential considerations for a plan might include communication channels, a timeline for interviews or user testing, and participant recruitment plans or resources. Future experiments at LC Labs might provide excellent opportunities to engage community managers in modeling early-stage volunteer community development and engagement tools.

STAGE 3: IMPLEMENTATION

Implementation

Objectives

Defined objectives for the implementation stage included:

- Build crowdsourcing application
- Build ML pipeline
- Track data flows and outputs
- Test and refine ML processes

.Goals

The goals for the Implementation stage focused primarily on ensuring users were supplied with broader contextual information about the collection, as well as providing context around machine learning goals.

Engaging	Ethical	Useful
● Volunteers have the opportunity to learn about the collection and related materials ● Volunteers are able to understand and track the progress of their contributions ● Volunteers are able to choose tasks and switch between different tasks	● Data provenance and accuracy of machine learning generated data is tracked ● Use of machine learning is clear and understandable to volunteers	● Accuracy of machine learning-generated data can be improved through ground-truth testing

.Challenges

Some of the main challenges during implementation were around the customization of crowdsourcing and ML technologies to meet user experience goals and generate accurate data.

To address these challenges we asked:

How do we…

- Achieve consistent results from varying collection content?
- Retrofit existing crowdsourcing platforms for new initiative?
- Refine ML processes to improve accuracy as more data is generated from crowdsourcing tasks?

- Communicate ML processes and interactions to volunteers in ways that are clear and digestible?

Humans

In order to achieve the goals and address the challenges of the Implementation phase:

We involve…

- **Machine learning experts**, who understand how to implement and refine ML processes
- **Software developers**, who understand how to build crowdsourcing platforms and architect data flows
- **Digital collection specialists**, who understand how to integrate digital collection content into crowdsourcing and ML pipelines
- **UX designers**, who understand how to present tasks and information to users in a clear and accessible way
- **Community managers**, who understand the wide range of volunteer needs for a crowdsourcing platform
- **Volunteers**, who contribute data to train and test ML processes and can offer feedback on crowdsourcing tasks

Feedback Mechanisms

In order to achieve the goals and address the challenges of this phase:

We learn from…

- **User testing** with volunteers to help improve the user experience of the crowdsourcing platform
- **Ground truth accuracy testing** to help improve ML processes throughout the course of the initiative
- **Workflow database** to track provenance and accuracy of all tasks, processes, and data for input on platform/pipeline improvement

Process

Crowdsourcing Platform Implementation

As mentioned above, the team selected the open-source Scribe platform for implementing the five crowdsourcing tasks as a prototype for the HITL experiment. While the team knew that using an existing platform would limit the ability to meet all user needs identified for the experiment, customizing and implementing an existing product would be more feasible within the limited time frame of the experiment than building from scratch.

Though Scribe has been used successfully for a number of crowdsourcing goals, including LC Labs's Beyond Words experiment, the team encountered a number of challenges in implementation. The first was in adapting an old codebase for use with current web frameworks. Scribe — initially released in 2015 — has not been actively maintained, and few recent forks exist that provide working code. Fortunately, AVP discovered a working Scribe instance at the Utrecht University Digital Humanities Lab[41] and inquired about borrowing their fork of the codebase.[42] Through their generosity, AVP team members were able to install and customize a working instance of Scribe without having to update the code. However, while this codebase was easier to get up and running, the Lab had made many customizations to the core application to accommodate their goals, so AVP developers needed to revert or change some of this code to meet the needs of the Yellow Pages experiment.

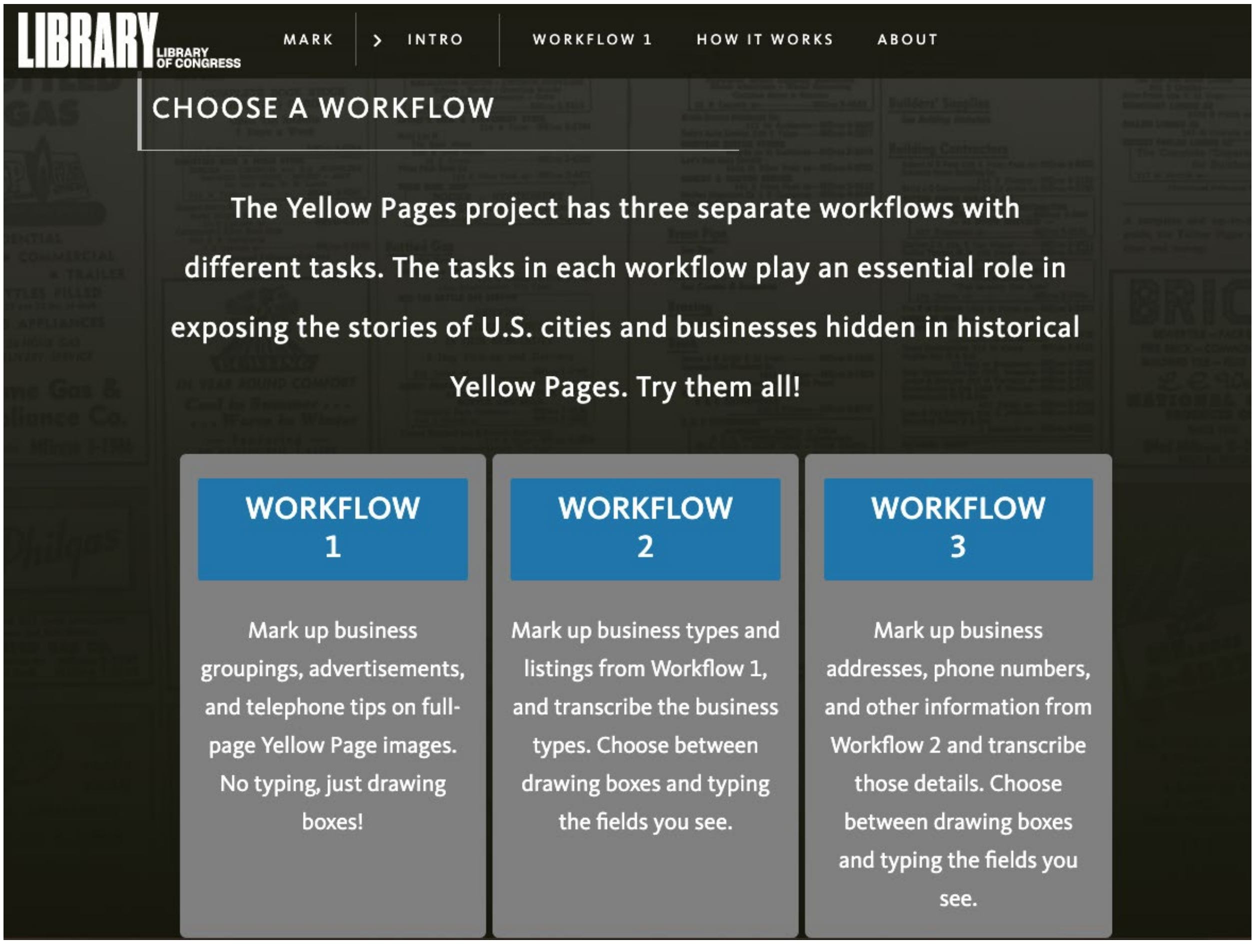
A much greater challenge was adapting Scribe for use with the five crowdsourcing tasks. The Scribe data model is flexible in how it allows for the design of multi-stage pipelines for data enhancement. Crowdsourcing workflows can be designed as a series of "mark," "transcribe," or "verify" tasks, each allowing for a variety of options and complexity. Crowdsourcing users can choose different types of tasks at any point in the pipeline, depending on their preference. As images are marked, transcribed, or verified, the crowdsourced data (such as a segment of an image or a text transcription) can be automatically piped into another task for further enhancement. This model offers many options for generating structured data from page images, but not without some limitations. Data generated from a "mark" task can only be piped into a "transcribe" or "verify" task, not another "mark" task. In the case of the Yellow Pages tasks, pages needed to be "marked" into segments, then each segment "marked" into business types and listings, which was not possible within the constraints of the Scribe model.

To get around this limitation without investing significant time in customizing the application, the AVP team decided to mimic a single platform by distributing the five tasks across three instances of Scribe, one for each "mark" task in the pipeline, and presenting them as separate "workflows." Piping data between instances required extracting crowdsourced data from one workflow instance and then reshaping and uploading it as new inputs for the next workflow instance. Fortunately, Scribe allows URLs for referencing image content to be crowdsourced, so AVP team members could leverage LC's IIIF image server[43] to provide the image inputs, rather than downloading, cropping, and moving images between instances. Image coordinates from the output of one workflow instance could be used to generate the IIIF image segments needed for the next. (Customized Scribe code and instructions for deploying Scribe as three instances can be found in Appendix F.)

[41] "CEMROL," Utrecht University Digital Humanities Lab, accessed July 1, 2021, https://cemrol.hum.uu.nl/#/.

[42] "Scribe (Utrecht University Digital Humanities Lab)," Github, accessed July 1, 2021, https://github.com/UUDigitalHumanitieslab/scribeAPI.

[43] "Image Services," Library of Congress, accessed July 1, 2021, https://www.loc.gov/apis/micro-services/image-services/.



CHOOSE A WORKFLOW

The Yellow Pages project has three separate workflows with different tasks. The tasks in each workflow play an essential role in exposing the stories of U.S. cities and businesses hidden in historical Yellow Pages. Try them all!

WORKFLOW 1

Mark up business groupings, advertisements, and telephone tips on full-page Yellow Page images. No typing, just drawing boxes!

WORKFLOW 2

Mark up business types and listings from Workflow 1, and transcribe the business types. Choose between drawing boxes and typing the fields you see.

WORKFLOW 3

Mark up business addresses, phone numbers, and other information from Workflow 2 and transcribe those details. Choose between drawing boxes and typing the fields you see.

Image 17. Three "workflows" across three instances of Scribe.

This workaround allowed the team to successfully deploy all five crowdsourcing tasks (screen captures of the deployed prototype can be viewed in Appendix P), but it limited the team's ability to present an engaging user experience.[44] This challenge is described in more detail later in this section, in "Crowdsourcing Interface Design and Usability."

Data Flow Tracking

Even though Scribe provides a MongoDB database[45] for storing the inputs and crowdsourced data generated for the crowdsourcing pipeline, an external datastore was necessary for tracking ML and crowdsourcing processes, including training data and accuracy scores, digital collection inputs for processes, and data outputs of processes, including image coordinates, OCR data, and entity values. Storing this data in a relational database allows for many different types of queries or structured exports of the ML outputs. It also helps to support ethical goals around transparency into ML processes, as accuracy scores can be stored for ground-truth testing, confidence scores for predictions can be tracked, when provided by a ML tool, and training or

[44] "Appendix P – Crowdsourcing Prototype Screen Captures." Library of Congress. Accessed November 24, 2021. https://github.com/LibraryOfCongress/hitl/tree/main/hitl-recommendations-report-appendices/Appendix%20P_%20Crowdsourcing%20Prototype%20Screen%20Captures.

[45] MongoDB, accessed July 1, 2021, https://www.mongodb.com/.

ground-truth data for any process can be followed all the way back to its collection source and originating crowdsourcing task. While the timeline for the HITL collaboration did not allow for experimentation with filtering ML outputs by accuracy scores or confidence levels, the database model supports such methods of controlling downstream data exports by data source or accuracy levels.

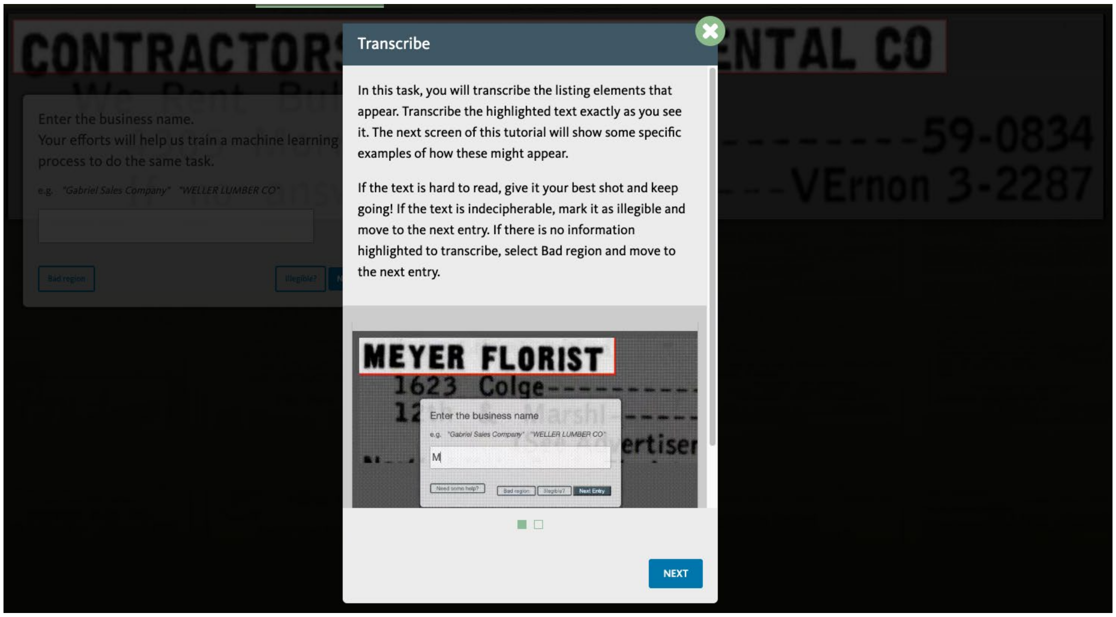
While the ML and crowdsourcing processes needed customizations based on collection content and desired structured data outputs, AVP team members designed this workflow database to be agnostic of ML and crowdsourcing systems so it could be reused for any cultural heritage human-in-the-loop endeavor. Scripts to manage data flows between the ML and crowdsourcing processes will need to be specific to those systems, but generalizable queries and programming frameworks could be built around functionalities common to all approaches, such as ground-truth testing or structured data output.

In building the workflow database, the core team first designed a high-level conceptual model encompassing the major entities in the domain of a human-in-the-loop initiative, including data sources, ML processes and versions, crowdsourcing tasks, and annotations (from machine learning or crowdsourcing). AVP team members then developed a data dictionary and then implemented the schema in a PostgreSQL database. Scripts for passing data between the database and ML and crowdsourcing processes were written to create a somewhat manual pipeline for the experiment, but these data flows could be further automated for a working production environment. (The workflow database diagram, data dictionary, initialization file, and backup file of all data generated for this HITL initiative are available in Appendices C-F.)

Crowdsourcing Interface Design and Usability

In considering the user experience of the Scribe crowdsourcing platform, AVP knew that there would be significant limitations to what level of customizations we could make to the active Scribe instance within the timeframe of the HITL initiative. Since the deployed prototype (Appendix P) was designed as a proof of concept and was not intended to be released in support of a public crowdsourcing effort, the team moved forward with less-than-ideal user flows and navigational structures in the interface design in order to provide a working platform for staff to scale up creation of ground-truth data to help train the ML processes.

In terms of design goals for the crowdsourcing platform, the team focused on how the interface might communicate and support transparency for volunteers around the ways in which the defined crowdsourcing tasks would interact with ML processes. The interface also needed to provide guidance to volunteers on how each task was intended to be approached and performed. Informational pages were created to provide experiment and collection context, including context around the defined crowdsourcing tasks and ML pipelines. In-app tutorials were also created to help prototype users understand how to approach each task as they were working to generate ground-truth data.



Transcribe

In this task, you will transcribe the listing elements that appear. Transcribe the highlighted text exactly as you see it. The next screen of this tutorial will show some specific examples of how these might appear.

If the text is hard to read, give it your best shot and keep going! If the text is indecipherable, mark it as illegible and move to the next entry. If there is no information highlighted to transcribe, select Bad region and move to the next entry.

Enter the business name.
Your efforts will help us train a machine learning process to do the same task.
e.g. "Gabriel Sales Company" "WELLER LUMBER CO"

Image 18: Screenshot of a Transcribe tutorial for workflow 3 open in the crowdsourcing prototype interface. See Appendix B below for a larger view of this graphic.

Although the prototype wasn't intended to be released broadly, the core team was still interested in understanding how users would experience the tasks supported by it. Specifically, would crowdsourcing volunteers be able to successfully accomplish the defined tasks? Would they enjoy the tasks, or find them overly repetitive? Would they be made adequately aware of triggering content that might surface during task completion?

To gather information around these questions, AVP asked Library core team staff who were using the prototype to generate ground-truth data for ML processes to respond to a survey about their experience completing Yellow Pages tasks. The surveys asked these users to indicate, generally through Likert scale[46] questions, how well they understood how to complete the tasks for each workflow, what they enjoyed most and least about the tasks, if they understood how the individual tasks related to the other experiment workflows, and other questions related to their experience working in the prototype.

Survey responses, found in Appendix K, indicated that staff users generally understood what was required of them to complete the tasks. And they mostly found the defined tasks enjoyable to complete. However, many did not feel they clearly understood how their work on a given task related to other experiment workflows, or how the defined tasks would interact with ML processes. Also, while tutorials were available to users, many did not read them, and expected clearer instructions to be embedded directly within the interface. These responses indicate that the overall navigation and design of the Yellow Pages Scribe instance did not lend itself to a particularly user-friendly or transparent experience for crowdsourcing volunteers. Were the prototype to be launched to a broader audience, more consideration and time would need to be spent addressing usability issues and concerns.

[46] "Likert scale," Wikipedia, accessed July 1, 2021, https://en.wikipedia.org/wiki/Likert_scale.

User Testing

Aside from gauging usability of the platform itself, the core team wanted to gather volunteer perspectives on the use of ML in libraries, and at the Library of Congress more specifically. Given that HITL is the first LC Labs experiment to explicitly focus on combining ML and volunteer-generated data, it was critical to explore participant expectations, perspectives, and understanding of the processes involved.

In order to evaluate potential user perspectives on interacting with ML processes, the core team wanted to conduct user testing focused on evaluating crowdsourcing volunteers' understanding of the defined interactions between HITL crowdsourcing tasks and the related ML processes those tasks would inform.

The core team first needed to identify and recruit five to ten power users of existing cultural heritage crowdsourcing platforms to participate in HITL user testing. Participant interest forms[47] were sent to identified crowdsourcing program leads and community managers to help surface potential participants. In early April 2021, HITL core team members reviewed responses to the interest form and reached out to potential recruits to schedule user tests.

Because there are restrictions around the types of incentives LC is able to offer volunteers, user testing participants were offered the opportunity to be named in report acknowledgements, and an opportunity to follow up and meet with HITL core team members to learn more about their work. In total, the HITLcore team recruited eight crowdsourcing volunteers to participate in facilitated, individual meetings organized around a structured agenda.

Given that the test goals centered around gathering perspectives on machine learning in libraries rather than on platform usability, AVP designed a test to:

1. identify user motivations for engaging in crowdsourcing activities
2. introduce participants to the HITL Yellow Pages experiment
3. introduce participants to the concept of "machine learning" and how HITL tasks were intended to interact with ML processes
4. gather user impressions on ML generally, and on the proposed crowdsourcing interactions with ML specific to the HITL initiative

The full user test plan and discussion guide are shared in Appendix I.[48]

The user tests kicked off with short interviews aimed at surfacing volunteer motivations for contributing to existing crowdsourcing efforts in cultural heritage. Then, in order to introduce users to the concepts and goals of the HITL experiment, AVP created three separate mockups of the crowdsourcing prototype in Miro[49], which served as the platform for the task-based portion of the user tests. To complete the tasks, participants were shown the mockups and

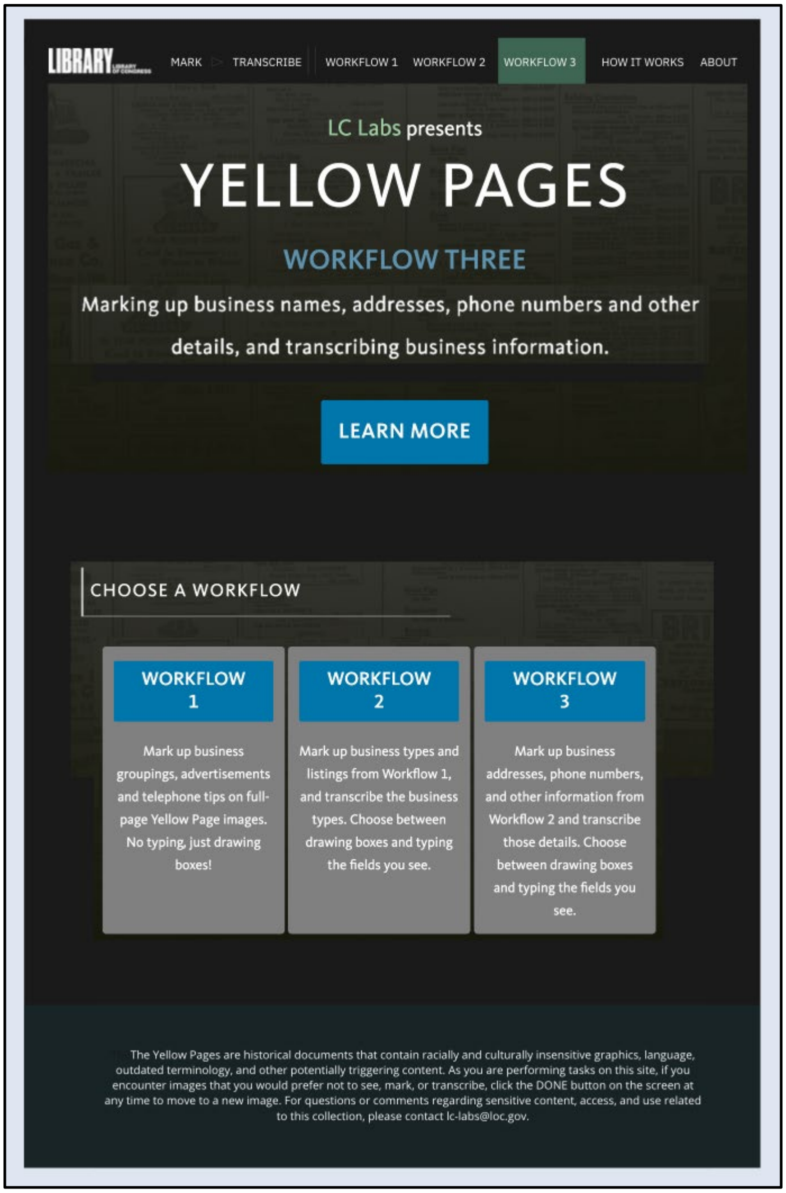
[47] "Help Design a New LC Labs Experiment!," Library of Congress, access July 1, 2021, https://www.research.net/r/9WCYFWJ.

[48] "Appendix I – Crowdsourcing Prototype User Testing Plan & Discussion Guide." Library of Congress, accessed November 24, 2021, https://github.com/LibraryOfCongress/hitl/blob/main/hitl-recommendations-report-appendices/Appendix%20I_%20Crowdsourcing%20Prototype%20User%20Testing%20Plan%20_%20Discussion%20Guide.pdf.

[49] Miro, https://miro.com.

asked to perform activities or respond to specific questions related to the information presented in each task.

For the first task, the Miro board displayed a mockup of the prototype landing page for HITL workflow 3.



LC Labs presents

YELLOW PAGES

WORKFLOW THREE

Marking up business names, addresses, phone numbers and other details, and transcribing business information.

CHOOSE A WORKFLOW

WORKFLOW 1

Mark up business groupings, advertisements and telephone tips on full-page Yellow Page images. No typing, just drawing boxes!

WORKFLOW 2

Mark up business types and listings from Workflow 1, and transcribe the business types. Choose between drawing boxes and typing the fields you see.

WORKFLOW 3

Mark up business addresses, phone numbers, and other information from Workflow 2 and transcribe those details. Choose between drawing boxes and typing the fields you see.

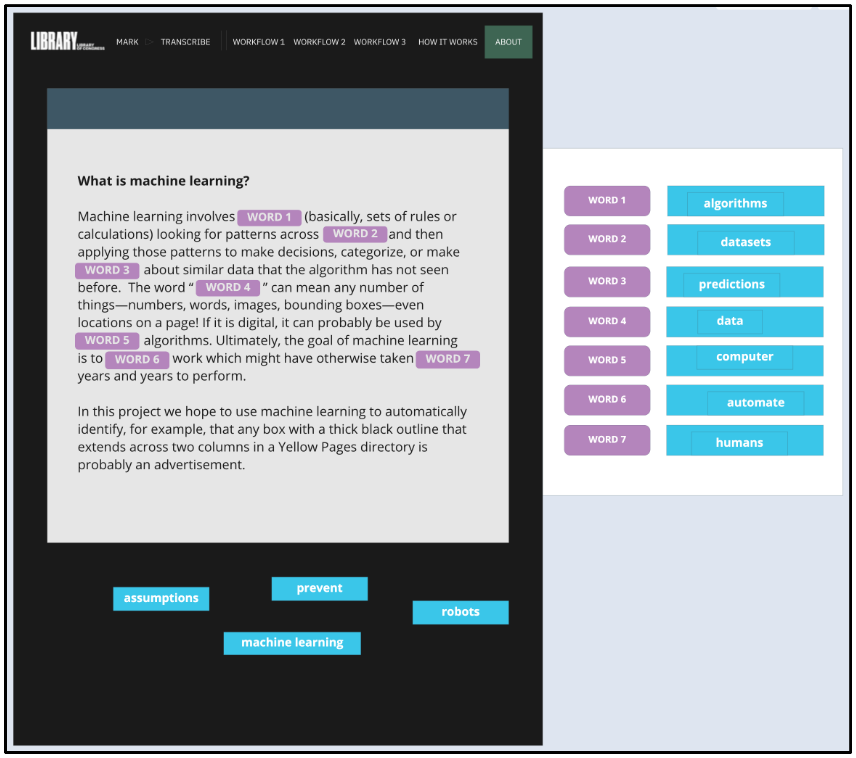
The Yellow Pages are historical documents that contain racially and culturally insensitive graphics, language, outdated terminology, and other potentially triggering content. As you are performing tasks on this site, if you encounter images that you would prefer not to see, mark, or transcribe, click the DONE button on the screen at any time to move to a new image. For questions or comments regarding sensitive content, access, and use related to this collection, please contact lc-labs@loc.gov.

Image 19. Screenshot of prototype mockup in Miro used to facilitate user test task 1.

Users were given an introduction to the Yellow Pages collection and an overview of the designed landing page for the crowdsourcing prototype. Once users were introduced to the overall experiment design, and had responded to some high-level questions about the mockup, the facilitator moved them to the second task.

To introduce participants to machine learning and how ML and crowdsourcing would interact in the Yellow Pages experiment, the team excerpted language written for the informational pages on the crowdsourcing prototype. User testing tasks 2 and 3 applied a technique called cloze testing, where participants are shown a portion of the excerpted text with keywords removed. The participants are then asked to replace the missing words in a way that completes and constructs meaning from the text.

Task 2 of the user test asked participants to complete a cloze test based on a paragraph from the crowdsourcing interface called, "What is Machine Learning."



What is machine learning?

Machine learning involves WORD 1 (basically, sets of rules or calculations) looking for patterns across WORD 2 and then applying those patterns to make decisions, categorize, or make WORD 3 about similar data that the algorithm has not seen before. The word " WORD 4 " can mean any number of things—numbers, words, images, bounding boxes—even locations on a page! If it is digital, it can probably be used by WORD 5 algorithms. Ultimately, the goal of machine learning is to WORD 6 work which might have otherwise taken WORD 7 years and years to perform.

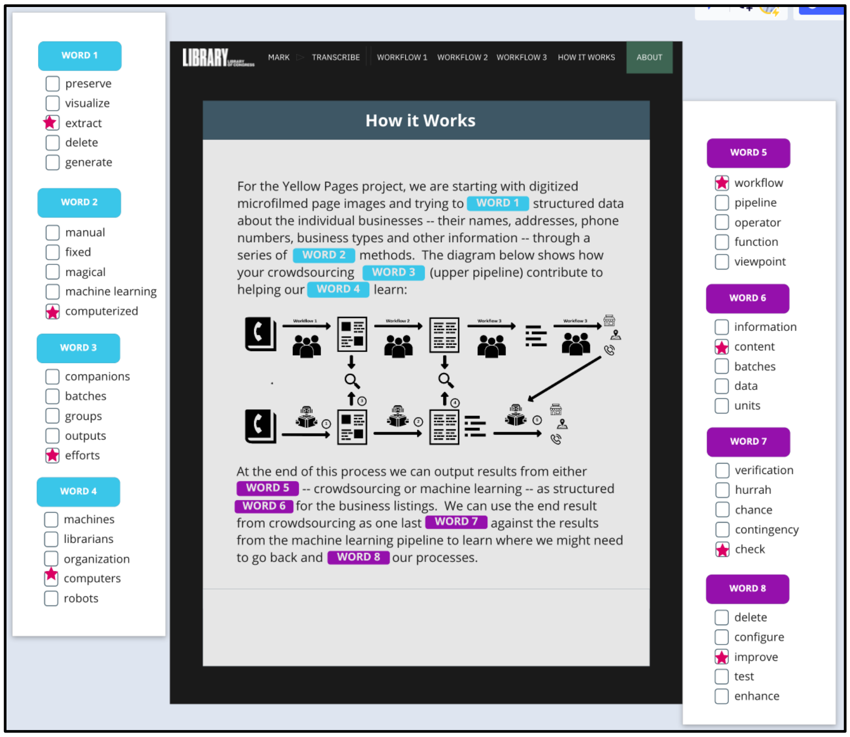
In this project we hope to use machine learning to automatically identify, for example, that any box with a thick black outline that extends across two columns in a Yellow Pages directory is probably an advertisement.

WORD 1	algorithms
WORD 2	datasets
WORD 3	predictions
WORD 4	data
WORD 5	computer
WORD 6	automate
WORD 7	humans

Image 20. Screenshot of a completed cloze test example from user test task 2. See Appendix B, Image 14 for a larger view of this graphic.

After completing the cloze test, participants were asked to explain their understanding of ML based on the completed paragraph, along with what benefits or concerns they might have about libraries, and the Library of Congress more specifically, using ML approaches.

From there, participants engaged in a second cloze test in task 3, in which they were asked to complete a paragraph describing at a very high level the intended interactions between crowdsourcing tasks and ML processes in the Yellow Pages HITL experiment.



WORD 1

- □ preserve
- □ visualize
- ☑ extract
- □ delete
- □ generate

WORD 2

- □ manual
- □ fixed
- □ magical
- □ machine learning
- ☑ computerized

WORD 3

- □ companions
- □ batches
- □ groups
- □ outputs
- ☑ efforts

WORD 4

- □ machines
- □ librarians
- □ organization
- ☑ computers
- □ robots

How it Works

For the Yellow Pages project, we are starting with digitized microfilmed page images and trying to WORD 1 structured data about the individual businesses -- their names, addresses, phone numbers, business types and other information -- through a series of WORD 2 methods. The diagram below shows how your crowdsourcing WORD 3 (upper pipeline) contribute to helping our WORD 4 learn:

At the end of this process we can output results from either WORD 5 -- crowdsourcing or machine learning -- as structured WORD 6 for the business listings. We can use the end result from crowdsourcing as one last WORD 7 against the results from the machine learning pipeline to learn where we might need to go back and WORD 8 our processes.

WORD 5

- ☑ workflow
- □ pipeline
- □ operator
- □ function
- □ viewpoint

WORD 6

- □ information
- ☑ content
- □ batches
- □ data
- □ units

WORD 7

- □ verification
- □ hurrah
- □ chance
- □ contingency
- ☑ check

WORD 8

- □ delete
- □ configure
- ☑ improve
- □ test
- □ enhance

Image 21. Screenshot of a completed cloze test example from user test task 3. See Appendix B below for a larger view of this graphic.

Once participants were able to articulate their understanding of the proposed interactions, they were asked a series of questions designed to surface their feelings about volunteering their time to crowdsourcing endeavors that interact with ML processes.

Full test results and raw data from the interviews are compiled and anonymized in Appendix J; a summary of high-level takeaways are summarized below.[50]

During the initial interviews, users expressed a wide range of motivations for participating in crowdsourcing endeavors:

- a classroom teacher expressed a drive to bring history alive for his students
- a librarian, furloughed during the pandemic, found volunteering for crowdsourcing initiatives an effective way to continue contributing to the cultural heritage sector while keeping her digital skills sharp
- an octogenarian stated that participating in crowdsourcing tasks improves her short-term memory, and has provided an opportunity to keep busy during the solitary months of the pandemic
- a history buff expressed developing a deep connection to the long-dead author of a journal collection she transcribes

[50] "Appendix J_ Crowdsourcing Prototype User Test Data - NO PII.xlsx." Library of Congress, accessed November 24, 2021, https://github.com/LibraryOfCongress/hitl/blob/main/hitl-recommendations-report-appendices/Appendix%20J_%20Crowdsourcing%20Prototype%20User%20Test%20Data%20-%20NO%20PII.xlsx

Each participant named specific and personal motivations for volunteering their time to crowdsourcing efforts, but most also expressed finding a broader sense of personal satisfaction and reward in giving their time to something that feels bigger than themselves.

During the tasks, five out of eight participants expressed that they felt ML offered libraries the potential to significantly improve or broaden access to library collections and materials at scale. While recognizing the potential benefits, test participants did surface concerns related to the potential for ML to introduce algorithmic biases or misidentification errors to library collections data.

In terms of participant reactions to interacting with ML processes in HITL crowdsourcing initiatives, six out of eight volunteers offered overall positive responses, indicating that it is worthwhile for LC to combine ML with volunteer contributions, and that they would be willing to volunteer for human-in-the-loop inititaitves. Four participants explicitly stated that knowing LC was taking this approach would positively impact their motivations for volunteering. Even volunteers who explicitly expressed distrust of ML and AI approaches indicated that knowing ML was involved in a crowdsourcing program would not deter them from volunteering, as long as the user experience of the platform was pleasing, and they found the tasks or content engaging.

Recommendations

Successful implementation of human-in-the-loop endeavors will require significant, ongoing investment in resources, as well as early and iterative testing of technical and crowdsourcing processes.

- **Commit to continuous staffing of key machine-learning and community experts throughout the initiative.**
  Creating successful human-in-the-loop pipelines will require significant, ongoing investment in operational resources. In order to implement and iterate on approaches to ML pipelines, human-in-the-loop initiatives need to be staffed with ML experts. ML processes cannot be realistically defined and created in a single pass, but will require ongoing refinement and iteration of processes throughout.

  Likewise, results from HITL user testing indicate that while it is important for libraries to clearly convey the intended interactions between human-generated data and ML processes to crowdsource users, volunteers are most likely to contribute if they have a good user experience on the platform, if they feel connected to the collection materials, if they feel connected to a broader community of volunteers, and if they feel their efforts help contribute to the greater good.

  For human-in-the-loop initiatives to succeed, sustaining volunteer engagement will be critical. Dedicated community management will be key to inviting and supporting the level of volunteer engagement necessary to support crowdsourcing pipelines.
- **Conduct ground-truth testing to improve accuracy and monitor risk in machine-learning processes.**

As the data generating activities progress, crowdsourcing output will grow, increasing the amount of ground-truth data that can be used to measure accuracy of ML processes. Given the significant amount of time needed to create ground-truth data, early assessment of ML processes will likely have been minimal or insufficient to predict a consistent level of accuracy, so continued ground-truth testing will be imperative to successful fine-tuning and improvement of processes.

Sample data in Appendix O shows just how rough initial ML outputs can be at the outset of an initiative with minimal testing to provide feedback on how a machine learning expert should adjust process parameters or trained models. As more varied collection content is introduced, ground-truth testing may reveal patterns in accuracy reflecting slight differences in layouts, fonts, or other differences in content that are less obvious to humans but detrimental to successful ML processing. Ground-truth testing to inform further refinement of processes or even development of separate streams of ML pipelines for different categories of content.

- **Iterate and improve upon crowdsourcing user experience through continued user testing.**
  In the same way that ML processes will need to be consistently reviewed and refined over the course of an experiment or initiative crowdsourcing user experience and engagement will also require ongoing testing and iteration. Crowdsourcing platforms should be carefully planned and designed for use by human volunteers. Ongoing user testing can help gauge and improve the user experience of a platform over time, and will help sustain the level of engagement and enjoyment for volunteers, both serving their interests and motivating them to participate.

STAGE 4: PRESENTATION AND SHARING

Presentation/ Sharing

Objectives

Defined objectives for the presentation/sharing stage included:

- Provide access to structured data
- Integrate data into discovery systems

Goals

The goals for Presentation/Sharing centered on designing successful experiences for collection end users and researchers.

Engaging	Ethical	Useful
● Users are presented with a variety of pathways to explore the content ● Collection data is presented through interesting and pleasing interface designs and experiences	● Provenance of data is communicated to library users ● Library users understand potential biases and incompleteness of data	● Users can discover data or collection content that meets their research needs ● Users can download and use data in tools they are familiar with

Challenges

One of the primary challenges to surface during the Presentation/Sharing phase centers around the difficulty of adequately and accurately conveying data provenance to researchers without distracting or overwhelming the user experience of the presentation interface.

To address these challenges we asked:

How do we…

- Convey provenance of data without overwhelming users?
- Communicate the incomplete and dynamic nature of data generated through large-scale ML processes to users?
- Integrate large volumes of data into discovery systems without diluting search results?

Humans

To achieve the goals and address the challenges of the Presentation and Sharing phase:

We involve…

- **Collection curators**, who understand collection contents and how they should be navigated
- **Reference librarians**, who understand researcher needs
- **Digital collection specialists**, who understand how to connect and display digital objects and metadata
- **UX designers**, who understand how to communicate and display data in a clear and accessible way
- **Library users**, who can offer feedback on website usability

Feedback Mechanisms

To achieve the goals and address the challenges of this phase:

We learn from…

- **User persona development** to help in understanding user needs and brainstorming interface functionality
- **Wireframing** to help imagine the potential of an interface to test with users
- **User testing** of wireframes with library users to help understand how interface design and the data driving it meets stated and implicit research goals

Process

Once the ML and crowdsourcing pipelines were planned and implemented, the core team began imagining how the derived data might be used to power an engaging and useful experience for end users of the Yellow Pages. This stage focused on centering user experience and researcher goals in the design of an imagined interface that would support researcher needs while clearly conveying data provenance to users.

User Personas and Wireframing

From the early stages of the HITL initiative, the team was writing and refining user stories for the Yellow Pages. During the Design stage of the HITL initiative, for example, user stories were instrumental in guiding data modeling decisions, driving the types of structured data that would be extracted or enhanced from the Yellow Pages to support researcher needs. Considering user needs and goals at that early phase helped the technical team understand what data structures would be needed to support and drive a successful end-user experience for the collection.

At the presentation and sharing phase , the AVP designers were revisiting those user stories to consider approaches to building an engaging and useful research experience for those same collection users.

Building on those early user stories, AVP designers developed ten distinct user personas for the Yellow Pages collection. The goal of developing the personas was to more deeply identify specific, *not generic*, Yellow Pages users who would become the focus of our research interface design.

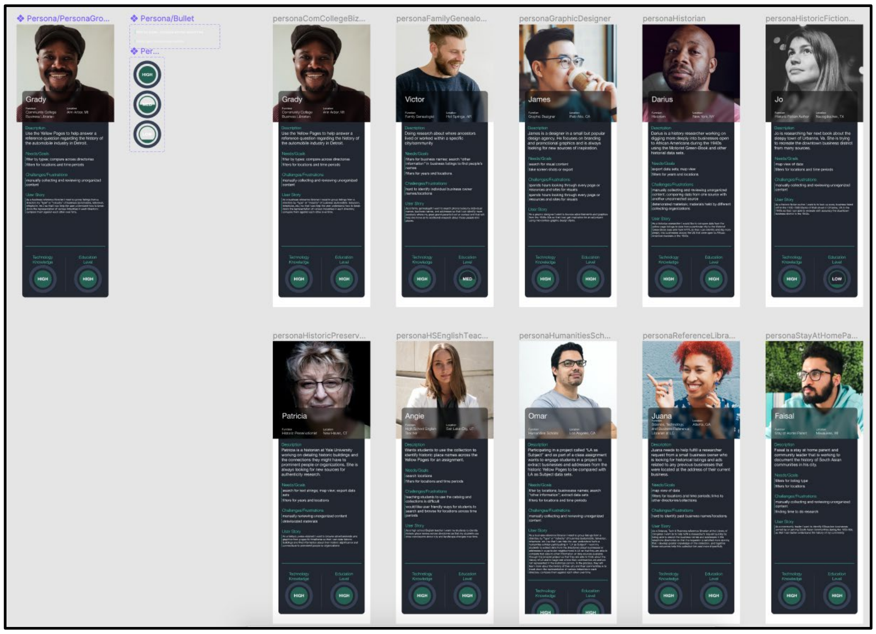
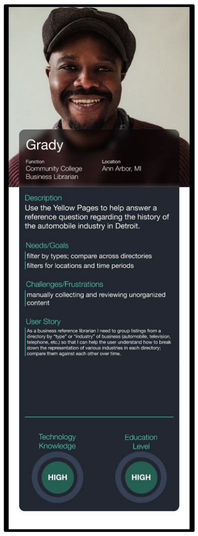


Image 22. Screenshot of persona designs in the Figma design tool. See Appendix B below for a larger view of this graphic.

Each persona was based on a different user story with expanded context and personal information to help personalize the goals and characteristics unique to each user-type. For example, an early user story about a reference librarian doing industry research, was transformed into a persona named Grady, a Community College Business Librarian with high technology and education levels, responding to a researcher request related to the history of the automobile industry in Detroit.



Grady

Function
Community College
Business Librarian

Location
Ann Arbor, MI

Description
Use the Yellow Pages to help answer a reference question regarding the history of the automobile industry in Detroit.

Needs/Goals
filter by types; compare across directories
filters for locations and time periods

Challenges/Frustrations
manually collecting and reviewing unorganized content

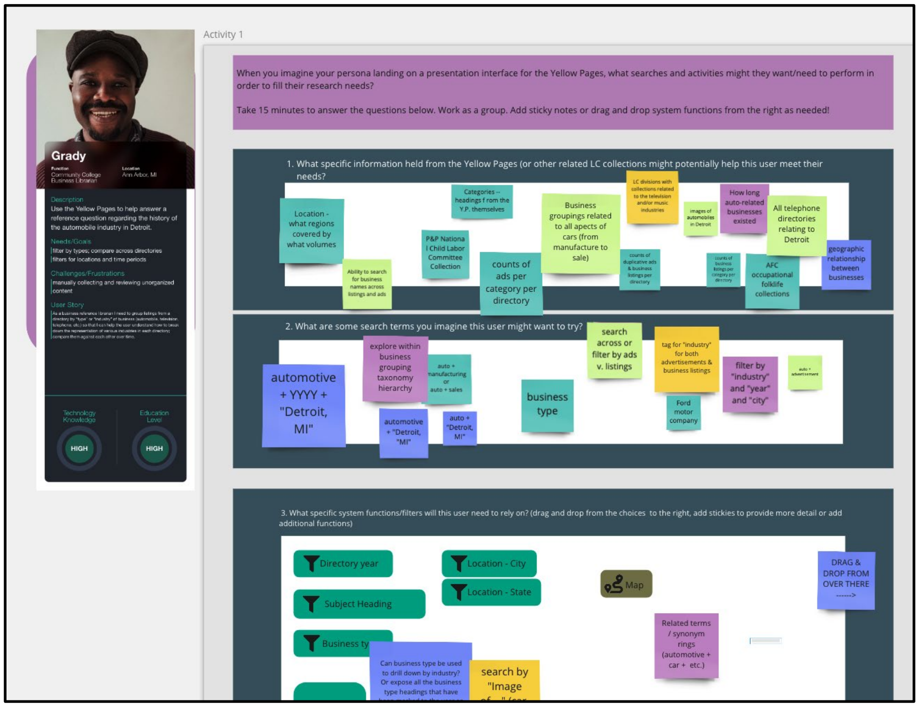
User Story
As a business reference librarian I need to group listings from a directory by "type" or "industry" of business (automobile, television, telephone, etc.) so that I can help the user understand how to break down the representation of various industries in each directory; compare them against each other over time.

Technology Knowledge: HIGH

Education Level: HIGH

Image 23. Screenshot of the Grady persona. See Appendix B for a larger view of this graphic.

Once define, individual personas like the one above were introduced to the core team in a second workshop facilitated by AVP. The team picked their top four personas — in this case, the personas team members felt were most representative of broad researcher needs — and worked in small groups to brainstorm how each of the selected personas might approach and explore a presentation site for the Yellow Pages.



Grady

Function
Community College
Business Librarian

Location
Ann Arbor, MI

Description
Use the Yellow Pages to help answer a reference question regarding the history of the automobile industry in Detroit.

Needs/Goals
filter by types; compare across directories
filters for locations and time periods

Challenges/Frustrations
manually collecting and reviewing unorganized content

User Story
As a business reference librarian I need to group listings from a directory by "type" or "industry" of business (automobile, television, telephone, etc.) so that I can help the user understand how to break down the representation of various industries in each directory; compare them against each other over time.

Technology Knowledge: HIGH

Education Level: HIGH

Activity 1

When you imagine your persona landing on a presentation interface for the Yellow Pages, what searches and activities might they want/need to perform in order to fill their research needs?

Take 15 minutes to answer the questions below. Work as a group. Add sticky notes or drag and drop system functions from the right as needed!

1. What specific information held from the Yellow Pages (or other related LC collections might potentially help this user meet their needs?

- Location - what regions covered by what volumes
- Ability to search for business names across listings and ads
- Categories -- headings f rom the Y.P. themselves
- P&P Nationa l Child Labor Committee Collection
- counts of ads per category per directory
- Business groupings related to all apects of cars (from manufacture to sale)
- counts of duplicative ads & business listings per directory
- LC divisions with collections related to the television and/or music industries
- images of automobiles in Detroit
- How long auto-related businesses existed
- counts of business listings per category per directory
- AFC occupational folklife collections
- All telephone directories relating to Detroit
- geographic relationship between businesses

2. What are some search terms you imagine this user might want to try?

- automotive + YYYY + "Detroit, MI"
- explore within business grouping taxonomy hierarchy
- auto + manufacturing or auto + sales
- automotive + "Detroit, "MI"
- auto + "Detroit, MI"
- business type
- search across or filter by ads v. listings
- tag for "industry" for both advertisements & business listings
- Ford motor company
- filter by "industry" and "year" and "city"
- auto + advertisement

3. What specific system functions/filters will this user need to rely on? (drag and drop from the choices to the right, add stickies to provide more detail or add additional functions)

- Directory year
- Subject Heading
- Business ty
- Location - City
- Location - State
- Map
- Can business type be used to drill down by industry? Or expose all the business type headings that have
- search by "Image
- Related terms / synonym rings (automotive + car + etc.)
- DRAG & DROP FROM OVER THERE ------>

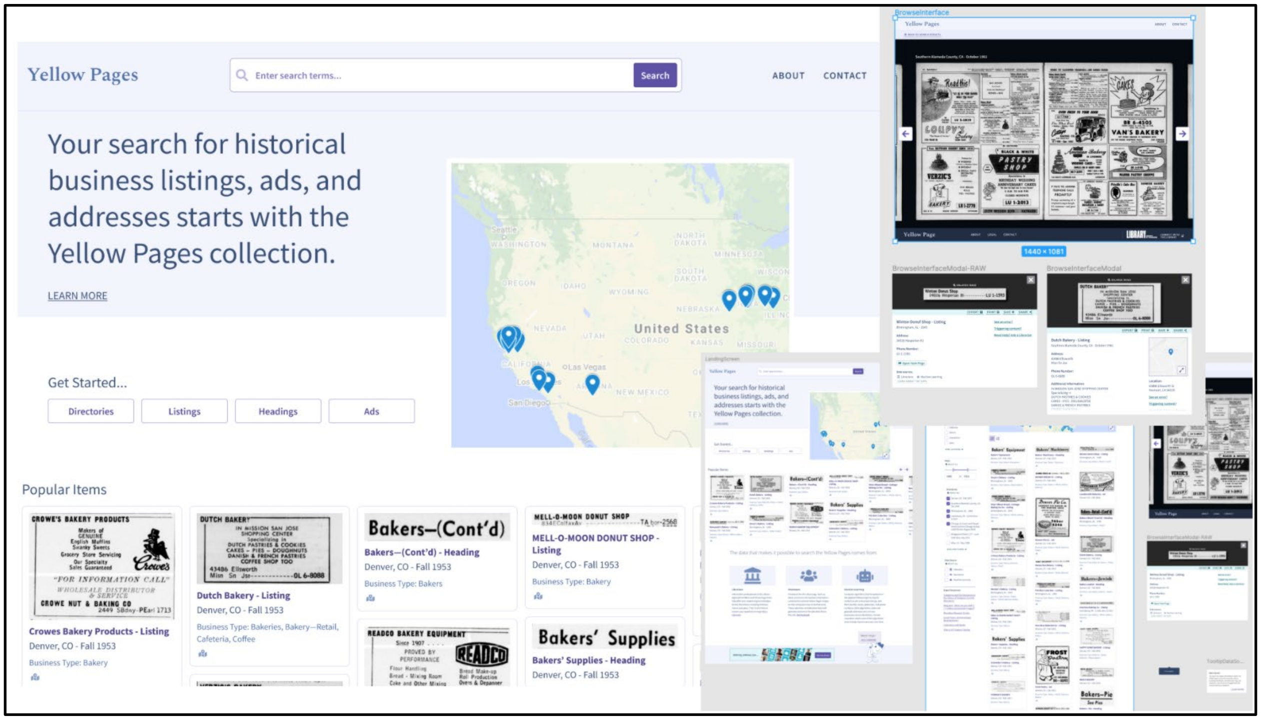
Image 24. Screenshot of workshop board in Miro. See Appendix B for a larger view of this graphic.

Takeaways from the workshop became the basis for the design of a wireframe mockup (Appendix L) of a Yellow Pages research interface, created in a collaborative design tool called

Figma[51]. The functionality imagined for the design also took into account the data model developed for the Yellow Pages experiment and considered how the derived data might inform the types of experiences Yellow Pages end users might expect to have when interacting with the collection.

The AVP designers also reviewed and attempted to incorporate features to help mitigate potential risks or biases likely to surface and impact users through the presentation interface. Even as early as the Collection Selection stage, the core team began identifying and defining potential risks and biases that might be surfaced through a human-in-the-loop experiment focused on the Yellow Pages. The risks and biases defined for end users of the Yellow Pages were reviewed again at this stage and used to inform mockup design.

The resulting mockup included a landing page for the collection, a search results screen, a map interface, and a browse experience for business listings and full-page scans from the digitized Yellow Pages collection. The overall design incorporated risk-intervention strategies, such as the ability for users to flag errors and triggering content within business listing records, as well as written and graphical information intended to convey data provenance in ways that would not overwhelm the user experience, or interrupt user goals.



Yellow Pages

Enter search terms...

Search

ABOUT

CONTACT

Your search for historical business listings, ads, and addresses starts with the Yellow Pages collection.

LEARN MORE

Get Started...

Directories | Listings | Headings | Ads

Popular Items

Crowes Bakery Products - Listing
Denver, CO - Fall 1953
Business Type: Bakery

Dutch Bakery - Listing
Denver, CO - Fall 1953
Business Type: Bakeries, Bakery—Retail, Cafeteria, Coffee

Bakers—(Cont'd) - Heading
Denver, CO - Fall 1953
Business Type: Bakers

MELL-O-MOON DONUT SHOP - Listing
Denver, CO - Fall 1953
Business Type: Bakery

Bakers' Supplies - Heading
Denver, CO - Fall 1953

Image 25. Screenshots of various wireframe pages in Figma. See Appendix B below for a larger view of this graphic.

User Testing

The wireframes were presented to end users via remote user test sessions conducted over Zoom in June 2021. The primary focus of user testing was to understand how well the imagined

[51] Figma, accessed July 1, 2021, https://www.figma.com/.

design would support user interests and research goals. The core team decided to conduct testing with researchers currently engaged in research supported by telephone directories, and other types of local history collection content. The team outreached via personal and professional networks in search of public historians, genealogists, public and county librarians, and other identified user types. Ten volunteers were identified and scheduled to participate in one-on-one, remote, user-testing sessions.

AVP designed a test plan (see Appendix M) that aimed to gather participant perspectives on how well the imagined interface design might support their research interests or needs, and to evaluate the effectiveness of the mockup in conveying data provenance.[52] The team also wanted to gather feedback on whether or not knowing a research site relies on crowdsourcing and ML-generated data was likely to impact the researchers' willingness to trust the information they discover on that site.

The user test sessions began with short interviews with each participant. The interviews were intended to surface each user's level of experience with similar research tools and products, and to hear about their particular research interests and goals. The facilitator also asked users about research sources they tend to trust implicitly, and if the Library of Congress was one of those sources.

During the initial interviews, users expressed a wide variety of research goals and interests that might bring them to a site like the imagined Yellow Pages site. For example,

- the head of a public branch library in Georgia described regularly searching city directories, as well as Ancestry.com, in support of research requests from local genealogists
- a public historian in New York shared about a failed attempt to gather information from scanned city directories in support of a study about George Gershwin
- a genealogist located in the Midwest conveyed successes searching online city directories from Montreal for clues about the lives of their ancestors
- a visual artist in Chicago described mining historical digital collection sites in search of images to reuse in their artwork

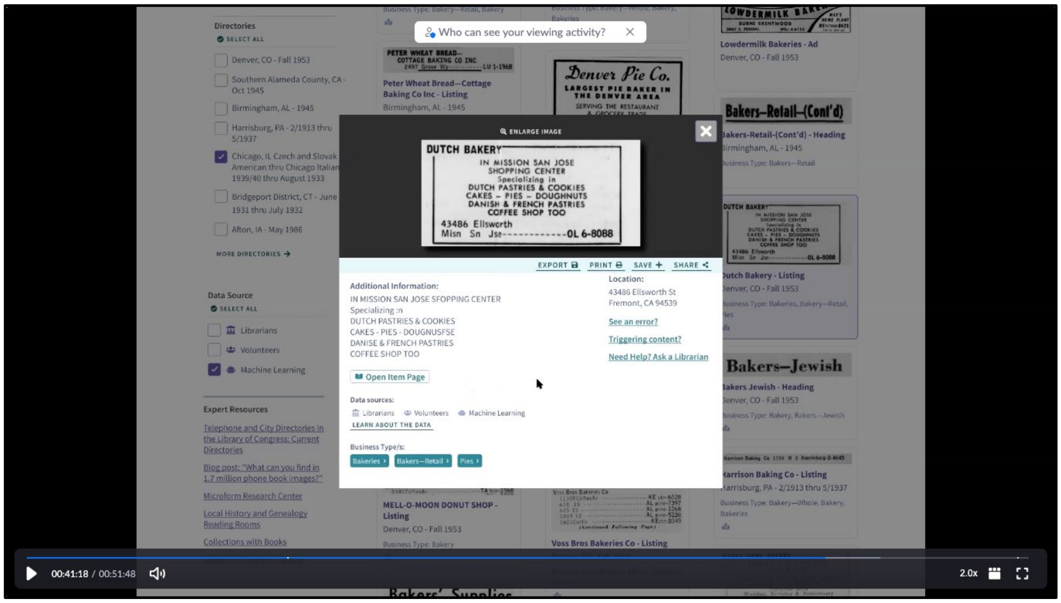
When asked to reflect on the trustworthiness of the sites they rely on for research, the majority of users expressed that the longevity of a source and the experience of finding quality, confirmable data multiple times in the same source helped build trust with a given site or collection. Many expressed an expectation that the historical records themselves will inevitably contain flaws and misleading information. All of the participants said they would tend to trust research sources hosted by the Library Congress.

Once the interviews were complete, the facilitator screen-shared and walked participants through the various components of the mockup. At certain points along the way, users were asked to provide feedback on what they liked about the proposed experience, and what they wished was different about the experience. Users were also prompted to reflect on concepts

[52] "Appendix M – Presentation Interface User Testing Plan _Discussion Guide." Library of Congress. Accessed November 24, 2021. https://github.com/LibraryOfCongress/hitl/blob/main/hitl-recommendations-report-appendices/Appendix%20M_%20Presentation%20Interface%20User%20Testing%20Plan%20_%20Discussion%20Guide.pdf.

presented via the mockup, especially related to data provenance. They were explicitly asked to share their responses to knowing the site would rely on ML and volunteer-generated data, and were asked to share any benefits or concerns raised when considering these approaches.

When asked about their perspectives on using ML and volunteers to generate data for the site, the majority of users expressed seeing the benefits of combining the two approaches. One user remarked that each approach — machine learning and crowdsourcing — “is likely to trip over different things, so if you’re using both of them, you’re more likely to get it right” (Appendix N, Raw Data tab). In general, users expressed appreciation that information about the data provenance was clearly communicated by the mockup, and did not feel that use of those approaches would detract from their ability to trust the site.



DUTCH BAKERY
IN MISSION SAN JOSE
SHOPPING CENTER
Specializing in
DUTCH PASTRIES & COOKIES
CAKES – PIES – DOUGHNUTS
DANISH & FRENCH PASTRIES
COFFEE SHOP TOO
43486 Ellsworth
Misn Sn Jse------------OL 6-8088

Additional Information:
IN MISSION SAN JOSE SFOPPING CENTER
Specializing :n
DUTCH PASTRIES & COOKIES
CAKES - PIES - DOUGNUSFSE
DANISE & FRENCH PASTRIES
COFFEE SHOP TOO

Location:
43486 Ellsworth St
Fremont, CA 94539

Data sources:
Librarians Volunteers Machine Learning

Business Type/s:
Bakeries › Bakers—Retail › Pies ›

Image 26. Screengrab of remote mockup test demo in Zoom. See Appendix B for a larger view of this graphic.

When presented with mockups of business listing records, viewers saw records that contained errors that might be introduced to the data through volunteer transcription or ML processes. They were asked for their reaction to seeing those errors, and whether or not viewing errors in search results and in item records would change their impression of the overall trustworthiness of the site. The expectation that historical data might be incomplete or imperfectly rendered was repeated by all participants. The fact that the interface offered opportunities to report errors from within item records was universally appreciated and called out as a feature users found particularly valuable. Several users indicated that the ability to report errors would add to their confidence in using the site.

When asked if they would be likely to use the Yellow Pages site, or another site built from human-in-the-loop processes in support of their research, users gave an average answer of 9.3, with 10 being the highest rating. The tests surfaced useful feedback about overall design choices, research pathways, and features and functionality that researchers expected to see on the Yellow Pages site. User testing data and responses are available in Appendix N.

Recommendations

Based on findings from the presentation and sharing stage, the team recommends drawing on early-stage user stories, identified risks and mitigation strategies, and conversations with users to help guide human-centered approaches for human-in-the-loop presentation interfaces and experiences.

- **Build holistic approaches to safe and ethical user experiences.**
  Bias and other risks are not only introduced to collection content during data generation processes. Collection source materials and discovery interfaces introduce risks to users just as ML and crowdsourced processes do. Consideration of potential risks to users and mitigation strategies in human-in-the-loop approaches should be one part of a broader, more holistic approach to building safe and ethical user experiences in libraries.

- **Design interfaces that support data in varied states of accuracy and completeness**.
  Given that human-in-the-loop approaches will require ongoing data generation and processing, the associated research interfaces will need to convey to users varying states of accuracy and completeness of the data presented. Presentation design should identify ways to communicate data provenance, varying levels of data processing across content, and offer opportunities for users to mitigate errors they find.

- **Include UX Designers in the development of research and user experiences.**
  To be transparent with end users about data provenance and completeness in human-in-the-loop initiativesrequires expert design approaches and techniques. The need to convey data provenance, complex interactions between volunteer and ML-generated data, and varying levels of metadata completeness to front-users will introduce design challenges. Including UX Designers in the development of personas, as well as early mockups and prototypes for human-in-the-loop interfaces, will support creation of ethical, engaging, and useful designs.

- **Test early prototypes and mockups with collection users and experts.**
  While this experiment did not intend to design or launch a fully functioning research site, gathering feedback from researchers on early designs was extremely valuable to the design. Researcher feedback highlighted issues and potential functionalities the core team and designers had overlooked in the mockups. For example, five out of ten researchers that participated in user testing called out the need for an advanced search option, which the wireframe did not include. Users also pointed out specific use cases for downloading the full data set, as well as selected portions of the data, both with and without images. The mockup, however, only offered data export at the listings level.

  Conducting interviews with current users of collections and content similar to the Yellow Pages collection, provided an opportunity for the team to gather a diversity of perspectives and use cases that broadened and helped expand on the user stories the team had previously defined. Interviews helped confirm and inform certain design decisions and choices. For example, users universally expressed an expectation to see errors in records related to historical documents and sources, and were not put off by knowing the data might be in various states of completeness. Similarly, the majority of users expressed appreciation for the ability to flag triggering content, although multiple users expressed wanting a clearly defined explanation of how that flag would be reviewed and used.

User testing with early-stage prototypes and mockups will provide valuable insights into design assumptions and decisions before staff and financial resources are invested in site development. Centering users at this stage will result in better, more engaging research sites and experiences.

CONCLUSIONS

MAJOR TAKEAWAYS

Human-in-the-loop approaches have the potential to be extremely powerful for maximizing access to LC's content at scale.
In the brief time that the core team was able to implement the crowdsourcing prototypes and initial ML pipeline, AVP and LC staff were able to generate fully structured data for 119 business listings across four books of Yellow Pages. While results have been untested and are still rife with errors, the ML pipeline was able to process all four phone books and generate data for around 15,000 business listings in roughly four days. This experiment demonstrates not only the great potential for improvable ML processes, but also how human knowledge and contributions will still be critical to ensuring accurate results. (Sample output data is available in Appendix O.)

Human-in-the-loop initiatives will require significant investment in staffing and resources.
Research and development activities in cultural heritage frequently borrow staff expertise for short stints from across the organization (or from outside the organization through residencies or short-term contracts) to build an experiment or demonstrate a proof of concept. These activities can show not only the promise within innovative ideas, but also the great potential in cross-functional collaboration. Such staffing models, however, cannot easily translate innovation to production, especially when staff time is already fully committed to other areas of the organization's mission.

Successfully operationalizing human-in-the-loop initiatives will require ongoing dedicated staffing from product managers/liaisons, ML experts, community managers, software engineers, and UX designers. This report also identifies many other staff roles across a library that are essential to informing the design and development at different stages. Before engaging in a human-in-the-loop endeavor, an organization should be ready to commit to (and support staff in) collaborating in this effort, which may require significant changes to organizational mission and culture.

There are ways to generalize human-in-the-loop approaches, however, there will not be a one-size-fits-all approach.
An initial goal of this HITL initiative was to find opportunities to re-use crowdsourcing and ML technology and data from the open-source and library communities. While the creation and sharing of general-use applications and collections-as-data is well-intentioned and has certainly contributed to innovations in many areas of cultural heritage, a reliance on reuse at the expense of an understanding of local users runs the risk of preferencing the design assumptions of a few technology and data creators over the real needs of an organization's target communities. Libraries can look for novel uses for technology solutions or existing datasets, but letting user needs for engaging activities, ethical experiences, and useful resources guide design will likely meet broader and more lasting appeal for innovations in cultural heritage.

This experiment revealed the challenges in tailoring an existing crowdsourcing platform for a set of engaging and ethical tasks supporting ML processes. The methods selected for the machine-learning pipeline were mostly able to use default models, but custom scripts had to be designed for the specific structure of Yellow Pages content. For the CRF process that required model training, even though data was available from a similar experiment with city directories in the NYPL Labs Space/Time Project, the data, drawn from images from an earlier time before phone numbers existed, was not similar enough to the Yellow Pages to produce the desired results. As discussed in the Implementation section, however, generalizing some common components of a human-in-the-loop initiative, such as a workflow database, can free up time and resources to spend on user research and platform and interface development custom to collection content and local community needs.

Included in the appendices of this report is the generalized framework for approaching a human-in-the-loop experiments and endeavors that speaks to objectives, goals, challenges, human needs, and feedback mechanisms that will be common to many cultural heritage organizations (Appendix A). While some details may differ case by case, it is the authors' hope that both the Library of Congress and organizations beyond will find value in the lessons learned through this experiment.

AREAS FOR FURTHER EXPLORATION

Ongoing user testing and iterative development of crowdsourcing and end-user platforms will significantly improve overall user engagement with and access to LC collections and content.
User testing of each prototype was conducted and did help inform the framework. Due to the limited nature of the HITL initiative, however, user feedback was not incorporated into fully functioning platforms. Future experiments may further investigate how ongoing user testing and iterative development processes can be incorporated into long-term plans. The HITL initiative team firmly believes that this approach will positively impact overall user experience for the humans engaged at every point in human-in-the-loop initiatives, and will ultimately improve broader user engagement with and access to Library collections and content.

Investigation of other methods of sharing human-in-the-loop data will benefit collection end users.

This experiment explored just one possible interface for presenting human-in-the-loop outputs to users, but there are many other possibilities for sharing data to be evaluated, including data dumps or APIs for sharing machine-actionable structured data with researchers or developers, integration of business listing data into LC digital collections pages, and opportunities for connecting data to related initiatives or collections.

Broader representation across teams will better surface and address potential risks and biases.

During the course of the HITL initiative, it became apparent that the core team was limited, especially in terms of risk identification and mitigation, by the homogenous make-up of its members, who were predominantly white, mostly female-identified, most with advanced degrees. One example of how this limitation surfaced was in the creation of user personas. AVP designers initially only created personas with medium- to high-levels of education. During our workshop, a core team member recognized an emerging trend to prefer personas with higher levels of education. It was only then that the team realized that none of the original personas had been designed to consider needs of less formally educated users. In this case, the team was able to catch and correct a specific bias introduced during the design process. However, the example highlights the dangers of limited representation on teams designing these approaches. The team composition is a central mechanism that can improve or prevent identification and mitigation of unconscious bias and other risks to humans that interact with Library collections. There is opportunity in future human-in-the-loop approaches to mitigate the introduction of implicit biases throughout the lifecycle of an initiative by increasing diversity and representation on collaborative teams.

ACKNOWLEDGEMENTS

ACKNOWLEDGEMENTS FROM LC LABS

We'd first like to thank our colleagues for their ongoing work to build the foundations of information and access to Library of Congress resources through description, digitization, and delivery and creating other essential paths for use. Without the work of our colleagues we would not be able to undertake these forms of experimentation; together we are taking steps to throw open the treasure chest, connect, and invest in our future.

We'd especially like to acknowledge:

- Director of General And International Collections Directorate Eugene Flanagan, Ron Bluestone, and Michele Sellars for their enthusiasm and support of this experiment and for empowering their staff to participate
- John Fenn, Nicki Saylor, and Rachel Trent for their participation and expertise shared to understand risks and biases in preliminary collection selection workshops
- Natalie Buda Smith and the User Experience Design team - Jamie Bresner, Amanda Perez, and Wendy Stengel - for sharing their expertise and improving these workflows and prototypes
- Trevor Owens and the By the People team - Lauren Algee, Carlyn Osborn, and Abby Shelton, as well as Head of Product Elaine Kamlley - for sharing their reflections on designing ongoing technical and programmatic approaches
- User testing participants for their time, feedback, and discussion - and all Library of Congress colleagues who helped with note taking in user research sessions
- LC Labs and Digital Strategy Directorate staff for their enthusiasm and co-designing energy throughout this experiment and Director of Digital Strategy Kate Zwaard for her support to pursue this work
- The core team -- Lauren Algee, Natalie Burclaff, Eileen Jakeway, Jaime Mears, Trevor Owens, Abbey Potter, and Leah Weinryb-Grohsgal -- and the AVP team for their willingness to engage deeply across the experiment design and evaluation, and for joining together in this unique opportunity to shape future possibility

ACKNOWLEDGEMENTS FROM AVP

Many thanks to Meghan Ferriter and everyone at the Library of Congress who touched this collaboration along the way.

Thanks go out to Sheean Spoel of the Digital Humanities Lab at Utrecht University for generously sharing their private fork of the Scribe codebase, which allowed the team to get a working instance of the crowdsourcing platform up and running quickly for the workflow prototypes.

Special thanks are also due to all of the individuals who participated in user testing for this experiment, including Judith Yellen, Maddie Tsurusaki, Mariana Ziku, Rosie Clark, Mia Ridge, Brad Sims, K. T. Vaughn, Jenn Smith, Sam Blickhan, Ellen Noonan, Tracy Seneca, Ellen Terrell, Justin Kau, Perida Mitchell, Judy Young, Melissa Olson, and Darshni Patel.

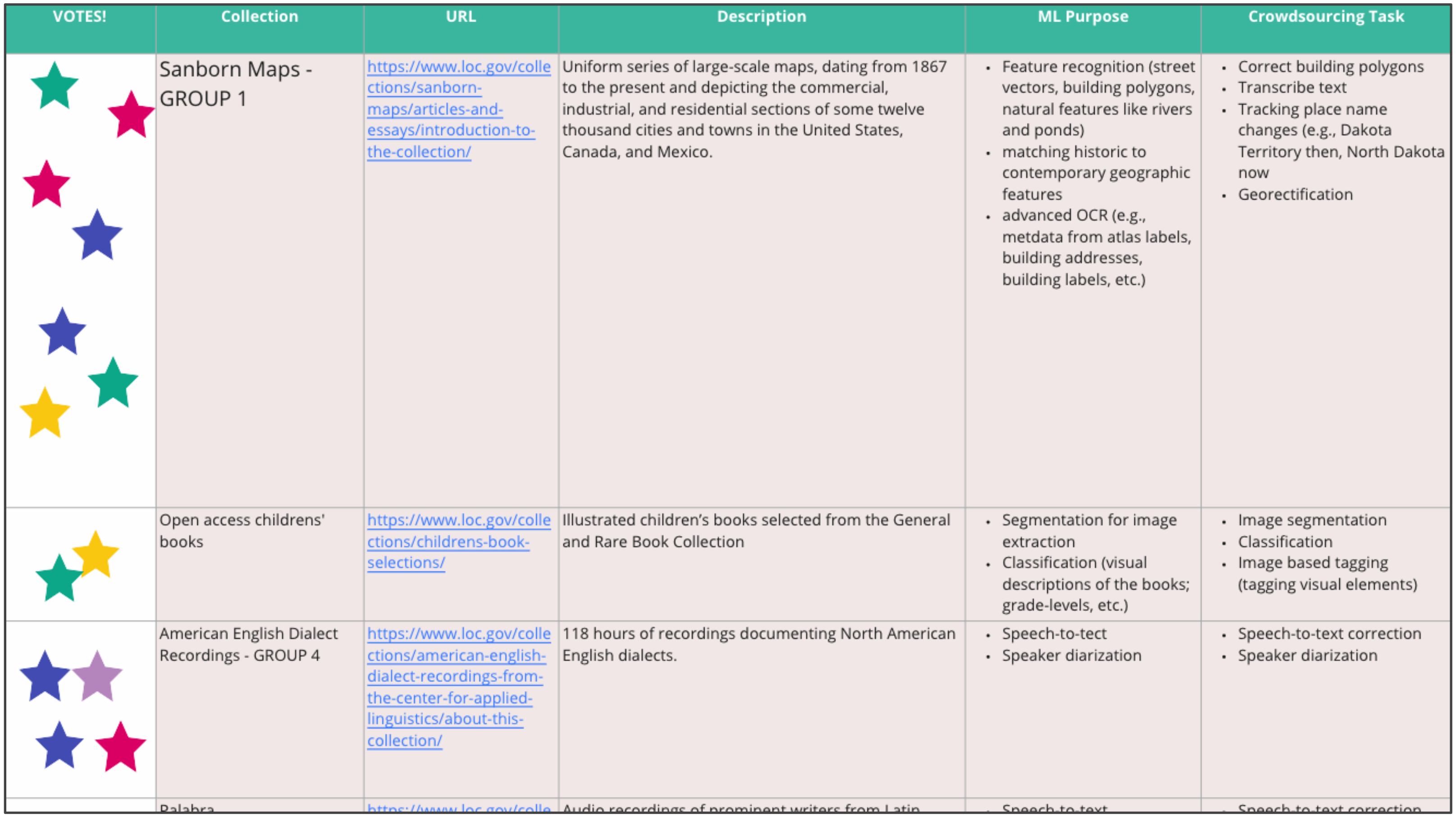
APPENDICES

Appendix A: Framework

Full Humans-in-the-loop Framework.

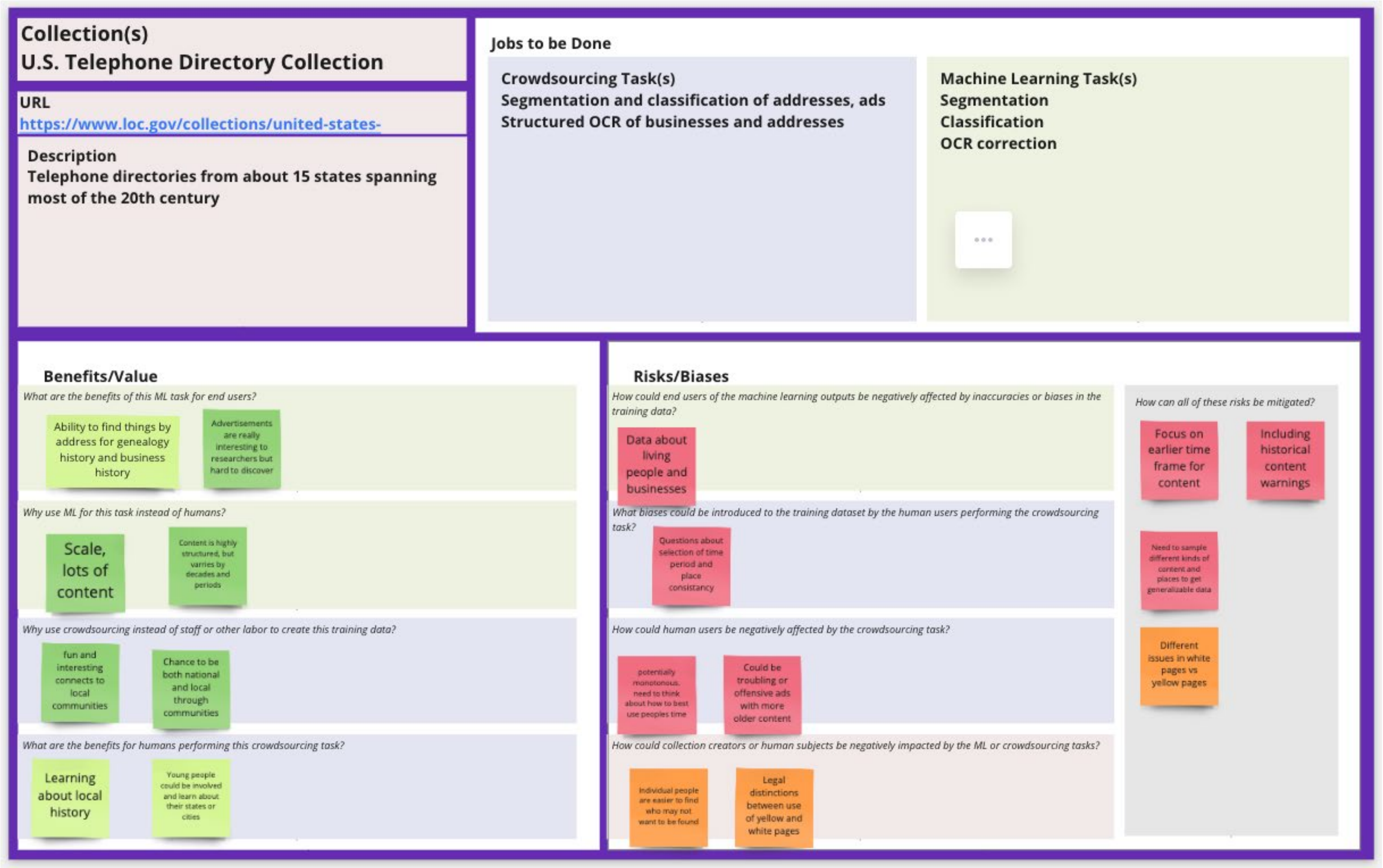
[see HITL GitHub repository]

Appendix B: Images



VOTES!	Collection	URL	Description	ML Purpose	Crowdsourcing Task
★★★★★★★	Sanborn Maps - GROUP 1	https://www.loc.gov/collections/sanborn-maps/articles-and-essays/introduction-to-the-collection/	Uniform series of large-scale maps, dating from 1867 to the present and depicting the commercial, industrial, and residential sections of some twelve thousand cities and towns in the United States, Canada, and Mexico.	• Feature recognition (street vectors, building polygons, natural features like rivers and ponds) • matching historic to contemporary geographic features • advanced OCR (e.g., metdata from atlas labels, building addresses, building labels, etc.)	• Correct building polygons • Transcribe text • Tracking place name changes (e.g., Dakota Territory then, North Dakota now • Georectification
★★	Open access childrens' books	https://www.loc.gov/collections/childrens-book-selections/	Illustrated children's books selected from the General and Rare Book Collection	• Segmentation for image extraction • Classification (visual descriptions of the books; grade-levels, etc.)	• Image segmentation • Classification • Image based tagging (tagging visual elements)
★★★★	American English Dialect Recordings - GROUP 4	https://www.loc.gov/collections/american-english-dialect-recordings-from-the-center-for-applied-linguistics/about-this-collection/	118 hours of recordings documenting North American English dialects.	• Speech-to-tect • Speaker diarization	• Speech-to-text correction • Speaker diarization
	Palabra	https://www.loc.gov/colle	Audio recordings of prominent writers from Latin	• Speech-to-text	• Speech-to-text correction

Appendix B. Image 3



Collection(s)
U.S. Telephone Directory Collection

URL
https://www.loc.gov/collections/united-states-

Description
Telephone directories from about 15 states spanning most of the 20th century

Jobs to be Done

Crowdsourcing Task(s)
Segmentation and classification of addresses, ads
Structured OCR of businesses and addresses

Machine Learning Task(s)
Segmentation
Classification
OCR correction

Benefits/Value

What are the benefits of this ML task for end users?

- Ability to find things by address for genealogy history and business history
- Advertisements are really interesting to researchers but hard to discover

Why use ML for this task instead of humans?

- Scale, lots of content
- Content is highly structured, but varries by decades and periods

Why use crowdsourcing instead of staff or other labor to create this training data?

- fun and interesting connects to local communities
- Chance to be both national and local through communities

What are the benefits for humans performing this crowdsourcing task?

- Learning about local history
- Young people could be involved and learn about their states or cities

Risks/Biases

How could end users of the machine learning outputs be negatively affected by inaccuracies or biases in the training data?

- Data about living people and businesses

What biases could be introduced to the training dataset by the human users performing the crowdsourcing task?

- Questions about selection of time period and place consistancy

How could human users be negatively affected by the crowdsourcing task?

- potentially monotonous. need to think about how to best use peoples time
- Could be troubling or offensive ads with more older content

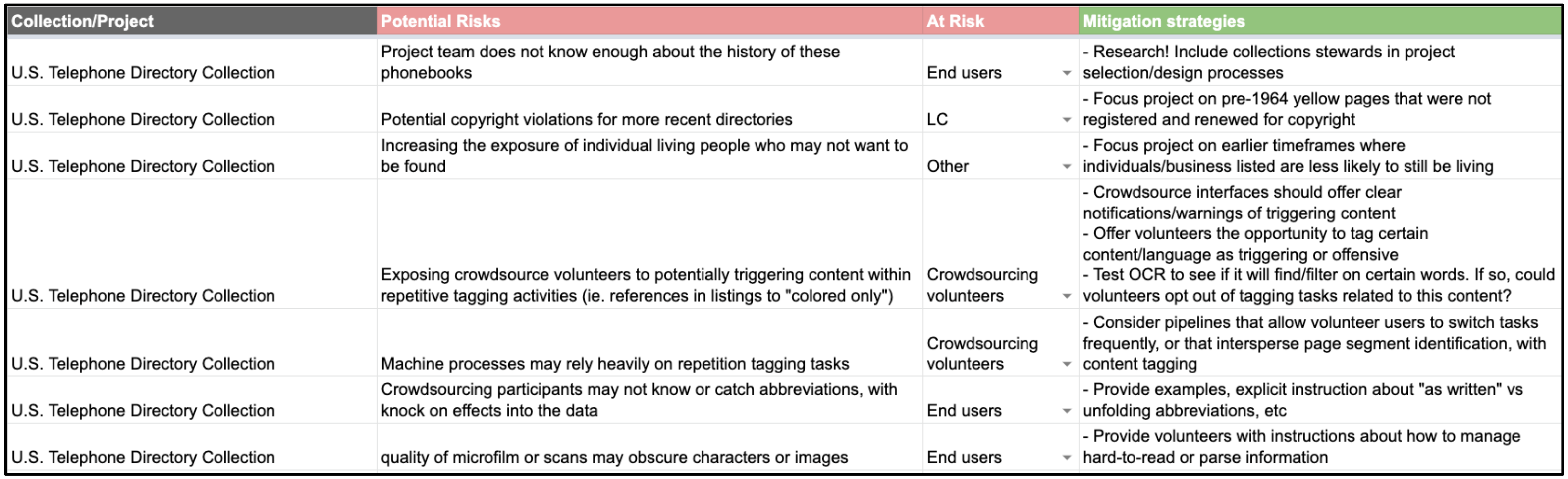
How could collection creators or human subjects be negatively impacted by the ML or crowdsourcing tasks?

- Individual people are easier to find who may not want to be found
- Legal distinctions between use of yellow and white pages

How can all of these risks be mitigated?

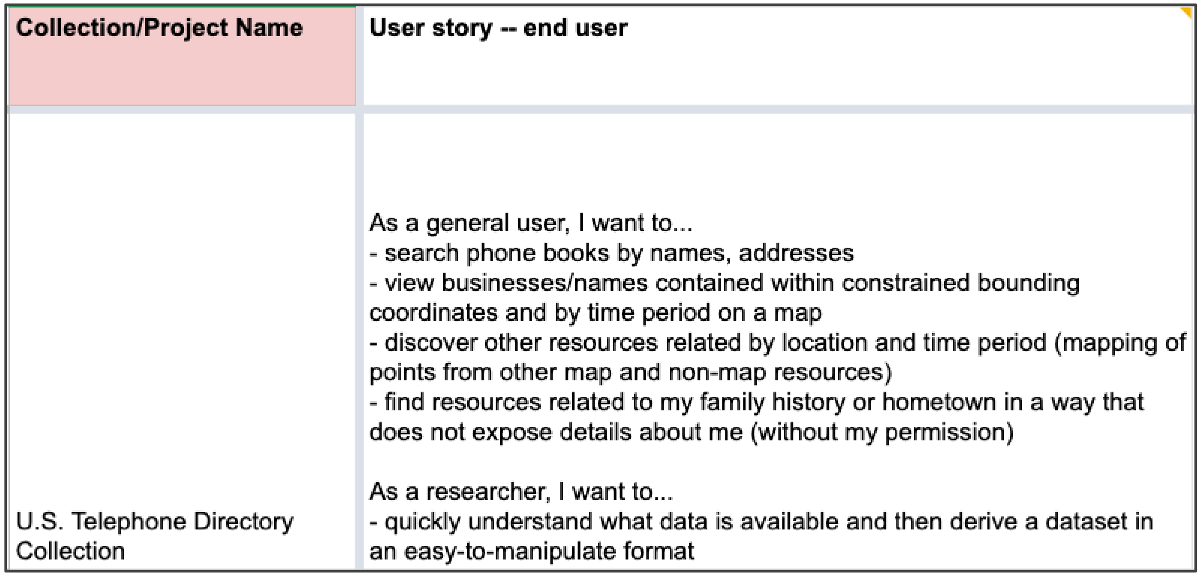
- Focus on earlier time frame for content
- Including historical content warnings
- Need to sample different kinds of content and places to get generalizable data
- Different issues in white pages vs yellow pages

Appendix B. Image 4



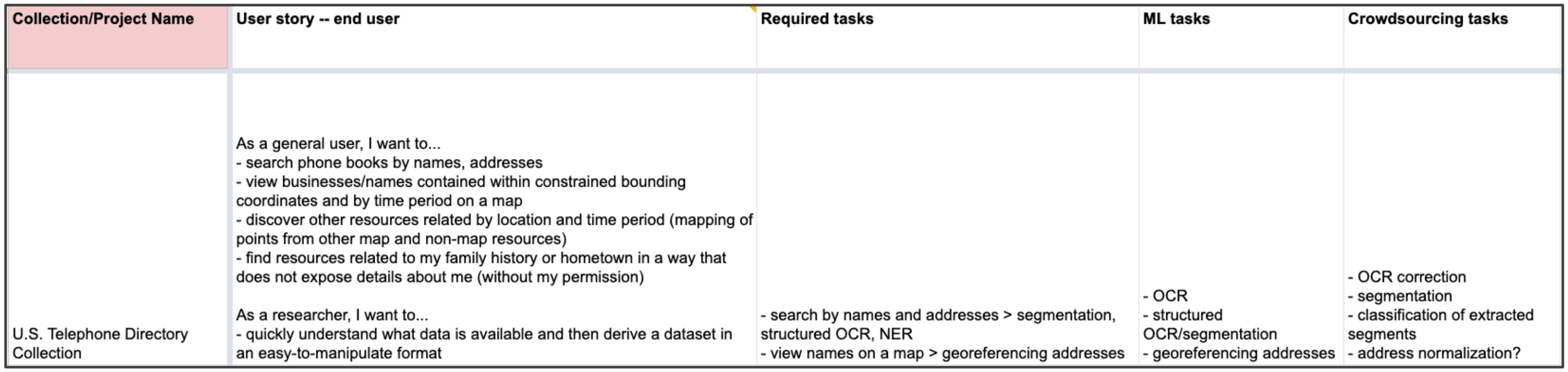
Collection/Project	Potential Risks	At Risk	Mitigation strategies
U.S. Telephone Directory Collection	Project team does not know enough about the history of these phonebooks	End users	- Research! Include collections stewards in project selection/design processes
U.S. Telephone Directory Collection	Potential copyright violations for more recent directories	LC	- Focus project on pre-1964 yellow pages that were not registered and renewed for copyright
U.S. Telephone Directory Collection	Increasing the exposure of individual living people who may not want to be found	Other	- Focus project on earlier timeframes where individuals/business listed are less likely to still be living
U.S. Telephone Directory Collection	Exposing crowdsource volunteers to potentially triggering content within repetitive tagging activities (ie. references in listings to "colored only")	Crowdsourcing volunteers	- Crowdsource interfaces should offer clear notifications/warnings of triggering content - Offer volunteers the opportunity to tag certain content/language as triggering or offensive - Test OCR to see if it will find/filter on certain words. If so, could volunteers opt out of tagging tasks related to this content?
U.S. Telephone Directory Collection	Machine processes may rely heavily on repetition tagging tasks	Crowdsourcing volunteers	- Consider pipelines that allow volunteer users to switch tasks frequently, or that intersperse page segment identification, with content tagging
U.S. Telephone Directory Collection	Crowdsourcing participants may not know or catch abbreviations, with knock on effects into the data	End users	- Provide examples, explicit instruction about "as written" vs unfolding abbreviations, etc
U.S. Telephone Directory Collection	quality of microfilm or scans may obscure characters or images	End users	- Provide volunteers with instructions about how to manage hard-to-read or parse information

Appendix B. Image 5



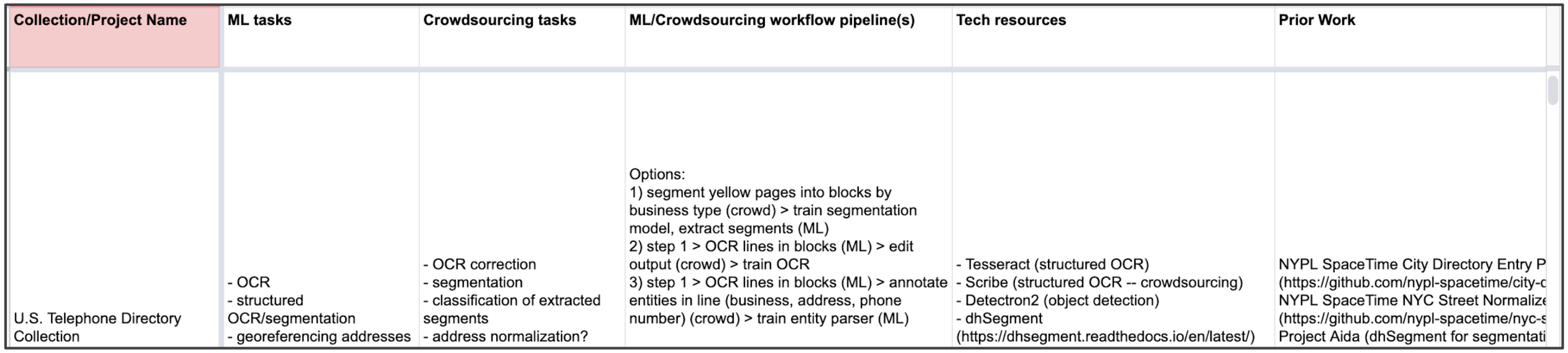
Collection/Project Name	User story -- end user
U.S. Telephone Directory Collection	As a general user, I want to... - search phone books by names, addresses - view businesses/names contained within constrained bounding coordinates and by time period on a map - discover other resources related by location and time period (mapping of points from other map and non-map resources) - find resources related to my family history or hometown in a way that does not expose details about me (without my permission) As a researcher, I want to... - quickly understand what data is available and then derive a dataset in an easy-to-manipulate format

Appendix B. Image 6



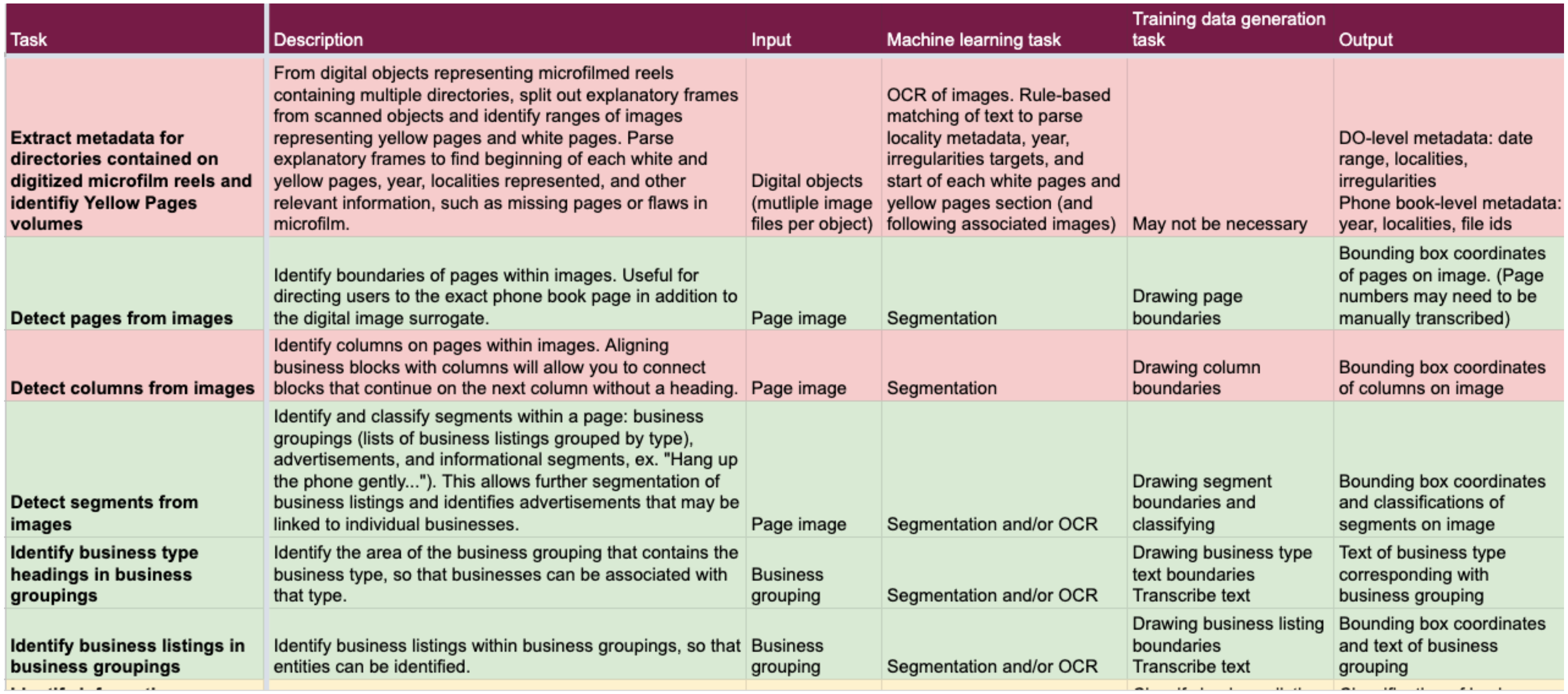
Collection/Project Name	User story -- end user	Required tasks	ML tasks	Crowdsourcing tasks
U.S. Telephone Directory Collection	As a general user, I want to... - search phone books by names, addresses - view businesses/names contained within constrained bounding coordinates and by time period on a map - discover other resources related by location and time period (mapping of points from other map and non-map resources) - find resources related to my family history or hometown in a way that does not expose details about me (without my permission) As a researcher, I want to... - quickly understand what data is available and then derive a dataset in an easy-to-manipulate format	- search by names and addresses > segmentation, structured OCR, NER - view names on a map > georeferencing addresses	- OCR - structured OCR/segmentation - georeferencing addresses	- OCR correction - segmentation - classification of extracted segments - address normalization?

Appendix B. Image 7



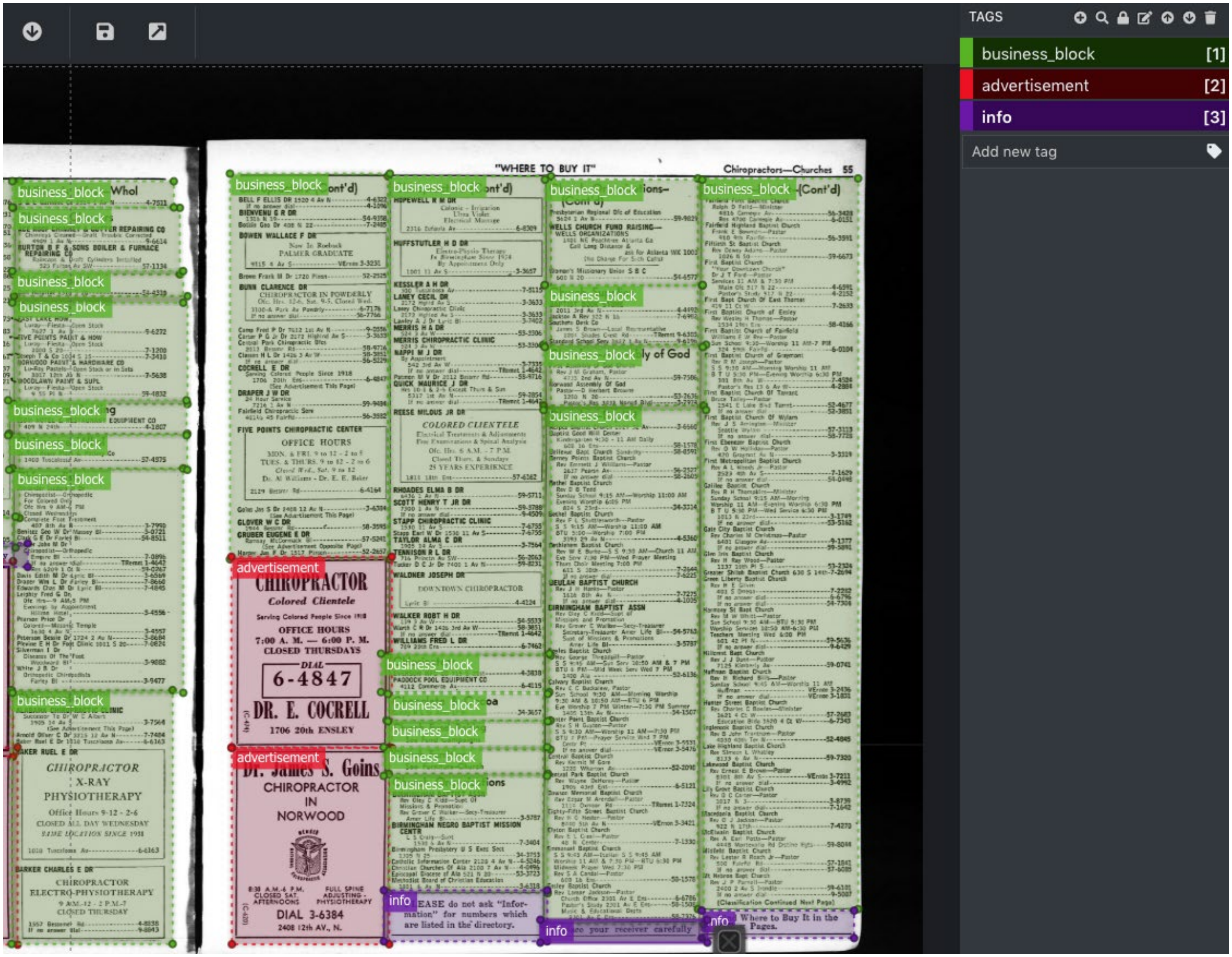
Collection/Project Name	ML tasks	Crowdsourcing tasks	ML/Crowdsourcing workflow pipeline(s)	Tech resources	Prior Work
U.S. Telephone Directory Collection	- OCR - structured OCR/segmentation - georeferencing addresses	- OCR correction - segmentation - classification of extracted segments - address normalization?	Options: 1) segment yellow pages into blocks by business type (crowd) > train segmentation model, extract segments (ML) 2) step 1 > OCR lines in blocks (ML) > edit output (crowd) > train OCR 3) step 1 > OCR lines in blocks (ML) > annotate entities in line (business, address, phone number) (crowd) > train entity parser (ML)	- Tesseract (structured OCR) - Scribe (structured OCR -- crowdsourcing) - Detectron2 (object detection) - dhSegment (https://dhsegment.readthedocs.io/en/latest/)	NYPL SpaceTime City Directory Entry P (https://github.com/nypl-spacetime/city-c NYPL SpaceTime NYC Street Normalize (https://github.com/nypl-spacetime/nyc-s Project Aida (dhSegment for segmentati

Appendix B. Image 8



Task	Description	Input	Machine learning task	Training data generation task	Output
Extract metadata for directories contained on digitized microfilm reels and identifiy Yellow Pages volumes	From digital objects representing microfilmed reels containing multiple directories, split out explanatory frames from scanned objects and identify ranges of images representing yellow pages and white pages. Parse explanatory frames to find beginning of each white and yellow pages, year, localities represented, and other relevant information, such as missing pages or flaws in microfilm.	Digital objects (mutliple image files per object)	OCR of images. Rule-based matching of text to parse locality metadata, year, irregularities targets, and start of each white pages and yellow pages section (and following associated images)	May not be necessary	DO-level metadata: date range, localities, irregularities Phone book-level metadata: year, localities, file ids
Detect pages from images	Identify boundaries of pages within images. Useful for directing users to the exact phone book page in addition to the digital image surrogate.	Page image	Segmentation	Drawing page boundaries	Bounding box coordinates of pages on image. (Page numbers may need to be manually transcribed)
Detect columns from images	Identify columns on pages within images. Aligning business blocks with columns will allow you to connect blocks that continue on the next column without a heading.	Page image	Segmentation	Drawing column boundaries	Bounding box coordinates of columns on image
Detect segments from images	Identify and classify segments within a page: business groupings (lists of business listings grouped by type), advertisements, and informational segments, ex. "Hang up the phone gently..."). This allows further segmentation of business listings and identifies advertisements that may be linked to individual businesses.	Page image	Segmentation and/or OCR	Drawing segment boundaries and classifying	Bounding box coordinates and classifications of segments on image
Identify business type headings in business groupings	Identify the area of the business grouping that contains the business type, so that businesses can be associated with that type.	Business grouping	Segmentation and/or OCR	Drawing business type text boundaries Transcribe text	Text of business type corresponding with business grouping
Identify business listings in business groupings	Identify business listings within business groupings, so that entities can be identified.	Business grouping	Segmentation and/or OCR	Drawing business listing boundaries Transcribe text	Bounding box coordinates and text of business grouping

Appendix B. Image 9

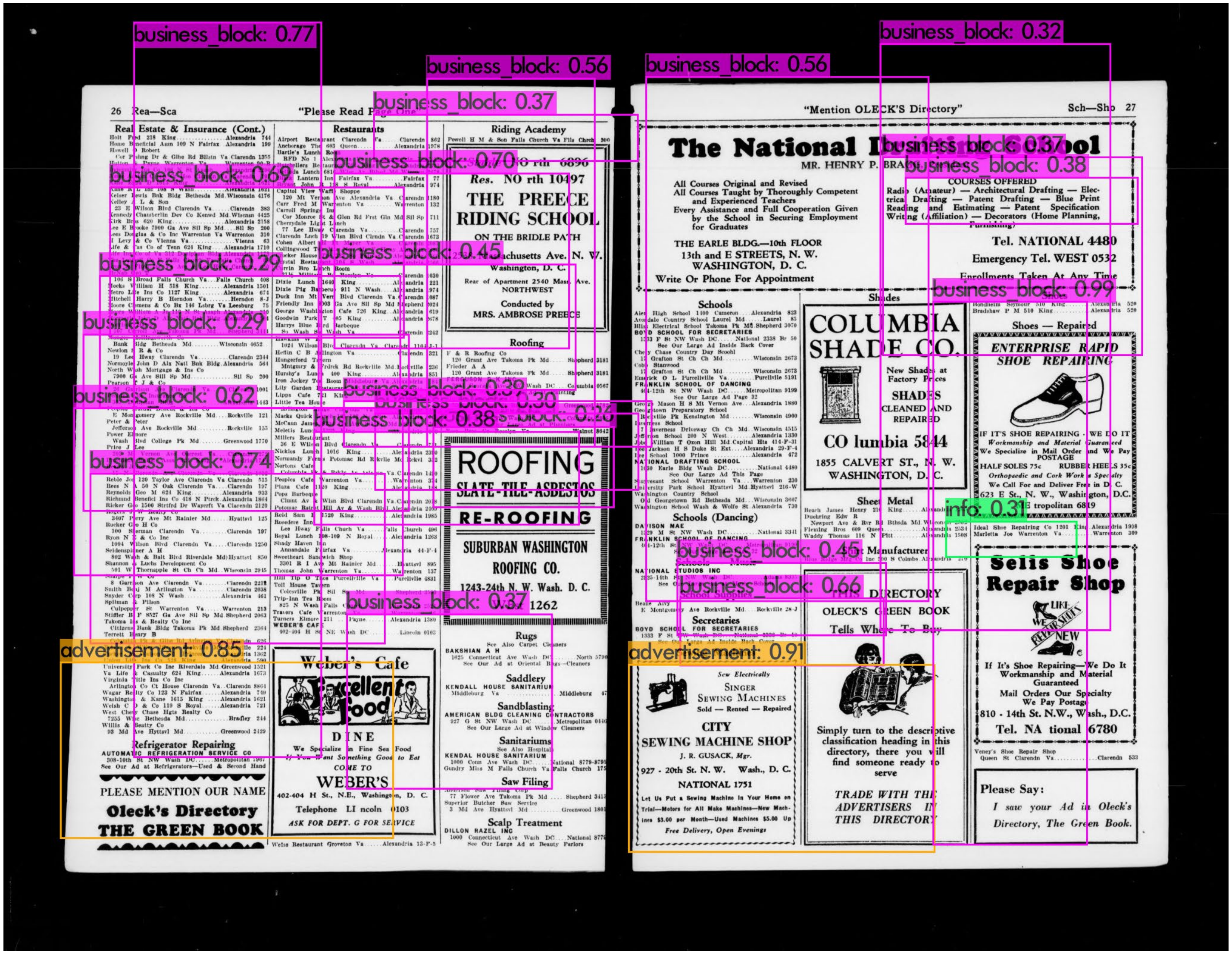


TAGS

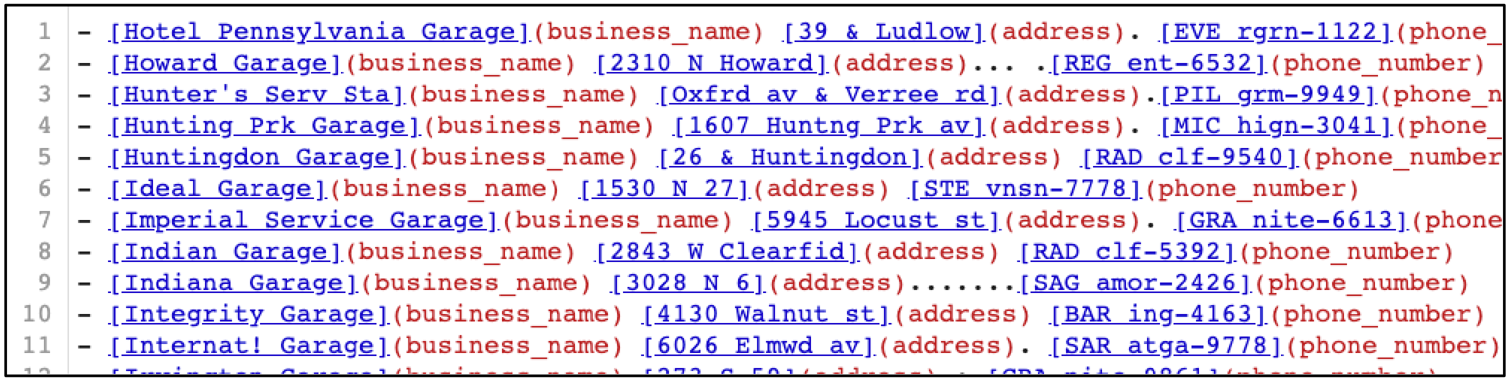
business_block	[1]
advertisement	[2]
info	[3]

Add new tag

Appendix B. Image 10

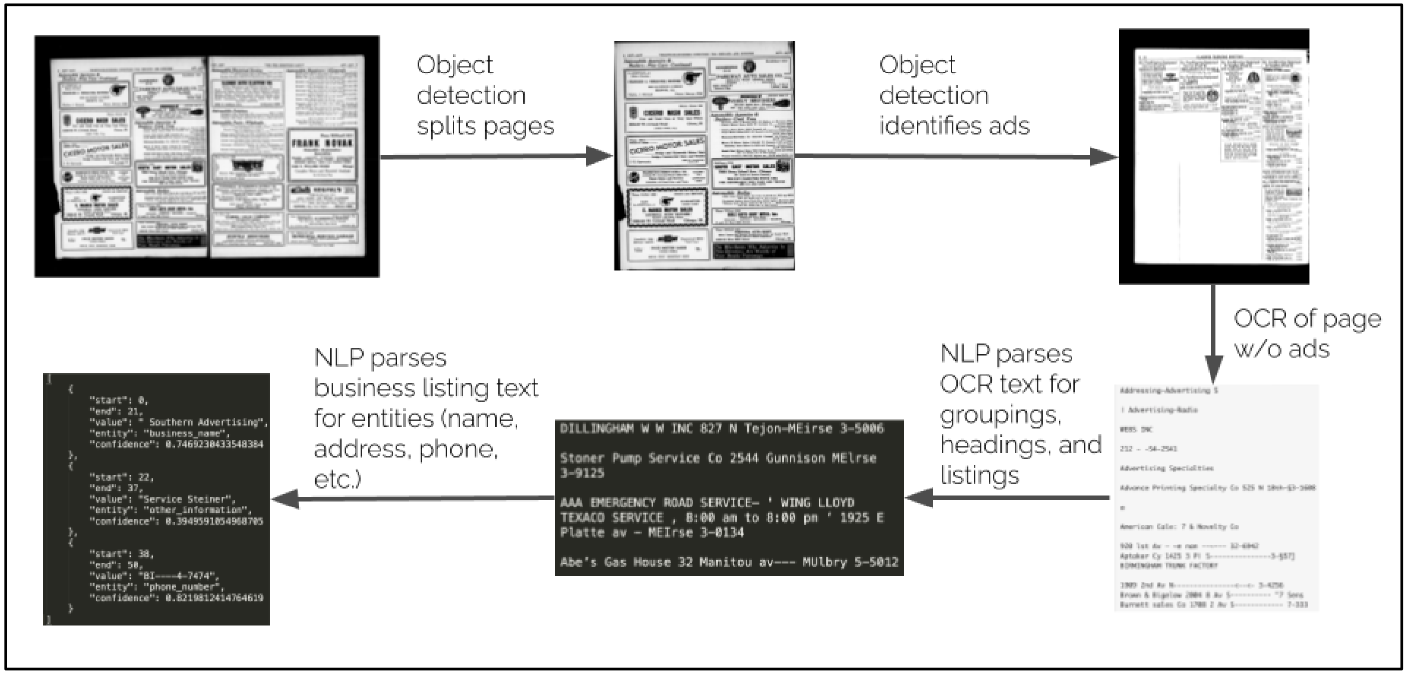


Appendix B. Image 11



```
- [Hotel Pennsylvania Garage](business_name) [39 & Ludlow](address). [EVE rgrn-1122](phone_
- [Howard Garage](business_name) [2310 N Howard](address)... .[REG ent-6532](phone_number)
- [Hunter's Serv Sta](business_name) [Oxfrd av & Verree rd](address).[PIL grm-9949](phone_n
- [Hunting Prk Garage](business_name) [1607 Huntng Prk av](address). [MIC hign-3041](phone_
- [Huntingdon Garage](business_name) [26 & Huntingdon](address) [RAD clf-9540](phone_number
- [Ideal Garage](business_name) [1530 N 27](address) [STE vnsn-7778](phone_number)
- [Imperial Service Garage](business_name) [5945 Locust st](address). [GRA nite-6613](phone
- [Indian Garage](business_name) [2843 W Clearfid](address) [RAD clf-5392](phone_number)
- [Indiana Garage](business_name) [3028 N 6](address).......[SAG amor-2426](phone_number)
- [Integrity Garage](business_name) [4130 Walnut st](address) [BAR ing-4163](phone_number)
- [Internat! Garage](business_name) [6026 Elmwd av](address). [SAR atga-9778](phone_number)
```

Appendix B. Image 12

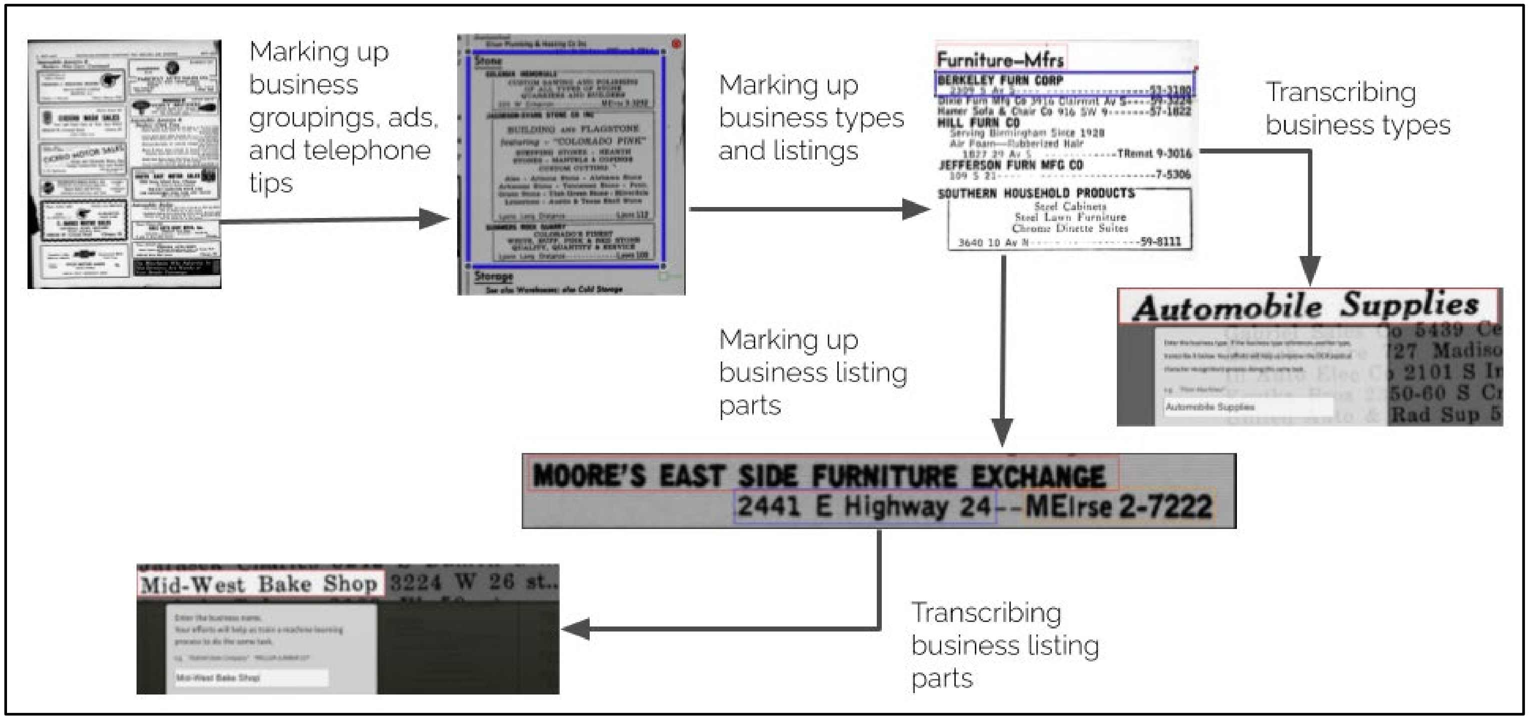

Object detection splits pages
Object detection identifies ads
OCR of page w/o ads
NLP parses OCR text for groupings, headings, and listings
NLP parses business listing text for entities (name, address, phone, etc.)

```
Addressing-Advertising S

| Advertising-Radio

REES INC

212 - -54-2541

Advertising Specialties

Advance Printing Specialty Co 525 N 10th—63-1600

e

American Calo: 7 & Novelty Co

920 1st Av - -e nam ------ 32-6942
Aptaker Cy 1425 3 Pl S---------------3-$37]
BIRMINGHAM TRUNK FACTORY

1909 2nd Av N-----------------+---+- 3-4258
Brown & Bigelow 2004 8 Av S---------- "7 Sons
Barrett sales Co 1708 2 Av S------------- 7-333
```

```
DILLINGHAM W W INC 827 N Tejon-MEirse 3-5006

Stoner Pump Service Co 2544 Gunnison MElrse
3-9125

AAA EMERGENCY ROAD SERVICE- ' WING LLOYD
TEXACO SERVICE , 8:00 am to 8:00 pm ' 1925 E
Platte av - MEIrse 3-0134

Abe's Gas House 32 Manitou av--- MUlbry 5-5012
```

```
[
    {
        "start": 0,
        "end": 21,
        "value": " Southern Advertising",
        "entity": "business_name",
        "confidence": 0.7469230433548384
    },
    {
        "start": 22,
        "end": 37,
        "value": "Service Steiner",
        "entity": "other_information",
        "confidence": 0.3949591054908705
    },
    {
        "start": 38,
        "end": 50,
        "value": "BI----4-7474",
        "entity": "phone_number",
        "confidence": 0.8210812414764619
    }
]
```

Appendix B, Image 13



Marking up business groupings, ads, and telephone tips

Marking up business types and listings

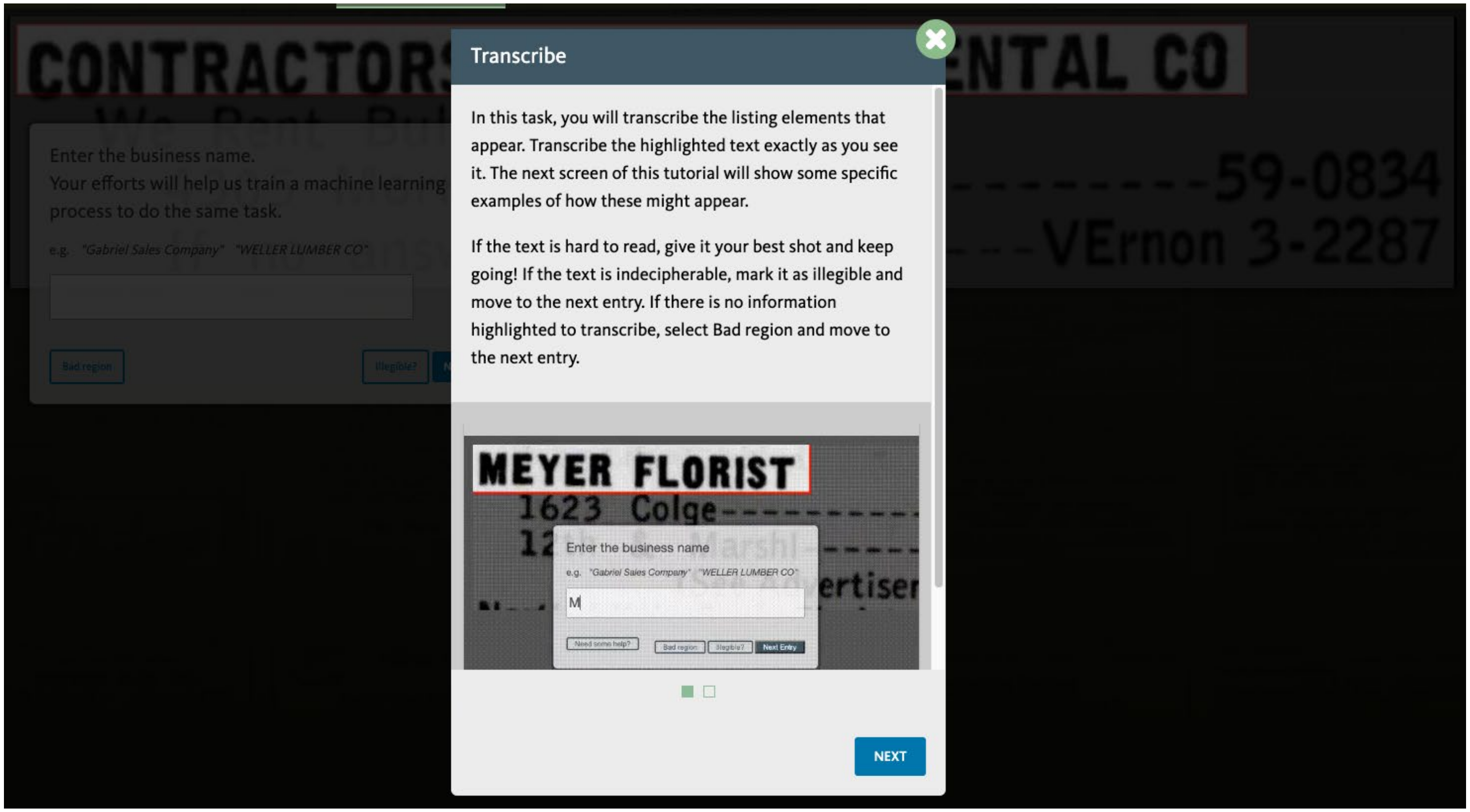
Transcribing business types

Marking up business listing parts

Transcribing business listing parts

Stone
Storage
Furniture–Mfrs
BERKELEY FURN CORP
HILL FURN CO
JEFFERSON FURN MFG CO
SOUTHERN HOUSEHOLD PRODUCTS
Automobile Supplies
MOORE'S EAST SIDE FURNITURE EXCHANGE
2441 E Highway 24--MElrse 2-7222
Mid-West Bake Shop 3224 W 26 st..
Mid-West Bake Shop

Appendix B. Image 14



Transcribe

In this task, you will transcribe the listing elements that appear. Transcribe the highlighted text exactly as you see it. The next screen of this tutorial will show some specific examples of how these might appear.

If the text is hard to read, give it your best shot and keep going! If the text is indecipherable, mark it as illegible and move to the next entry. If there is no information highlighted to transcribe, select Bad region and move to the next entry.

Enter the business name.
Your efforts will help us train a machine learning process to do the same task.

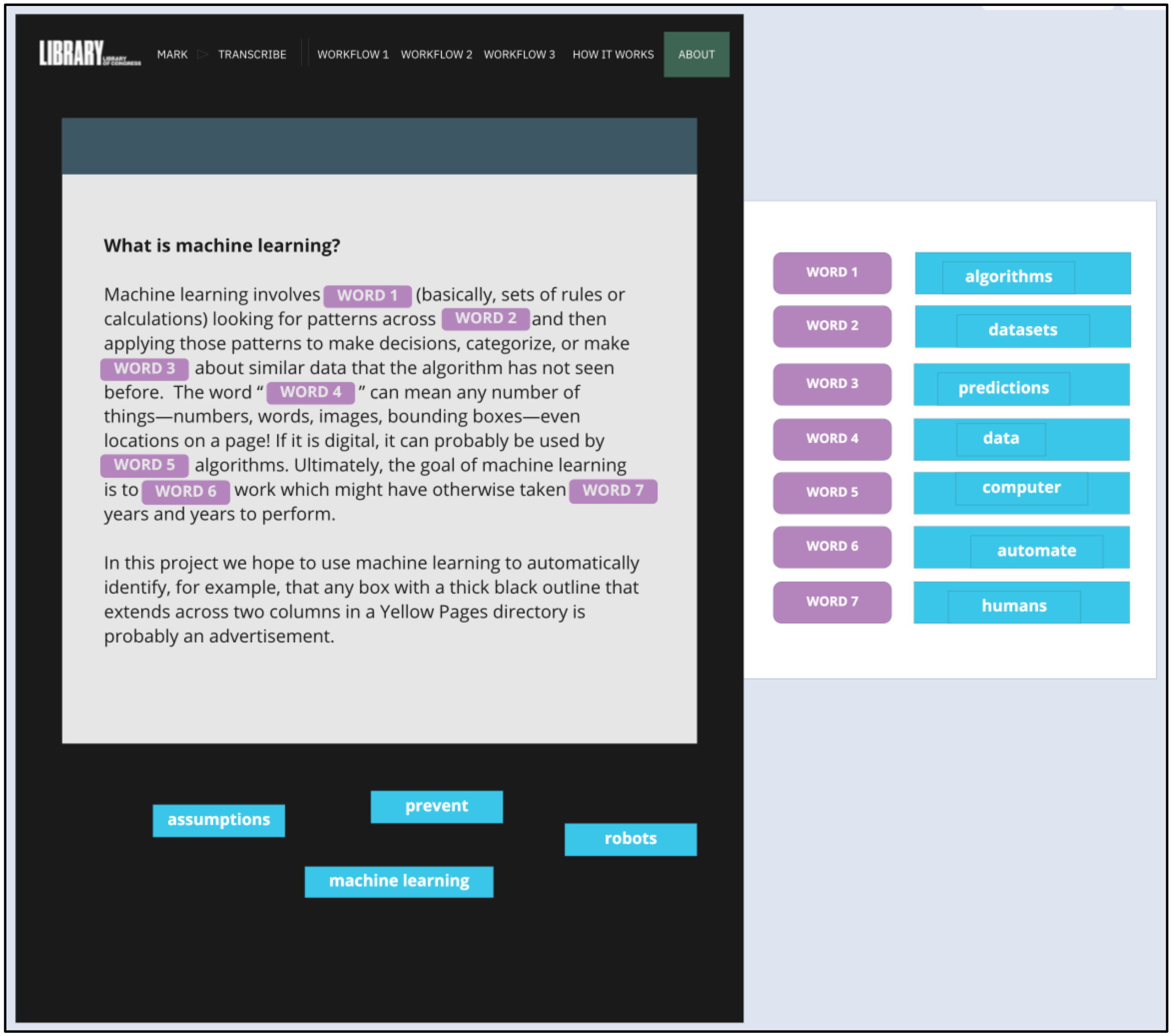
e.g. "Gabriel Sales Company" "WELLER LUMBER CO"

Enter the business name

e.g. "Gabriel Sales Company" "WELLER LUMBER CO"

M

Appendix B. Image 18



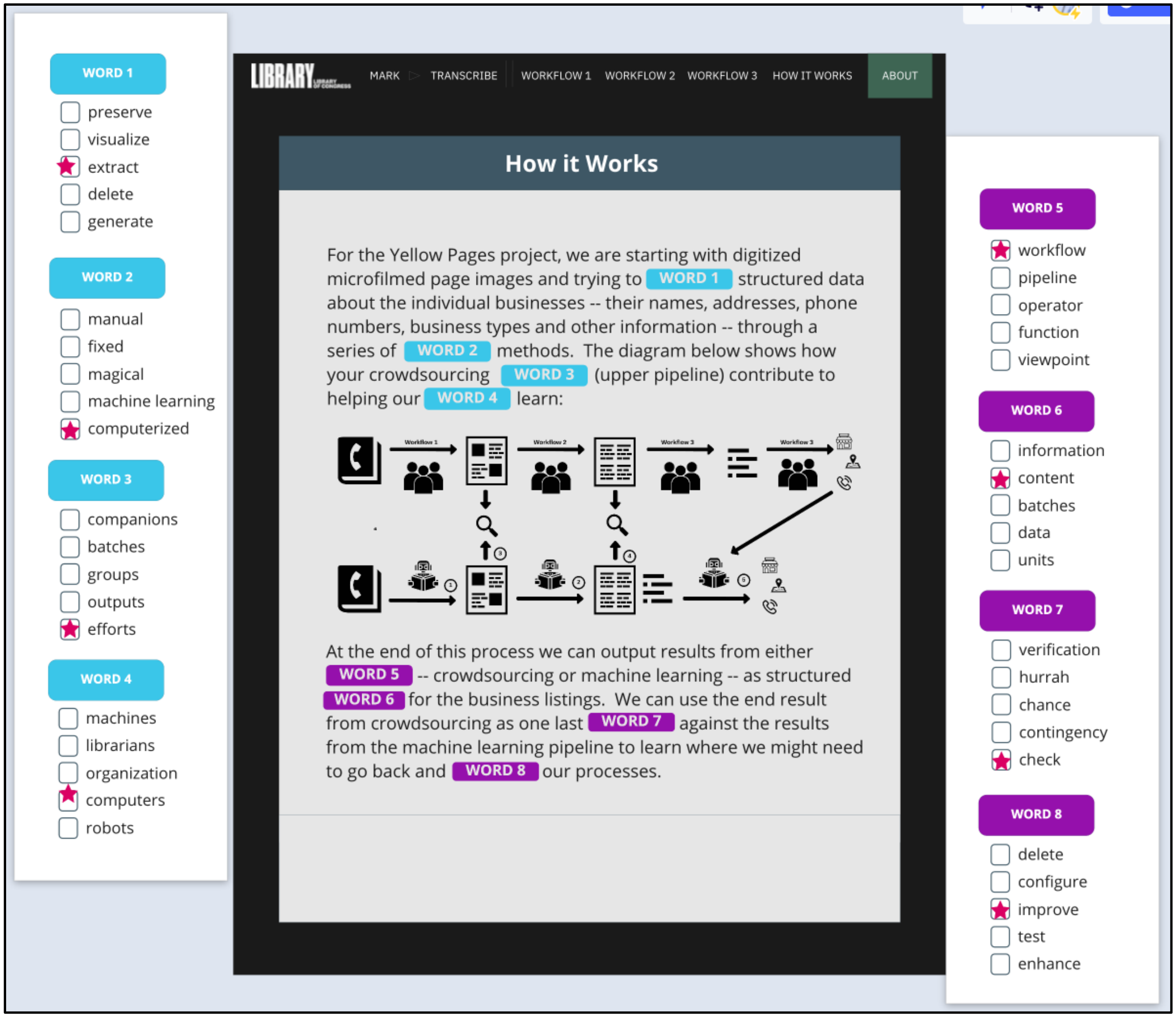
What is machine learning?

Machine learning involves WORD 1 (basically, sets of rules or calculations) looking for patterns across WORD 2 and then applying those patterns to make decisions, categorize, or make WORD 3 about similar data that the algorithm has not seen before. The word " WORD 4 " can mean any number of things—numbers, words, images, bounding boxes—even locations on a page! If it is digital, it can probably be used by WORD 5 algorithms. Ultimately, the goal of machine learning is to WORD 6 work which might have otherwise taken WORD 7 years and years to perform.

In this project we hope to use machine learning to automatically identify, for example, that any box with a thick black outline that extends across two columns in a Yellow Pages directory is probably an advertisement.

WORD 1	algorithms
WORD 2	datasets
WORD 3	predictions
WORD 4	data
WORD 5	computer
WORD 6	automate
WORD 7	humans

Appendix B. Image 20



WORD 1

□ preserve
□ visualize
★ extract
□ delete
□ generate

WORD 2

□ manual
□ fixed
□ magical
□ machine learning
★ computerized

WORD 3

□ companions
□ batches
□ groups
□ outputs
★ efforts

WORD 4

□ machines
□ librarians
□ organization
★ computers
□ robots

How it Works

For the Yellow Pages project, we are starting with digitized microfilmed page images and trying to **WORD 1** structured data about the individual businesses -- their names, addresses, phone numbers, business types and other information -- through a series of **WORD 2** methods. The diagram below shows how your crowdsourcing **WORD 3** (upper pipeline) contribute to helping our **WORD 4** learn:

At the end of this process we can output results from either **WORD 5** -- crowdsourcing or machine learning -- as structured **WORD 6** for the business listings. We can use the end result from crowdsourcing as one last **WORD 7** against the results from the machine learning pipeline to learn where we might need to go back and **WORD 8** our processes.

WORD 5

★ workflow
□ pipeline
□ operator
□ function
□ viewpoint

WORD 6

□ information
★ content
□ batches
□ data
□ units

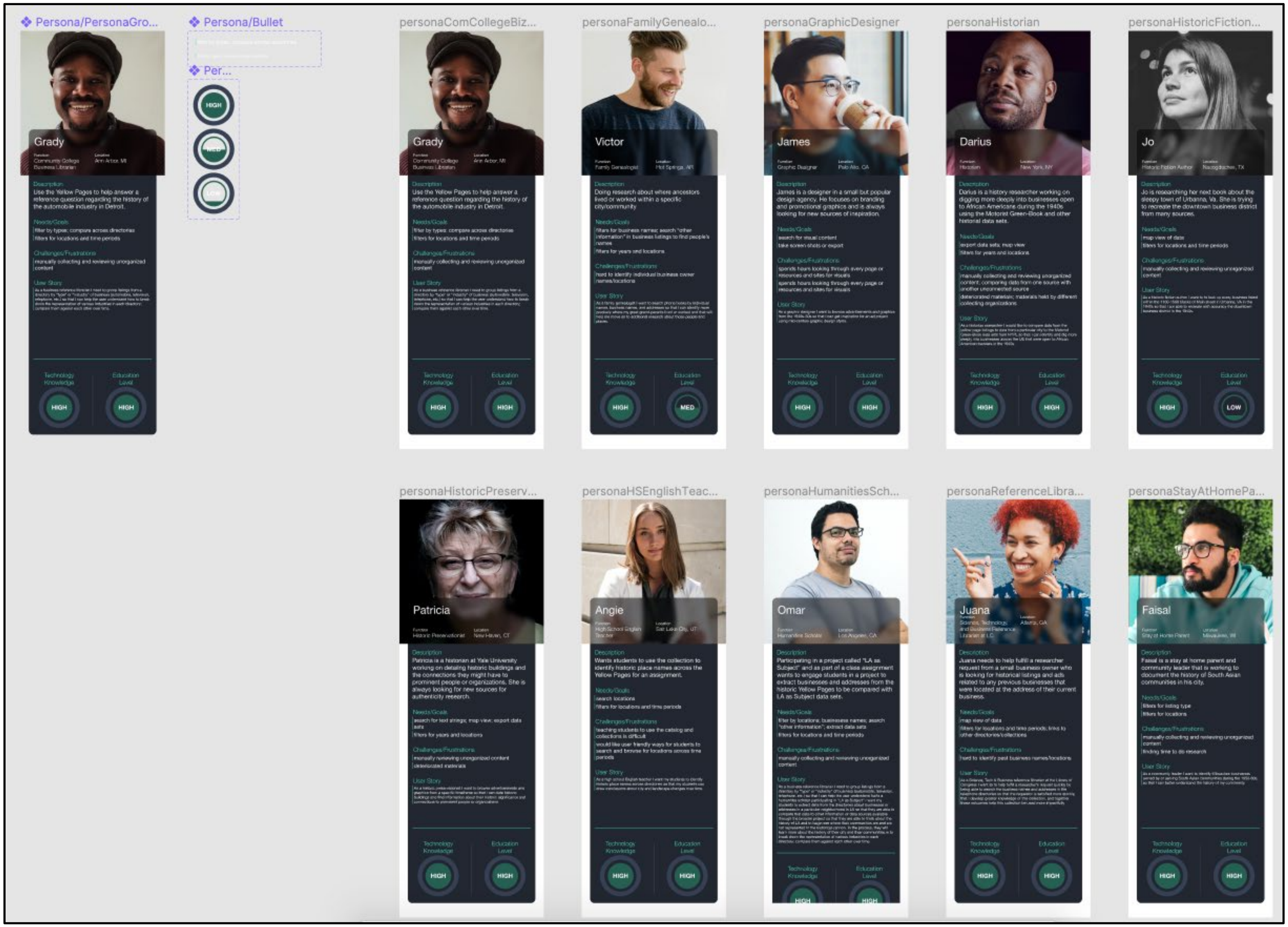
WORD 7

□ verification
□ hurrah
□ chance
□ contingency
★ check

WORD 8

□ delete
□ configure
★ improve
□ test
□ enhance

Appendix B, Image 21



Persona/PersonaGro...

Grady

Function
Community College Business Librarian

Location
Ann Arbor, MI

Description
Use the Yellow Pages to help answer a reference question regarding the history of the automobile industry in Detroit.

Needs/Goals
filter by types; compare across directories
filters for locations and time periods

Challenges/Frustrations
manually collecting and reviewing unorganized content

User Story
As a business reference librarian I need to group listings from a directory by "type" or "industry" of business (automobile, television, telephone, etc.) so that I can help the user understand how to break down the representation of various industries in each directory; compare them against each other over time.

Technology Knowledge
HIGH

Education Level
HIGH

Persona/Bullet

[illegible]

Per...

HIGH

MED

LOW

personaComCollegeBiz...

Grady

Function
Community College Business Librarian

Location
Ann Arbor, MI

Description
Use the Yellow Pages to help answer a reference question regarding the history of the automobile industry in Detroit.

Needs/Goals
filter by types; compare across directories
filters for locations and time periods

Challenges/Frustrations
manually collecting and reviewing unorganized content

User Story
As a business reference librarian I need to group listings from a directory by "type" or "industry" of business (automobile, television, telephone, etc.) so that I can help the user understand how to break down the representation of various industries in each directory; compare them against each other over time.

Technology Knowledge
HIGH

Education Level
HIGH

personaFamilyGenealo...

Victor

Function
Family Genealogist

Location
Hot Springs, AR

Description
Doing research about where ancestors lived or worked within a specific city/community

Needs/Goals
filters for business names; search "other information" in business listings to find people's names
filters for years and locations

Challenges/Frustrations
hard to identify individual business owner names/locations

User Story
As a family genealogist I want to search phone books by individual names, business names, and addresses so that I can identify more precisely where my great grand-parents lived or worked and that will help me move on to additional research about those people and places.

Technology Knowledge
HIGH

Education Level
MED

personaGraphicDesigner

James

Function
Graphic Designer

Location
Palo Alto, CA

Description
James is a designer in a small but popular design agency. He focuses on branding and promotional graphics and is always looking for new sources of inspiration.

Needs/Goals
search for visual content
take screen shots or export

Challenges/Frustrations
spends hours looking through every page or resources and sites for visuals
spends hours looking through every page or resources and sites for visuals

User Story
As a graphic designer I want to browse advertisements and graphics from the 1940s-50s so that I can get inspiration for an ad project using mid-century graphic design styles.

Technology Knowledge
HIGH

Education Level
HIGH

personaHistorian

Darius

Function
Historian

Location
New York, NY

Description
Darius is a history researcher working on digging more deeply into businesses open to African Americans during the 1940s using the Motorist Green-Book and other historical data sets.

Needs/Goals
export data sets; map view
filters for years and locations

Challenges/Frustrations
manually collecting and reviewing unorganized content; comparing data from one source with another unconnected source
deteriorated materials; materials held by different collecting organizations

User Story
As a historian researcher I would like to compare data from the yellow page listings to data from a particular city to the Motorist Green-Book data set from NYPL so that I can identify and dig more deeply into businesses across the US that were open to African American travelers in the 1940s.

Technology Knowledge
HIGH

Education Level
HIGH

personaHistoricFiction...

Jo

Function
Historic Fiction Author

Location
Nacogdoches, TX

Description
Jo is researching her next book about the sleepy town of Urbanna, Va. She is trying to recreate the downtown business district from many sources.

Needs/Goals
map view of data
filters for locations and time periods

Challenges/Frustrations
manually collecting and reviewing unorganized content

User Story
As a historic fiction author I want to be able to look up every business listed within the 1100-1500 blocks of Main street in Urbanna, VA in the 1940s so that I am able to recreate with accuracy the downtown business district in the 1940s.

Technology Knowledge
HIGH

Education Level
LOW

personaHistoricPreserv...

Patricia

Function
Historic Preservationist

Location
New Haven, CT

Description
Patricia is a historian at Yale University working on detailing historic buildings and the connections they might have to prominent people or organizations. She is always looking for new sources for authenticity research.

Needs/Goals
search for text strings; map view; export data sets
filters for years and locations

Challenges/Frustrations
manually reviewing unorganized content
deteriorated materials

User Story
As a historic preservationist I want to browse advertisements and graphics from a specific timeframe so that I can date historic buildings and find information about their historic significance and connections to prominent people or organizations.

Technology Knowledge
HIGH

Education Level
HIGH

personaHSEnglishTeac...

Angie

Function
High School English Teacher

Location
Salt Lake City, UT

Description
Wants students to use the collection to identify historic place names across the Yellow Pages for an assignment.

Needs/Goals
search locations
filters for locations and time periods

Challenges/Frustrations
teaching students to use the catalog and collections is difficult
would like user friendly ways for students to search and browse for locations across time periods

User Story
As a high school English teacher I want my students to identify historic place names across directories so that my students can draw conclusions about city and landscape changes over time.

Technology Knowledge
HIGH

Education Level
HIGH

personaHumanitiesSch...

Omar

Function
Humanities Scholar

Location
Los Angeles, CA

Description
Participating in a project called "LA as Subject" and as part of a class assignment wants to engage students in a project to extract businesses and addresses from the historic Yellow Pages to be compared with LA as Subject data sets.

Needs/Goals
filter by locations; businesses names; search "other information"; extract data sets
filters for locations and time periods

Challenges/Frustrations
manually collecting and reviewing unorganized content

User Story
As a business reference librarian I need to group listings from a directory by "type" or "industry" of business (automobile, television, telephone, etc.) so that I can help the user understand [illegible] As a humanities scholar participating in "LA as Subject" I want my students to extract data from the directories about businesses or addresses in a particular neighborhood in LA so that they are able to compare that data to other information or data sources available through the broader project so that they are able to think about the history of LA and to begin see where their communities are and are not represented in the historical canon. In the process, they will learn more about the history of their city and their communities in to break down the representation of various industries in each directory; compare them against each other over time.

Technology Knowledge
HIGH

Education Level
HIGH

personaReferenceLibra...

Juana

Function
Science, Technology, and Business Reference Librarian at LC

Location
Atlanta, GA

Description
Juana needs to help fulfill a researcher request from a small business owner who is looking for historical listings and ads related to any previous businesses that were located at the address of their current business.

Needs/Goals
map view of data
filters for locations and time periods; links to other directories/collections

Challenges/Frustrations
hard to identify past business names/locations

User Story
As a Science, Tech & Business reference librarian at the Library of Congress I want to be able to help fulfill a researcher's request quickly by being able to search the business names and addresses in the telephone directories so that the requestor is satisfied more quickly; that I develop greater knowledge of the collection, and together these resources help this collection be used more [illegible].

Technology Knowledge
HIGH

Education Level
HIGH

personaStayAtHomePa...

Faisal

Function
Stay at Home Parent

Location
Milwaukee, WI

Description
Faisal is a stay at home parent and community leader that is working to document the history of South Asian communities in his city.

Needs/Goals
filters for listing type
filters for locations

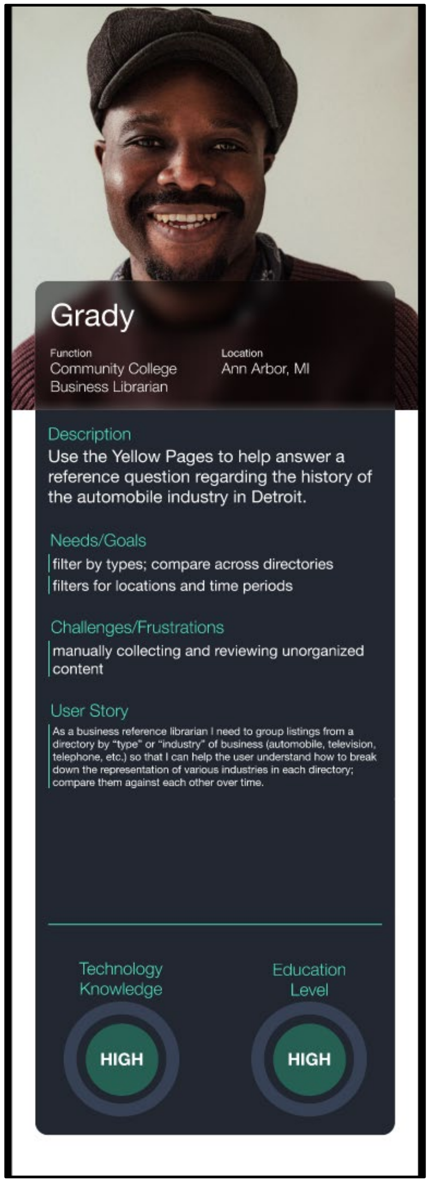
Challenges/Frustrations
manually collecting and reviewing unorganized content
finding time to do research

User Story
As a community leader I want to identify Milwaukee businesses owned by or serving South Asian communities during the 1950-60s so that I can better understand the history of my community.

Technology Knowledge
HIGH

Education Level
HIGH

Appendix B. Image 22



Grady

Function
Community College Business Librarian

Location
Ann Arbor, MI

Description
Use the Yellow Pages to help answer a reference question regarding the history of the automobile industry in Detroit.

Needs/Goals
filter by types; compare across directories
filters for locations and time periods

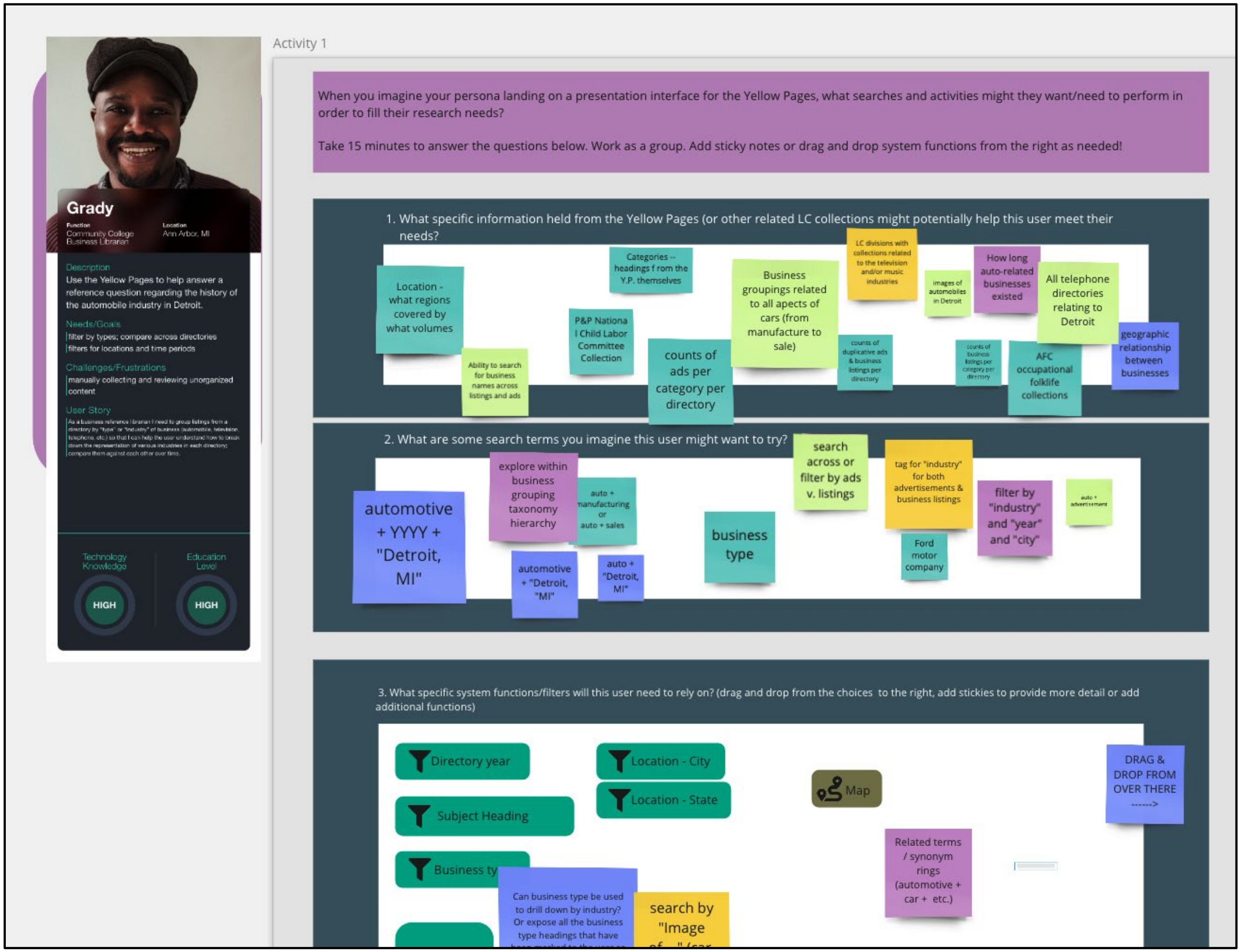
Challenges/Frustrations
manually collecting and reviewing unorganized content

User Story
As a business reference librarian I need to group listings from a directory by "type" or "industry" of business (automobile, television, telephone, etc.) so that I can help the user understand how to break down the representation of various industries in each directory; compare them against each other over time.

Technology Knowledge: HIGH

Education Level: HIGH

Appendix B. Image 23



Grady

Function
Community College Business Librarian

Location
Ann Arbor, MI

Description
Use the Yellow Pages to help answer a reference question regarding the history of the automobile industry in Detroit.

Needs/Goals
filter by types; compare across directories
filters for locations and time periods

Challenges/Frustrations
manually collecting and reviewing unorganized content

User Story
As a business reference librarian I need to group listings from a directory by "type" or "industry" of business (automobile, television, telephone, etc.) so that I can help the user understand how to break down the representation of various industries in each directory; compare them against each other over time.

Technology Knowledge: HIGH

Education Level: HIGH

Activity 1

When you imagine your persona landing on a presentation interface for the Yellow Pages, what searches and activities might they want/need to perform in order to fill their research needs?

Take 15 minutes to answer the questions below. Work as a group. Add sticky notes or drag and drop system functions from the right as needed!

1. What specific information held from the Yellow Pages (or other related LC collections might potentially help this user meet their needs?

- Location - what regions covered by what volumes
- Ability to search for business names across listings and ads
- Categories -- headings f rom the Y.P. themselves
- P&P Nationa l Child Labor Committee Collection
- counts of ads per category per directory
- Business groupings related to all apects of cars (from manufacture to sale)
- LC divisions with collections related to the television and/or music industries
- counts of duplicative ads & business listings per directory
- images of automobiles in Detroit
- How long auto-related businesses existed
- counts of business listings per category per directory
- All telephone directories relating to Detroit
- AFC occupational folklife collections
- geographic relationship between businesses

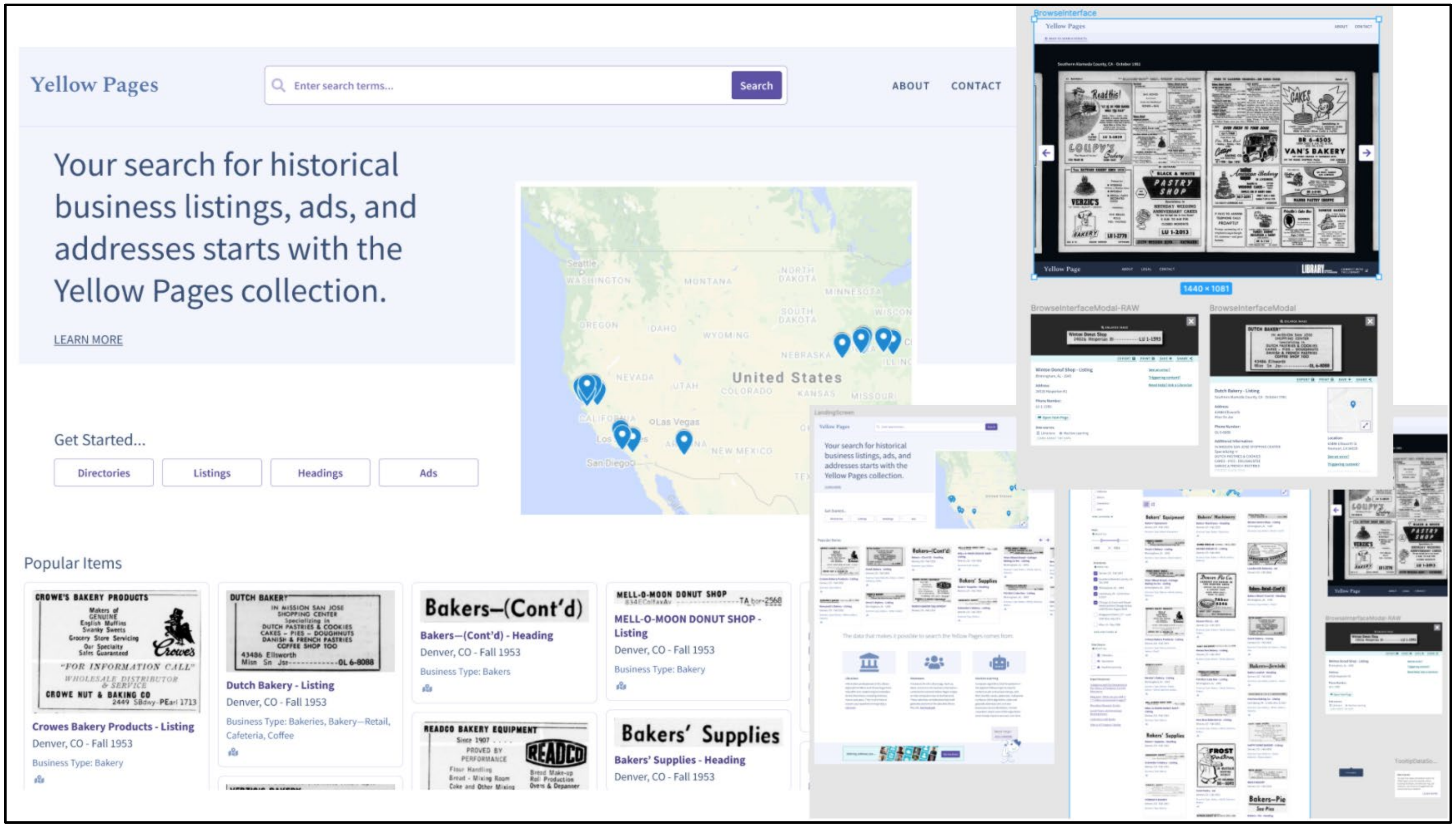
2. What are some search terms you imagine this user might want to try?

- automotive + YYYY + "Detroit, MI"
- explore within business grouping taxonomy hierarchy
- auto + manufacturing or auto + sales
- automotive + "Detroit, "MI"
- auto + "Detroit, MI"
- business type
- search across or filter by ads v. listings
- tag for "industry" for both advertisements & business listings
- Ford motor company
- filter by "industry" and "year" and "city"
- auto + advertisement

3. What specific system functions/filters will this user need to rely on? (drag and drop from the choices to the right, add stickies to provide more detail or add additional functions)

- Can business type be used to drill down by industry? Or expose all the business type headings that have
- search by "Image
- Related terms / synonym rings (automotive + car + etc.)
- DRAG & DROP FROM OVER THERE ------>

Appendix B. Image 24



Yellow Pages

Enter search terms...

Search

ABOUT

CONTACT

Your search for historical business listings, ads, and addresses starts with the Yellow Pages collection.

LEARN MORE

Get Started...

Directories

Listings

Headings

Ads

Popular Items

Crowes Bakery Products - Listing
Denver, CO - Fall 1953
Business Type: Bakery

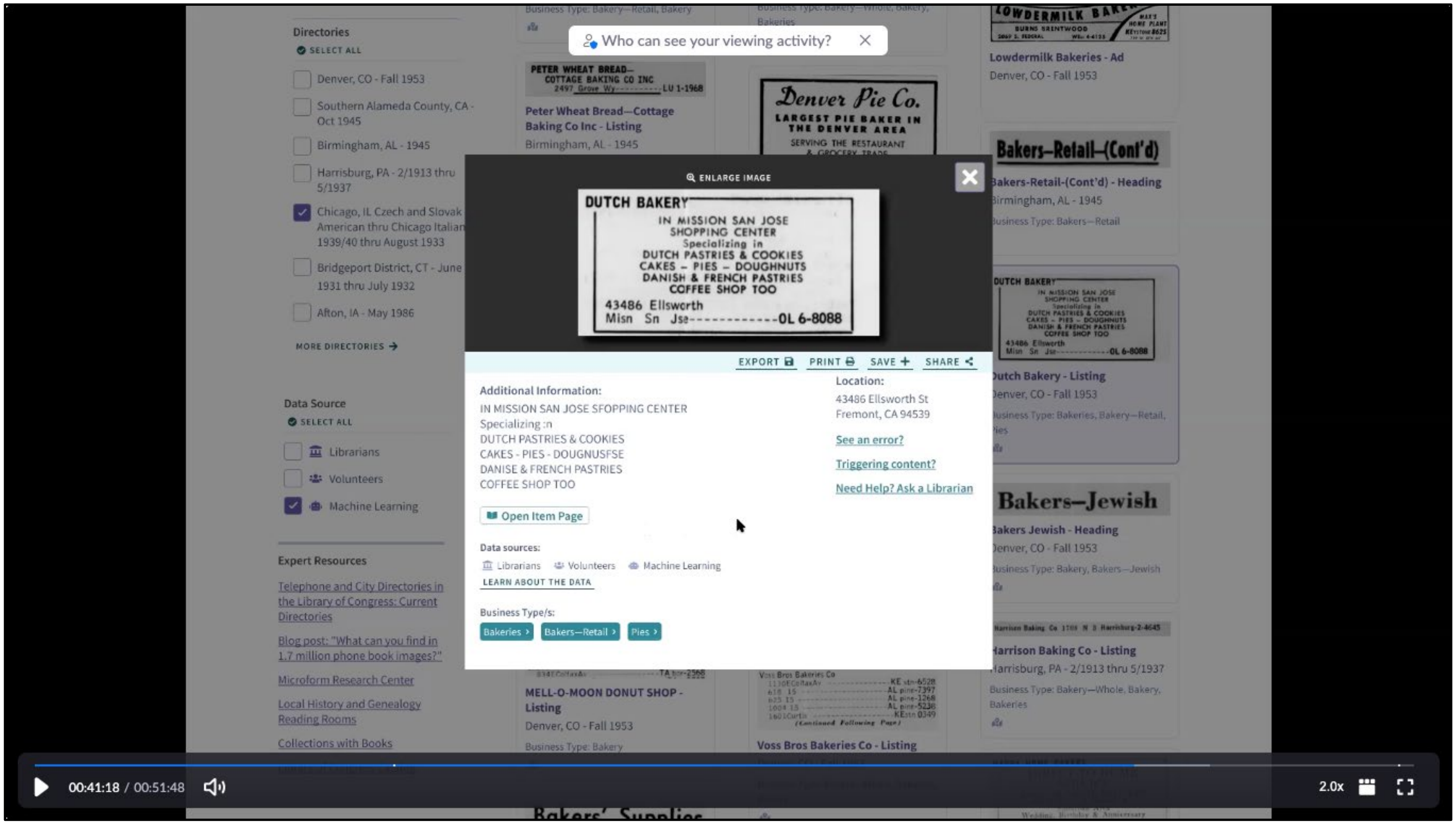
Dutch Bakery - Listing
Denver, CO - Fall 1953
Business Type: Bakeries, Bakery—Retail, Cafeteria, Coffee

Bakers—(Cont'd) - Heading
Denver, CO - Fall 1953
Business Type: Bakers

MELL-O-MOON DONUT SHOP - Listing
Denver, CO - Fall 1953
Business Type: Bakery

Bakers' Supplies - Heading
Denver, CO - Fall 1953

Appendix B. Image 25



Who can see your viewing activity?

Directories

SELECT ALL

□ Denver, CO - Fall 1953

□ Southern Alameda County, CA - Oct 1945

□ Birmingham, AL - 1945

□ Harrisburg, PA - 2/1913 thru 5/1937

☑ Chicago, IL Czech and Slovak American thru Chicago Italian 1939/40 thru August 1933

□ Bridgeport District, CT - June 1931 thru July 1932

□ Afton, IA - May 1986

MORE DIRECTORIES →

Data Source

SELECT ALL

□ Librarians

□ Volunteers

☑ Machine Learning

Expert Resources

Telephone and City Directories in the Library of Congress: Current Directories

Blog post: "What can you find in 1.7 million phone book images?"

Microform Research Center

Local History and Genealogy Reading Rooms

Collections with Books

ENLARGE IMAGE

DUTCH BAKERY
IN MISSION SAN JOSE
SHOPPING CENTER
Specializing in
DUTCH PASTRIES & COOKIES
CAKES – PIES – DOUGHNUTS
DANISH & FRENCH PASTRIES
COFFEE SHOP TOO
43486 Ellsworth
Misn Sn Jse------------OL 6-8088

EXPORT PRINT SAVE SHARE

Additional Information:
IN MISSION SAN JOSE SFOPPING CENTER
Specializing :n
DUTCH PASTRIES & COOKIES
CAKES - PIES - DOUGNUSFSE
DANISE & FRENCH PASTRIES
COFFEE SHOP TOO

Open Item Page

Data sources:
Librarians Volunteers Machine Learning

LEARN ABOUT THE DATA

Business Type/s:
Bakeries › Bakers—Retail › Pies ›

Location:
43486 Ellsworth St
Fremont, CA 94539

See an error?

Triggering content?

Need Help? Ask a Librarian

Peter Wheat Bread—Cottage Baking Co Inc - Listing
Birmingham, AL - 1945

Lowdermilk Bakeries - Ad
Denver, CO - Fall 1953

Bakers-Retail-(Cont'd) - Heading
Birmingham, AL - 1945
Business Type: Bakers—Retail

Dutch Bakery - Listing
Denver, CO - Fall 1953
Business Type: Bakeries, Bakery—Retail, Pies

Bakers Jewish - Heading
Denver, CO - Fall 1953
Business Type: Bakery, Bakers—Jewish

Harrison Baking Co - Listing
Harrisburg, PA - 2/1913 thru 5/1937
Business Type: Bakery—Whole, Bakery, Bakeries

MELL-O-MOON DONUT SHOP - Listing
Denver, CO - Fall 1953
Business Type: Bakery

Voss Bros Bakeries Co - Listing

00:41:18 / 00:51:48 2.0x

Appendix B. Image 26

Appendix C: Yellow Pages data model

Business rules

The Library of Congress U.S. Telephone Directory **Collection** contains **Phone Books** digitized from microfilm that may contain either **White Pages** (individual listings) and/or **Yellow Pages** (business listings). Digitized microfilm **Images** represent 1-2 phone book **Pages**, microfilm technical targets, or frames of explanatory material (such as indicators for where the White Pages or Yellow Pages sections start, where material is missing, or metadata for the original object).

Pages of the Yellow Pages are usually divided into 2-4 **Columns**, though sometimes advertisements may span several columns, often creating shorter columns. Columns usually include **Groupings** of **Business Listings**, **Advertisements**, or **Tips/Information** about using the phone book. Business groupings are organized by business or service type displayed in a larger font above the listings. Business listings may be **Informational Listings** that give additional information about the business, sometimes in the form of an advertisement, with graphical elements. Information listings are usually set apart from standard listings with boxes or bounded by horizontal lines.

Business listings include information about the business, including the name, address, and phone number and, in some cases, additional information such as hours, specific services offered, or date the business was established. Some businesses, such as hotels, restaurants, and hospitals, may include indicators of racial segregation or culturally offensive imagery.

Businesses listed in the Yellow Pages may have one or more listings under the business type headings that apply to them. They may have one or more addresses and phone numbers that reference one or more business listings. They may also be associated with advertisements outside of business listing blocks with additional information about the business.

Entity definitions

LC Collection

An LC Collection is an online Library of Congress digital collection that follows the Library of Congress digital object model for consistently identifying and addressing digital objects over the Internet.

Item

An LC digital object--member of an LC Collection that is made up of one or more digital assets. In the Yellow Pages model, this represents a sequence of digitized assets representing one or more microfilmed phone books.

Phone Book

A directory of individuals or businesses (published in the United States).

White Pages (subclass)

A phone book containing listings of individuals.

Yellow Pages (subclass)

A phone book containing listings and advertisements of businesses, organized by business or service.

Image

An LC digital asset contained within an item, representing a frame of microfilm. Digitized microfilm Images represent 1-2 phone book pages, microfilm technical targets, or frames of explanatory material (such as indicators for where the White Pages or Yellow Pages sections start, where material is missing, or metadata for the original object).

Page

A single page (front or back) of a phone book. There are usually two pages on an image, but not always.

Column

A column of information on a phone book page usually spanning from top to bottom but sometimes interrupted by an advertisement spanning multiple pages. A column may contain any type of segment.

Segment

An area of information on a page or in a column, separated into a semantic category.

Business Grouping (subclass)

A segment of business listings grouped by business type displayed in a larger font above the listings.

Advertisement (subclass)

A textual or textual/graphical segment of information that may exist outside of a block or be contained within it. May also be an informational listing within a block of business listings/

Telephone Tip (subclass)

A tip for using the telephone/telephone directory or another piece of helpful information from the phone book publisher.

Business Listing

Information about a business, such as name, address, phone number, and services listed alphabetically within a block.

Informational Listing (subclass)

A business listing that gives additional information about the business, sometimes in the form of an advertisement, with graphical elements. Information listings are usually set apart from standard listings with boxes or horizontal lines.

Business

A business is a local service or business listed in the Yellow Pages. Businesses may have one or more listings under the business type headings that apply to them. They may have one or more addresses and phone numbers that reference one or more business listings. They may also be associated with advertisements outside of business listing blocks with additional information about the business.

Data dictionary

Definitions and datatypes of the entities listed above. Only properties relevant to this Humans-in-the-Loop initiative are included.

Property	Definition	Datatype
LC Collection		
identifier	LC identifier for the collection, for use in API calls. (‘usteledirec’ is identifier for Telephone Directory collection)	string
Item		
identifier	LC identifier for the item, for use in API calls.	string
title	Title of the item	string
dates	Date range of published dates of the phone books in the item	string
locations	Geographic locations of the phone books in the item	[string]
Phone Book		
title	Title of the phone book	string
year	Year the phone book was published	date
numImages	Number of images on which the phone book is contained	integer

numPages	Number of pages in the phone book	integer
Yellow Pages (subclass of Phone Book)		
title	Title of the phone book (inherited from Phone Book)	string
year	Year the phone book was published (inherited from Phone Book)	date
numImages	Number of images on which the phone book is contained (inherited from Phone Book)	integer
numPages	Number of pages in the phone book (inherited from Phone Book)	integer
startImage	Item image where this Yellow Pages starts	integer
endImage	Item image where this Yellow Pages ends	integer
pageStart	Page number in the phone book where this Yellow Pages starts	string?
pageEnd	Page number in the phone book where this Yellow pages ends	string?
location	Name of the city, town, or area that this phone book covers	string
geographic location	A controlled term for the city and state of this phone book	string
publisher	Publisher of the phone book	string
Image		
identifier	LC identifier for the image, for use in API calls	string
filename	LC filename of the image	string
file type	File type of the image	string

sequence	Sequence of the image within the Yellow Pages	integer
Page		
identifier	Identifier given to the page in relationship to all pages in the Yellow Pages	string
pageNum	Page number as listed on the page	string
sequence	Sequence of the page within the Yellow Pages	integer
imageCoords	Coordinates of the page within the image	coordinates
Column		
identifier	Unique identifier given to the column	string
name	Name given to the column based on its layout in relationship to other columns on the image (page? both?)	string
image	Image the column appears on	Image (identifier)
imageCoords	Coordinates of the column within the image	coordinates
page	Page the column appears on	Page (identifier)
Segment		
identifier	Unique identifier given to the segment	string
image	Image the column appears on	Image (identifier)
imageCoords	Coordinates of the segment within the image	coordinates
columnCoords	Coordinates of the segment within the column	coordinates
page	Page the segment appears on	Page (identifier)

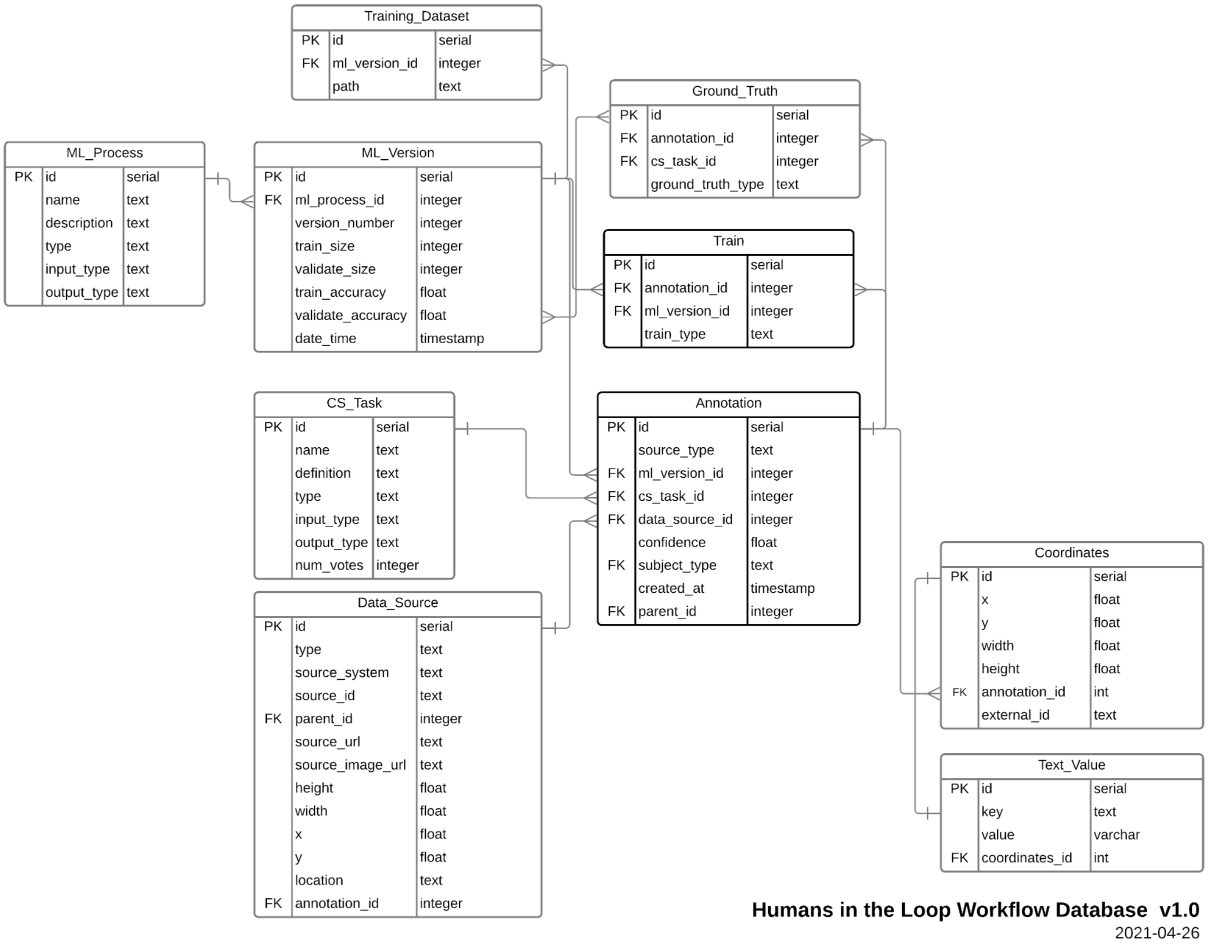
Business Grouping (subclass of Segment)		
identifier	Unique identifier given to the segment	string
image	Image the column appears on	Image (identifier)
imageCoords	Coordinates of the segment within the image	coordinates
column	Column the segment appears on	Column (identifier)
columnCoords	Coordinates of the segment within the column	coordinates
page	Page the segment appears on	Page (identifier)
businessType	Transcription of header of block	string
Advertisement (subclass of Segment)		
identifier	Unique identifier given to the segment	string
image	Image the column appears on	Image (identifier)
imageCoords	Coordinates of the segment within the image	coordinates
column	Column the segment appears on	Column (identifier)
columnCoords	Coordinates of the segment within the column	coordinates
page	Page the segment appears on	Page (identifier)
listing	Business listing the advertisement is associated with	Listing
ocr	OCR of the advertisement	OCR
Info (subclass of Segment)		
identifier	Unique identifier given to the segment	string
image	Image the column appears on	Image (identifier)
imageCoords	Coordinates of the segment within the image	coordinates

column	Column the segment appears on	Column (identifier)
columnCoords	Coordinates of the segment within the column	coordinates
page	Page the segment appears on	Page (identifier)
ocr	OCR of the info box	OCR
Business Listing		
identifier	Unique identifier given to the listing	string
image	Image the listing appears on	Image (identifier)
image coordinates	Coordinates of the listing within the image	coordinates
segment	Segment the listing appears in	Segment (identifier)
segment coordinates	Coordinates of the listing within the segment	coordinates
ocr	OCR of the listing	OCR
text	Plain text of the listing	string
annotations	Tagged parts of the listing, including name, address, phone number, other properties as needed along with OCR coordinates of teach	object
business type	The heading of the business grouping the business appears under	string
Informational Listing (subclass of Business Listing)		
identifier	Unique identifier given to the listing	string
image	Image the listing appears on	Image (identifier)
image coordinates	Coordinates of the listing within the image	coordinates
segment	Segment the listing appears in	Segment (identifier)
segment	Coordinates of the listing within the	coordinates

coordinates	segment	
ocr	OCR of the listing	OCR
annotations	Tagged parts of the listing, including name, address, phone number, other properties as needed along with OCR coordinates of each	object
Business		
identifier	Identifier given to the business (unique to the specific instance of Yellow Pages)	string
name	Normalized name of the business	string
address	Normalized address(es) of the business	[string]
phone number	Normalized phone number(s) of the business	[string]
listing	Associated business listings	[Business Listing (identifier)]
advertisements	Associated advertisements	[Advertisement (identifier)]
Business types	List of business types the business' listings appears under	[string]
Other properties as needed		

Appendix D: Workflow database ER diagram

Entity-relationship diagram for the workflow database.



Training_Dataset

PK	id	serial
FK	ml_version_id	integer
	path	text

Ground_Truth

PK	id	serial
FK	annotation_id	integer
FK	cs_task_id	integer
	ground_truth_type	text

ML_Process

PK	id	serial
	name	text
	description	text
	type	text
	input_type	text
	output_type	text

ML_Version

PK	id	serial
FK	ml_process_id	integer
	version_number	integer
	train_size	integer
	validate_size	integer
	train_accuracy	float
	validate_accuracy	float
	date_time	timestamp

Train

PK	id	serial
FK	annotation_id	integer
FK	ml_version_id	integer
	train_type	text

CS_Task

PK	id	serial
	name	text
	definition	text
	type	text
	input_type	text
	output_type	text
	num_votes	integer

Annotation

PK	id	serial
	source_type	text
FK	ml_version_id	integer
FK	cs_task_id	integer
FK	data_source_id	integer
	confidence	float
FK	subject_type	text
	created_at	timestamp
FK	parent_id	integer

Coordinates

PK	id	serial
	x	float
	y	float
	width	float
	height	float
FK	annotation_id	int
	external_id	text

Data_Source

PK	id	serial
	type	text
	source_system	text
	source_id	text
FK	parent_id	integer
	source_url	text
	source_image_url	text
	height	float
	width	float
	x	float
	y	float
	location	text
FK	annotation_id	integer

Text_Value

PK	id	serial
	key	text
	value	varchar
FK	coordinates_id	int

Humans in the Loop Workflow Database v1.0

2021-04-26

Appendix E: Workflow database data dictionary

Definitions and constraints for entities, properties, and relationships in the workflow database.

[see HITL GitHub repository]

Appendix F: Code repository

Data, code, and other design and development artifacts of the HITL initiative.

Contents include:

- *crowdsourcing-data-flow-scripts:* Python scripts for managing data flow between Scribe and the workflow database
- *machine-learning-scripts:* Python scripts for initializing the PostgreSQL workflow database and running the machine learning pipeline
- *sample-output-data:* sample structured data for business listings and Python code for generating it from the workflow database
- *scribe-hitl:* a version of the Scribe platform customized for the HITL initiative
- *workflow-database:* documentation and initialization scripts for the workflow database

[see HITL GitHub repository]

Appendix G: Machine learning and Crowdsourcing data flows

Database inputs and outputs for ML and crowdsourcing pipelines.

Machine learning pipeline

ML1. Split images into pages — OpenCV algorithm

- Download page images from LC and relevant metadata
    - Write to database Data_Source table:
        - Parent digital object:
            - name: [title/label]
            - type: digital object
            - source_system: lc
            - source_id: [item id]
            - source_url: [url]
        - Digital object images
            - name: [filename]
            - type: digital object image
            - source_sytem: lc
            - source_id: [page id (item id _ zero-filled page number)]
            - source_url: [url]
            - source_image_url: [iiif url]

      - location: [local location of image]
      - height: [height]
      - width: [width]
      - parent_id: [digital object db id]
- Input page images
- Output coordinates (x, y, width, height)
  - Write to database Annotation table:
    - source_type: digital object image
    - ml_version_id: [db id of the ML version used to detect pages]
    - data_source_id: [db id from Data_Source of the image]
    - confidence: [if available]
    - created_at: [timestamp]
    - subject_type: page
  - Write to database Coordinates table:
    - annotation_id: [id from the Annotation]
    - x
    - y
    - width
    - height

ML2. Identify advertisements in pages — OpenCV

- Download yp pages from LC IIIF with coordinates from above
  - If using page images, see CS1 below for writing page images to Data_Source table
- Input pages
- Run script (includes pre-processing of images)
- Output coordinates of ads
  - Write to Annotation table:
    -
- QC against ground truth from C1

ML3. OCR of pages — Tesseract

- Blank out advertisements in downloaded pages
- Run OCR on downloaded yp pages
  - Write to Data_Source table
- Output OCR text/coordinates
  - Write to Annotation table:
    - Subject_type: page ocr
    - Confidence: Mean word level confidence
    - Ml_version_id: Ocr version
    - Data_source_id: Page data source
    - location<new>: Local/Shared path

- Create new coordinates: Page coordinates

ML4. Parsing OCR for business groupings and listings

- Input annotations with subject_type of “page ocr”
- Create new data source
  - Name: file name
  - Type: page ocr
  - Source_system: ml
  - Source_id: null
  - Parent_id: page data_source_id
  - Height, width, x, y: same as page data source
  - Location: location from annotation
  - annotation_id : id from page ocr annotation
- Output identified business groupings and listings within, include coords and text
  - Business Grouping:
    - Write to Annotation table:
      - source_type: page OCR
      - ml_version_id: [db id of the ML version used]
      - data_source_id: [db id from Data_Source of the page OCR]
      - confidence: null
      - created_at: [timestamp]
      - subject_type: business grouping
    - Write to database Coordinates table:
      - annotation_id: [id from the Annotation]
      - x
      - y
      - width
      - Height
  - Listings/business grouping types:
    - Write to Annotation table:
      - source_type: page OCR
      - ml_version_id: [db id of the ML version used for parsing script]
      - data_source_id: [db id from Data_Source of the page OCR]
      - confidence: mean confidence of listing word-level OCR
      - created_at: [timestamp]
      - subject_type: “business listing” OR “business type”
      - parent_id<new>: Business grouping annotation id
    - Write to database Coordinates table:
      - annotation_id: [id from the Annotation]
      - x
      - y
      - width
      - Height

        - Write to database Text_Value table:
            - id
            - key: “business listing” or “business_type_text”
            - value: [text of the listing or grouping type]
            - coordinates_id: [corresponding id from Coordinates]
- QC against ground truth from C1 (grouping)

ML5. CRF NLP

- (Training from crowdsourcing)
- Run NLP on each business listing (separate CRF for business grouping type?)
- Output structured data for listing associated with parent business grouping heading
    - Write to Annotation table:
        - source_type: business listing OR business grouping type
        - ml_version_id: [db id of the ML version used]
        - Data_source_id: ?
        - confidence: [if available]
        - created_at: [timestamp]
        - subject_type: structured business listing OR structured business grouping type
    - Write to Coordinates table (for each entity):
        - annotation_id: [id from the Annotation]
        - x
        - y
        - width
        - Height
    - Write to Text_Value table (for each entity):
        - id
        - key: [respective entity type, e.g. “business name”]
        - value: [text value]
        - coordinates_id: [corresponding id from Coordinates]

OCR advertisements (optional)

Crowdsourcing pipeline

CS1. Segment pages into business groupings, ads, telephone tips

- Input pages from IIIF server based on coordinates from ML 1
    - Select all pages for a specific phone book
    - Write to Data_Source table (if not already in table):
        - name: parent id + ‘_a’ or ‘_b’ (a=left, b=right)
        - type: page
        - source_system: lc

        - source_id: [do page id (item id + zero-filled page number)]
        - source_url: [image url]
        - source_image_url: [iiif url with region from x, y, w, h]
        - location:
        - x: [x]
        - y: [y]
        - height: [height]
        - width: [width]
        - parent_id: [image db id]
        - annotation_id: Annotation id from which page coordinates were derived
    - Create group csv files for workflow 1

CS2. Segment business listings from business groupings

- Output segment coordinates from CS1 to use for C2 and as ground truth for ML 2 and ML 4
    - Write to Annotation table (if mongo id doesn’t exist in Coordinates table):
        - source_type: page
        - cs_task_id: [db id of the CS task]
        - data_source_id: [db id from Data_Source of the page]
        - created_at: [timestamp]
        - subject_type: [advertisement, business grouping, or telephone tip]
    - Write to Coordinates table (if mongo id doesn’t exist in Coordinates table):
        - annotation_id: [id from the Annotation]
        - x
        - y
        - width
        - height
        - external_id: [mongo id]
- Input business groupings from IIIF server based on coordinates from C1 and/or ML4
    - Select all business groupings for a specific phone book
    - Write to Data_Source table:
        - name: parent page id + ‘_’ + x + ‘_’ + y
        - type: business grouping
        - source_system: lc
        - source_id: [do page id (item id + zero-filled page number)]
        - source_url: [do page url]
        - source_image_url: [iiif url with region from x, y, w, h]
        - location:
        - height: [height]
        - width: [width]
        - x: [x]
        - y: [y]
        - parent_id: [page db id]

        - annotation_id: Annotation id from which page coordinates were derived

CS3-4. Identify and transcribe entities in business listings

- Output coordinates for business listings from CS2 to use for C3 and as ground truth for ML4 (business listings)
    - Write to Annotation table
        - source_type: business grouping
        - cs_task_id: [db id of the CS task]
        - data_source_id: [db id from Data_Source of the page]
        - created_at: [timestamp]
        - subject_type: [business listing or business type]
    - Write to Coordinates table:
        - annotation_id: [id from the Annotation]
        - x
        - y
        - width
        - height
        - external_id: [mongo id]
    - Write to TextValue table (if subject_type is business type):
        - key: [business type or business type reference]
        - value: [transcription value]
        - coordinates_id: [id from the Coordinates table]
        - external_id: [mongo id]
    - Write to TextValue table (if subject_type is business listing):
        - key: [business name, address, phone number, graphic, see advertisement, or other information]
        - value: [transcription value]
        - coordinates_id: [id from the Coordinates table]
        - external_id: [mongo id]
- Input business listings from IIIF server based on coordinates from C2 and/or ML4
- Output training data in spacy markdown format for ML5 training

Appendix H: Collection Candidate Evaluation Sheet

HITL Collection Candidates evaluation spreadsheet including separate tabs with defined user stories, risks and mitigation strategies, and Yellow Pages tasks.
[see HITL GitHub repository]

Appendix I: Crowdsourcing Prototype User Testing Plan & Discussion Guide

User testing plan and guide for the HITL Crowdsourcing Prototype.
[see HITL GitHub repository]

Appendix J: Crowdsourcing User Testing Data

Compiled feedback from user testing of the HITL Crowdsourcing Prototype.
[see HITL GitHub repository]

Appendix K: Crowdsourcing Prototype Usability Survey Questions & Responses

Responses to usability survey questions from staff users adding ground-truth data to the crowdsourcing prototype.
[see HITL GitHub repository]

Appendix L: Presentation Interface Wireframes

Still image captures of the HITL presentation interface wireframes from original mock-up.
[see HITL GitHub repository]

Appendix M: Presentation Interface User Testing Plan & Discussion Guide

User testing plan and guide for the HITL presentation interface wireframes.
[see HITL GitHub repository]

Appendix N: Presentation Interface User Test Data

Compiled feedback from user testing of the HITL Crowdsourcing Prototype.
[see HITL GitHub repository]

Appendix O: Sample Structured Data

Structured output data for business listings generated by ML and crowdsourcing processes for the four Yellow Pages directories used in the experiment.
[see HITL GitHub repository]

Appendix P: Crowdsourcing Prototype Screen Captures

Screenshots and recordings of the crowdsourcing prototype workflows.
[see HITL GitHub repository]