ABOUT THE COLLECTION

The Library of Congress's digitized U.S. Telephone Directory Collection represents the following states and localities: Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, the District of Columbia, Florida, Georgia, Hawaii, Iowa, Maryland, Pennsylvania, and the city of Chicago. The dates of the directories span most of the 20th century. The Library's United States telephone directory collection consists of 8,327 digitized reels of microfilm; of these, about 3,500 are presented in the digital collection. The remainder of the collection may be requested from the Microform Reader Services (LJ 139).

This project and website use selected Yellow Pages from the U.S. Telephone Directory Collection as the basis for research, so you might see mid-century directories from your home state as you are tagging and enhancing the data on this site. These are just a small sample of the full collection of over 8,000 digitized reels of microfilm!

Learn more about the collection by connecting with a specialist through the Library's Ask A Librarian service.

ABOUT THE PROJECT

Through this website and the overarching project, called Humans in the Loop (HITL), the Library of Congress is looking for new ways to generate and enrich metadata across its vast digital collections, beginning with the Yellow Pages.

In recent years, the Library has experimented with crowdsourcing and machine learning methods separately, only to discover challenges specific to each approach. Crowdsourcing does not work as an approach for all data enrichment tasks at scale, while machine learning models need significant amounts of training to achieve acceptable measures of accuracy. Designing crowdsourcing tasks that create training data for machine learning methods could be a solution to both of these challenges, potentially enabling human users to improve machine learning processes at scale.

HITL builds on the foundation of the Library's success with past crowdsourcing and machine learning initiatives, and seeks to address the challenges inherent in each approach. Using the Yellow Pages as a test case, HITL is modeling, testing, and evaluating various crowdsourcing and machine learning methods that aim to ethically enhance usability, discovery, and user engagement with the Library's digital collections.

One of the HITL project outputs is this website! This site includes experimental workflows that we're asking users to test directly through tagging and transcribing of the Yellow Pages.

WHAT IS CROWDSOURCING?

Crowdsourcing is the practice of engaging a group of people for a common goal — in this case, enhancing metadata about the Library's US Telephone Directory Collection through the tasks of marking up, labeling, tagging, and transcribing.

As a crowdsourcing volunteer, the data you create in the activities on this site will be used in a few different ways. Some of the "mark" tasks will help verify data created by machine learning processes and improve their accuracy, and others will be used as training data to teach a computer how to identify patterns in text. In either case, your work will be used to help generate and enrich the data that drives access to the Library's collections.

"Transcribe" tasks will be used to help check the results of machine learning processes, especially in recognizing the text on Yellow Pages directories; your transcriptions will also create training data for computers to interpret text.

WHAT IS MACHINE LEARNING?

Machine learning involves algorithms (basically, sets of rules or calculations) looking for patterns across datasets and then applying those patterns to make decisions, categorize, or make predictions about similar data that the algorithm has not seen before. The word "data" 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 machine learning algorithms. Ultimately, the goal of machine learning is to automate work which might have otherwise taken humans 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.

WHAT DO WE HOPE TO LEARN?

The overarching goal of this experiment is to better understand potential approaches to designing machine learning and crowdsourcing tasks for data enrichment projects while offering interesting and engaging activities to volunteers. By combining technology with human annotators, we also hope to reveal the ways in which machine learning relies on human subjectivity and decision-making rather than objective, or neutral, classification. Ultimately, through this work with the Yellow Pages collection, we hope to identify a replicable framework for creating other engaging, ethical, and useful projects.

We appreciate your participation!

Last Update 2021-04-16T15:46:15.371Z

SCRIBE

This project is built using Scribe: document transcription, crowdsourced.

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.