Digital Libraries, Intelligent Data Analytics, And Augmented Description: A Demonstration Project

A COLLABORATORY BETWEEN THE LIBRARY OF CONGRESS AND THE IMAGE ANALYSIS FOR ARCHIVAL DISCOVERY (AIDA) LAB AT THE UNIVERSITY OF NEBRASKA, LINCOLN, NE

Liz Lorang (faculty)
Leen-Kiat Soh (faculty)
Yi Liu (PhD student)
Chulwoo Pack (PhD student)

Aida

Funding

Project awarded by the Library of Congress under notice ID 030ADV19Q0274, “The Library of Congress – Pre-processing Pilot”

Period of performance: July 16-to November 8, 2019

Introduction

Collaborative research project between the Library of Congress and the Aida digital libraries research team at the University of Nebraska

5-month demonstration project with the following goals:

- **Develop and investigate the viability and feasibility of textual and image-based data analytics approaches to support and facilitate discovery**
- **Understand technical tools and requirements for the Library of Congress to improve access and discovery of its digital collections**
- **Enable the Library of Congress to plan for improved applications and technical capacity as well as future innovations**

Aida

Participants

University of Nebraska-Lincoln

Elizabeth Lorang Senior Adviser

Leen-Kiat Soh Senior Adviser

Yi Liu Research Associate and Developer

Chulwoo (Mike) Pack Research Associate and Developer

Ashlyn Stewart Research Assistant

Library of Congress

Meghan Ferriter Chief (Acting) LC Labs/Senior Innovation Specialist

Abbey Potter Senior Innovation Specialist

Jaime Mears Senior Innovation Specialist

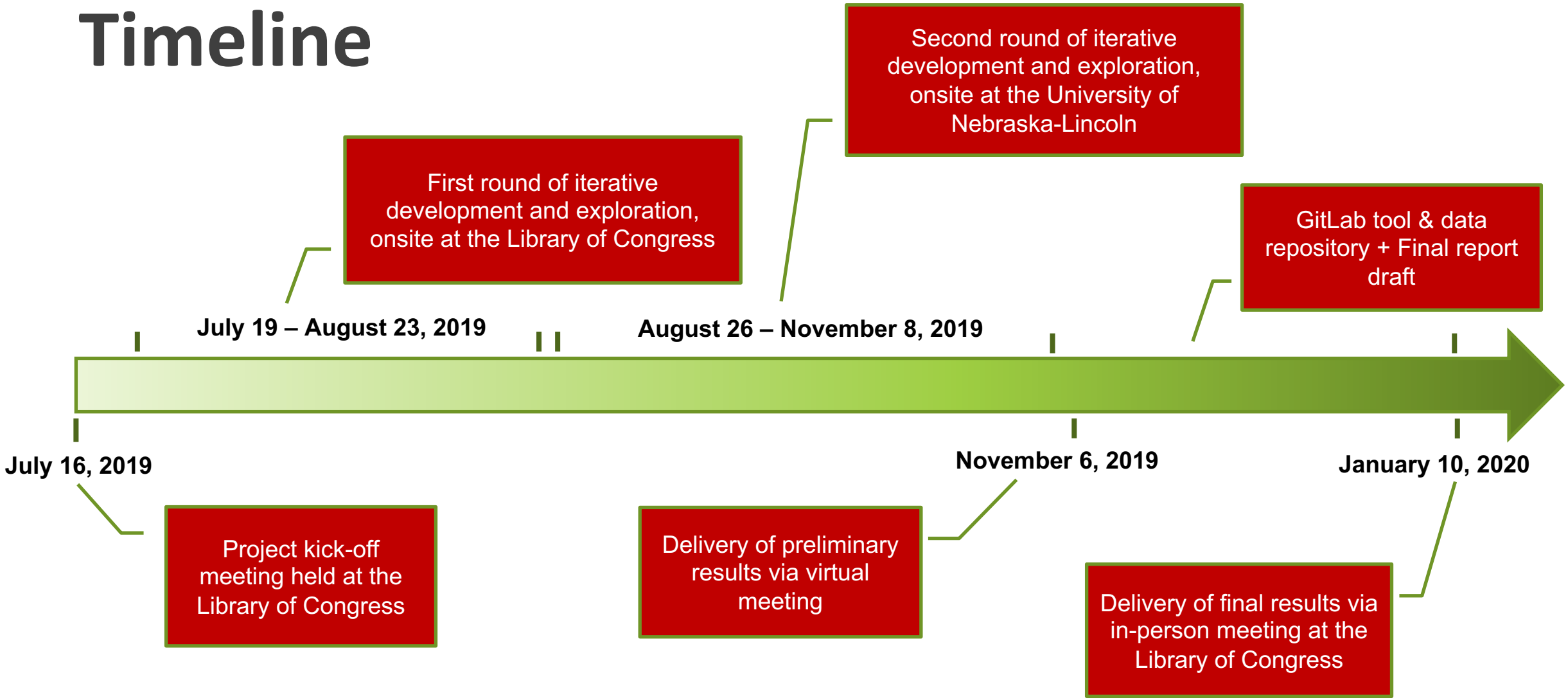
Eileen Jakeway Innovation Specialist

Tong Wang Senior IT Specialist, OCIO

Lauren Algee Senior Innovation Specialist

Victoria Van Hyning Senior Innovation Specialist

Timeline


Second round of iterative development and exploration, onsite at the University of Nebraska-Lincoln
First round of iterative development and exploration, onsite at the Library of Congress
GitLab tool & data repository + Final report draft
July 19 – August 23, 2019
August 26 – November 8, 2019
July 16, 2019
November 6, 2019
January 10, 2020
Project kick-off meeting held at the Library of Congress
Delivery of preliminary results via virtual meeting
Delivery of final results via in-person meeting at the Library of Congress

Demonstration Project Design & Approach

We anchored our work around two areas:

(1) extracting and foregrounding **visual content** from **Chronicling America** (chroniclingamerica.loc.gov) through a variety of techniques and approaches and

(2) applying a series of **image processing and machine learning** methods and techniques to minimally processed manuscript collections featured in **By the People** (crowd.loc.gov).

- Collections already deemed **significant** by the Library of Congress and because they had a degree of ground-truthing work already completed as well as associated domain expertise and use experiences
- Benefit of generating **rich** and **varied metadata**, so that the Library might explore the ways in which more robust metadata allow for alternative points of entry into the materials and the opportunity for researchers to pursue questions of varying nature

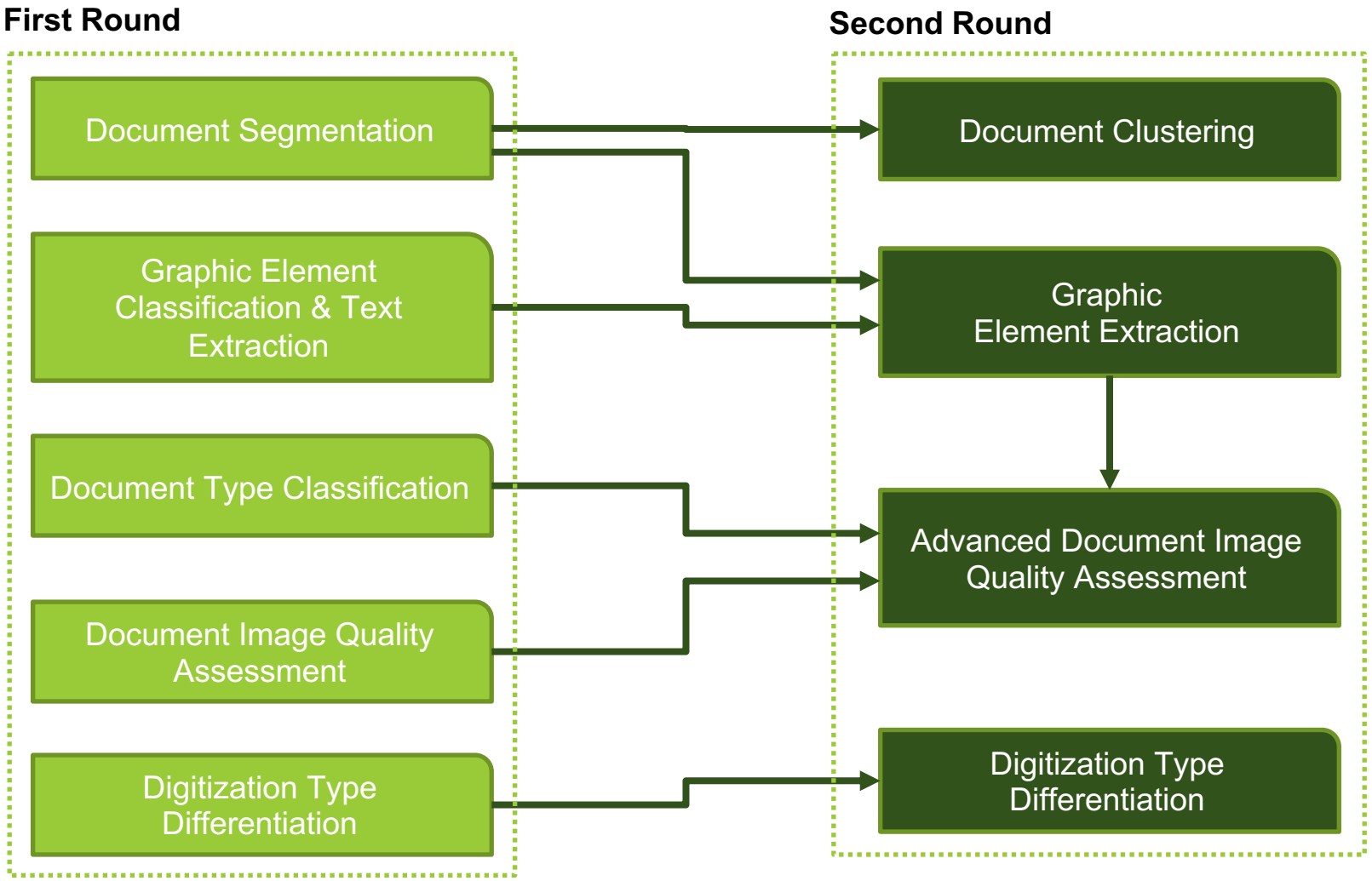
Demonstration Project Design & Approach 2

Ultimately, we designed a series of **explorations** that allowed us to investigate a range of issues and challenges related to machine learning and the Library's collections

- Developed through an **iterative** process and in **regular consultation** with members of the Library of Congress staff
- Through that process, some explorations **merged**, others **concluded more quickly** than others, and areas of inquiry **seeded in one exploration began to sprout in others** as well
- Individually, the explorations pursued particular **technical** and **collections-oriented questions**

We also used the explorations as points of entry into and paths to reflection on larger issues, questions, and challenges for machine learning and cultural heritage (**Discussion and Recommendations**)

The Explorations



Aida

First-Round Explorations

	Selected Potential Applications					
	Metadata generation (structural, descriptive, etc.)	Graphical content extraction	Influence decision-making for human and/or machine processing	Faceted data for end-users or researchers in search and discovery interface	Ground truth and benchmark sets for machine learning and image analysis projects competitions	Understanding collections
Document Segmentation	✓	✓		✓	✓	
Graphic Element Classification and Text Extraction	✓	✓		✓	✓	
Document Type Classification	✓		✓	✓	✓	✓
Document Image Quality Assessment	✓		✓	✓	✓	✓
Digitization Type Differentiation	✓		✓	✓	✓	✓

Aida

Second-Round Explorations

	Selected Potential Applications					
	Metadata generation (structural, descriptive, etc.)	Graphical content extraction	Influence decision-making for human and/or machine processing	Faceted data for end-users or researchers in search and discovery interface	Ground truth and benchmark sets for machine learning and image analysis projects competitions	Understanding collections
Document Clustering	✓		✓	✓	✓	✓
Figure/Graph Extraction	✓	✓		✓	✓	
Advanced Document Image Quality Assessment	✓		✓	✓	✓	✓
Digitization Type Differentiation	✓		✓	✓	✓	✓

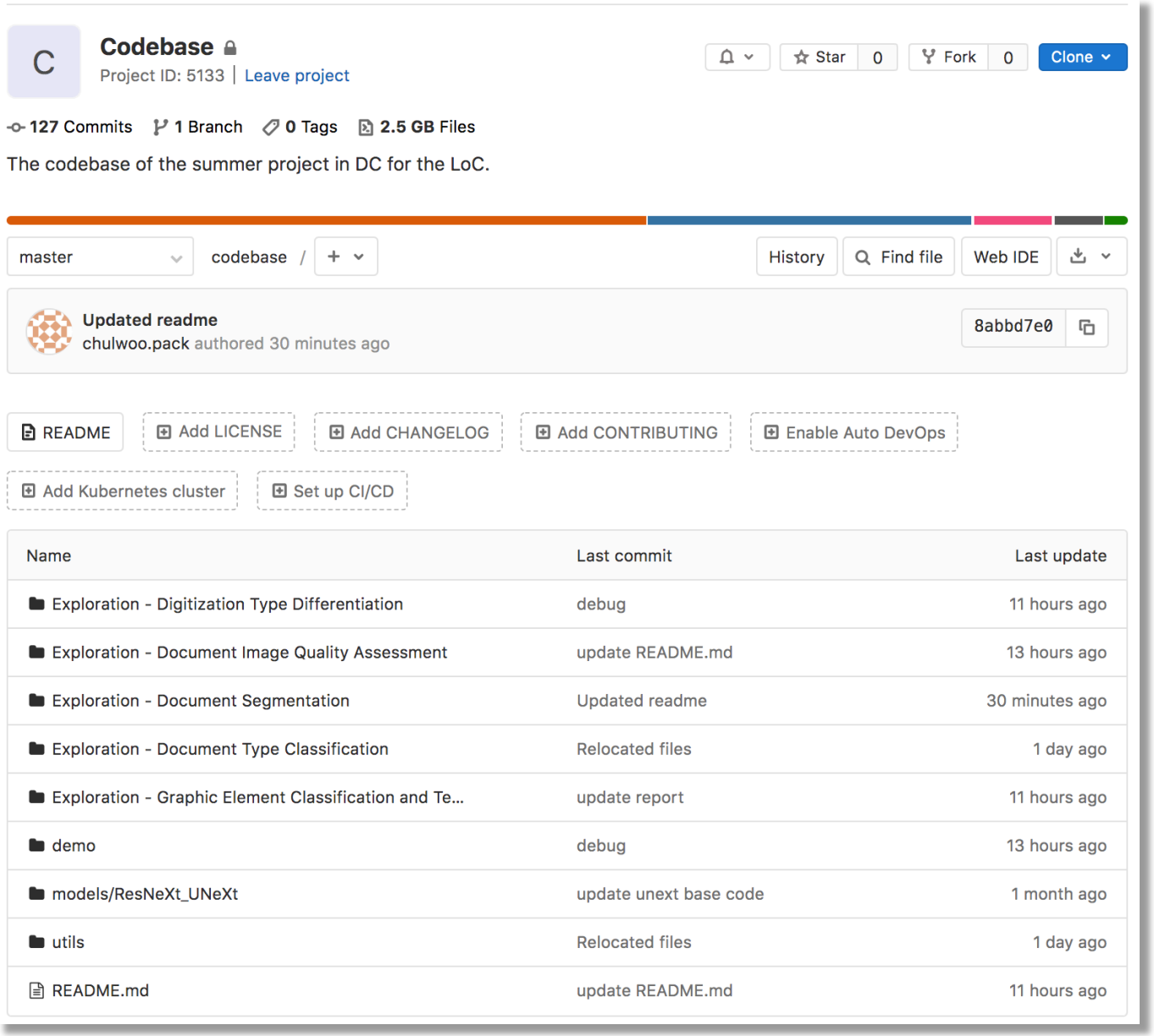
GitLab Repository

Reports, code, data

Documentation of code, data, and exploration projects

Aida

GitLab Repository



Aida

GitLab Repository

README.md

Introduction

The University of Nebraska-Lincoln's (UNL) Aida digital libraries research team and the Library of Congress (LC) collaborated on a "summer of machine learning" in 2019 to explore machine learning techniques for extending the accessibility of digital collections. The UNL team developed a number of prototype explorations over multiple iterations to investigate a range of questions and issues related to the digital materials, the LC's collections, and to machine learning practices in cultural heritage organizations. The UNL team employed a variety of machine learning approaches such as back-propagation neural network-based classifiers and deep learning approaches, including convolutional neural networks. More specifically, these projects involve VGG16, ResNeXt, dhSegment, and a fusion network combining ResNeXt and U-Net.

This repository includes the code developed and used across the team's explorations.

Getting Started

These instructions will get you a copy of the project up and running on your local machine for development and testing purposes.

Prerequisites

For Exploration - Digitization Type Differentiation, the required software systems and libraries are:

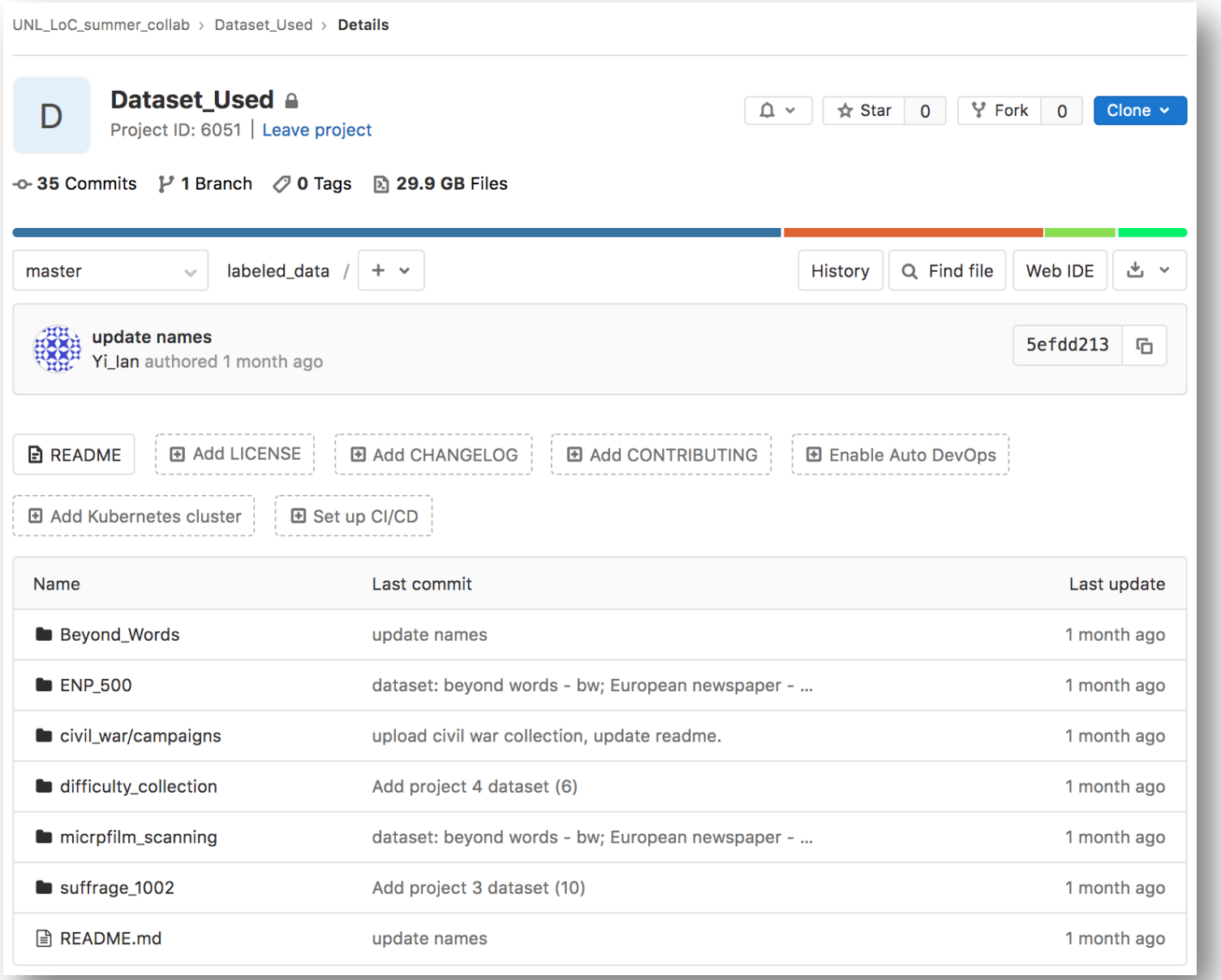
- Anaconda >= 4.3
- Python >= 3.6
- TensorFlow 1.13
- CUDA 10.0 [if training on GPU]
- imageio >= 2.5
- pandas >= 0.24.2
- shapely >= 1.6.4
- scikit-learn >= 0.20.3
- scikit-image >= 0.15.0
- opencv-python >= 4.0.1
- tqdm >= 4.31.1
- sacred 0.7.4
- requests >= 2.21.0
- click >= 7.0

For Exploration - Graphic Element Classification and Text Extraction and Exploration - Digitization Type Differentiation, the required software systems and libraries are:

- Python 3.7
- MXNet 1.5
- CUDA 10.0 [if training on GPU]
- Matplotlib 3.1.1

Aida

GitLab Repository



Aida

GitLab Repository

Name	Last commit	Last update
Beyond_Words	update names	1 month ago
ENP_500	dataset: beyond words - bw; European newspaper - ...	1 month ago
civil_war/campaigns	upload civil war collection, update readme.	1 month ago
difficulty_collection	Add project 4 dataset (6)	1 month ago
micrpfilm_scanning	dataset: beyond words - bw; European newspaper - ...	1 month ago
suffrage_1002	Add project 3 dataset (10)	1 month ago
README.md	update names	1 month ago

README.md

Introduction

Datasets used in the collaboration projects of the University of Nebraska-Lincoln and the Library of Congress.

Datasets used in:

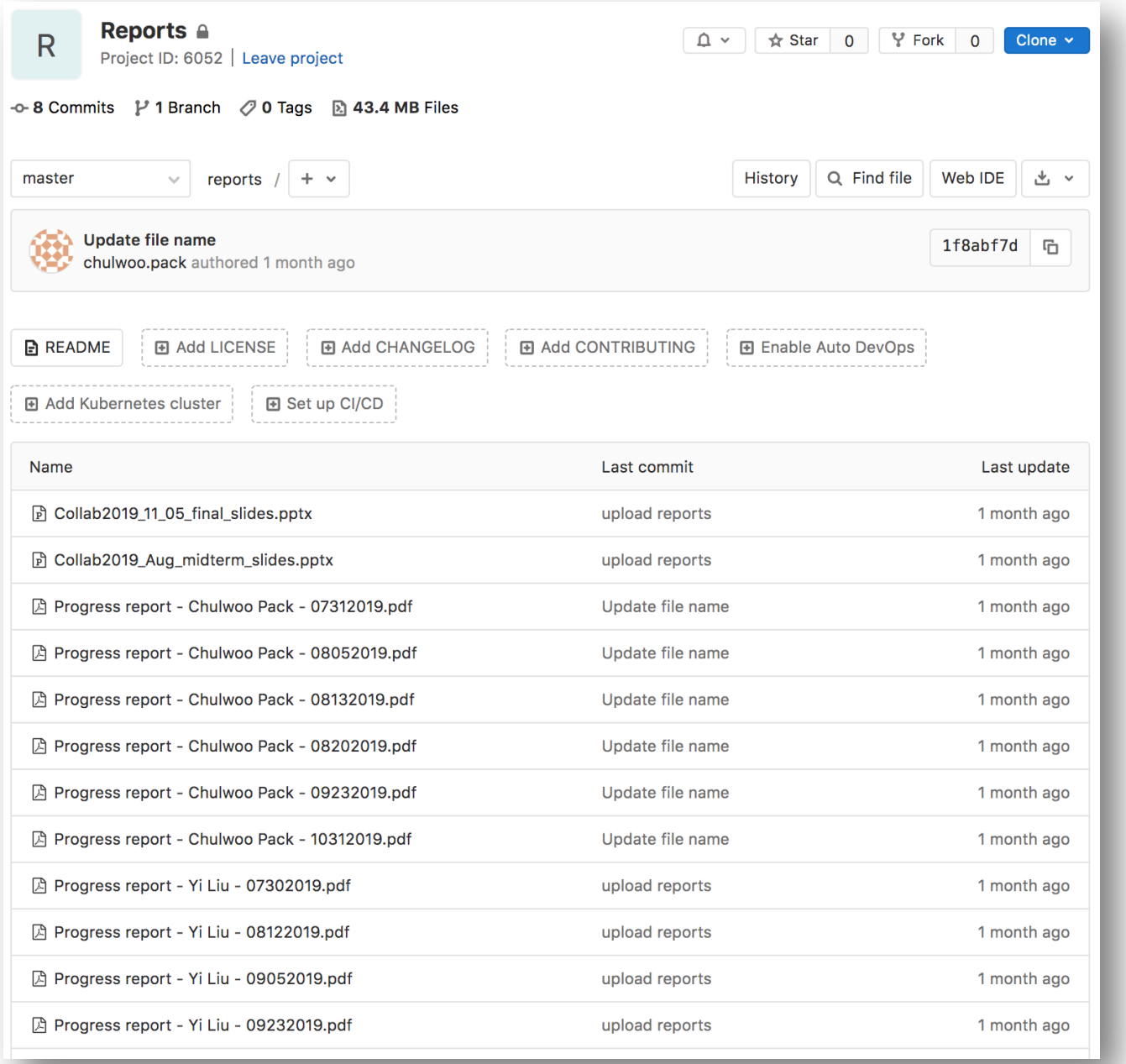
- Project 1: [ENP_500]
- Project 2: [ENP_500] [Beyond_Words]
- Project 3: [suffrage_1002] (collected from LoC Suffrage campaign)
- Project 4: [civil_war] [difficulty_collection] (collected from LoC mauscript/mixed material)
- Project 5: [micrpfilm_scanning] (part of the Civil War collection)

Acknowledgments

- The groundtruth of the datasets "micrpfilm_scanning" and "suffrage_1002" are created by the Aida Team of the University of Nebraska-Lincoln.

Aida

GitLab Repository



Name	Last commit	Last update
Collab2019_11_05_final_slides.pptx	upload reports	1 month ago
Collab2019_Aug_midterm_slides.pptx	upload reports	1 month ago
Progress report - Chulwoo Pack - 07312019.pdf	Update file name	1 month ago
Progress report - Chulwoo Pack - 08052019.pdf	Update file name	1 month ago
Progress report - Chulwoo Pack - 08132019.pdf	Update file name	1 month ago
Progress report - Chulwoo Pack - 08202019.pdf	Update file name	1 month ago
Progress report - Chulwoo Pack - 09232019.pdf	Update file name	1 month ago
Progress report - Chulwoo Pack - 10312019.pdf	Update file name	1 month ago
Progress report - Yi Liu - 07302019.pdf	upload reports	1 month ago
Progress report - Yi Liu - 08122019.pdf	upload reports	1 month ago
Progress report - Yi Liu - 09052019.pdf	upload reports	1 month ago
Progress report - Yi Liu - 09232019.pdf	upload reports	1 month ago

Aida

GitLab Repository

Progress report - Yi Liu - 10292019.pdf	upload reports	1 month ago
README.md	Update file name	1 month ago

README.md

Introduction

The real-time slides and reports along with the projects.

[Collab2019_Aug_midterm_slides.pptx]

The slides of the on-site end-of-summer presentation -- First Iteration (August 2019)

[Collab2019_11_05_final_slides.pptx]

The slides of the wrap-up presentation -- Second Iteration (October 2019)

[Progress report - Yi Liu - 07302019.pdf]

This is the first iteration of Project 2. It purposes the U-NeXt model for the figure/graph extraction task.

[Progress report - Yi Liu - 08122019.pdf]

This is the first iteration of Project 4. It purposes a deep learning model to subjectively assess the document image quality.

[Progress report - Yi Liu - 09052019.pdf]

This is the first iteration of Projects 4 and 5. It contains the evaluation of the image quality of Civil War collection using the objective DIQA code (https://git.unl.edu/unl_loc_summer_collab/codebase/tree/master/project4) . And It contains the performance of the first iteration project 5.

[Progress report - Yi Liu - 09232019.pdf]

This is the second iteration of Project 5. It has a comprehensive evaluation of Project 5.

Aida

Brief Discussions on Explorations

For details, audience is referred to our presentation made on November 6, 2019

Also, final report identifies **guiding questions**; outlines and describes our **approaches**, techniques, and methods; presents high-level **results** and **analysis**; and offers **ideas** toward future development and/or potential applications

In the following slides, we briefly summarize the goals and questions for each exploration

Exploration: Document Segmentation

The **goal** of this exploration was to see if **we could localize textual zones, figures, layout borders, and tables and then identify image-like components in historic newspaper pages**

- Newspaper page images presented through Chronicling America are not zoned or segmented below the page level
- Content within a newspaper page is also not identified or classified by genre, type, or other features

Guided by **questions**:

- How might we use image zoning and segmentation to generate additional information about newspaper pages in the Chronicling America corpus?
- Could image zoning and segmentation be used to pull out graphical content from Chronicling America newspapers?
- How might ML projects draw on ground truth or benchmark data already generated through crowdsourcing efforts?

Exploration: Graphic Element Classification & Text Extraction

Initial goal of this exploration was to **find, localize, and classify figures, illustrations, and cartoons present in historical newspaper page images**; and **extract any text from the content**

By its second iteration, this exploration focused on **fine-tuning of the identification of graphical content** in historic newspaper page images and **the distinction of graphical content regions from textual content regions**

Guided by **questions**:

- How might we use image zoning and segmentation, and text extraction from graphical regions, to generate additional information about newspaper pages in the Chronicling America corpus?
- Could image zoning and segmentation be used to pull out graphical content from Chronicling America newspapers?
- What benefits do different types or approaches to zoning and segmentation have for various information tasks?
- What strategies might be necessary to deal with rare content types in the training and evaluation of machine learning systems?

Aida

Exploration: Document Type Classification

This exploration pursued whether we could **effectively distinguish among handwritten, printed, and mixed (both handwritten and printed) documents** within a collection of minimally processed manuscript materials at the Library of Congress

Guided by **questions**:

- What features might be useful for influencing processing pipelines, for generating additional metadata, or for distinguishing among materials?
- How viable might large-scale indexing of documents be, for certain types of criteria? To what level of performance could we meta-tag document images?
- Would a deep learning model that had shown remarkable performance for natural scene images also show promising performance for document images?
- Or, to be more precise, would a feature extractor trained with millions of natural scene images also capably extract useful features for document images?

Exploration: Digitization Type Classification

The goal of this exploration was to **distinguish among digital images created by digitization from different source types**

- items digitized from an original document item and those digitized from a microform reproduction of an original item

Guided by **questions**:

- What features might be useful for influencing processing pipelines, for generating additional metadata, or for distinguishing among materials?
- How viable might large-scale indexing of documents be, for certain types of criteria?
- To what level of performance could we meta-tag document images?
- Who might benefit from the ability to facet or search according to this criterion—digitization source—and how that might information might be made available?

Exploration: Document Image Quality Assessment (DIQA) & Advanced DIQA

This exploration set out to **analyze the quality of document images in minimally processed manuscript collections** based on a variety of criteria with the goal of **using information about image quality to inform future processes**

Guided by **questions**:

- How might we distinguish among materials that most need human intervention and those materials that might be well-suited to machine approaches? When might materials be best suited to a combined approach?
- Could image quality assessments be useful in compiling ground truth and benchmarking sets in some capacity? Might such features be useful further downstream for users, to be able to facet for difficulty, for example?
- How might metadata about image quality of document images enrich understanding of individual items and of collections and corpora?
- To what extent can quality be computationally assessed, and might it help to better understand overall visual attributes of a dataset?

Aida

Exploration: Document Clustering

This exploration extended from the initial documentation segmentation exploration and applied clustering to document images. Drawing on our work in other explorations, we wondered **whether document images clustered together share similar visual features recognizable to human observers**

Guided by **questions**:

- Would page images with graphical content cluster?
- Could we discern other clustering features?
- Could such clusters be useful in decision-making, for metadata generation, or other processes?

Discussion

The explorations touched upon types of investigations to be pursued with machine learning and the information that can be gleaned from and about digitized materials, the collections in which they sit, and about organizational and institutional practices and beliefs

Through these explorations, we developed a heightened awareness of the number of possibilities and challenges, both those **social** and **technical**, as well as of their scale

Aida

Discussion | Social

Processing image and textual data with existing machine learning platforms and programs is increasingly accessible (e.g., lower barrier to entry)

This perceived simplicity, however, **hides significant complexity, nuance, assumptions and decision-making, and labor**

Furthermore, this perceived simplicity **has the potential to mask the implications of machine learning-generated knowledge**

Discussion | Social 2

Domains considering implementing machine learning must **engage deeply and critically with the technology**, what it does, and what it means

For cultural heritage digital libraries, now is a **critical moment** to grapple with epistemologies of machine learning and the knowledge it structures, shapes, and appears to codify

Machine learning in digital libraries should be committed to, in the words of Thomas Padilla, "**responsible operations**"

Aida

Discussion | Social 3

Early in this demonstration project, Meghan Ferriter framed a range of different types of machine learning explorations and their outcomes

These included machine learning in the Library of Congress for **description**, **discovery**, and **delight**

- Each has the potential to help people see materials from new angles, to peruse them in alternative ways, and to begin to frame additional questions and ways of thinking
- Each foregrounds different values and carries with it a different set of requirements and responsibilities

Discussion | Social 4

Building on Ferriter's "three Ds," we add "**deployment**" and "**debate/dialogue**."

- As a **community of practice** and as **communities of researchers**, what do we expect from projects and applications that proceed with these—and other—purposes in mind?
- Perhaps most critically, for any project that is about **large-scale deployment**, or a **deployment of machine learning** that may have significant implications for reasons beyond scale, what expectations do we hold as to what such projects must do, consider, make transparent?
- What **contexts** must we be able to see and understand?

Aida

Discussion | Technical

Computational access to the Library of Congress's digital objects is relatively straightforward

- Access via the Library's API and other bulk download options
- This *collections-as-data* approach is an important layer for machine learning
- ***However***, we depended on our inside access to people at the Library in order to make sense of some of the data

There is need for **additional levels of documentation and/or to new types of reference support needed** in the Library of Congress as it facilitates emergent areas of research with its digital collections

*Note: We anticipate that the **Library's Mellon-funded project, Computing Cultural Heritage in the Cloud**, will advance thinking and conversations on these topics*

Aida

Discussion | Technical 2

Machine learning approaches also require **accurate ground truth data** from which to learn and validate

In our explorations, even when it seemed we could utilize existing Library of Congress data as ground truth information, **ground truth data proved challenging**

- We had to create ground truth sets ourselves or turn to externally available datasets that provided the type/nature of ground truth information needed

Not a criticism of the Library's efforts or of individuals' labor and effort over time

The **bibliographic information and collections-centered metadata previously pursued in libraries is a limited vision** of what will be needed for machine learning applications and new areas of research

Discussion | Technical 3

Machine learning models developed and trained on other types of ground truth sets skew toward the contemporary and born-digital

- ***not readily transferable*** to digitized historical materials that are typically noisy and of lesser quality

Existing datasets for competitions that focus on historical documents are relatively small

- ***not comprehensive of the range of materials in collections*** as large and diverse as those in cultural heritage institutions

Discussion | Technical 4

The challenges around ground truth connect with other questions that surfaced across many of our explorations:

- How might **data created by users** via the Library of Congress's crowdsourcing projects be used as ground truth data?
- What **size** of ground truth and training sets are necessary for different purposes?
- Are ground truth data created for one purpose **transferrable** for other purposes?
- What happens when we attempt to **extrapolate** from ground truth created for one purpose to another? Or when there isn't a direct match between ground truth data and output data?
- Etc.

Discussion | Technical 5

We wondered about **the interplay of human expertise and processes and machine knowledge and processes**

- What human-computer processes might be viably and validly **adopted and operationalized** as, say, part of a daily routine?
- What human-computer approaches are viable and valid in terms of effectiveness and efficiency in order to address issues of **scalability**?
- What value might there be in **cross-learning, loop-learning, and cross-processing**, where machines learn from humans, humans respond to and adapt understanding based on machine learning, and this looped learning informs processes and decision-making?
- Rather than seeing machine learning as an end, how can the Library of Congress **embed and value critique across such a system**, so that **both human and machine assumptions are routinely tested**?

Discussion | Technical 6

Furthermore, to facilitate **effective and efficient human-computer interaction** …

- What are the **foundational data and metadata** needed and required to facilitate **cross-learning and cross-processing**?
- What is the place for **data-science paradigms**, where problems or issues are derived **bottom-up**—are surfaced through the collections and feature analysis—rather than top-down?

Recommendations

As the largest library in the world, **the Library of Congress is uniquely situated to play a leadership role in advancing the theory and practice of machine learning in the cultural heritage sector**

With that in mind, we have two top-level recommendations for the Library as it moves forward in its efforts to “throw open the treasure chest,” “connect,” and “invest in our future.” :

- that the Library focus the weight of its machine learning efforts and energies on **social and technical infrastructures** for the development of machine learning in cultural heritage organizations, research libraries, and digital libraries
- that the Library invest in continued, ongoing, **intentional explorations and investigations of particular machine learning applications to its collections**

Aida

Recommendations 2

What we do ***not*** recommend at this time is the broad application of machine learning to the Library's digital collections with an eye toward broadly making claims about the materials or restructuring access to them

- On a very practical level, such broad application would be premature due to the challenges with ground truth data and validation

We advise ***against*** a "more product, less process" approach to machine learning applications

- The ways in which ML-generated knowledge stands to influence decision-making is too powerful to adopt such an approach, or make such a commitment, at this nascent stage

Aida

Recommendations 3

People are central to all of the recommendations

- None of the recommendations imagine a library without information professionals and experts
- Any future for machine learning in libraries will require an **investment in people** with many types of expertise
- A best-case future for machine learning in cultural heritage organizations is that the people who work in them are able to bring ***even more*** of their experience and expertise to bear

Recommendations | Infrastructure

We recommend that **the Library dedicate itself to a range of infrastructure projects** that will create a strong foundation for machine learning in the profession and field, particularly as applied to historical cultural heritage materials

- **Educative infrastructures**
- **Platforms for conversations**
- **Pathways for gathering and delivering machine learning models and verifiable learning data** that extend beyond individual projects
- **Pathways for bringing together cross-domain researchers**

Recommendations | Infrastructure 2

1. Develop a **statement of values or principles** that will guide how the Library of Congress pursues the ***use, application, and development of machine learning*** for cultural heritage
2. Create and scope **a machine learning roadmap** for the Library that looks both ***internally*** to the Library of Congress and its needs and goals and ***externally*** to the larger cultural heritage and other research communities
3. Focus efforts on developing **ground truth sets and benchmarking data** and making these easily available

Recommendations | ML Applications

We recommend that **explorations** are

- framed and understood as **intellectual endeavors** rather than being output-driven and
- **collaborations** among computer scientists, developers, and information professionals, drawing in other participants and stakeholders

We also encourage the Library of Congress to be careful in the **presentation of machine learning generated data**

- particularly when that data might be read or experienced by others as uncontested knowledge or fact about cultural heritage materials, and also with care and concern about what is absent as well as what is present

Recommendations | ML Applications 2

1. Join the Library of Congress's emergent efforts in machine learning with its existing expertise and leadership in **crowdsourcing**
   - Combine these areas as "**informed crowdsourcing**" as appropriate
2. **Sponsor challenges** for teams to create additional metadata for digital collections in the Library of Congress. As part of these challenges, require teams to engage across a range of social and technical questions and problem areas
3. Continue to create and support **opportunities for researchers to partner** in substantive ways with the Library of Congress on machine learning explorations

Recommendations | Alignment w. Digital Strategy

Digital Strategies	Recommendations on Infrastructure	Recommendations on ML Applications
maximizing use of content	✓	
supporting emerging styles of research	✓	✓
welcoming other voices	✓	✓
driving momentum in our communities	✓	✓
cultivating an innovation culture	✓	✓
ensuring enduring access to content	✓	
building toward the horizon	✓	✓

Aida

Recommendations | Alignment w. *Responsible Operations*

Strategies	Sub-Strategies	Statement of Vision	Roadmap of ML	Ground-Truthing & Benchmarking	ML + Crowd-sourcing Efforts	Sponsoring Challenges	Research Partnerships
Committing to Responsible Operations	Managing Bias	✓	✓	✓		✓	
	Transparency, Explainability, Accountability	✓	✓			✓	
	Distributed Data Science Fluency	✓	✓				
Workforce Development	Investigating Core Competencies		✓		✓		
	Committing to Internal Talent		✓				
Description & Discovery	Enhancing Description at Scale			✓	✓	✓	
Shared Methods and Data	Shared Development and Distribution of Training Data			✓	✓		
	Shared Development and Distribution of Methods					✓	✓
Sustaining Interprofessional & Interdisciplinary Collaboration						✓	✓

Padilla, Thomas. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. Dublin, OH: OCLC Research. 2019.

Conclusion

This demonstration project—via its explorations, discussion, and recommendations—has shown **the potential of machine learning** toward a variety of goals and use cases, and it has argued that **the technology itself will *not* be the hardest part of this work**

The hardest part will be the myriad challenges to undertaking this work in ways that are **socially and culturally responsible**, while also **upholding responsibility to make the Library's materials available** in ***timely and accessible ways***

The Library of Congress is in a remarkable position to advance machine learning for cultural heritage organizations, through its size, the diversity of its collections, and its commitment to digital strategy

Many Thanks

We sincerely thank the team at the Library of Congress for this collaboration. This project would not have been possible without their insights, expertise, dedication, patience, and collegiality. It's been a privilege to learn more about the Library of Congress, get the opportunity to see behind the scenes, and build this relationship. We are especially grateful for the six weeks that the Library and the team hosted Yi and Mike and for making them feel welcome, including them as part of the team, and fostering so many remarkable learning opportunities.

Aida

Additional Details

Recommendations | Infrastructure 3

1. A **statement of values or principles**

Example questions to address:

- If units within the Library seek to apply machine learning to collections, under what principles and values should that work proceed?
- What are the expectations around transparency and explainability, both for internal and external audiences, for example?
- Or around confronting problematic historical knowledge and knowledge structures in training data?

Recommendations | Infrastructure 4

2. A **machine learning roadmap**

Example questions to address:

- What are the Library's goals and objectives in each of the investigation areas?
- Will it pursue all of the areas or prioritize particular areas?
- With regard to the Library's goals and objectives, are there investigations areas that the Library would add?

Recommendations | Infrastructure 5

3. Ground truth sets and benchmark data

- allow researchers—including cultural heritage professionals, computer scientists, and developers—to focus their energies and research, development, and analysis, rather than on creating one-off, niche datasets
- create the possibility of more rapid development around particular problem domains

Creating and distributing ground truth sets will foreground the **significance of metadata**, including **technical**, **structural**, and **descriptive**

- **Descriptive of the content of the historical materials**, including metadata about what is depicted and represented as well as how
- **Descriptive of the properties of the image**, including features such as digitization source, contrast, skew, noise, range effect, complexity

Aida

Recommendations | Infrastructure 6

3. Ground truth sets and benchmark data

3.1. Development of DocuNet

- We recommend the Library of Congress **develop, or partner in developing, DocuNet**
  - an image database of historical documents with accompanying taxonomic and typological metadata
- Features or characteristics important to a DocuNet are
  - **ground truth** (e.g., document types, coordinates of article regions, etc.);
  - **openness** (e.g., accessibility);
  - **diversity** and **balance** (e.g., different document types should be comprehensively covered and equally distributed); and
  - **clear objectives** (e.g., segmentation, classification, clustering, etc.)

Recommendations | Infrastructure 7

3. Ground truth sets and benchmark data

3.2. Pursuit of Low-Cost Ground-Truthing

- We also recommend that the Library explore options for, and contribute to efforts to **advance, low-cost ground-truthing**
  - Having subject matter experts hand-label data is expensive and is a barrier to machine learning
- Instead, the Library could pursue **heuristics-based models**
  - Computers use human-created clues to label data points using heuristic rules, constraints, distributions, and/or variances of the dataset
  - Less accurate than item-by-item expert-labeled ground truth, ***but it may produce effective machine learning systems***

Recommendations | ML Applications 3

1. Joining Library's ML and Crowdsourcing Efforts

Through its By the People application and campaigns, and other earlier efforts, the Library of Congress has established a strong portfolio of **crowdsourcing** experience

We see significant potential in bringing together **machine learning** and **crowdsourcing** efforts:

- E.g., joining these areas, even in a limited way, would allow the Library to research cross-learning and looped learning.
- In a hypothetical project, members of the crowd might receive labeled data from a model; users then revise the labels, and the model improves its predictions based on those revisions; with each successive iteration, the model improves further

Recommendations | ML Applications 4

2. Sponsoring Challenges

The purpose of this recommendation is multipart:

1. To see what **types of metadata** researchers/teams might produce
   - What metadata is of interest to them?
2. To encourage the creation of particular **types of metadata**, including through an expanded sense of what **descriptive metadata** might include and what is of descriptive value
3. To anchor critical engagement with **core problems**, such as of **bias** in the data and in what may be produced, as inseparable from technical development
4. To emphasize, underscore, and champion that **cross-disciplinary, community-centered** and **community-engaged** development (responsible ML)

Recommendations | ML Applications 5

3. Opportunities for Research Partnerships

We recommend that the Library see **formal collaborations** as central to taking this machine learning work forward

- We have benefitted in significant ways from the additional levels of access to Library staff with this demonstration project and the formal collaboration afforded

We recommend that some measure and shape of formal collaboration **opportunities be part of the Library's support for both machine learning explorations and larger social and technical infrastructures**

Text Extraction from Figure/Graph | Preliminary Results

- Performance on detecting texts in newspaper figure/graph is good
- Texts location is recorded

Text Lines
- 6 text lines
- { "x0": 62, "y0": 608, "x1": 135, "y1": 588, "x2": 143
- { "x0": 188, "y0": 33, "x1": 312, "y1": 31, "x2": 313,
- { "x0": 331, "y0": 31, "x1": 423, "y1": 30, "x2": 423,
- { "x0": 116, "y0": 34, "x1": 166, "y1": 33, "x2": 166,
- { "x0": 405, "y0": 755, "x1": 470, "y1": 757, "x2": 47
- { "x0": 475, "y0": 756, "x1": 531, "y1": 757, "x2": 53

Document Type Classification | Datasets

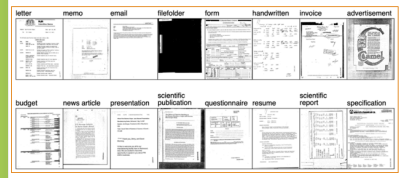


Figure 9. Example document images from each 16 different classes in *RVL_CDIP* dataset

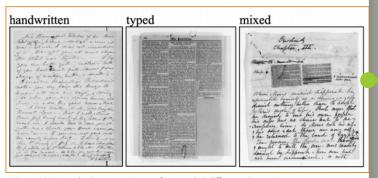
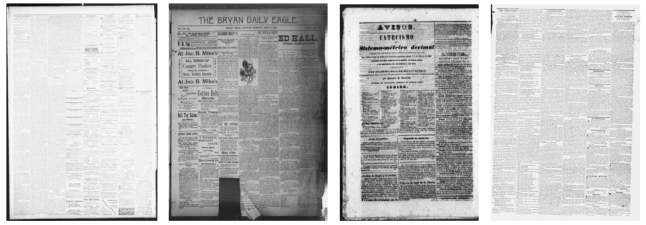


Figure 10. Example document images from each 3 different classes in *suffrage_1002* dataset

Objective Quality Assessment | Examples



Contrast | Range-effect | Bleed-through | Skewness

Project 1. Document Segmentation

Objectives | Find and localize *Figure/Illustration/Cartoon* presented in an image
Applications | metadata generation, discover-/search-ability, visualization, etc.

Project 2.1. Figure/Graph Extraction

Objectives | Find and localize *Figure*/Graph in a document image
Applications | Graph retrieval, document segmentation based on content type

Project 2.2. Text Extraction from Figure/Graph

Objectives | Extract texts from figure/graph
Applications | Metadata generation, OCR for figure/graph caption

Project 3. Document Type Classification

Objectives | (1) Classify a given image into one of *Handwritten/Typed/Mixed* type; (2) Classify a given image into one of *Scanned/Microfilmed*
Applications | metadata generation, discover-/search-ability, cataloging, etc.

Project 4. Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People
Applications | Providing quality scores for machine reading on four criteria: (1) *skewness*, (2) *contrast*, (3) *range-effect*, and (4) *bleed-through*

Project 5. Digitization Type Differentiation: Microfilm or Scanned

Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*
Applications | Metadata generation, pre-processing policy selection

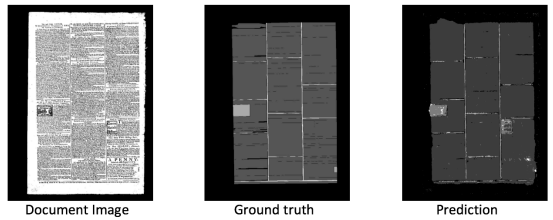
Document Segmentation | Dataset

European Historical Newspapers (ENP)
- Total of 57,339 image snippets in 500 pages
  - All pages have multiple snippets
- Issues
  - Data imbalance
    - Text: 43,780
    - Figure: 1,452
    - Line-separator: 11,896
    - Table: 221

Figure 4. Example of image (left) and ground-truth (right) from ENP dataset. In the ground-truth, each color represents the following components: (1) black: background, (2) red: text, (3) green: figure, (4) blue: line-separator, and (5) yellow: table.

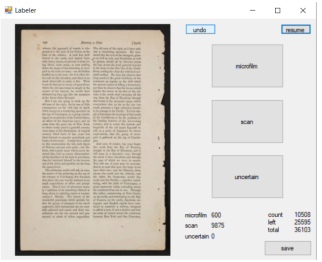
Figure/Graph Extraction | Preliminary Results

- Transfer parameters from pre-trained ResNeXt101 64x4d
- Trained on ENP dataset

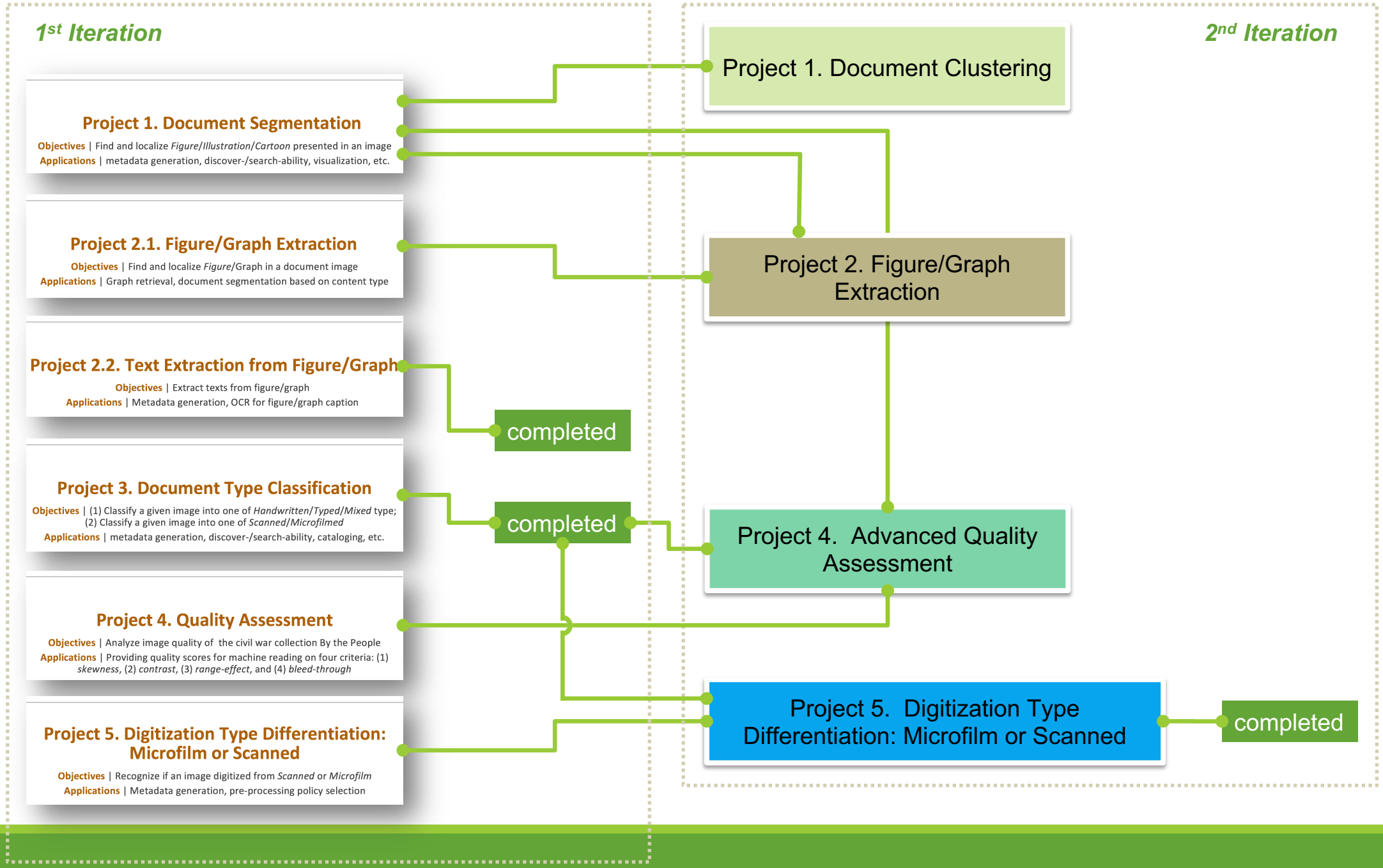


Document Image | Ground truth | Prediction

Digitization Type Differentiation | Datasets



- **Rough estimate**: Based on 10,508 images that was processed, ***ratio of images from microfilm to scanned materials is about 1:16***


1st Iteration
Project 1. Document Segmentation
Objectives | Find and localize Figure/Illustration/Cartoon presented in an image
Applications | metadata generation, discover-/search-ability, visualization, etc.
Project 2.1. Figure/Graph Extraction
Objectives | Find and localize Figure/Graph in a document image
Applications | Graph retrieval, document segmentation based on content type
Project 2.2. Text Extraction from Figure/Graph
Objectives | Extract texts from figure/graph
Applications | Metadata generation, OCR for figure/graph caption
completed
Project 3. Document Type Classification
Objectives | (1) Classify a given image into one of Handwritten/Typed/Mixed type; (2) Classify a given image into one of Scanned/Microfilmed
Applications | metadata generation, discover-/search-ability, cataloging, etc.
completed
Project 4. Quality Assessment
Objectives | Analyze image quality of the civil war collection By the People
Applications | Providing quality scores for machine reading on four criteria: (1) skewness, (2) contrast, (3) range-effect, and (4) bleed-through
Project 5. Digitization Type Differentiation: Microfilm or Scanned
Objectives | Recognize if an image digitized from Scanned or Microfilm
Applications | Metadata generation, pre-processing policy selection
2nd Iteration
Project 1. Document Clustering
Project 2. Figure/Graph Extraction
Project 4. Advanced Quality Assessment
Project 5. Digitization Type Differentiation: Microfilm or Scanned
completed

2nd Iteration

Project 1: Document Segmentation

Objectives | Find and localize *Figure/Illustration/Cartoon* presented in an image
Applications | metadata generation, discover-/search-ability, visualization, etc.

Project 2: Figure/Graph Extraction

Objectives | Find and localize *Figure/Illustration/Cartoon* presented in an image
Applications | metadata generation, discover-/search-ability, visualization, etc.

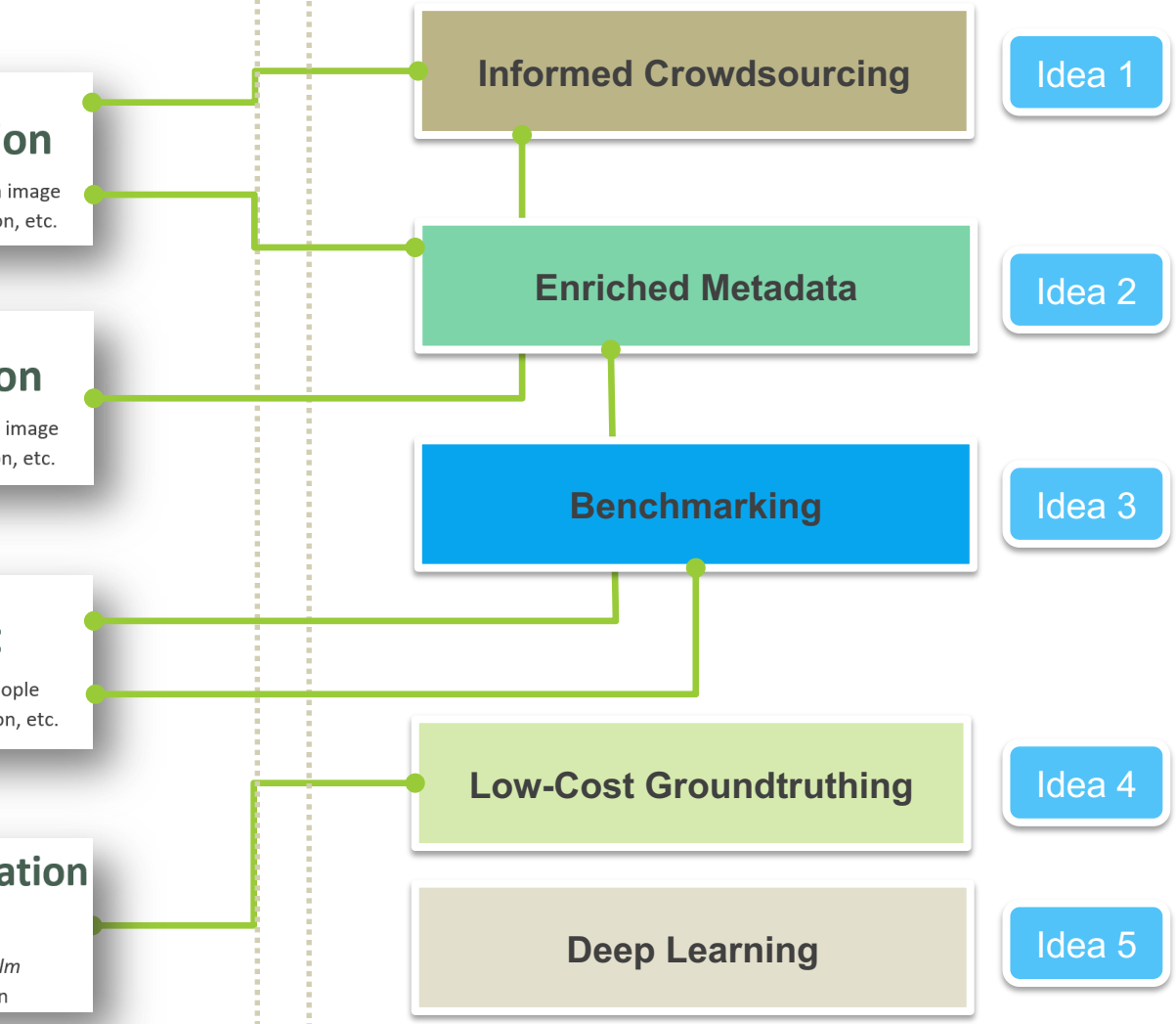
Project 4: Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People
Applications | metadata generation, discover-/search-ability, visualization, etc.

Project 5: Digitization Type Differentiation Microfilm or Scanned

Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*
Applications | Metadata generation, pre-processing policy selection

Future Direction



Aida

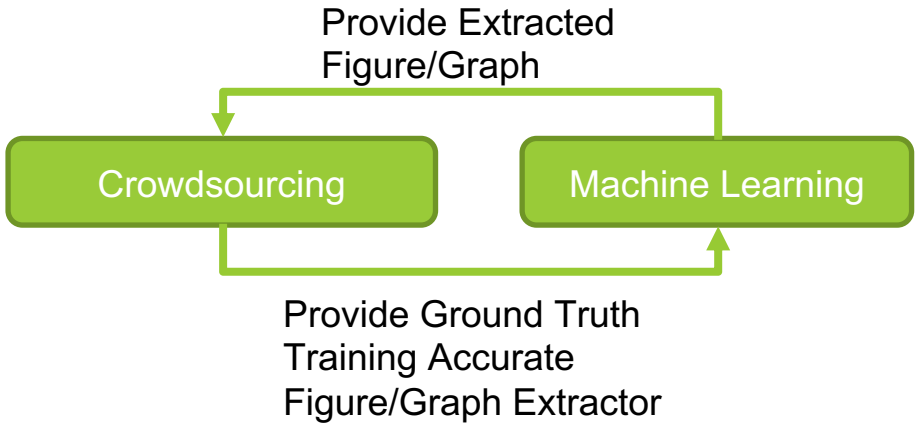
Idea 1

Informed Crowdsourcing

Objectives | Allow machine learning models to cumulatively improve their performance

Motivations | The need for an effective ground truthing approach for hard tasks

- **With informed crowdsourcing**, a loop-based system could be built to improve our U-NeXt models
  - Crowd-sourcing operations receive labeled data from the U-NeXt model, users revise labels, the U-NeXt model improves its predictions based on revision, and repeats



Aida

Idea 2

Enriched Metadata

Objectives | Improve accessibility and searchability of digital libraries

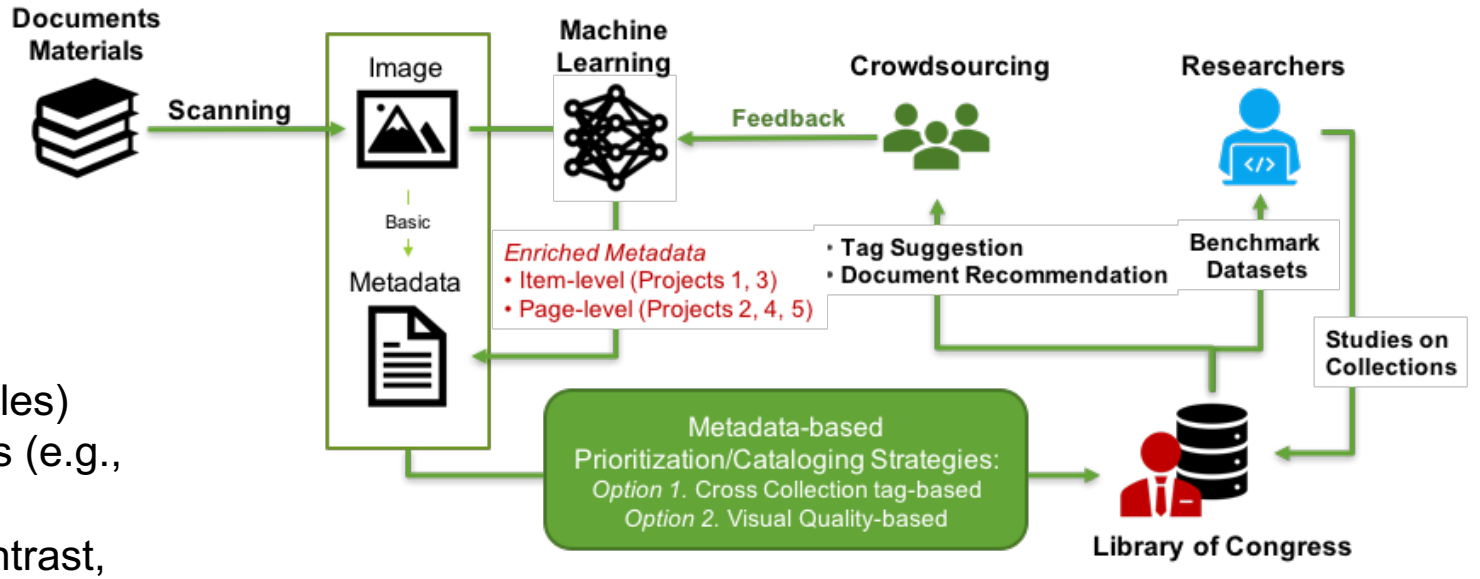
Motivations | The need for enriched any-level searchability

Basic metadata

- Image resolution
- Generated data/time
- Poor quality OCR

Enriched metadata

- Keywords tagged by crowdsourcing
- High quality OCR
- Structural information (e.g., location of articles)
- Logical relationships between substructures (e.g., reading-order)
- Objective/subjective visual quality (e.g., contrast, noise, range effects)



Aida

Idea 3

Benchmark Datasets

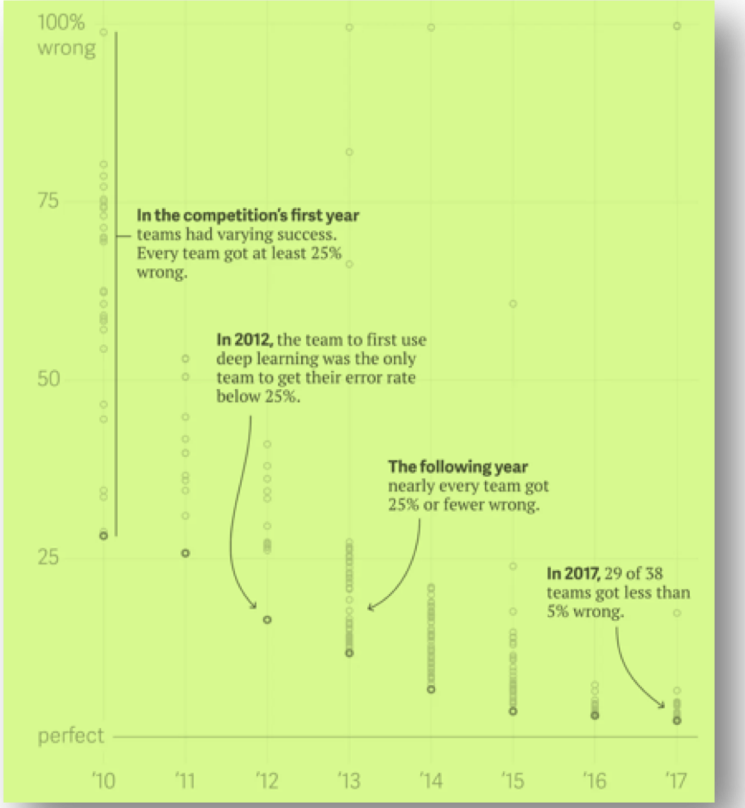
Objectives | Create standard databases to evaluate approaches

Motivations | A shared database can encourage systematic rigorous research towards finding better approaches

Why not "DocuNet"?

ImageNet

- ImageNet is a large-scale ***natural scene*** image dataset
- ImageNet Challenge boosts image and vision research field vastly



Aida

Idea 4

Low-Cost Groundtruthing

Objectives | Build ground truth for machine learning models in a low-cost fashion

Motivations | Having subject matter experts' hand-label data is expensive

Weak supervision

- Computers label data using heuristic rules, constraints, distributions, or/and invariances of the dataset
- Instead of having experts to hand-label data, only need to consult an expert on *how* to label data
- *Example*: Snorkel: A system for programmatically building training datasets using a labeling program based on heuristic rules

Aida

Applying Deep Learning

Objectives | Apply deep learning models to analyze documents in digital library

Motivations | Different deep learning models appropriate for different tasks

Task Type	Task Properties	Suitable Models	Examples
Document layout analysis	Need pixel-level understanding	U-shaped models e.g., dhSegment, U-NeXt	Project 2
Document categorization	Need page-level recognition	Convolutional neural networks e.g., ResNet, ResNeXt	Projects 3 and 5
Audio/video understanding	Sequential data understanding	Recurrent neural networks	

Is There Labeled Data?	Learning Scheme	Examples
Yes	Supervised Learning	Projects 2, 3 and 5
No	Unsupervised Learning	Projects 1 and 4