Explorations in Data Innovations:

Can Machine Learning Support Data Catalog Development?

June 2024

Authors

Siobhan Mills De La Rosa, Luke Patterson, and Mason Miller
American Institutes for Research

Albert Liu
Insight Policy Research

Submitted to

U.S. Department of Labor
Chief Evaluation Office
200 Constitution Avenue, NW
Washington, DC 20210

Project Officer

Dr. Janet Javar

Submitted by

Westat Insight
1310 North Courthouse Road
Suite 880
Arlington, VA 22201

Project Director

Dr. Allison Hyra

Disclaimer

This report was prepared for the U.S. Department of Labor (DOL), Chief Evaluation Office by Westat Insight and its partners, under contract number 1605DC-18-A-0018. The views expressed are those of the authors and should not be attributed to DOL, nor does mention of trade names, commercial products, or organizations imply endorsement of same by the U.S. Government.

Acknowledgments

The report was prepared for the U.S. Department of Labor Chief Evaluation Office (CEO). The authors received invaluable support, guidance, and contributions from Giolerny Altamirano-Rayo, Kelly Gleason, Tara Martin, and Janet Javar at the CEO. The authors thank Allison Hyra and Breanna Wakar from Insight Policy Research, who provided constructive feedback throughout the project.

Suggested Citation

Mills De La Rosa, S., Patterson, L., Miller, M., & Liu, A. (2024). *Explorations in data innovations–Can machine learning support data catalog development?* Westat Insight and American Institutes for Research for U.S. Department of Labor Chief Evaluation Office.

Contents

Tables

Executive Summary

The Chief Evaluation Office of the U.S. Department of Labor (DOL CEO) is committed to using innovative tools to meet the Department's research, evaluation, and data analytics needs. In December 2021, DOL CEO commissioned the Westat Insight and American Institutes for Research® (AIR®) study team to explore potential opportunities to use machine learning methods to facilitate the automated data collection of labor-relevant data. Between May 2022 and December 2023, the study team worked with experts in machine learning, web scraping, and labor-related data to understand how DOL CEO could use machine learning approaches to automate data collection efforts. Specifically, the team explored options to use machine learning to create and maintain a public-facing, labor-related data catalog, which would serve as a use case for automated data collection in general. In this effort, the study team—

- Explored the relevant literature
- Piloted a manual process to assemble data that would support a data catalog to inform potential options for automation
- Developed options for automating each step of the data catalog process
- Consulted a technical working group (TWG) of computer science experts to solicit their feedback on the proposed options and automated data collection in general

This brief describes lessons learned from this exploration. In general, the study team found that, at this time, machine learning methods may not be the best tools to create, populate, and update a labor-relevant data catalog. The study team identified several challenges, including the following:

- Some data sources are especially difficult to scrape and may not be well suited for an automatic data collection process.
- DOL CEO may need additional operational capacity to carry out its ambitious vision.
- Scraping a large number of diverse data sources may make using automated methods more challenging.

Key Findings

- Insights from the pilot study and feedback from TWG members suggest that DOL may not be able to achieve its ambitious vision to automate the data catalog process at this time.
- Artificial intelligence (AI) is a rapidly evolving field. Literature suggests there may be many opportunities to support automated data collection in the future, including the potential use of generative AI (Cherradi et al., 2023; Yarlagadda, 2017).
- Federal agencies might explore and use existing tools to meet their needs.
- When using machine learning to produce public-facing products, Federal agencies may need to use a mix of staff with different skill sets, including data scientists, website developers, experts in cloud computing, and subject matter experts.

Based on the study team's experience, the ideal conditions to automate such a process may call for federal agencies to—

- Invest in the staff and computing capacity required for complex machine learning efforts.
- Track new innovations in automation technologies—such as generative artificial intelligence (AI)—and explore how they may be used to meet agency goals.
- Draft agency-specific guidance for teams using machine learning.
- Foster ongoing relationships with machine learning experts who can serve as thought partners in future machine learning efforts.

In addition, the field, including researchers and data collectors, may help facilitate the success of future efforts to automate data catalog creation by—

- More consistently reporting on data and metadata using standardized templates
- Establishing data quality standards to help data catalog creators decide which datasets should be included in data catalogs.

The remainder of this brief describes the data catalog use case, efforts to design automated options to support the development and maintenance of a potential data catalog, challenges in applying these methods, and opportunities for the future.

A. Using Machine Learning to Automate Data Collection—The Data Catalog Use Case

The Chief Evaluation Office (CEO) of the U.S. Department of Labor (DOL) is interested in making data sources and data initiatives in the employment and training field more accessible to researchers (DOL, 2021). Through its Administrative Data Research and Analysis project, DOL CEO asked the Westat Insight and American Institutes for Research teams to explore options to create a publicly accessible data catalog that would help labor researchers more easily understand the full range of available labor-related datasets they could potentially use to answer pressing questions on labor-related policies and programs. DOL CEO envisioned a data catalog that could host information on a range of labor-related data sources and their metadata. Data catalogs are organized inventories of datasets that enable researchers and other members of the public to scan potentially available data and quickly understand what data are included in those datasets and how to potentially access them. DOL CEO was also interested in exploring how machine learning, a type of artificial intelligence that relies on using data and algorithms to imitate human behavior, could be used to automate the creation and ongoing maintenance of the data catalog.

> ***Machine learning*** is a branch of artificial intelligence (AI) that uses the development, training, validation, and deployment of algorithmic models to mimic human behavior.

DOL CEO also envisioned the creation and ongoing maintenance of a public-facing data catalog as the use case to explore how machine learning could be used to enhance data sharing and access at DOL. In having to propose a set of automation options for a specific process (in this case, identifying, scraping, and categorizing potential datasets and/or their metadata for the data catalog), the study team had to consider how operational, legal, or other constraints would affect the feasibility of each option proposed.

1. Using a Data Catalog as a Use Case

DOL CEO envisioned a potential data catalog serving as a "one-stop shop" to make researchers aware of a wide range of labor-related datasets and how to access them. DOL CEO anticipated that such a data catalog would be publicly available on its website and include a large range of labor-related datasets structured by topic area. In discussions, DOL CEO indicated the data catalog was to include the following components:

- Individual- and/or organization-level data that could be used to replicate existing analyses or conduct new ones to inform the field
- Publicly available data and, to the extent possible, information on restricted-use data and how to access them
- Numerical and text data

> ***Restricted-use data*** are data or datasets that are not publicly available and that can only be accessed through specific processes as local, State, and Federal laws allow.

As envisioned, the catalog would include information on each dataset and its metadata, which describe information about each dataset, such as the dataset's name, creator and/or institutional owner, upload date, and tags describing its content. Table 1 contains a sample of metadata tags the study team would have sought to collect using automated methods and published in the data catalog. Possible metadata tags were informed by Project Open Data Metadata Schema (Open Schema) guidelines.[1] To the extent possible, the study team planned to use tags from this schema so the catalog would align with other Federal efforts.

Table 1. Sample Metadata Tags for Potential DOL Data Catalog

Tag(s)	Definition
Title	Name of the data source
publisher and bureauCode	Publishing entity that created the data source, along with the Federal agency and bureau, if relevant
landingpage	Location of information about the data or initiative (e.g., web page, report)
description and key word	Data elements available in the data source or initiative, the population of interest, the nature of the data, and the type of information source
license and rights	Accessibility of the data or initiative (e.g., public, free but restricted, proprietary), including reusability considerations
dataQuality	Data quality assessment following the Federal Committee on Statistical Methodology Framework on Data Quality

Sources: Project Open Data Metadata Schema (DCAT-US Schema v.1.1) guidelines, https://resources.data.gov/resources/dcat-us/; Federal Committee on Statistical Methodology Framework on Data Quality, https://nces.ed.gov/FCSM/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf

The study team assumed a large variety of data sources would include labor-relevant data. The study team anticipated that possible sources of labor-relevant data might include (a) academic journals; (b) data from corporations or private entities; (c) existing data repositories; (d) local, State, and Federal government agencies; (e) public-facing web pages with relevant text data; and (f) research and

[1] See https://resources.data.gov/resources/dcat-us/

evaluation projects that had posted their analytic files. In exploring the use case, the study team did not have strong preferences for some sources of data over others.

2. Creating a Data Catalog: Overview of the Manual Data Catalog Pilot

To understand which activities could be automated, the study team first conducted a manual pilot of the process for identifying, collecting, and tagging datasets for inclusion in the catalog (referred to as the "data catalog process" in the rest of this brief). To do this, the study team created search terms relevant to a potential topic area—reentry populations.[2] Search terms included reentry, prisoner reentry, employment and reentry, and prisoner reentry and employment. The study team limited its search to sources published in or after 2010. We then:

> The ***reentry populations*** topic area explores issues related to individuals returning to the workforce after involvement in the criminal justice system.

- Searched a variety of data sources for relevant datasets, including academic journals[3] such as *American Journal of Criminal Justice* and *Journal of Correctional Education* and data repositories such as the National Archive of Criminal Justice Data
- Tracked information about potential data sources and datasets, including their location, dataset name, probable sample size, whether datasets were directly available, and relevance to the proposed data catalog

Piloting the data catalog creation process enabled the study team to explore the likely set of data sources and datasets for inclusion in the catalog, as well as the steps needed to create, populate, and update the data catalog over time. Ultimately, the study team screened 60 data sources to identify possible datasets or leads to datasets for inclusion in a potential data catalog. Through this search, the study team identified 37 datasets that were relevant to the topic of reentry populations and could potentially be included in a labor-related data catalog. Most of these datasets were identified through existing data repositories. Though articles in academic journals often referenced potentially relevant datasets, journals did not typically include those data for public review. Some of the datasets identified in journal articles are publicly available datasets. Others are proprietary or restricted use and could not reasonably be accessed by the study team. Across all data sources, the study team found substantial variation in how data and metadata were reported across and within potential data sources.

[2] In consultation with DOL CEO, the study team selected reentry populations as our topic area of interest for the use case because searches would likely contain a wide range of publicly available and restricted-use data across a diverse range of data sources, including academic journals, data repositories, and administrative data held by State and local governments. The potential diversity of data encouraged the team to create automation options that could capture data and metadata available in a variety of environments, reflecting the real-world change of automating this kind of data collection.

[3] The study team searched for datasets or leads to datasets in the following academic journals: *American Journal of Criminal Justice, Behavioral Sciences & the Law, Contemporary Justice Review, Criminal Behaviour and Mental Health, Journal of Correctional Education, Journal of Experimental Criminology, Punishment & Society, Social Justice, The Journal of Research in Crime and Delinquency,* and *Theoretical Criminology*.

B. Findings From the Manual Pilot—Data Catalog Activities That Could Potentially Be Automated

> **Application programming interface** is a software interface that enables pieces of software to communicate with each other. Some websites' APIs make it significantly easier for computer programmers to obtain data from those websites.

The manual process pilot provided valuable insights on how to create a data catalog, which steps in the process could likely be automated, and what challenges the study team might need to consider when designing automated options. For example, the study team found that data sources related to the potential topic area varied widely in scope, administration, and accessibility. Topic-agnostic data sources such as journal repositories offered application programming interfaces (APIs)—software interfaces that facilitate automated data collection and make websites much easier to scrape.[4] These data sources often contained the same datasets and metadata as smaller data sources with a more focused scope, suggesting these repositories could be used to find much of the data of interest to DOL CEO. Since the study team found that datasets were frequently duplicated across data sources, we anticipate that it would be important for an automated solution to be able to standardize metadata[5] and intelligently deduplicate datasets. These insights guided the study team's thinking when designing possible automated options, described in section C.

Based on the pilot process, the study team identified four main tasks from the manual data catalog process that could potentially be automated:

1. Identify potential sources of datasets. DOL envisioned including a diverse set of data sources, which would need to be identified from the web. The study team would need to identify potential sources relevant to specific topic areas. Systematic review efforts have had success automating the source identification portion of their processes, which suggests this might be a key area for automation (Marshall & Wallace, 2019; van Dinter et al., 2021).[6]
2. Identify potential datasets for inclusion in the catalog. Once potential data sources have been identified, the study team would need to search each data source to identify specific, potentially relevant datasets that meet the criteria for inclusion. Systematic review efforts have also had success in using automation techniques to classify particular articles as relevant or not relevant to their efforts based on a subset of labeled data (Marshall & Wallace, 2019; van Dinter et al., 2021). The study team believed many of the same principles used to categorize journal articles for systematic review efforts could be applied to dataset classification.
3. Scrape data (as available) and metadata and populate tags with appropriate information about the dataset. Once datasets had been flagged as appropriate to include in the catalog, the study

[4] During the pilot phase, the study team identified two large topic-agnostic data repositories that included relevant datasets: data.gov and Harvard Dataverse. Though the team did not explore it for the pilot effort, Google's dataset search is likely another topic-agnostic data repository that could be used to create a labor-relevant data catalog.

[5] Each dataset's metadata are unique, and all datasets in a shared data catalog may not report the same set of metadata. To report a standard set of metadata across datasets, the study team would need to identify a common or standard set of metadata shared by most datasets and apply metadata tags to all datasets accordingly. If datasets do not include information on standardized tags in their metadata, study team members or automated solutions would need to review datasets (if available) and apply tags appropriately.

[6] Several systematic review efforts have used automated technologies to facilitate study screening. For example, a systematic review may initially identify 1500+ potentially relevant studies. Before the rise of automated technologies, a trained individual would need to review each study manually and determine its appropriateness for inclusion in the review effort (also known as screening). Today, using automated technologies, systematic review teams need to screen a fraction of the potentially eligible studies because automation tools can identify patterns in study eligibility from the initial set of studies screened by trained staff and predict which additional studies will be eligible.

team would need to scrape that dataset and its metadata and populate the Open Schema tags appropriately. The literature did not suggest that this step would be easily automated. However, the study team wanted to explore the possibilities in automating this step because it would likely be labor intensive for human coders to complete.

4. Populate a web-based catalog with new entries. Once tags had been populated, the study team would need to populate the data catalog regularly. DOL CEO envisioned updating the catalog over time as new datasets became available, which would entail determining whether and how frequently each dataset had to be updated, establishing a schedule for identifying updates, and posting updated data and metadata to the catalog. The study team explored efforts to automate this step to reduce burden for staff maintaining a potential data catalog and make the catalog more sustainable in the long run.

The study team proposed a range of options to automate each of these tasks in the data catalog process. These options are described at a high level in the next section and in more detail in Appendices A and B.

Web scraping (or scraping) refers to the process of collecting data from websites using machine learning tools (Mills De La Rosa et al., 2021).

C. Options for Automating the Data Catalog Process

Based on the study team's technical knowledge of machine learning methods and the likely data sources and datasets of interest, the study team explored and designed a number of potential options for automating each step of the data catalog process identified during the manual pilot. In general, these options fell into one of two approaches: a tailormade automated data catalog approach and an API-focused automated data catalog[7] approach. Each approach is described in more detail below:

- **Tailormade approach.** In the tailormade approach, the study team would build a series of customized scrapers for each website to be explored for potentially relevant data as well as a series of algorithms to identify and categorize potentially relevant data and metadata. Using this approach, the study team would need to build a unique series of scrapers for each data source website that was attuned to that website's structure, click patterns, and any antiscraping technologies the website deploys.
- **API-focused approach.** Using the API-focused approach, the study team would focus on building scrapers and algorithms only for potentially relevant websites that included both (a) topic area–relevant data and (b) an API to facilitate automated data collection. The study team's experience from the manual pilot suggests that few data sources meet both criteria.[8]

Although both the tailormade and API-focused approaches are potentially useful for automating the data catalog process, they have key differences:

- **The tailormade option pursues a much larger and wider variety of data sources than the API-focused option.** The tailormade approach purposely focuses on gathering the largest and most diverse set of datasets possible. By contrast, the API-focused approach limits itself to only those

[7] Both approaches were designed before the release of ChatGPT in 2022 and similar generative AI tools. As described in the opportunities section, generative AI may play an important role in making an automated data catalog operationally feasible.
[8] As noted above, just two data sources, Harvard's Dataverse (https://dataverse.harvard.edu/) and data.gov, that we identified in the manual pilot met these criteria.

data sources that (a) contain labor-relevant datasets and (b) have APIs available for easy data extraction.

- **The tailormade approach requires a much more computational and labor-intensive data source identification strategy than the API-focused approach.** As noted in table 2, the tailormade approach relies on a series of scrapers and algorithms to identify potential data sources and datasets, scrape plain-text data related to both, and assess the data's relevance to the task (i.e., how likely each is to contain labor-related data). Building, deploying, and maintaining a series of datasets is computationally and labor intensive. In addition to programmers who build, train, deploy, and revise code, the tailormade solution requires a large number of staff to label training datasets that will be used to train each algorithm designed. The API-focused approach has a less intense data source identification strategy. The study team's experience from the pilot process suggests that few data sources are both labor-related and have APIs to facilitate data extraction. Therefore, staff would rely on the subject matter expertise of content area experts to create a list of likely data sources that meet both criteria. This would require no computational time and far fewer staff hours.

Using Algorithms to Automate the Data Catalog Process

To some extent, both the tailormade and API-focused approaches used algorithms to automate the data catalog process. When using algorithms, Federal agencies or their contractors would need to do the following:

1. **Create well-labeled datasets** that could train algorithms on which data sources or datasets are most relevant to the topic area of interest. This would represent a substantial investment because well-labeled datasets for this activity may not exist. Substantial staff time would be dedicated to labeling data sources and datasets relevant or not relevant.
2. **Check the relevance of a subset of algorithm-produced results** to determine whether the algorithm is performing well.
3. **Revise and retrain the algorithm** until the algorithm's results reasonably meet agencies' or contractors' anticipated outputs. In most cases, it may not be possible for an algorithm to perform perfectly all the time. Agencies and/or contractors would need to determine an acceptable level of accurate performance for each task performed by each algorithm.

- **The API-focused approach targets data sources for which dataset extraction will be easier using automated methods.** APIs are software interfaces that enable pieces of software to communicate with one another. Many website APIs contain features for each data extraction of key data points, including .csv datasets and their metadata. Extracting data via API would enable Federal agencies to take advantage of data extraction tools that website owners have built and incorporated into their websites. Although the study team would need to extract data through API, they would not need to develop web scrapers customized to each website that require human-like navigation and automated keystrokes. The study team would also likely have an easier time tagging metadata from datasets identified through the API-focused approach because many APIs include these metadata as a standard part of their data extraction. In contrast, the tailormade approach would require customized web scrapers for each web page, which would be more complex and time-intensive to create.

Table 2 describes the high-level steps the study team identified as necessary to implement the tailormade approach and the API-focused approach for each step in the data catalog process. Appendices A and B provide more detail on how to implement each approach and a complete list of all the automation options considered across the two approaches. An important initial step to implement

either approach would be working with DOL CEO to identify topic areas of interest and search term definitions. By identifying the topic area and identifying search terms, the study team:

1. defines the scope of the search to be conducted. Doing so allows the team to appropriately determine whether particular data sets are eligible for inclusion and helps to narrow the search to a defined set of parameters, and

2. helps data catalog teams identify potential data sources for exploration to those that are most relevant to the topic area of interest or that are identified by using the selected search terms.

Table 2. Two Proposed Approaches for Automating the Data Catalog Process

Data Catalog Process Step	Tailormade Approach	API-Focused Approach
Step 1: Identify potential data sources	▪ Identify search engine APIs to use for web crawling.* ▪ Build web-crawling code to conduct regular internet searches for websites that may have labor-relevant datasets. ▪ Flag websites for dataset extraction based on keywords and identifiers from web page text.	▪ Use subject matter expert recommendations and web searches using search terms, and manually identify possible data sources. ▪ Identify sources with mature API frameworks for data extraction. ▪ Flag sources for dataset extraction.
Step 2: Identify potential datasets	▪ Construct an algorithm that uses keywords and source format features to identify data sources. ▪ Use algorithms to check all candidate websites for possible datasets. Web crawlers will extract plain text from each web page of candidate data sources' websites. Algorithms will then scan text for keywords and other dataset identifiers.	▪ Manually check each candidate website for relevant data sources. ▪ Check each data source to verify API availability. ▪ Flag data sources that meet these criteria for data/metadata extraction.
Step 3: Scrape data and metadata and populate tags	▪ Build a scraper for each relevant website, which involves manual programming customized to each data source. ▪ Identify credentials and click patterns required to navigate to web page with data sources. ▪ Identify HTML path to objects that store data sources, and program clicks/keystrokes. ▪ Test and run scraper to extract data from the website. ▪ Update data catalog with extracted data and metadata.	▪ Build a scraper for each relevant website, which involves manual programming customized to each data source. ▪ Identify API endpoints and obtain credentials for them. ▪ Create extraction code that requests data and metadata from appropriate API endpoints. ▪ Test and run scraper to extract data from the website. ▪ Update data catalog with extracted data and metadata.
Step 4: Populate a web-based catalog with new entries	▪ Set up servers to run each scraper regularly for each website flagged. ▪ Have servers report on each run's success. ▪ If scrapers fail, troubleshoot to resolve the issue. ▪ Scrapers update the catalog with new data and metadata.	▪ Set up servers to run each scraper regularly for each website flagged. ▪ Have servers report on each run's success. ▪ If scrapers fail, troubleshoot to resolve the issue. ▪ Scrapers update the catalog with new data and metadata.

* Web crawling refers to the process of using a bot to download and index content from the internet. Web crawlers can download and index a large volume of internet content that can then be parsed for relevancy.

Though it is possible to use either of these approaches for this data catalog automation use case, Federal agency staff or contractors may pick one over the other based on their specific needs and preferences. For example, if Federal agency staff or contractors (1) wanted to capture the largest possible number of topic-relevant datasets and (2) had the capacity to build and maintain a series of scrapers and algorithms, they might use the tailormade approach.

For this data catalog automation use case, the study team would recommend using the processes described under the API-focused approach rather than the tailormade approach for two reasons:

1. During the manual pilot, the study team found that data sources that had APIs included most of the datasets found in more narrowly topic-specific data sources that did not have APIs. This finding suggests that many of the topic-relevant datasets could be identified by using topic-agnostic data sources that include APIs to substantially ease the burden of data extraction.
2. The API-focused approach is far less labor-intensive and may be more sustainable over time because data sources' web developers will update APIs over time to reflect website changes.

Though it is technically possible to use a combination of tailormade and API-focused approaches to automate the data catalog, the study team does not think it makes operational sense to mix and match processes from the two approaches. Given limited resources and staff capacity, the study team does not recommend engaging in both the tailormade and API-focused approaches because the pilot experience suggests that many of the topic-specific datasets likely to be identified through the tailormade approach would also likely be identified using the API-focused approach, limiting the value add of executing the tailormade approach. If Federal agency staff or contractors wanted to implement both approaches, they might first conduct automation attempts using the API-focused approach and only engage in the tailormade approach to capture specialty data sources with datasets not captured using the API-focused approach.

D. Lessons Learned and Opportunities for the Future

DOL CEO encouraged the study team to consider an ambitious set of options to automate as much of the data catalog development process as possible. DOL CEO also asked the study team to establish a technical working group (TWG) to solicit members' feedback on the feasibility of each option in DOL CEO's likely operating environment. Based on the TWG's feedback in a virtual meeting in 2023, the manual data catalog process, and subsequent conversations with DOL CEO, the study team determined that using machine learning methods to automate some or all parts of the data catalog process is not currently feasible. This section describes the challenges of using machine learning methods for automating the data catalog process at this time and opportunities for future work to support this process.

Technical Working Group

To better understand the feasibility of different automation options, DOL CEO convened a TWG consisting of two experienced computer science professors from top-25 research universities that publish at the forefront of machine learning and automation-supported decision making. The study team sent TWG members an options memo and asked them to provide feedback on proposed options in a 2-hour virtual meeting in May 2023.

1. Lessons Learned

Exploring options to automate the data catalog process revealed some potential challenges DOL CEO and other Federal agencies might face when attempting to automate data collection using web scraping. Three lessons learned follow:

Lessons Learned

1. Some data sources are especially difficult to scrape using automated methods.
2. DOL CEO may need additional operational capacity to achieve its ambitious vision.
3. Scraping a large number of diverse data sources may make it more challenging to use automated methods.

Lesson 1: Some data sources are especially difficult to scrape and may not be well suited for an automatic data collection process

The data sources explored for this effort presented a number of scraping challenges, described in detail below.

- **Some data sources, like academic journals, do not consistently or reliably report information on the data used to generate findings.** According to the study team's pilot experience, even when data are described, academic journals do not typically make the data used to generate findings available to individuals who access journal articles. Article authors do not typically provide enough information about the data to populate tags.
- **Some data sources are not formatted for scraping.** Web developers often do not design web pages with data extraction in mind, instead prioritizing user experience and visual presentation over data accessibility. These priorities often make it more challenging and labor intensive to create scrapers because the structure and organization of such web pages are not optimized for easy data extraction. Federal agency staff or contractors who develop scrapers to support automated data collection techniques would have to adapt their scraping techniques to account for web developers' focus on user experience and presentation, which would add complexity and time-consuming steps to the scraping process. In addition, popular computing information sharing sites, including Stack Overflow, note that some website developers take additional steps to make data difficult for scrapers to access; such methods include embedding data in complex structures, loading data dynamically through JavaScript, or protecting data with antiscraping measures like CAPTCHAs or rate limiting. Federal or contractor staff attempting to scrape data from websites that use these features may need to alter workflows, such as scraping in smaller batches, to extract data from these sites. Finally, web pages containing data sources may change frequently over time, making it difficult for scrapers to maintain consistency as website layouts, structures, and URLs evolve. The changing nature of many data sources means teams that hope to use scrapers to automate data collection would need to maintain and update scrapers to respond to website changes on an ongoing basis (Dogucu & Cetinkaya-Rundel, 2021).

How Do Websites Prevent Scraping?

Some website developers take additional steps to prevent scraping, including—

- Embedding data in complex structures
- Loading data dynamically through JavaScript
- Protecting data with antiscraping measures such as CAPTCHAs or rate limiting

Federal or contractor staff attempting to scrape data from websites that use these features may need to alter workflows, such as scraping in smaller batches, to extract data from these sites.

- **Many organizations that host restricted-use data do not describe these data or the process to access them in sufficient detail on public-use websites.** In many cases, organizations that collect and/or host restricted-use data have a legal or business obligation to keep data private and protect datasets' contents. This may mean these organizations do not sufficiently describe (a) whether particular datasets can be accessed and by whom, (b) the procedures for accessing data, and (c) the specific variables the datasets contain. This may limit the ability of Federal agency or contractor staff to describe the data and metadata in sufficient detail to populate a data catalog. The restricted-use nature of these datasets also means they are not readily available for the Federal agency or contract staff who are compiling the data catalog to access, assess, and describe the data and metadata for the data catalog.

Lesson 2: DOL CEO and other Federal agencies interested in similar efforts may need additional operational capacity to carry out an automated data collection and categorization effort of this size and scope

TWG members said that, in their expert opinion, an automation effort of this size and scope would take considerable resources to implement successfully. In part, this is because the envisioned effort is essentially a very large data collection task with a robust data source and dataset identification strategy. Efforts to identify relevant data would also inadvertently identify a large number of slightly relevant datasets that would need to be screened out as poor fits for the catalog. Making sense of such "noisy" data is not uncommon in the information technology field, but sense-making efforts often try to leverage existing training sets to minimize the level of effort to the extent possible. To the study team's and DOL CEO's knowledge, no existing training datasets would facilitate the data catalog effort. Based on likely operating constraints, described in more detail in table 3, TWG members indicated that DOL CEO would likely need significant computing and staff resources to pursue the proposed automation approaches, especially the tailormade approach.

Table 3. Likely Operating Environment for Automated Data Catalog Process Solutions

Category	Assumptions
Data catalog architecture	The high-level architecture of the data catalog would have to consist of— ■ A front-end website where public users could browse and query the data catalog ■ A back-end database that stores the metadata ■ Maintenance programming scripts/manual procedures for creating and maintaining the data catalog
Data catalog operations	■ A cloud-based server would have to host the architecture, data catalog, any datasets directly downloadable from the catalog, and any automated solutions; it also would have to execute any automated scripts for creating/maintaining the catalog
Practical considerations	■ Approximately 0.25 full-time equivalent staff would have to be available to support automation efforts ■ Existing resources for the development and maintenance of automated options might be limited

Lesson 3: Scraping a large number of diverse data sources could make using automated methods more challenging

The study team originally proposed the tailormade approach to DOL CEO and the TWG as an approach that would attempt to collect the largest possible set of labor-relevant data. During the pilot study, which occurred concurrently with automation option development, the study team noticed significant variation in data sources that may make executing the tailormade approach challenging. For example, across data sources, there was substantial variation in—

- The ease with which the study team could identify and scrape information about potentially relevant datasets from each data source
- Each data source's website structure, click patterns, and approaches to documentation information about datasets
- The existence of APIs to facilitate automated data collection from data sources
    - APIs significantly reduce the burden for automating data collection because they enable researchers to extract relevant data and metadata without writing additional scraping scripts. The study team anticipated that most data sources, including academic journals, local and State governments, web pages of many corporate entities, and public-facing labor-relevant web pages, would not have APIs to facilitate the automated collection of data or metadata.

Is Web-Scraping Labor Intensive?

Depending on the scope of the effort, it can be. If DOL CEO had elected to build a data catalog using the tailormade approach, it would have had to build and maintain customized scrapers for tens to hundreds of websites. Each scraper would need to—

- Account for each website's underlying structure and code.
- Mimic human-like navigation of websites.
- Follow website-specific click patterns to each potential data set to be included.
- Overcome obstacles to scraping developed by website owners.
- Be maintained over time as small and large changes to the website caused scrapers to break.

To build and maintain each scraper, DOL CEO would have needed to dedicate sufficient staff time to build each scraper and conduct maintenance tasks regularly.

Based on the likely variation and dearth of APIs among the likely data sources of interest, the study team anticipated needing to build made-for-purpose web scrapers for each data source. Each scraper would need to be configured and maintained over time. If a substantially large number of potential datasets was found, Federal agency or contractor staff might have to build and maintain tens to hundreds of scrapers for a particular topic area.[9] In short, variation in data sources was likely to be a substantial driver of the level of effort and cost. TWG feedback suggested that the level of customization required to address this variation and implement the tailormade approach would likely exceed DOL CEO's existing computing and staff resources.

[9] The number of datasets identified will vary by the prevalence of available data by topic area. For study authors, substantially large may refer to tens to hundreds of data sets that may all need to be screened for inclusion in the data catalog. For example, counties, states, and the federal government all collect data on reentry populations including jail and prison data, unemployment insurance wage records, public benefit receipt, and more. Under the use case used for this report, we could imagine that a robust search of potentially relevant data for reentry populations would generate hundreds of publicly-available and restricted use datasets.

2. Opportunities for the Future

Although the TWG, DOL CEO, and the study team reached a consensus that machine learning techniques were not currently an effective means to conduct large-scale automated data collection and categorization for a given topic of interest, DOL CEO and other Federal agencies have many opportunities to use machine learning to support their data-related missions in the future. This section describes potential opportunities for interested Federal agencies to use machine learning to facilitate automated data collection, based on DOL CEO's example.

Opportunity 1: Make additional investments in staff and computing capacity

Federal agencies require sufficient computing power and staff capacity to successfully complete many machine learning projects. However, TWG feedback suggests agencies may need to make sufficient additional investments to carry out large-scale, exploratory web scraping; automated data collection; and categorization activities like those envisioned in this this case.

Opportunities for the Future

1. Make additional investments in staff and computing capacity.
2. Track cutting-edge methods and explore how they could be used to meet Federal agencies' research on policies, programs, and practices.
3. Establish agency-specific systems, data documentation standards, and operational guidance that staff and contractors can use to guide agencies' machine learning projects.
4. Establish ongoing relationships with machine learning experts who can inform DOL's ongoing strategic decisions on AI and machine learning applications.

Staff capacity: Initial conversations with DOL CEO staff suggested they could commit 0.25 full-time equivalent staff to support this use case to automate processes to create a proposed data catalog. Based on TWG member feedback and the expertise of the study team, table 4 describes the types of staff the study team anticipates Federal agencies would likely need, either internally or through a contractor, to implement the data catalog using machine learning methods. The specific number of full- or part-time staff needed at each level to support any machine learning project would depend on the size and scope of the specific project an agency was attempting to implement.

Table 4. Estimated Staffing Needs for Automating a Publicly Available Data Catalog Process

Staff	Role and responsibilities
Senior data scientist	Manage the overall project, develop the workplan and process design, work with software and cloud computing engineers to assess computing needs, troubleshoot problems, and mentor junior staff.
Midlevel data scientist	Write code to scrape data sources and datasets, train algorithms and assess their performance, review scrape results and revise models as needed, elevate challenges to experienced data scientists and act based on their guidance.
Midlevel subject matter expert	Establish a coding protocol for labeling training data as relevant or not relevant to a particular topic; review junior staff work and provide guidance for edits. These staff do not need to be data scientists but should have a very good knowledge of the subject matter represented in the training data and be familiar with working with data.

Staff	Role and responsibilities
Junior subject matter expert staff	Label training data as relevant or not relevant to a particular topic area. These staff do not need to be data scientists but should have a very good knowledge of the subject matter represented in the data and be familiar with working with data.
Cloud computing engineer	Assist senior data scientists in assessing computing needs, set up and maintain servers that support automated options, troubleshoot problems as they arise.
Software engineer and website developers	Build and maintain the public-facing data catalog tool.

Computing capacity: Not all machine learning applications are computationally intensive. Less complex machine learning projects that have discrete goals and rely on preexisting analytic packages can often be performed on ordinary laptop or desktop computers with no special augmentations. However, for more exploratory tasks that require deploying a series of algorithms to search, identify, and scrape data, agencies may need to invest in additional cloud computing infrastructure. Specific cloud computing needs and costs will depend on the scope of the specific project Federal agencies or contractors undertake. An experienced cloud computing engineer can help Federal agencies or their contractors scope out their specific cloud computing needs and make a reasonable estimate of the cost associated with the estimated additional computing capacity.

Opportunity 2: Track cutting-edge methods and explore how they could be used to support Federal agencies' research on policies, programs, and practices

Machine learning methods are evolving rapidly. New innovations can mean that tasks that were very burdensome a few months ago become much more operationally feasible with new methods (Jordan & Mitchell, 2015). New methods may also result in substantial improvements in accuracy, which is important to Federal research agencies' missions of delivering high-quality data, research, and insights to the field (Pugliese et al., 2021). By tracking new methodological developments, Federal agencies may explore how new innovations make machine learning activities that were previously difficult or produced less accurate results possible or of better quality. The following two examples of burgeoning innovations—generative AI and the Department of Health and Human Services' (HHS) AI Use Case Inventory— may shed light on how Federal agencies could conduct more large-scale, automated data collection efforts.

Generative AI may make some steps of the data catalog process easier to automate than was previously possible. Generative AI tools, known as large language models (LLMs) like ChatGPT, are large neural networks that can process and generate human-like language (Kasneci et al., 2023). The greatest potential value of LLMs is their ability to help automatically identify relevant information from unstructured text (Liu et al., 2023). In the case of a data catalog, LLMs might substantially reduce the burden related to finding appropriate data sources and datasets (see text box for more information). Reducing this burden may make automating the data catalog process more operationally feasible and cost effective for DOL CEO and similar agencies. Pilots and test use cases could help Federal agencies or their contractors determine the extent to which generative AI could facilitate the automation of the data catalog process. Generative AI is a relatively new technology, and agencies planning to use it may want to carefully consider whether its use aligns with their guidance on AI use.

Existing federal applications of machine learning and AI may provide inspiration for how other **F**ederal agencies might use machine learning to answer important policy, program, and practice questions in the future. HHS recently released its Artificial Intelligence Use Cases Inventory (HHS, 2024) covering AI projects implemented in fiscal years 2022 and 2023. This inventory provides information on a variety of Government AI projects and use cases, the office that commissioned the use case, and a short description of what the AI application does. These use cases could serve as rich sources of inspiration for many Federal projects and as a starting point for staff in other Federal agencies to connect with staff who have successfully implemented these projects.

HHS AI Use Cases Inventory

The HHS AI Use Case Inventory provides a high-level summary of 164 use cases of AI tools in the Federal Government. Select projects include chatbots, approaches to facilitate working with large datasets, natural language processing to support comment review, and more. The inventory also includes contact information for agencies overseeing each AI effort. Find out more about this inventory at https://www.hhs.gov/about/agencies/asa/ocio/ai/use-cases/index.html.

Opportunity 3: Establish the agency-specific systems, data documentation standards, and operational guidance that staff and contractors can use to guide agencies' machine learning projects

Although machine learning methods may have great potential to help Federal agencies answer important questions about policies, programs, and practices, these methods are complex, have specific data requirements for insights to be useful, and often entail a substantial investment by agencies. Although machine learning experts have focused much of their effort on methods that make sense of large volumes of data (Pugliese et al., 2021), establishing data quality and documentation standards will be essential to ensure insights are meaningful and can meet specific agency needs. TWG members suggested that DOL CEO develop data quality and documentation standards to help achieve the goal described in its use case—a data catalog that could help researchers answer important questions about labor-related programs and policies. Many Federal agencies, including DOL, have established data enterprise strategies that prioritize shared data quality and documentation standards (DOL, 2022). These strategy documents, as well as any specific guidance developed in their wake, should be widely shared with Federal agency staff and contractors looking to use machine learning methods. Chief data officers within Federal agencies may also provide insights on how existing policies can be adapted for machine learning applications. Finally, the evolving nature of machine learning methods and the legal environment in which they operate suggest that agency staff and contractors will need agency-specific guidance about how to use these methods to inform their real-time decision making on projects. Based on the study team's suggestions, the following steps describe how Federal agencies might move toward establishing standards that will facilitate future use of machine learning in their agencies.

President's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

In October 2023, President Biden released an Executive Order establishing safety and privacy standards for Federal agencies when using AI methods, including machine learning, in the course of their work. The order calls on (a) the developers of AI systems affecting national security, national economic security, or national health to share the results of their safety testing and other critical information with the U.S. Government; and (b) the National Institute of Standards and Technology (NIST) to develop standards, tests, and tools to ensure AI systems are "safe, secure, and trustworthy."

The order also encourages Federal agencies to—

- Prioritize efforts to develop, strengthen, and use privacy-preserving methods in AI systems
- Develop additional guidance for agency staff, contractors, and others on—
  - The use of AI
  - Privacy guidelines and privacy-preserving techniques for AI
  - Addressing algorithmic discrimination to ensure algorithm use does not create or strengthen patterns of discrimination

Finally, the Executive Order encourages efforts to mitigate the risks of AI to workers, calling for—

- The development of principles and practices that will minimize the risks and maximize the benefits of AI to workers
- A report on the possible impacts of AI usage on workers

Source: The White House, 2023

Establish approaches and standards to document data uniformly across each agency's sphere of influence. First, Federal agencies' research arms could work with each of their subagencies to establish a uniform structure for organizing available information on datasets coming from an agency's own project, program, or evaluation work. Next, Federal agencies could identify nontraditional data sources that might be leveraged for agency use and work with data owners to encourage their adoption of standardized data documentation approaches. Opportunities for such an effort seem especially feasible with other Federal, State, and local agencies that also want to understand other organizations' data more easily through standardized data reporting. By ensuring clear and consistent data standards, this effort will contribute to the equitable access of critical datasets for all.

Establish guidance on how Federal agency staff and contractors should navigate the evolving legal environment surrounding machine learning methods. Because machine learning methods are new compared with traditional data collection and analysis methods, the Federal Government and many States, counties, and cities have yet to legislate their use. Case law is also evolving, and what is permitted today may not be in the future (Krotov & Silva, 2018; Sobel, 2022). To ensure agency staff and its contractors act in ways that align with an agency's understanding of this evolving legal environment, Federal agencies could issue guidance to staff and contractors on the following key topics:

- **Use of web scraping.** Web scraping has been the subject of legal disputes in recent years. In a series of high-profile court cases between HiQ Labs and LinkedIn between 2018 and 2022, the U.S. Ninth Circuit Court of Appeals upheld HiQ Labs' ability to scrape publicly available data from LinkedIn's website, despite antiscraping provisions in LinkedIn's user agreement (Sobel, 2022). Although the established case law supports web scraping for now, recent legislative interest in AI and web scraping at the State and Federal levels suggests the legal issues related to web scraping may not be settled for the long term (Harris, 2023). Federal agency staff and contractors may benefit from

agency-specific guidance on how and when web scraping and other machine learning methods can be used appropriately in agency projects.

- **Privacy protections.** Machine learning methods often work by analyzing and identifying patterns in large volumes of data that may not be well understood. Given the volume of data these methods use, it may be challenging for teams using machine learning methods to ensure no identifying information[10] is contained in (a) datasets used to train algorithms or (b) algorithmically produced results. To protect the privacy of all individuals, Federal agencies could release easy-to-understand guidance on the importance of ensuring individuals' privacy when using machine learning methods. National Institute of Standards and Technology (NIST), the White House, and others have recently released high-level guidance related to privacy protections when Federal agencies deploy machine learning projects. Each Federal agency is in a different place in developing its own agency-specific guidance. Especially where guidance does not exist or is in earlier stages of development, agency-specific privacy-related guidance could be contextualized specifically for machine learning applications and include specific instruction on the tools and methods staff and contractors should be using to determine whether AI and machine learning algorithms and outputs meet Federal standards for anonymization, access control, censoring, and identification susceptibility. This guidance could build on Federal agencies' existing privacy regulations, including DOL's (DOL, n.d.). Federal agencies could also release guidance on special protections developers of AI and machine learning algorithms should use to limit the risk of a third party obtaining the identities of program participants because public AI and machine learning tools may increase the likelihood of data exposure (Song et al., 2017; Yeom et al., 2018). With respect to the data catalog, DOL CEO staff and contractors could take steps to ensure the original datasets included in the catalog remove or anonymize personally identifiable information. This might include searching metadata for variable names that could potentially be used for participant identification. The quality of these reviews can only be as thorough as the metadata allow; Federal staff and contractors will have a harder time making thorough privacy checks if metadata are scant or incomplete.

Establish guidelines for identifying and addressing bias that may be present in machine learning algorithms and their results. AI systems are emerging in numerous sensitive settings, where they play a pivotal role in decision making (Liu et al., 2023).[11] In the case of the data catalog, machine learning methods may have over- or underselected certain types of datasets, potentially biasing the types of research or insights generated based on the datasets available in the catalog. As Federal agencies expand the use of machine learning in their research and project work, AI experts recommend agencies take steps to ensure that using machine learning methods does not inadvertently introduce sources of

[10] Though it is important to ensure personally identifying information (PII) is not included in any datasets included in a data catalog, even information that is not personally identifying could help someone reidentify individuals in a particular dataset. Bad actors could potentially reidentify individuals in a dataset by using a combination of non-personally identifying data. For example, imagine the data catalog included a dataset that included survey responses of researchers in a niche field that is dominated by men. By combining location, gender, age, and a general knowledge of researchers in the field, bad actors could potentially link survey responses to the small number of researchers who are women in this field. These responses could be linked to individual researchers if there are only a few in each state by using information on their age.

[11] Examples of AI-use in decision-making include using algorithms to determine creditworthiness in mortgage lending, making diagnoses in the medical field, and predicting market trends and investment decisions in finance.

bias toward specific groups or populations (Mehrabi, 2021). Two efforts could help address the types of bias that may appear in machine learning applications:

- Facilitate greater awareness of the types of bias that can be present in AI or machine learning approaches as well as potential methods to combat this bias. The National Institute of Standards and Technology (NIST) identifies three broad categories of AI bias that need to be managed (NIST, 2022):
  - *Systemic bias* may be found within datasets fed into AI systems or in the organizations that make decisions within the AI lifecycle.
  - *Computational and statistical biases* can be found within datasets and modeling processes; they are often the result of nonrepresentative samples.
  - *Human-cognitive biases* focus on how the outputs of AI systems are thought about in relation to the other factors relevant to decision making.

  Methods to address bias in machine learning are evolving. Federal agency staff and contractors will need to tailor approaches to addressing bias to the machine task and data being used. The NIST risk management framework recommends using experts in data collection and modeling to help Federal agency staff and contractors detect bias in machine learning models and generate appropriate approaches to mitigate these biases to the extent possible. To do so, experts and support staff will likely need to conduct a thorough review of machine learning–produced results to assess whether the results are accurate—and for all subgroups.

AI Risk Management Framework

NIST has created an AI Risk Management Framework designed to be a practical guide on how to make AI applications safer for a large group of audiences. The framework includes information on the types of risk present when using AI; factors to consider when designing safe AI; and approaches to encourage the design, deployment, and use of safe AI over time. To learn more about the framework, visit https://www.nist.gov/itl/ai-risk-management-framework.

- Provide guidance to staff and contractors on how to identify bias in machine learning applications and what to do when they detect bias. In the study team's assessment, to ensure all agency-sponsored machine learning projects are attempting to identify bias, Federal agency staff could provide guidance to staff on the following:
  - At what stages and how frequently in the development process Federal agency staff and contractors are required to check for bias.
  - Preferred methods to check for and mitigate bias in agency-sponsored machine learning applications. This guidance could include a menu of possible identification and mitigation options and guidance on the appropriate use of each method.
  - What staff should do when they detect bias and are unable to remove it. This guidance might include engaging with program staff, institutional review boards, and experts in machine learning ethics to think through the ramifications of continuing to use a machine learning application and whether other, unbiased alternatives exist. Ultimately, this guidance will be less technical, but it will establish a decision-making process that aligns with the agency's values and existing regulations.

Opportunity 4: Establish ongoing relationships with machine learning experts who can inform DOL's ongoing strategic decisions on AI and machine learning applications

A substantial challenge for this project was identifying machine learning experts with the right knowledge of automated data collection procedures, labor-specific content knowledge, and availability to serve as a TWG member. Although DOL CEO has traditionally identified and engaged methodological subject matter experts on an as-needed basis, DOL CEO and other Federal agencies might consider establishing longer term and ongoing relationships with a small number of machine learning experts who are at the intersection of machine learning methods and subject matter expertise. Establishing such relationships would enable Federal agencies like DOL CEO to access machine learning expertise more easily on an ongoing basis. It would also help methods experts develop a thorough understanding of the needs and concerns of their specific agencies to ensure they can make recommendations accordingly. Such ongoing relationships may help Federal agencies consider larger machine learning issues beyond the context of specific projects and build a more robust machine learning infrastructure across agencies and work.

National Artificial Intelligence Advisory Committee

NIST established the National Artificial Intelligence Advisory Committee (NAIAC) in 2023. The NAIAC comprises 35 experts in computer science, AI, social sciences, privacy, and AI ethics. These experts meet regularly to discuss important AI topics of the day and guide NIST's thinking on U.S. AI competitiveness, trustworthy AI systems, preparing the U.S. workforce for AI technology, and coordinating AI development across agencies. More information about the NAIAC, including its most recent recommendations, can be found at https://ai.gov/naiac/.

E. Conclusion

This study was an exciting exploration of whether and how machine learning methods might be used to automate the creation, population, and ongoing maintenance of a public-facing, labor-focused data catalog. Ultimately, DOL CEO, the study's TWG, and the study team determined that automating the data catalog process is not feasible at this time for this use case. However, an investigation these methods identified a list of challenges and opportunities for future consideration as DOL continues to explore its use of machine learning methods.

Appendix A. Options for Automating the Data Catalog Process Using the Tailormade Approach

The study team proposed a range of potential options to automate each task within the data catalog process, which are described in detail in this appendix. Options fell into two categories: options that used machine learning automation (MLA) (Hutter et. al, 2019) and options that used rules-based automation (RBA) (De Ley & Jacobs, 2011). The study team also provided information on how the data catalog process would be completed using manual options as a kind of counterfactual when considering MLA and RBA options. Table A.1 provides a high-level description of these categories and examples of each. Not all option types were viable for every task.

Table A.1. Option Types for Automating the Data Catalog Process

Option Type	Definitions	Example
Machine learning automation	These options employ the development, training, validation, and deployment of machine learning models to automate components of the data catalog creation and maintenance process. Machine learning automation is often not applicable to parts of this process. We have not included MLA options for each proposed step.	Develop a machine learning algorithm to detect links to data sources within a web page.
Rules-based automation	Like machine learning options, these options employ use of computer automation. However, an algorithm is not trained to identify patterns independently and return the desired result. RBA options do not include training an algorithm using labelled data.	Develop a web scraper for a specific website. The scraper navigates and extracts data/files deterministically by navigating set paths within the website's HTML code and taking predetermined actions.
Manual	These options rely solely on the use of human effort to perform tasks. No computer automation is employed.	Have staff review websites by hand for relevant data sources and save the data through human inputs.

The remaining tables describe the options for completing each task using MLA, RBA, and manual options.

Task 1: Identify Potential Data Sources

This task identifies websites that house datasets of potential interest or, in the case of restricted-use data, information on potential datasets of interest for the data catalog. Table A.2 describes the options for this task.

Table A.2. Task 1 Options to Identify Potential Data Sources, by Automation Type

Machine Learning Automation	Rules-Based Automation	Manual
▪ Identify areas of interest and their definitions for search purposes. ▪ Identify search engine offered application programming interfaces (APIs) to use to conduct systematic reviews of websites using web crawling. ▪ Build web crawling code to regularly conduct searches of the internet for websites that may have datasets DOL is interested in. ▪ Manually read and tag a large number of websites as containing relevant data sources. This will be used as the training dataset for the machine learning model. Repeat this process for each new topic area. ▪ Construct code to turn a website's text, links, and source code into model features. ▪ Train a machine learning algorithm to classify whether web pages contain a data source from the manually labeled websites. ▪ Apply machine learning algorithm to all websites flagged by the web crawler and flag websites to search for data sources with it (for task 2).	▪ Identify areas of interest and their definitions for search purposes. ▪ Identify search engine APIs to use to conduct crawling. ▪ Build web crawling code to regularly conduct searches of the internet for websites that may have datasets DOL is interested in. ▪ Flag websites for dataset extraction using keywords/other identifiers from web page text (for task 2).	▪ Identify areas of interest and their definitions for search purposes. ▪ Manually scan search engines for candidate sources using keywords. ▪ Flag sources for dataset extraction (for task 2).

Task 2: Identify Potential Datasets for Inclusion in the Catalog

Once websites had been identified as containing potentially relevant datasets, the next task would have involved searching the websites for datasets, determining their relevance, and flagging relevant datasets for inclusion in the catalog. Table A.3 describes options for identifying datasets.

Table A.3. Task 2 Options to Identify Potential Datasets for Inclusion in the Catalog, by Automation Type

Machine Learning Automation	Rules-Based Automation	Manual
▪ Manually read and tag several hundred web pages as having or not having a dataset and, if there is a dataset, whether the dataset is relevant. ▪ Construct code to turn a website's text, links, and source code into model features. ▪ Train a machine learning algorithm to identify whether a certain web page contains a dataset. ▪ Train a second machine learning algorithm to detect whether a dataset is relevant to DOL. ▪ Apply machine learning algorithm to all websites collected in task 1 and flag all web pages that contain datasets of interest.	▪ Construct an algorithm that uses keywords, presence of certain kinds of links (e.g., common data source formats such as .csv) as identifiers of data sources. ▪ Check all candidate websites for keywords and other identifiers. This involves crawling through each web page on the website, extracting the plain text from the website's source code, and scanning the plain text for the keywords and other identifiers in the algorithm.	▪ Manually check each candidate website for relevant data sources and identify which web pages have them for data/metadata extraction (task 3).

Task 3: Scrape Data and Metadata About the Datasets, and Tag Datasets

After datasets were identified for inclusion, the next task would have been to extract the dataset's metadata for inclusion in the catalog. If possible, the actual datasets would have been extracted from the website and made available in the catalog. This step also would have included tagging the datasets according to the appropriate schema distinctions. Table A.4 describes options for scraping data and metadata and tagging datasets.

Table A.4. Task 3 Options to Scrape Data and/or Metadata About the Datasets and Tag Datasets, by Automation Type

Machine Learning Automation	Rules-Based Automation	Manual
Machine learning automation is not feasible for this task because there is no specific question or action to train a model to perform or not perform.	▪ Build a scraper for each website flagged. This step consists of manual programming customized to each data source. ▪ Identify credentials and click patterns required to navigate to web page with data sources. ▪ Identify HTML path to objects that store data sources. ▪ Program click and keystrokes. ▪ Test and run scraper to extract data from the website. ▪ Update data catalog with extracted data and metadata.	▪ Manually extract data and record relevant metadata in the data catalog.

Task 4: Populate a Web-Based Catalog with New Entries

Datasets would have been updated over time, and the study team anticipated new datasets of interest would have become available. To keep the catalog up to date, procedures to rescan the internet for updated or new datasets would have been necessary. Table A.5 describes options for identifying and updating the catalog with new entries.

Table A.5. Task 4 Options to Update a Web-Based Catalog With New Entries by Automation Type

Machine Learning Automation	Rules-Based Automation	Manual
Machine learning is not an appropriate technique to use for this task because there is no specific question or action to train a model to perform or not perform.	▪ Set up servers to regularly run each scraper for each website flagged. ▪ Have servers report on the success of each run. ▪ If the scrapers fail, have programmers check each scraper and troubleshoot to resolve the issue. ▪ The scrapers will then update the data and metadata in the data catalog.	▪ Create a schedule that specifies when each data source is anticipated to be updated. ▪ Have staff available to manually check the data source at each specified time. ▪ Have staff manually transport the updated data and metadata to the data catalog.

Considerations for Automated Options

When contemplating which option may be preferable to accomplishing each task, the study team asked TWG members to consider the strengths and limitations of each approach and the level of effort associated with each approach. Table A.6 highlights some likely implications for each option.

Table A.6. Important Considerations for Automated Options

Machine Learning Automation (MLA)	Rules-Based Automation (RBA)	Manual
▪ For tasks 1 and 2, MLA would involve a complex implementation process that required creating multiple training datasets and multiple algorithms. Doing so would likely require a significant level of effort (Razno, 2019). Text-based models often have longer runtimes because of the large number of features and parameters associated with natural language processing. – In some cases, models may have struggled to correctly identify relevant information because of semantic nuance, especially with terms not often uniformly defined, as is typical in workforce and labor topics. – The study team would need to use subject matter experts to correctly tag training data and assess model performance. ▪ MLA would not be a feasible option for tasks 3 and 4 because there is no specific question or action to train a model to address.	▪ Would require a simpler implementation approach than MLA, largely because RBA approaches would not have required the study team to create training data or train models. ▪ Keywords may have been insufficient to accurately identify websites of interest. ▪ Because web scrapers would have to be tailored to each website, the RBA approach would require significant programmer time to build and maintain multiple scrapers. ▪ Programming costs would be higher if the study team primarily pulled data from sources without application programming interfaces. ▪ Websites change over time, and scrapers can be sensitive to even minor changes. DOL would need to invest resources in routine scraper maintenance, especially as websites updated their formatting and content over time. ▪ Feasibility testing would help determine this approach's appropriateness for the task.	▪ Labor-intensive implementation would require a large number of junior staff to— – Search, identify, and assess appropriate data sources and datasets. – Enter metadata information into a database of data catalog information. ▪ Would require regular staff trainings to encourage interrater reliability and a quality control process to ensure staff are identifying appropriate datasets. ▪ Would require careful documentation across a large group of staff. ▪ DOL's systematic review efforts provide a potential template for manual coding efforts that do not exist for MLA or RBA approaches.

Appendix B. Options for Automating the Data Catalog Using the API-Focused Approach

An alternative to the tailormade approach is an API-focused automation approach. Using an API-focused approach, the study team would limit the search to a limited set of data repositories that (1) include some topic-relevant data and (2) have APIs to support data extraction. By using the data sources most amenable for automated approaches, the study team may have built and demonstrated the value of an initial data catalog in a cost-feasible way. This appendix describes a potential process for building such a catalog. To use this approach, the study team would need to label training datasets and train models for each new topic area. However, the study team would be able to preserve the scraping code for each website, perhaps with a limited number of tweaks.

Step 1: Search Accessible and Expansive Catalogs

Under this approach, the study team would prioritize searching existing data repositories with topic area–relevant information that also had search APIs. Data catalogs that cover an expansive set of topic areas and datasets frequently draw from topic-specific data repositories, usually maintained by a government agency, nonprofit, research university, or combinations of these.[12] Examples of such catalogs include catalog.data.gov and the Dataverse project by Harvard. The proposed automated search workflow would identify any of these topic-specific catalogs that more expansive catalogs integrate. This approach would allow the study team to build a list of more topic-specific repositories to be assessed for inclusion over time. If topic-specific repositories are found to include more relevant datasets than the expansive catalog includes, the study team could assess whether it makes sense to build a scraper for that repository. Decisions to build scrapers for topic-specific repositories would be made on a case-by-case basis and consider whether the topic-specific repository (1) has an API to facilitate data extraction, (2) has datasets pertaining to multiple labor topics of interest, and (3) addresses a data gap not well-filled by other data sources.

Step 2: Label Search Results for Training Dataset

Once search results are collected, standardized, and deduplicated, a subject matter expert would determine whether each result qualifies for inclusion in the catalog. The ultimate goal of this effort is labor-saving because the qualification labels will serve as the training dataset for the machine learning model described in Step 3. This model would use the labels provided by the subject matter expert to estimate the relevance of any new datasets loaded into the model, whether newly published or uncovered for the first time by the integration of an additional repository.

Step 3: Train Natural Language Processing Model on Relevancy Labels

Once a sufficiently sized batch of search results are labeled as relevant or not relevant, a set of topic-agnostic algorithms would extract natural language processing features from search result metadata and fit machine learning predictive models on them to predict relevance. Decision tree models are efficient at identifying text terms in metadata that strongly correlate with relevance. Relative to the human

[12] In the context of ex-offender reentry, both expansive catalogs link extensively to the National Archive of Criminal Justice Data, sponsored by the National Institutes of Justice and maintained in part by the University of Michigan.

development of a rules-based approach, this machine learning application would save time because modern algorithms vastly outperform the human brain in pattern recognition on large sets of structured data (Kowsari et al., 2019).

Experience from the pilot shows many search results are not datasets generated from a program evaluation but are administrative data that do not feature a program or policy intervention. Though these administrative datasets can constitute a large share of search results, they are usually deemed less relevant by the search result ranking algorithms embedded in each catalog. A trained machine learning model, however, can determine their irrelevance even more quickly by keying in on terms such as "aggregated," "census," and "State-level," as well as the absence of text-based program data correlates such as "evaluation," "participants," or "treatment."

The minimum number of datasets required for effective prediction is not known in advance. The minimum number required depends on a variety of factors, including the richness of metadata, complexity of the topic, terminological similarity to other topics, and catalog-dependent dispersion of topics returned by the search terms. The study team was not concerned about the lack of enough search results for model training, however. Based on the pilot, the study team anticipated expansive catalogs to return hundreds of search results that could have been labeled by a subject matter expert within a few hours.

Step 4: Integration of New Datasets

Once a predictive model is trained and tuned on a labor-relevant topic, the system can accept new metadata entries and made predictions on their relevance. This process can occur on newly published datasets or on preexisting but previously unseen datasets introduced by a new scraper.

Appendix C. Works Cited

Cherradi, M., Bouhafer, F., & El Haddadi. (2023). Data lake governance using IBM-Watson knowledge catalog. *Scientific African*, *21*(September 2023). https://www.sciencedirect.com/science/article/pii/S2468227623003101.

De Ley, E., & Jacobs, D. (2011). Rules-based analysis with JBoss Drools: Adding intelligence to automation. In *Contributions to the Proceedings of ICALEPCS 2011* (pp. 790-793). https://accelconf.web.cern.ch/ICALEPCS2011/papers/wepks008.pdf.

Dogucu, M., & Çetinkaya-Rundel, M. (2021). Web scraping in the statistics and data science curriculum: Challenges and opportunities. *Journal of Statistics and Data Science Education*, *29*(Supp. 1), S112–S122. https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1787116.

DOL (U.S. Department of Labor) (n.d.). *Department of Labor Privacy Program.* https://www.dol.gov/general/privacy.

DOL. (2022). *Enterprise data strategy.* https://www.dol.gov/sites/dolgov/files/Data-Governance/DOL-Enterprise-Data-Strategy-2022.pdf.

Federal Committee on Statistical Methodology. (2020). *A Framework for Data Quality.* https://nces.ed.gov/FCSM/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf.

Federal Housing Finance Agency. (2022). *AB-2022-02: Artificial intelligence/machine learning risk management.* https://www.fhfa.gov/SupervisionRegulation/AdvisoryBulletins/AdvisoryBulletinDocuments/Advisory-Bulletin-2022-02.pdf.

General Services Administration IT Modernization Center of Excellence. (2023). *AI guide for government: A living and evolving guide to the application of artificial intelligence for the U.S. Federal Government.* https://coe.gsa.gov/ai-guide-for-government/.

Government Accountability Office. (2024). *Artificial Intelligence.* https://www.gao.gov/artificial-intelligence.

Harris, L. (2023). *Artificial intelligence: Overview, recent advances, and considerations for the 118th Congress*. Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R47644.

HHS (U.S. Department of Health and Human Services). (2024). *Department of Health and Human Services: Artificial Intelligence Use Cases Inventory*. https://www.hhs.gov/about/agencies/asa/ocio/ai/use-cases/index.html.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning.* Springer Cham.

Jordan, M., & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://www.science.org/doi/10.1126/science.aaa8415.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103(102274). https://www.sciencedirect.com/science/article/abs/pii/S1041608023000195?via%3Dihub.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Krotov, V., & Silva, L. (2018). *Legality and ethics of web scraping* [Paper presentation]. 24th Americas Conference on Information Systems, New Orleans, Louisiana. https://www.researchgate.net/profile/Vlad-Krotov/publication/324907302_Legality_and_Ethics_of_Web_Scraping/links/5aea622345851588dd8287dc/Legality-and-Ethics-of-Web-Scraping.pdf.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology, 1*(2), 100017.

Marshall, I., & Wallace, B. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews, 8*(163), 1–10.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. A (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.

Mills De La Rosa, S., Greenstein, N., Schwartz, D., & Lloyd, C. (2021). *Machine Learning in Workforce Development Research: Lessons and Opportunities.* Abt Associates.

National Artificial Intelligence Advisory Committee. (n.d.). *National AI Advisory Committee.* https://ai.gov/naiac/.

NIST (National Institute of Standards and Technology). (2022). AI Risk Management Framework: Second draft. https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

Project Open Data. (n.d.). DCAT-US Schema v1.1 (Project Open Data metadata schema). https://resources.data.gov/resources/dcat-us/.

Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: Global trends, research directions, and regulatory standpoints. Data Science and Management, 4, 19–29. https://www.sciencedirect.com/science/article/pii/S2666764921000485?via%3Dihub.

Sobel, B. (2022). A new common law of web-scraping. *Lewis & Clark Law Review, 25(*1), 147–207. https://law.lclark.edu/live/files/31605-7-sobel-article-251pdf.

Song, C., Ristenpart, T., & Shmatikov, V. (2017). *Machine learning models that remember too much.* CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas, Texas. https://dl.acm.org/doi/proceedings/10.1145/3133956.

The White House. (2023). *Executive Order on the safe, secure, and trustworthy development and use of artificial intelligence*. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology, 136*, 106589.

Yarlagadda, R. (2017). AI automation and it's future in the United States. *International Journal of Creative Research Thoughts, 5*(1), 382–389.

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. *IEEE 31st Computer Security Foundations Symposium, 268–282*. https://ieeexplore.ieee.org/document/8429311.