Implementing Data Collection and Analysis Pipelines in the Social Media Archive at ICPSR

Introduction

The Inter-university Consortium for Political and Social Research's (ICPSR's) Social Media Archive (SOMAR) proposes an Implementation project to build streaming data collection and analysis pipelines. Our project addresses objectives 3.2 and 3.3, listed under "Goal 3: Improve the ability of libraries and archives to provide broad access to and use of information and collections." We request $429,360 from IMLS and will provide $429,360 in cost share. Our project will generate datasets and data services that democratize researchers' access to social media data and enable researchers to build and deploy machine learning models in a virtual data enclave (VDE). We will produce code and documentation that empower other archives to adapt our tools for their technical infrastructure and unique data challenges.

Project Justification

How to protect individuals represented in data while enabling data reuse, especially at scale and with computational analysis tools, are fundamental questions facing data archives. Archives are wrestling with a deluge of data from sensors, administrative records, social media and other digital communication tools, and scientific research. In this project, SOMAR implements pipelines for archiving streaming data and facilitating their analysis with computational methods. While SOMAR focuses on data from social media, the collection and analysis tools we will develop will be useful for other types of streaming, large-scale, sensitive data. SOMAR developed the first cloud-based elastic virtual data enclave (VDE) in a large archive, and this project extends the VDE's capabilities and utility to more users with various levels of computational ability.

Social media data are incredible resources for science. For example, they have been used to study political communication, public health messages, social support, and to predict changes in the labor market. Researchers face myriad challenges when working with social media data, some due to the people and platforms involved in its creation and others due to its scale. We have designed a process for collecting, indexing, and accessing social media data that addresses the challenges and regulations they present. In this section, we review our collection development policy, storage and indexing infrastructure (i.e., the SOMAR Data Lake), access mechanisms (i.e., the virtual data enclave; VDE), and analysis tools.

ICPSR has been facilitating access to sensitive research data for more than 60 years and has stable business practices and policies that SOMAR also follows. ICPSR's most common VDE is a Windows-based virtual desktop infrastructure that cannot handle the scale of social media data efficiently. SOMAR's Linux- and cloud-based VDE is designed to work with large-scale data while following ICPSR's strict data custody and privacy protection procedures. SOMAR will follow existing ICPSR policies around restricted data, including an application process, disclosure risk review, and data use agreements. What is new and necessary for SOMAR that this proposal enables are the data collection and analysis mechanisms crucial for social media data research.

Social media data present challenges for archives because of their size, the ways researchers remix and retrieve them, and how they are regulated. Collecting, storing, and distributing social media data require archives to reconsider fundamental notions such as original order, provenance, and custody. We propose a technical infrastructure and related analysis tools that enable SOMAR to preserve data in multiple formats and collections, to capture transformations made to data (e.g., preprocessing steps common in computational research), and to connect users with data even when we cannot keep data in our custody (e.g., platforms won't allow data to leave their servers; privacy regulations limit data distribution).

Social Media Data for Research

Social media are implicated in many of contemporary society's most pressing issues, from influencing public opinion, to organizing social movements, to identifying economic trends. Increasing researchers' capacity to understand the dynamics of such social, behavioral, and economic phenomena will depend on reliable, curated, discoverable, and accessible social media data. Research has shown strong and consistent evidence that data sharing, both formal and informal, increases research productivity across a wide range of publication metrics and that formal data archiving yields the greatest returns on investment with an increased number of publications resulting when data are archived (Pienta et al., 2010; Piwowar et al., 2007). Thus, SOMAR will encourage more science based on the analysis of social media data.

Social media data are widely used by researchers and are ripe for archiving. For instance, social scientists use social media data to study a range of topics such as economic and consumer behavior (Antenucci et al., 2014; Asur & Huberman, 2010), cultural differences (Hochman & Schwartz, 2012), social capital (Ellison et al., 2014; Gil de Zúñiga et al., 2012), feminist and anti-racist movements (Brock, 2012; Dixon, 2014; Freelon et al., 2016), political activism (Boulianne, 2015; Freelon, 2015; Roback & Hemphill, 2013), and the impact and reach of research (Haustein et al., 2016; Thelwall et al., 2013). SOMAR encourages new efforts on fundamental research questions and will allow disparate communities using social media data to learn from each other.

Analyzing Social Media Data

Social media data are large in scale, especially compared to traditional social science data sources such as surveys. Data sets such as a day of tweets or a year of Reddit posts are too large to analyze manually, both in disk size and in number of observations. Computational and statistical tools, such as topic modeling, component analysis, and unsupervised classification, help researchers make sense of large datasets. However, machine learning models can also present privacy challenges because they can store and leak data used to train them. Running advanced models also requires significant computing resources, such as multiple graphic processing units (GPUs) and high-performance computing (HPC) clusters. Computing resources and skills are not equitably distributed among researchers, and their distribution impacts researchers and research questions that engage with social media data. More equitable and inclusive means to ensure access to data are essential for broad and diverse data use. Our existing users have asked to use machine learning models to analyze large datasets, and therefore we focus on HuggingFace integration as a first step towards making this possible. HuggingFace is a widely used platform for sharing artificial intelligence, machine learning, and natural language processing models and datasets.

Target Group and Beneficiaries

SOMAR users are researchers from academic and other non-profit organizations who need to use social media data in their research. For instance, our current users are studying the US 2020 federal elections, political speeches from the US and India, and the Israel-Hamas and Ukraine-Russia wars. ICPSR has been a leader in digital curation for decades, and the practices and technologies that we adopt are often used by other archives.

Project Work Plan

SOMAR will develop data collection and analysis pipelines to address the pressing needs listed above. SOMAR's Project Management and Engineering staff will collaborate to produce accessible end-user tools for collection and analysis. We divide our work into two phases:

1. *Collect and store social media data*. In year 1, SOMAR will focus on data collection efforts by partnering directly with platforms, advocating publicly for data access, and collecting data for research. The project management and engineering teams will collaborate to set up an active data collection and storage system, which will access and store publicly visible social media data. We begin with platforms such as Stack Overflow and Meta. Our collection tools will adhere to Web archiving standards as they collect data and metadata. Project Management will curate metadata and dataset files, making datasets discoverable and accessible to researchers through SOMAR's website. Engineering will create a "data lake" that enables cross-platform and linked data research.
2. *Improve SOMAR collection search and analysis*. In Year 2, the project management and engineering teams will focus on building systems for enhancements within SOMAR's secure virtual data enclave (VDE). Enhancements include data linkage, word embedding, auto indexing, and novel measure generation, particularly the ability for researchers to apply existing machine learning models to data in the archive. Engineers will also configure OpenSearch and a related end-user graphical user interface within the secure data enclave. The OpenSearch system will provide users with a no-code mechanism for searching across datasets to create bespoke samples that include observations from multiple datasets and platforms. These

tools enable researchers to analyze data at scale, study phenomena across social media platforms, and perform longitudinal and comparative analyses that are not currently possible.

SOMAR will provide access to the resources through its existing cloud-based VDE. Instead of creating local downloads or copies of data, researchers log into the enclave remotely and perform their analysis on SOMAR's computing resources. SOMAR's enclave approach helps prevent data leakage and ensures each project can access the right computing resources (e.g., GPUs, right-sized storage). SOMAR creates individualized VDE instances that contain only the minimum data needed for the approved research and suitable computing and storage resources for analysis. Furthermore, SOMAR protects the privacy of the individuals represented in the data by conducting manual risk review of all results before researchers can export them from the enclave. After the Implementation project ends, SOMAR will continue data collection, data enhancement development, archive maintenance, and data user support. As a project of ICPSR, SOMAR's ongoing maintenance is supported by a consortium of more than 825 research institutions with more than 60 years of experience in data archiving and services. In 2019, IMLS recognized ICPSR with the National Medal for Museum and Library Service.

SOMAR's Virtual Data Enclave

SOMAR has three main reasons for using a virtual data enclave (VDE) to enable researchers to analyze data in our collections: protecting privacy, following data sharing rules and regulations, and enabling research replication. Access to data in the VDE depends on the approval of an application for specific data, signed restricted data use agreements (RDUAs), and robust data security plans.

Protecting Privacy

Social media data are generated by individuals and groups and reach audiences through platforms such as Facebook, X (formerly known as Twitter), and LinkedIn. Much of the data individuals generate using social media is technically or legally "public" (meaning anyone online can see it). However, individual users and regulatory areas recognize that making something public does not imply that using the data for research is acceptable, especially when the data is sensitive or connected to an identifiable individual. For example, the European Union's General Data Protection Regulation specifically addresses the tension between research and privacy by describing acceptable research exemptions and practices (e.g., including safeguards, balancing risks and rewards). The United States does not have a similar privacy law at the federal level, but public conversations and research on social media wrestle with this tension between individual privacy and the public benefits of data analysis (Fiesler & Proferes, 2018; Hemphill et al., 2022; Klassen & Fiesler, 2022; Vitak et al., 2016).

The SOMAR VDE protects the individuals represented in the data through both technical and social mechanisms. Restricted data are accessed through the VDE in cases when removing potentially identifying information would significantly impair the analytic potential of the data, or in cases where data have highly sensitive personal information and cannot be shared as a publicly available file. The VDE is isolated from the user's physical computer, restricting the user from downloading files or parts of files to their physical computer or even to any other Internet-connected destination. Administrators are the only users with technical permissions to bring data or code into or out of the VDE. Data leave the VDE only after rigorous disclosure risk reviews (DRR).

ICPSR has stable business practices for reviewing restricted data analysis output for disclosure risk, and SOMAR will follow the same exacting standards that ICPSR follows (and has for 60 years). When researchers complete their VDE onboarding, they are provided with guiding documentation, including instructions on preparing and submitting requests to remove the output from the VDE and the criteria for DRR approval. SOMAR uses a user support ticketing system to manage and track requests for restricted data output review. When a request is sent, a designated SOMAR team member reviews the request and materials in the VDE. If modifications to the output are necessary, SOMAR staff work with the researcher to address these changes and provide recommendations. Any confidential information is shared only within shared documents inside the researcher's secure environment. Once an output request is approved, SOMAR staff release the approved materials to researchers through a digital airlock. Finally, each correspondence between the VDE user and

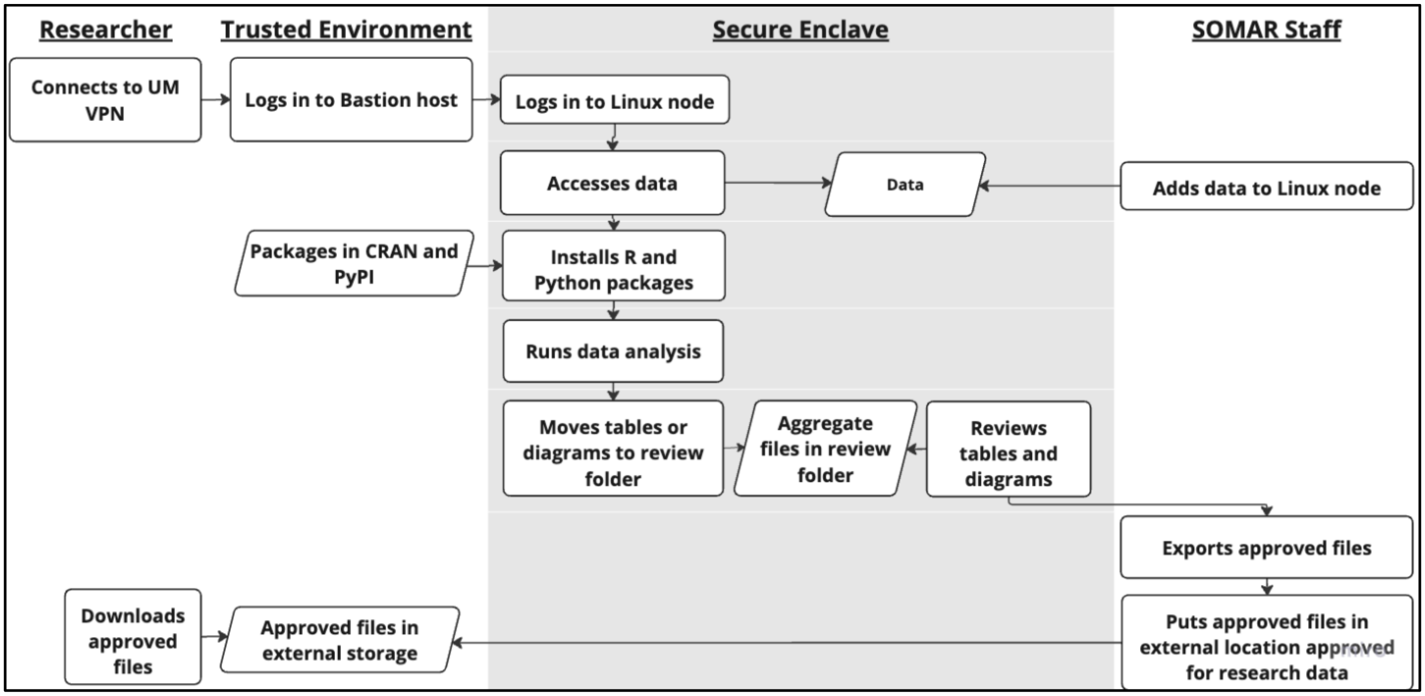

Researcher
Trusted Environment
Secure Enclave
SOMAR Staff
Connects to UM VPN
Logs in to Bastion host
Logs in to Linux node
Accesses data
Data
Adds data to Linux node
Packages in CRAN and PyPI
Installs R and Python packages
Runs data analysis
Moves tables or diagrams to review folder
Aggregate files in review folder
Reviews tables and diagrams
Exports approved files
Downloads approved files
Approved files in external storage
Puts approved files in external location approved for research data

Figure 1. Overview of the SOMAR VDE

SOMAR includes a reminder that their RDUA prohibits removing or transcribing results without approval from designated SOMAR staff. A diagram of the workflow for disclosure review is provided in supplementary materials.

Socially and legally, the ICPSR RDUAs govern user behavior and provide severe and enforceable consequences for users who violate the agreement. For example, the RDUA includes the following language about investigators' responsibilities: "no attempt will be made to identify Private Person(s), no Restricted Data of Private Person(s) will be published or otherwise distributed, the Restricted Data will be protected against Deductive Disclosure risk." RDUAs are institution-level agreements that can be signed only by legal representatives of investigators' institutions with signatory authority. We have provided a copy of the RDUA in the supplementary documents.

Complying With Data Sharing Rules and Regulations

Social media platforms have various rules, often called "terms of service" (TOS), that govern access to their data. Many TOS prohibit data sharing, requiring that individuals collect their own datasets. Requiring users to generate their own datasets restricts participation in science that depends on social media data; only researchers with the necessary computational skills and resources can collect and manage data. The sharing prohibition also limits replicability and validity checks. Common data resources that ensure broad, perpetual, and consistent data access are necessary for science and related activities that depend on social media.

The VDE's data egress limitations allow SOMAR to support data analysis without sharing the data directly. Data never leave SOMAR's custody; only the results of data analysis can leave the VDE and only after manual review and egress by an ICPSR administrator. This structure allows SOMAR to comply with some platforms' TOS while facilitating science and limiting individual disclosure risk. When we enter agreements with platforms, as we have with Meta, the VDE is a key part of the agreements. Platforms are willing to work with SOMAR mainly because of our VDE and the data protection mechanisms we have in place.

The European Parliament's Digital Services Act (DSA) (Digital Services Act, 2021) requires that very large online platforms (VLOPs) provide access to vetted researchers (see Article 40). The mechanisms for providing this access are still under development, and SOMAR is working directly with Meta and other platforms to provide this access. The European Digital Media Observatory (EDMO) has a working group responsible for piloting processes and systems for Article 40 compliance. Dr. Rebekah Tromble, the director of that working group, is also a member of the SOMAR advisory board and provides direct feedback to the SOMAR team about what researchers need, how platforms can comply, and what EU officials want to see. The SOMAR VDE is designed to allow platforms to deposit data in compliance with Article 40 and for researchers to analyze those data securely.

Facilitating Research Replication

SOMAR's common data and analysis resources make it easier to replicate research. We use common scripts and CloudFormation templates to create VDEs and can easily generate an enclave that matches the specifications of the enclave used for any given research project. The VDE is primarily a Jupyter environment, which encourages researchers to create re-runnable notebooks. The VDE can be configured to handle the dependencies and requirements specified in these notebooks, ensuring that the environment used to generate the analysis can be faithfully reconstructed.

When users access the Meta Content Library API, we archive the queries they send so that the queries can be rerun. This ensures that researchers can find differences in data collected from the API at different times and compare queries and result sets across time and projects.

Restricted Data Application Process

To grant access to restricted data in the VDE, SOMAR requires researchers to follow control conditions. Specifically, researchers must complete their application processes through the SOMAR application platform and agree to follow strict legal and electronic requirements to maintain data confidentiality. These applications include but are not limited to the following requirements:

- Investigator information: Name, email address, institutional affiliation, credentials, and research and coding experience. In most cases, the researcher investigator must have a terminal degree and be affiliated with an academic, research, nonprofit institution.
- Collaborator information: Name, email address, and institutional affiliation. The collaborator must be affiliated with the investigator institution or obtain a separate data use agreement.
- Research description and justification: Brief explanation of the research project for which the data will be used and clear information about why these specific data are requested.
- Institutional signatory information: Name, title, and email address.
- Documentation including signed and dated RDUAs, researchers' resumes or CVs, and supporting publications of research experience.

After researchers submit their applications, SOMAR staff reviews and verifies the information provided. The amount of time it takes to review an application depends on whether and how many changes are needed before approval. SOMAR staff work with the requester until all required information is provided and the request can be approved. It takes about 2-4 weeks for an application to be approved after submission. Approved researchers receive onboarding instructions and materials upon notification of application approval.

Social media data collection

Collecting Data

SOMAR's primary method of data collection is through deposit agreements with platforms and researchers. The TOS that nearly all platforms have in place severely restrict researchers' abilities to access data directly and to share it with others. Therefore, SOMAR works directly with platforms to secure access for research. The Meta Content Library API proxy (Meta Platforms, Inc., 2023) and the datasets from the US 2020 Study datasets (*U.S. 2020 Facebook and Instagram Election Study*, 2023) are examples of SOMAR's work in this space. We are in conversation with TikTok and other platforms to enter similar agreements that would make platform data widely available through SOMAR. PI Hemphill and the project manager will work to secure these agreements; we have budgeted for their time on these engagements, including managing legal contracts.

The DSA is another mechanism for adding to SOMAR's collection. When complying with the DSA, platforms must generate datasets. SOMAR is working with EDMO and platforms to become the archive where these datasets are stored and accessed. This ensures that more researchers can access common datasets and that the platforms do not handle access. Instead, Digital Services Coordinators will handle researcher and project vetting, and SOMAR could handle data management and access. SOMAR is currently working with Meta and EDMO on DSA-compliant access mechanisms.

Lastly, SOMAR will collect data from platforms that allow it. For instance, Stack Exchange makes “dumps” of its data available on its website and through the Internet Archive. These large dumps require extensive computing resources to parse and analyze. SOMAR will work to create subsamples of data that are more manageable for individual researchers and will host the complete dataset in the VDE, where it can be analyzed using the new tools described below under Social Media Data Analysis. The SOMAR engineering team will handle ingesting Stack Exchange data.

Storage and Indexing System - The SOMAR Data Lake

Social media data platforms expose data in many different formats, and researchers are seeking collections that cross platforms, media, and time horizons. Cloud-based data solutions that are common in the private sector, such as extract, transform, and load (ETL) services, data lakes and warehousing, on-demand observation-level indexing, are not yet common in digital archives.

SOMAR proposes to facilitate cross-platform and linked data research by structuring our storage as a “data lake” and developing technical infrastructure that is more flexible and secure than existing solutions for archives. A data lake is a central storage system for housing data, structured or not, from multiple sources in a single place. It differs from a database in that the data in the lake do not need to be formatted in a special way or stored in tabular format. It can be stored in its “raw” form. The SOMAR data lake will hold CSV documents and JSON documents, for instance, and they will each have their own columns and fields (i.e., data from Facebook has “reactions”, comments from Reddit have “karma”). The data lake approach allows us to store data from multiple platforms in a single resource and to store multiple versions of the data (e.g., raw, processed) efficiently. Many SOMAR users have the computational skills to work with raw social media data, but our aim is to diversify the group of users who can benefit from our data. One way we can do so is to create transformed datasets that are subsets of larger datasets (e.g., only the US politicians in the Politweets data), or that have already been preprocessed using common techniques (e.g., creating word vectors, stemming words in preparation for natural language processing) and are ready for analysis. The diagram below shows SOMAR’s technical components, and we describe each briefly in the following sections. We use various Amazon Web Services (AWS) services to manage components and processes.

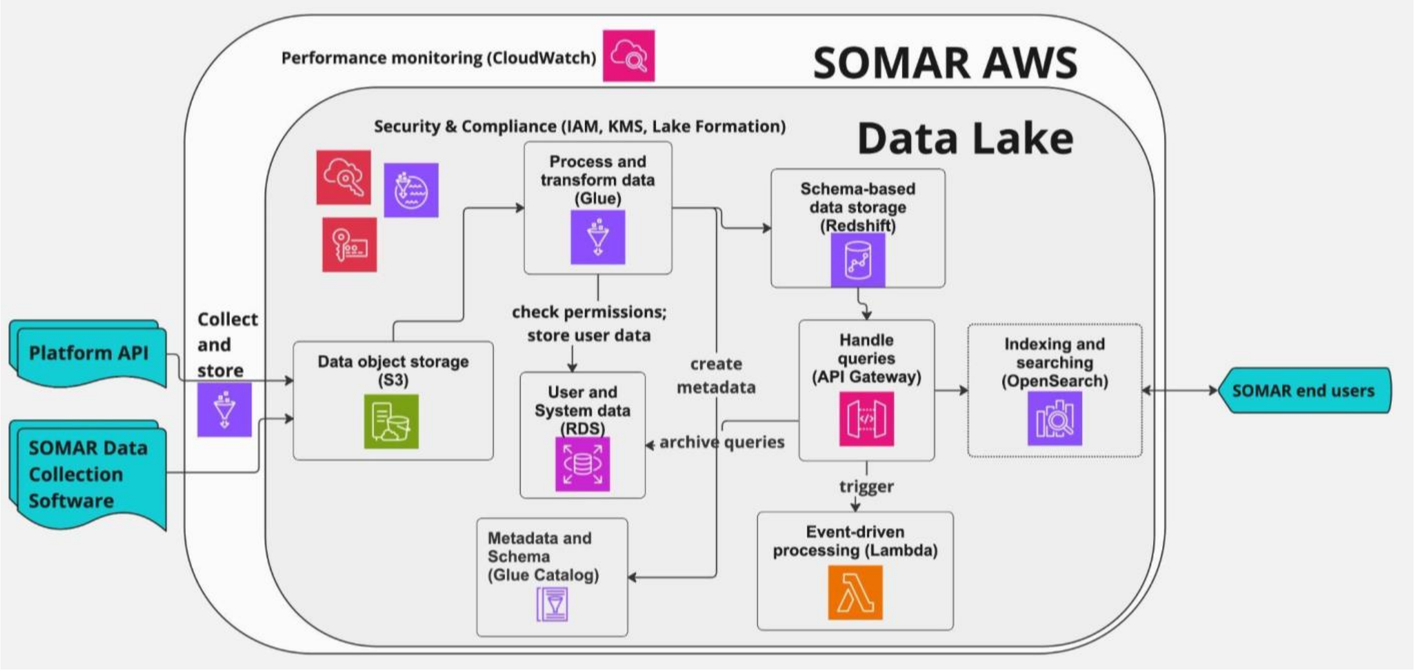

SOMAR AWS
Performance monitoring (CloudWatch)
Data Lake
Security & Compliance (IAM, KMS, Lake Formation)
Platform API
SOMAR Data Collection Software
Collect and store
Data object storage (S3)
Process and transform data (Glue)
check permissions; store user data
User and System data (RDS)
create metadata
archive queries
Metadata and Schema (Glue Catalog)
Schema-based data storage (Redshift)
Handle queries (API Gateway)
trigger
Event-driven processing (Lambda)
Indexing and searching (OpenSearch)
SOMAR end users

Figure 2. Overview of SOMAR’s cloud-based architecture. Each icon shows a specific AWS service, and the text describes the role of that service. Labeled arrows show what travels between services or how they interact with one another. For instance, the API Gateway sends data about user queries to the RDS.

Data Storage

We will store our social media data primarily in Amazon S3. S3 stores data of arbitrary formats as objects and scales to meet our size demands. It also allows for “cold” storage for files that are accessed infrequently, helping to control storage costs. We will also be able to implement automated backup and recovery procedures built directly into S3.

We store data on system use, including permissions, in Amazon Relational Databases (RDS). For example, data about which users are approved for access to which datasets are stored in RDS.

Data Processing and Transformation

AWS Glue for ETL Jobs: AWS Glue will be the primary service for cleaning, transforming, and enriching data within the data lake. AWS Glue helps us transform raw data into usable data. It is a fully managed ETL service that automates time-consuming data preparation tasks for analytics (e.g., reformatting nested JSON into flat CSV files for analysis in Stata, and creating sentence vectors from text data to prepare for machine learning analysis). Glue is highly scalable and serverless and integrates seamlessly with Amazon S3, Amazon RDS, and Amazon Redshift.

Python for Data Processing: We will use Python in conjunction with AWS Glue for writing custom ETL scripts. Python's rich ecosystem of data processing libraries (e.g., Pandas (The pandas Development Team, 2020)) makes it an excellent choice for complex data transformation tasks. Using Python directly within Glue jobs adds flexibility and power to our data processing capabilities. Python’s broad use within the computational social science community also means that many of SOMAR’s users understand Python code and will be able to inspect our scripts to understand what, if any, transformations we have made to date. Using S3 for storage means that we can always make raw data available if users prefer, but we can also save some users time by performing common transformation and cleaning tasks in Python.

AWS Lambda for Event-Driven Processing: AWS Lambda will be used to execute event-driven data processing tasks (e.g., parsing data on ingest and archiving API queries). This serverless computing service allows us to run code in response to triggers such as changes in data in an S3 bucket or updates in a database. For instance, we can trigger a new metadata drafting task in Jira whenever new data appear in an S3 bucket so that we know when to start curatorial work on data that our collectors retrieve. Lambda functions can be written in Python and integrated with AWS Glue and other AWS services, providing a highly scalable and efficient way to process and transform data on the fly. We currently use Lambda functions to manage the API Proxy of the Meta Content Library and have experience ensuring their efficiency and security.

Data Cataloging

Central Place for All Data Info: We are using AWS Glue Catalog to store, annotate, and share metadata for SOMAR’s data. Every social media platform structures its data differently, and creating static metadata for streaming collections is challenging. By using AWS Glue Catalog, we make it possible to document individual platforms’ data and to describe streaming collections. We can also create metadata at different stages of dataset’s lifecycle, from its “raw” state at the platform to its various states of transformation within SOMAR (e.g., after it is indexed, after users have combined multiple platforms’ data). This approach helps ensure that we can capture the provenance of data in our systems and transparently describe any processing steps.

API Gateway Proxy for Data Access

The API Gateway Proxy acts as the go-to point for accessing the data. Using an API Gateway Proxy to manage data access allows us to implement strict and specific security measures and to return minimal data on request. Users will not need to merge or subset data returned; they will receive only the data they requested. This ensures that they have the smallest datasets necessary for their work, minimizing the networking and computing overhead of doing their analyses. SOMAR staff will prepare extensive documentation for using the API and provide one-on-one hands-on support for users new to using APIs.

Security and Compliance

AWS offers two primary mechanisms to protect data and control access: AWS Identity and Access Management (IAM) and AWS Key Management Service (KMS).

Social media data analysis

Redshift

Amazon Redshift serves both data warehousing and analysis functions for SOMAR. By "warehousing" we mean commingling data from multiple sources. Data in S3 will be stored in virtual folders and buckets, but the data are not connected to any other data in S3. Redshift carries out the connections. We use Glue Catalogs to create data schemas (e.g., a Facebook schema, a Stack Exchange schema) and to set specific permissions on data in S3. Redshift then allows users to query the schemas that their credentials allow (by sending queries through the API Gateway). We can also create cross-platform and aggregated schemas to enable users to search across datasets housed in S3. We expect that we will have schemas for data that we have transformed (e.g., stemmed for natural language processing), and we can expose them through Redshift.

Users will be able to query Redshift through the API Gateway, but SOMAR will also provide datasets in CSV and JSON formats that users can access directly through Python, R, and Stata. Which datasets we create will depend on the queries users are making. We are capturing and reviewing user data queries and requests and will use those data to prioritize datasets for creation. For instance, users of the politweets data have requested a subsample of only current US members of congress. While users can get that sample from the Politweets data, doing so requires them to first access, then sample, a large (many GBs) dataset. We could easily produce a smaller, US-only dataset once it was stored in S3, indexed through the Glue Catalog, and its schema was exposed to Redshift. We anticipate other datasets like "all user-generated comments about the Superbowl" or "any post mentioning legislation about TikTok" that could be useful for multiple researchers but are not yet available.

OpenSearch

We will explore Amazon OpenSearch Service (formerly Amazon Elasticsearch Service) as a mechanism for indexing and searching text data within the data lake. OpenSearch offers several advantages that may make it useful for SOMAR users: a) it uses JSON, which is the format in which most platform APIs provide data; b) indices can be created and recreated when data description needs change; and c) it scales efficiently. OpenSearch represents data as indices. Indices, in turn, are sets of documents that have one or more fields. Fields are like variables and are represented as nested key, value pairs in JSON. For example, the JSON snippet in Table 1 shows a "public_metrics" object that has variables "retweet_count", "reply_count", "like_count", and "quote_count".

```
"public_metrics" : {
        "retweet_count": 8,
        "reply_count": 2,
        "like_count": 39,
        "quote_count": 1
 }
```

Table 1. Example JSON snippet from a Twitter Response Object

Our development environment included data from Gab, Reddit, and Twitter; those data are stored in indices called somar_gab, somar_reddit, and somar_twitter. Each index has a different collection of fields because the metadata and data provided by each platform differ. For instance, Reddit has no "quote" or "retweet" option, but Twitter does. Using OpenSearch as our storage system allows us to add and update indices as we expand to support data from new platforms.

We experimented with relational databases and Elasticsearch systems to find whether either approach had computational, environmental, or user experience advantages. We found that Elasticsearch implementations had higher upfront costs but lower ongoing costs; Elasticsearch also provided more flexibility in text searches, the most common search SOMAR users conduct (Hemphill et al., 2023).

HuggingFace Integration

Our goal in integrating with HuggingFace is to allow secure access and application of artificial intelligence, machine learning, and natural language processing models within the VDE. For instance, users analyzing Politweets (Panda et al., 2023) have requested Sentence-BERT sentence transformers (Reimers & Gurevych,

2019) to construct embeddings for politicians' tweets to do comparative analyses. The Sentence-BERT authors have released their model through HuggingFace. Using their model saves researchers time and expense of training their own transformers. HuggingFace also integrates with Transformers-interpret (Pierse, 2021) to help users understand how models make predictions and to better explain their findings.

Using artificial intelligence (AI), machine learning (ML) and natural language processing (NLP) techniques to analyze data is becoming a normal practice in science that depends on social media data. Integrating with HuggingFace is the first step toward supporting this type of analysis. It is like providing SPSS or Stata to social scientists working with smaller datasets using statistical techniques these software support (e.g., multivariate regression). Without HuggingFace integration, users must write, train, test, and compute their own models, which takes tremendous computational skill and power and costs researchers time and money.

Our HuggingFace implementation carries the following requirements:

- Secure model hosting on Hugging Face with proper access controls.
- Configure VDE tools to integrate with Hugging Face's transformers library.
- Establish a secure API key management system.
- Create usage guidelines and example scripts to apply classifiers to VDE datasets.
- Train and support VDE users on using Hugging Face models.
- Monitor and audit tools to oversee and improve the use of classifiers within the VDE.

The SOMAR Team

Principal investigator (PI) Libby Hemphill will commit 20% FTE of her time in year 1 and 18.5% FTE in year 2 to provide project leadership and direction to SOMAR staff. She will also lead partner engagement and SOMAR's efforts to secure data use agreements with social media platforms. At 49.03-50% FTE each year, Project Management staff will plan, monitor, and communicate project deliverables. They will also prepare reports and coordinate with Engineering staff to confirm plans of action and milestones, including quality assurance tests to verify the completion of each product deliverable. Engineering staff will commit 60% FTE in year 1 and 85% FTE in year 2 to build, implement, test, and maintain the technical developments. Project Management and Engineering staff will also create documentation for researchers to help them navigate the secure data enclave.

Progress Tracking and Evaluation Plan

The SOMAR team uses Jira Project Management to support project planning, task assignments, and progress monitoring, fostering real-time collaboration. We use Jira's built-in reporting to facilitate our performance measurement. Jira will allow us to efficiently organize tasks, prioritize them, and adhere to project timelines. Using Jira's hierarchical structure (epics, stories, and tasks), we break the work down into manageable units, aligning with our goals for developing streaming data collection and analysis pipelines. Data we keep in Jira includes prioritization, due dates, and time tracking, and all team members are responsible for updating Jira with their progress. Regular team reviews, at minimum monthly, help us assess our progress against project aims and confirm the completion of key milestones. Additionally, these regular reviews support our team's continuous learning and improvement efforts.

User feedback will be solicited through short surveys distributed to SOMAR users at regular intervals; user experience evaluations of our user-facing data search and analysis tools; from expert collaborators; and when working with researchers in user support correspondences, professional meetings, and presentations. The short surveys will be administered at the beginning, midpoint, and end of the project to researchers who have used SOMAR data and analyzed them in the Virtual Data Enclave. Based on the specifications and feedback from the SOMAR community, enhancements will be prioritized and incorporated as time allows. Just as the tools continuously learn from the data with which they are working, the SOMAR team will be constantly learning from the users of their tools and resources.

Diversity Plan

Dismantling barriers to data access and analysis are the primary goals of this proposal. Shared data resources play a pivotal role in expanding participation in research. SOMAR levels the playing field to ensure that individuals and institutions can access and analyze social media data regardless of computational proficiency and local resources. The PI is a queer cis woman, and her team includes individuals with disabilities and members of minoritized ethnic groups. The primary goals of the proposal and composition of the team underscore SOMAR's commitment to diversity and underrepresented voices.

Project Results

This Implementation project addresses researchers' demands for large-scale, automated analysis of born-digital data without compromising the privacy or confidentiality of the individuals represented in the data. It will produce valuable data resources, including social media datasets, improved search and indexing, and AI, ML, and NLP analysis capabilities. All data resources will be provided through SOMAR's existing VDE.

We will also create software code and training materials and conduct outreach activities to ensure the utility and awareness of our data resources.

Adaptable, Usable Deliverables

We will archive our software code through Zenodo. Zenodo is an open access repository based out of CERN that focuses on software code and toolkits. In accordance with the FAIR principles, Zenodo follows the standard of issuing a DOI (digital object identifier) for all deposited objects, as well as any related materials. This unique persistent identifier makes the code permanently Findable, Identifiable, and Retrievable on the Web. We will use Zenodo for both scripts stored in GitHub and models distributed on HuggingFace.

The SOMAR team will be working beginning in year one to build awareness of upcoming developments. As the system builds make their way through SOMAR's project management process and are ready to release, SOMAR staff will create webinars, YouTube videos, and open content to inform the social media data community about the developments. The team will produce one online seminar in year one to discuss plans and upcoming implementations and to encourage input or questions from the audience. In year two, SOMAR will produce two webinars: 1) one to announce, describe, and discuss the benefits of the finished products to the community; and 2) another online seminar to host a question-and-answer session or to receive feedback from the community. SOMAR will use the feedback gained through these interactions to continuously improve their systems, and they will also work with key stakeholders of the social media data community to ensure that the final products integrate seamlessly with existing resources and are responsive to existing, emerging, or changing needs.

Sustainability

We are developing a comprehensive industry action plan to build a consortium of social media companies to help sustain this effort, both in terms of data access and financial support. Meta has funded the development of key SOMAR infrastructure, and we are now working with Meta to help build a consortium of industry partners. SOMAR and ICPSR work closely with the Corporate and Foundation Relations team at the Office of University Development (at the University of Michigan) and the Development team at the Institute for Social Research to approach foundations and individual donors whose mission aligns with this work.

SOMAR's VDE is also a pilot for the new ICPSR Cloud-Based Research Environment (COBRE). COBRE is funded by the National Science Foundation under a Mid-scale Research Infrastructure award. Its development will begin in 2025 or 2026 and includes multiple years of financial support.

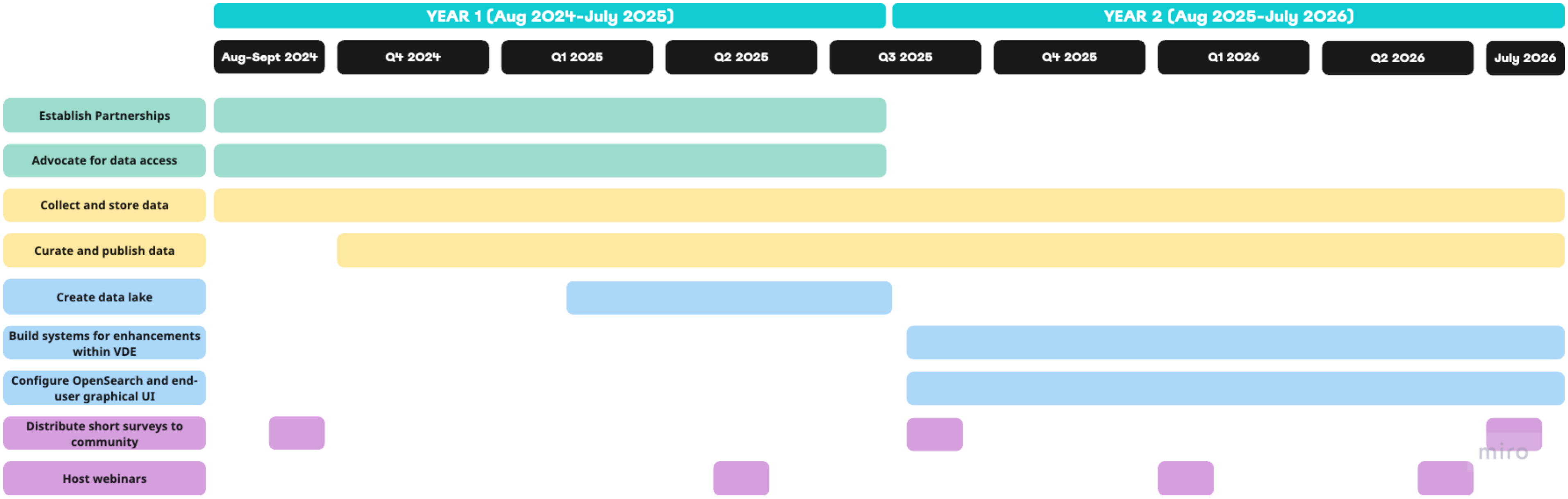
In addition to our SOMAR-specific fundraising activities, we rely on the ICPSR membership for project support. ICPSR was established in 1962 to serve social scientists worldwide by providing a central repository and dissemination service for computer-readable social science data, training in basic and advanced techniques of quantitative social analysis, and resources that support the use of advanced computer technology by social scientists. ICPSR maintains the world's oldest and largest archive of research and instructional data for the social and behavioral sciences. ICPSR is supported by more than 800 member institutions – universities, colleges, government agencies, and other organizations. In 2019, ICPSR was awarded a National Medal of Service by the Institute of Library and Museum Services. ICPSR is committed to SOMAR's continued success.

Schedule of Completion

We have divided the work across two phases.

In phase 1 (August 1, 2024-July 31, 2025), SOMAR will focus on data collection efforts through partnering directly with platforms, advocating publicly for data access, and collecting data for research. Also, Project Management and Engineering teams will collaborate to establish an active data collection and storage system, which will access and store publicly visible social media data from platforms such as Stack Overflow and Meta. By halfway through phase 1, Engineering will start to create a “data lake” that enables cross-platform and linked data research. Throughout both phases, Project Management will curate metadata and dataset files, making datasets discoverable and accessible to researchers through SOMAR's website.

In phase 2 (August 1, 2025-July 31, 2026), Project Management and Engineering teams will focus on building systems for enhancements (e.g.,data linkage, word embedding, auto indexing, and novel measure generation) within SOMAR’s secure virtual data enclave (VDE). Engineering staff will also configure OpenSearch and a related end-user graphical user interface within the secure data enclave. Further, at the beginning, midpoint, and end of the proposed timeline, SOMAR will administer short surveys to the research community to evaluate user experience with SOMAR resources, such as the user-facing data search and analysis tools. Finally, the team plans to host three webinars between May 2025 and July 2026.



Digital Products Plan

Type	Availability	Access and Rights	Sustainability
Data processing software *Programming language:* Python *Documentation Standard:* PEP 8 (or later) *File Formats:* PY, IPYNB, JSON, YML	*Public Websites:* GitHub Zenodo *Delivery strategy:* Code will be made available after it's gone through rigorous review by SOMAR, ICPSR, UMSI, and/or ARC staff Publicly available through standard web browsers; manifests will include requirements and dependencies	*License:* MIT Permissible licenses, widely used among software developers Our goal is to provide usable software to other archives and research data users; we hope others will build on our software and release under similarly permissive licenses. We will test our models extensively to ensure that they do not expose (or leak) any underlying data. Our other software will not implicate privacy concerns or cultural sensitivities.	*Preservation*: Zenodo ensures the files will be available perpetualy. Our requirements files will indicate the software requirements necessary to run the software and will indicate the configuration of the servers on which the code was originally developed. *Maintenance*: We will actively maintain SOMAR software as long as we have engineering resources to do so. We will rely on ICPSR membership and ICPSR's Computing and Network Services to support engineering after the grant program.
Data analysis software *Programming language:* Python *Documentation Standard:* PEP 8 (or later) *File Formats:* PY, IPYNB, JSON, YML			
Machine learning models *Programming language:* Python *Documentation Standard:* PEP 8 (or later) *File Formats:* PY, PKL, HDF5, PT	*Public Websites:* HuggingFace Zenodo *Delivery strategy:* Code will be made available after it's gone through rigorous review by SOMAR, ICPSR, UMSI, and/or ARC staff Publicly available through standard web browsers; manifests will include requirements and dependencies		
Webinars and related materials *Webinars:* Captioned YouTube Videos coupled with copies of applicable slides *Standard:* ADA Compliance per federal standards *File formats:* ● YouTube: MP4/MPEG4 ● Slides: PDF, PPTX	*Public Websites:* SOMAR's website (socialmediaarchive.org) ICPSR's website and YouTube channel *Delivery strategy:* All SOMAR webinars, applicable slides, and related materials will be made available through after they have undergone ADA Compliance review and captioning (as applicable) processes	*License, Terms of Use:* ● *Webinars:* YouTube Terms of Use ● *Related Materials:* YouTube Terms of Use, ICPSR's Terms of Use Our goal is to make all our webinars and related materials publicly available and that data users will use these items for educational or scientific purposes	*Preservation*: Having access to ICPSR's resources ensures that SOMAR materials will be available perpetually whether on ICPSR's or SOMAR's website and remain ADA-compliant with the help of ICPSR's Membership and Communications Team. *Maintenance:* SOMAR and ICPSR staff will actively the

Related Materials include: Guides for data users and associated publications *Standard:* ADA Compliance per federal standards *File formats*: PDF, PPTX, Downloadable web items (e.g., Google Documents)	Publicly available through standard web browsers.		resources and ensure they remain publicly available to the audience. We will continue to rely on ICPSR membership and staff resources for any necessary maintenance support after the grant program.