CriticalMAAS Phase 1 research plan – Macrostrat TA4 team

Daven Quinn, Shanan Peters, Shivaram Venkataraman, and Brian Bockelman

October 13, 2023

Abstract

This document presents Macrostrat's research plan for Phase 1 of the CriticalMAAS project. Macrostrat's main goal is to provide the highest-possible quality geologic basemap for CMA workflows, integrating outputs from TA1 and TA2 with other NGMDB, USGS, and external data sources. This basemap will be usable across scales and project areas, with consistent API-driven data access patterns usable by TA3 workflows. To solve alignment problems that have been documented for past CMA workflows, we will prioritize the linking of geological data into a consistent entity framework, driven by geologic unit matching across data sources. We will also work with other CriticalMAAS performers to build feedback capabilities into their data synthesis workflows, especially for TA1 and TA2.

Contents

1 Introduction

The primary objective of our TA4 activities is to adapt and leverage the Macrostrat and xDD data systems for the CriticalMAAS workflow. This places emphasis on the geological data that are central to TA3 modelling pipeline. Our specific immediate objectives are to *(1)* build a geological map ingestion and harmonization system for TA1 output that can rapidly augment the more than 300 geological maps that are already available in Macrostrat and tailoring the map data access points (APIs and tile servers) to conform with TA3 requirements, and *(2)* establish an xDD corpus and document annotation and distribution pipeline that can faciliate TA2 data extraction tasks. We are attempting to facilitate integration of TA1 and some of TA2 outputs by focusing on geological units that appear in maps, geologic columns, and the literature, the goal being to augument map/column units with additional geological data that can be used in modelling steps. This initial thrust is, therefore, largely focused on integrating data extraction and assimilation pipelines that are directed towards TA3. Human-in-the-loop interfaces for assessing, annotating, and editing the data from TA1 and TA2 will also be developed as the data flow pipelines are established. Here we report on our Phase 1 research plan to accomplish these objectives.

1.1 Success criteria for Macrostrat TA4 system

In addition to fulfilling the goals set out in the Phase 1 Evaluation Plan, we seek to establish a system that can support CriticalMAAS performers in response to the following challenges:

- Can we provide the data produced by TA1 and TA2 to TA3 (and other TA4 performers), on demand and en masse?
- Can we augment and standardize these datasets to provide relevant geological information, particularly lithology and geological unit properties?

These guiding questions are intended to focus our work on the key challenges that face Macrostrat, in conversation with the goals, activities, and expertise of the other TA4 performers.

1.2 Adjustments in response to project integration

We have adapted our approach somewhat in light of initial integrations with other performers and discussions with other TAs.

First, we will plan to integrate with capabilities led by other TA4 performers:

- Jataware will produce some end-to-end solutions particularly for georeferencing and page-level evaluation of map data objects. This may reduce the need for Macrostrat-led user interfaces in these domains (see Sec. 2.4).
- MTRI will work with TA3 to ensure that geologic mapping data provided by Macrostrat in vector-tile format can be filtered and subset according to the needs of model pipelines and operators (Sec. 2.1)

Second, based on descriptions of bottlenecks in CMA process workflows (e.g., *Lawley et al, 2022*), and the expected structure of TA1 and TA2 outputs, we forecast that a major problem will be assembing a geologic dataset that is sufficiently well-characterized and standardized to be useful across scales and study areas to extract fairly specific CMA-relevant information. Given this expected challenge, we will devote extra effort to internally characterizing and harmonizing geologic map units, in order to provide appropriately queryable TA1 and TA2 datasets. This activity has therefore been upgraded to a major thrust of our software development plan (Sec. 2.3).

2 Software development plan

Our software development plan will prioritize aggregating TA1-2 data to support TA3 workflows, providing geologic data in appropriate formats for CMA, and augmenting the attributes of geologic data to form harmonized, multiscale products that support CMA workflows. As part of these efforts, we will build data-providing infrastructure, APIs, and user-facing human in the loop (HITL) interfaces. Overall progress is tracked in the `UW-Macrostrat/criticalmaas` GitHub repository, with individual software components tracked in the repositories discussed below (an index of all repositories discussed is available in Sec. 6).

This plan is organized around several lines:

1. Provide geologic datasets to CriticalMAAS performers
2. Build a system for geologic map integration
3. Characterize and link geologic entities
4. Build HITL interfaces

This is broadly similar to the Tasks 1-3 developed in our initial proposal, but with "linking geologic entities" extracted to a top-level task to align with our new understanding of its critical importance in the context of this project (Sec. 2.3).

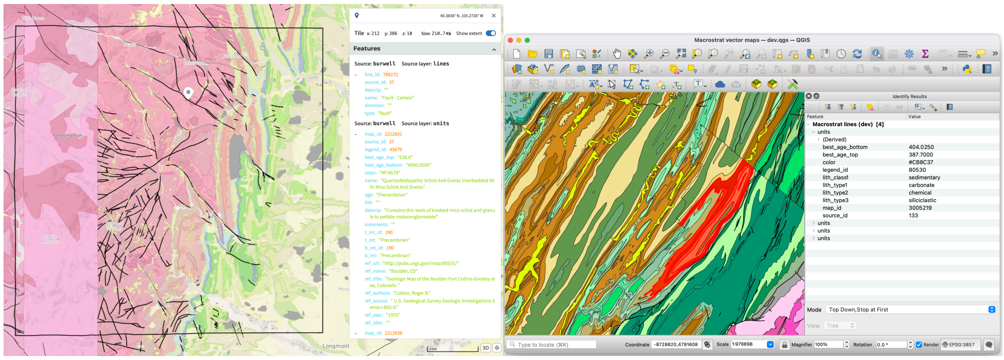
2.1 Providing geologic datasets for CriticalMAAS performers

Macrostrat's key task for supporting CriticalMAAS is to provide harmonized geologic datasets over stable APIs to other CriticalMAAS performers. The most critical task is to provide these datasets to TA3, but we will also provide them to TA1 and TA2 to support feedback. Additionally, we will support the activities and HITL interfaces of other TA4 performers with stable APIs (e.g., for geologic and raster data tiles) atop shared TA4 data repositories.

Geologic map data

Tiled Macrostrat output, API-available in vector-based tile format has gained agreement from MTRI (TA4) and SRI (TA3) that it has the requisite structure and properties to be used as a base for CMA workflows. This dataset is usable on the web, in analytical pipelines, and in QGIS and other GIS software (see Fig. 1). We will continue to refine this output and provide it to TA3. Our key goals for this dataset are to:

- Improve the structure of tileserver output to better support querying by TA3 (ex., by adding ability to filter by lithology).
- Improve API capabilities for querying and filtering by relevant data fields
- Integrate attribute types discussed by Lawley et al. (2022) and others, such as age ranges, paleolatitude, and vetted lithologic classes.



(a) Tiled Macrostrat output in Macrostrat's web interface

(b) Attributed Macrostrat map in QGIS

Figure 1: Different views of Macrostrat's tiled output API, showing its multiscale nature and use in multiple environments.

Much of this work relies on the curation of well-attributed geologic map units (Sec. 2.3). The key codebase for this work, as well as for raster data provision, is the `UW-Macrostrat/tileserver` repository.

Mineral site data

Site-based geologic data must also be forwarded to TA3 CMA pipelines, in a way that allows it to be generalized across geologically relevant areas (e.g., through intersection with mapping data), filtered for the specific CMA task at hand, and validated based on source material (e.g., mine reports and USGS publications).

We have successfully validated serving point datasets relevant to CMA through publicly accessible APIs, including the MRDS dataset (Fig. 2). We will additionally explore approaches to linking mine-site data closely to geologic context, and thereby enabling spatial and geological time/unit filtering. The key codebase for these capabilities is the `DigitalCrust/weaver` GitHub repository. This software will forward mineral site data to TA3 analytical pipelines and TA4 HITL interfaces (ex., MTRI and/or EIS QGIS plugins).

To synthesize and validate mine site data received from TA2, we will need to develop data pipelines and HITL interfaces that forward users to the specific document sources that underly mine-site and mineral-system datasets (see Sec. 2.4).

Macrostrat maintains links to other point datasets that may be useful to forward to TA3, such as USGS legacy geochemical data. TA2 also plans to compile datasets from other existing structured data sources. We will work with TA2 performers to ensure that these datasets are available and properly contextualized for TA3. If more site-based datasets must be integrated, we will work with TA3 to identify and integrate them into the system, which will make them readily avaiable on demand for modelling tasks.

Raster datasets

Raster datasets can generally be integrated directly into TA3 workflows. However, compositing raster datasets across scales, and making the same datasets available across performer teams, presents data-integration challenges that can be supported by TA4. We have validated and will maintain key capabilities to store and serve raster datasets, to support map feedback and CMA workflows. Our systems will be based around storage of raster datasets as "Cloud-Optimized GeoTIFFs" (COGs), which

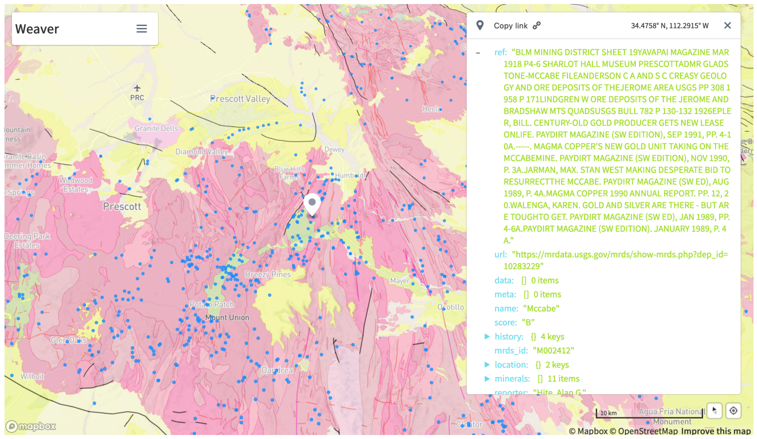


Figure 2: Macrostrat MRDS data layer, showing basic capabilities for point data provision. This interface is publicly available on Macrostrat's development website.

allow efficient use of raster datasets in networked environments. We will also provide indexing and tiling services over these datasets. Raster datasets will be integrated into feedback user interfaces (e.g., Fig. 4) and made available for TA1-3 pipelines and validation. The key codebase for this work is the `UW-Macrostrat/raster-cli` GitHub repository. We will seek to integrate these capabilities with other TA4 systems at the Month 3 hackathon.

The Macrostrat infrastructure platform

The core of the Macrostrat system consists of the databases and infrastructure that hosts the above data capabilities. In order to maintain and extend these APIs and data-provision systems, we will invest in the design and structure of underlying Macrostrat systems.

The key codebases for this work are Macrostrat's infrastructure and command-line interface GitHub repositories. These repositories are, for now, private, due to their role in orchestrating system components on specific infrastructure systems. Much of this configuration will be made public as it is augmented and vetted for security. For the final CriticalMAAS system, an end-to-end implementation encompassing the full set of capabilities will be published.

Literature extractions for TA2 support

Our integration with the xDD document library allows us to provide literature artifacts ready for TA2 extractions, both over USGS documents and the broader geologic literature. We will provide datasets relevant to CriticalMAAS (both USGS-sourced and otherwise) using existing data interfaces (e.g., the xDD API) and tools (the `UW-COSMOS/COSMOS` entity extraction pipeline). This work will be supported by HITL interfaces over documents (Sec. 2.4). Creation of these pipelines is already well underway as one of the first deliverables promised in our project plan (Sec. 4.3). xDD systems can potentially play an important role in surfacing documents and providing context within them for TA2 extractions.

We will continue to integrate with TA2 to build capabilities to access and manipulate the literature corpus of documents from USGS and other sources, potentially developing them further with capabilities needed for CMA. Since USGS documents are broadly in the public domain, TA2 performers will have an opportunity to follow all extractions back to their full source material; this is usually encumbered by publisher agreements in the case of other literature sources, such as Elsevier, Wiley and the like. However, specific information in these source documents can be surfaced and integrated into knowledge bases, provided that the code to locate and extract the information is run within UW's CHTC environment and the output conforms to the expectations of publisher agreements (e.g., extractions constitute a derived data product, such as a list of entities and their relations, and not original unaltered content beyond short snippets of context).

2.2 Geologic map integration pipeline

As a first step towards HITL interfaces to standardize geological map information (e.g., legend data, line types, etc.) from TA1 outputs, we will seek to improve the speed and interactivity of Macrostrat's vector data ingestion pipeline. This system moves from GIS data inputs, like Geodatabases, Shapefiles, and ArcInfo files, to the standardized layers that drive Macrostrat APIs. The system will be adapted to support TA1 outputs, which requires it to work efficiently over heterogeneous and variably attributed data.

This pipeline will operate over:

- NGMDB geologic maps already published in vector form but not yet ingested into Macrostrat
- TA1 outputs (which will be provided in a representation analogous to existing vector datasets, per TA1 output schemas; Sec. 4.1)
- Paired vector/raster map datasets (in conjunction with Macrostrat's raster pipeline; Sec. 2.1, to facilitate TA1 training tasks)

Geologic map ingestion relies on both GIS data manipulation (and in the case of TA1 performers, image analysis), and on geological expertise. Geological decisions include splitting unit ages from stratigraphic names, descriptions, and lithological information in legend text, which must in many cases must be done manually. For CriticalMAAS, it will be use useful to allow geologic expertise to be applied without the need for data manipulation, as that will allow geologists to more readily participate in data curation.

Data ingestion pipelines will be supplemented with web-based interfaces for metadata and map extraction editing (Sec. 2.4). These interfaces will allow geologists to interactively manipulate map data and metadata, to support both the ingestion of accurate maps. This will potentially also allow linking to structured data describing geologic map units synthesized from other data sources (such as map pamphlets and geologic literature). These pipelines are described in Sec. 2.3.

Several GitHub-hosted codebases are relevant to this task:

- `UW-Macrostrat/map-integration` will hold map ingestion/harmonization command-ine interfaces and web app.
- `UW-Macrostrat/python-libraries` holds GIS-oriented Python libraries used across projects.
- `UW-Macrostrat/cli` holds map harmonization scripts. These are currently private but will be made public as they are vetted for sensitive information.

2.3 Geologic entity characterization

Geologic entities are discrete packages of rock (e.g., mapped bedrock map units) that can be characterized by a variety of attributes, including lithostratigraphic names, lithologies, environmental/tectonic context of formation, and geochemistry. Macrostrat maintains a database of geologic entities represented in, for example, regional geologic columns that can be used to further characterize geologic map units. However, Macrostrat's entity characterization is not of adequate completeness to fully support CMA workflows. This will be rectified by a combination of leveraging outputs from TA1-2, augment-

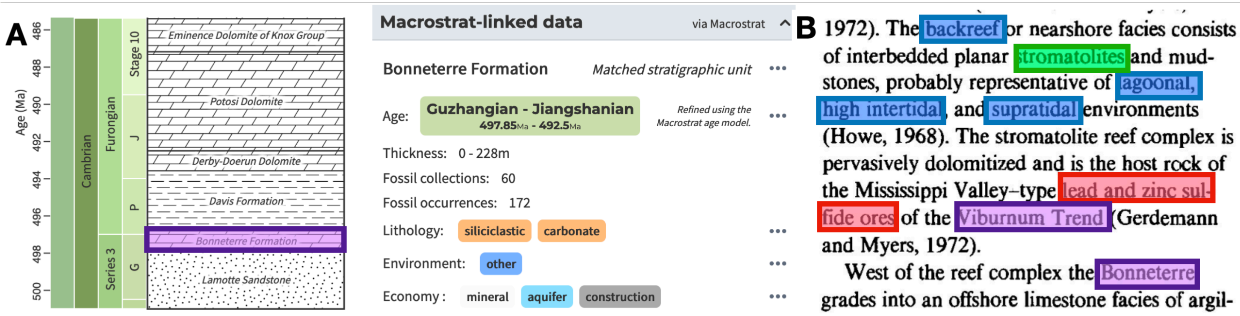


Figure 3: *(a)* Starting user interface and *(b)* potential additional extractions for CMA-focused entity canonicalization tasks

ing those outputs with our own AI-assisted literature synthesis, and building new HITL interfaces for curating information that better charaterizes the properties of rock units that are relevant to CMA.

We are taking two separate approaches to this problem:

1. Finding new descriptors of existing entities by, for example, reading publications.
2. Finding new entities not currently represented in the database and incorporating them.

These two approaches will help correct several deficiencies of Macrostrat's current representation of geologic entities, including more depauperate data on non-sedimentary units and a relatively general characterization of lithologies in heterogeneous units. The overarching goal of these activities is to produce more specific representations of geologic entity properties, particularly lithology, that are amenable to use in TA3 modelling workflows.

This system will be developed and described in the `UW-Macrostrat/macrostrat-xdd` GitHub repository. Feedback tools for evaluating extraction quality will be built atop Macrostrat's map interface (Sec. 2.4).

2.4 HITL interfaces for TA1-2 pipeline support

We will develop several types of HITL feedback interface to support TA1 and TA2 pipelines. We have already produced several prototypes demonstrating our approaches to these UI tasks (Sec. 4.2). In general, we will seek to keep our approach, focused predominantly on map synthesis interfaces, distinct from that of Jataware and MTRI. We will also develop document-based interfaces to support TA2 workflows as necessary and in conjunction with Jataware. Jataware's map-projection system will be required to add projection information to TA1 outputs prior to ingestion into Macrostrat's systems. Likewise, MTRIs QGIS plugin will be a key interface through which TA3 manipulates our vector-tile geologic mapping outputs into binary and probability-based prospectivity rasters.

Map-based feedback

The main HITL interfaces produced by Macrostrat will operate in a map-based environment, and will be designed to support the following tasks:

1. Evaluation of and correction geologic map feature extractions and legend information (Sec. 2.2)
2. Evaluation and correction of TA2 mine-site extractions, and linking to geologic context (Sec. **??**)
3. Improving attributes describing geologic units synthesized from map descriptions and literature (Sec. 2.3)

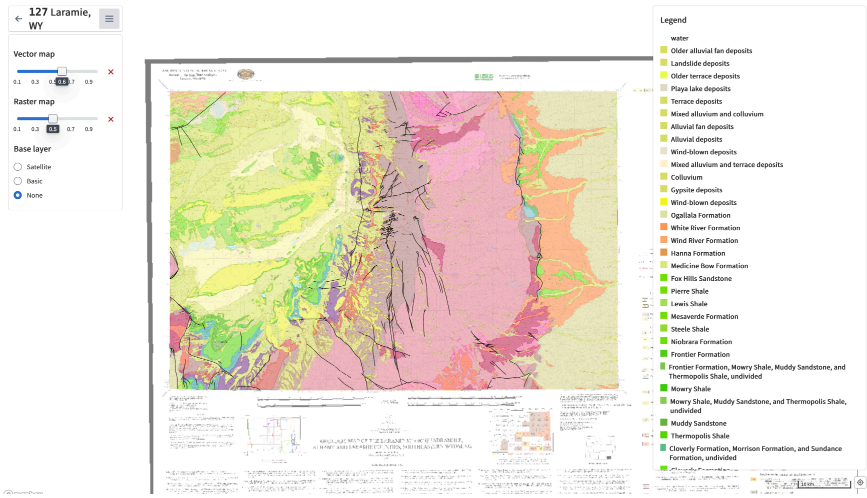


Figure 4: Map interface showing both vector and raster datasets for the same geologic map, in an interface with synthesized legend information. This demo interface was produced as a prototype for future feedback interfaces and is available on Macrostrat's staging website.

These interfaces will be built on top of Macrostrat's existing web interfaces, which are world-class examples of user interfaces for geologic data. Interfaces already demoed for presenting raster, vector, and site data (e.g., Fig. 4 and Fig. 2) will be adapted to support these integrations. Several prototypes have already been demonstrated (Sec. 4.2).

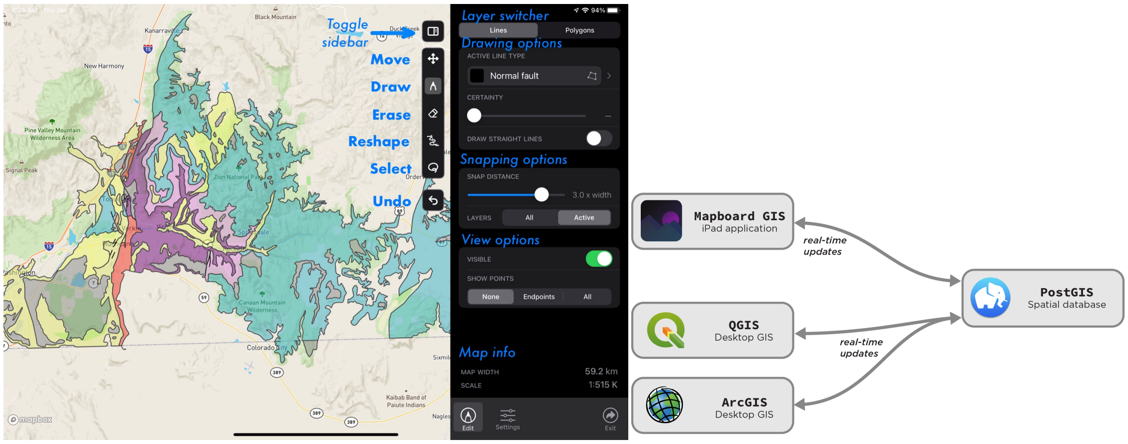
The key codebases for this work are the `UW-Macrostrat/web` and `UW-Macrostrat/web-components` GitHub repositories. The components housed in these repositories will be used to develop HITL interfaces across different subsystems (e.g., for map integration; Sec. 2.2).

Map extraction editing

While initial extractions demonstrated by TA1 pipelines are impressive, it is likely that human intervention will be required to produce GIS datasets of map points, lines, and polygons that are suitable for downstream use. We will produce a system that solves the topology of a TA1 geologic map and creates a representation that can be rapidly edited. This system will be based on the Mapboard GIS system, which combines the purpose-built **Mapboard GIS** iPad application for pen-based mapping with a PostGIS-based topology management system (Fig. 5). This topology management system will allow geologists to rapidly edit geologic maps, both via iPad streaming digitizing and using standard GIS platforms (QGIS and ArcGIS). This system will support both pipeline feedback to TA1 performers and the final production of high-fidelity, topologically correct geologic mapping datasets that can be integrated into Macrostrat and passed to TA3. The topology engine is an open-source component housed in the `Mapboard/topology-manager` repository. This engine will be supplemented with data migration scripts and management APIs that will be part of the Macrostrat system deliverable (Sec. 2.1) for CriticalMAAS Phase 1.

Document-based feedback interfaces

The potential also exists for Macrostrat to contribute to the development of document-based feedback interfaces. xDD has produced visualization interfaces for page-based document annotations (Fig. 6) in support of the COSMOS pipeline (Sec. 2.1). These components, housed in the



(a) *Mapboard GIS* map interface, an iPad app optimized for drawing geological maps

(b) Tethered mode for the Mapboard topology manager, which allows topological editing of geologic maps in both standard and purpose-built GIS environments

Figure 5: Mapboard GIS interface and GIS system design

`UW-COSMOS/cosmos-visualizer` GitHub repository, can be adapted to support TA2-supporting user interfaces. However, given our primary focus on map-based feedback interfaces, we will likely seek to defer to Jataware's work in this domain.

3 Targets for hackathon events

Month 3 hackathon

- Containerized Macrostrat system that supports basic capabilities, running on CHTC infrastructure.
- A prototype import pipeline for geologic maps that can assimilate GIS datasets, including TA1 outputs (Sec. 2.2)
- A basic feedback interface for map legend correction atop this integration system (Sec. 2.4), with raster and vector data layers (Fig. 4)
- A demonstration of topological map editing using Mapboard GIS (Sec. 2.4) and QGIS.
- A `criticalmaas` document set in xDD with links to the original source documents, as well as COSMOS and snippets APIs (Sec. 4.3)

Month 6 hackathon

- Automated pipeline for ingesting TA1-2 datasets into Macrostrat systems (Sec. 2.2).
- HITL interface prototype for geologic map legend synthesis atop this pipeline (Sec. 2.2 and Sec. 2.4).
- Prototype APIs to provide model feedback to TA1-2 performers
- Demonstrated pipeline for geologic map unit characterization from xDD documents (Sec. 2.1) to augment Macrostrat geologic units (Fig. 3), with a feedback tools based on Macrostrat's map interface (Sec. 2.4)
- Pipeline for Macrostrat vector data through MTRI interfaces to TA3 performers, for evaluation by USGS users (Sec. 2.1)

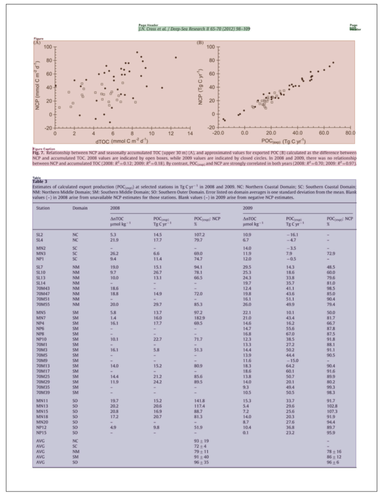


Figure 6: COSMOS image tagger user interface, which is an option for adaptation into HITL systems

- Data models and pipelines for ingesting TA1-2 datasets, and user interfaces for providing feedback on these datasets.

Month 9 hackathon

- Stable system to stage TA1-2 datasets into Macrostrat systems and track their status
- HITL tools for map legend curation (Sec. 2.2) and geologic unit characterization (Sec. 2.3) accessible to USGS staff behind web login
- Usable system for USGS staff to edit geologic map extractions from Mapboard and GIS software (Sec. 2.4)
- Prototype APIs to provide TA2 mineral site data to TA3
- Macrostrat output tile datasets, attributed and projected for TA3 use (Sec. 2.1)

Phase 1 Base evaluation

- APIs to provide map extraction corrections and feedback to TA1 pipelines
- End-to-end system for synthesizing geologic unit data from xDD literature extractions and TA1-2 output, with feedback interfaces for regional geological experts (Sec. 2.3)
- APIs to provide mineral site data to TA3 that are linked to geologic context, filterable, and can be traced to source documents (Sec. **??**)
- Demonstration of providing geologic datasets to TA3 with the level of attribution necessary for accurate CMA workflows (Sec. 2.1)
- Demonstration of TA1 maps moving through Macrostrat ingestion pipeline (Sec. 2.2) to TA3 data provision (Sec. 2.1) with data feedback, curation, and handling entirely by USGS staff

4 Current activities

In the first two months of CriticalMAAS, the Macrostrat team has made substantial progress towards the goals of Phase 1, in line with its proposed work plan (Sec. 5). We have made progress along several lines.

4.1 System and interaction design

The Macrostrat team has been a major contributor to the design of CriticalMAAS data schemas (`DARPA-CRITICALMAAS/schemas`), for harmonizing TA1-3 datasets and ensuring their interoperability. We have worked closely with the other TA4 teams and TA1-3 performers to ensure that the data schemas are specific and well-designed; we have also been an advocate for including geologic data objects in the schemas, to support our pursuit of linked geologic data objects.

4.2 Prototype HITL interfaces

We have made substantial progress towards the development of HITL interfaces in support of TA1-2 pipelines. Macrostrat's main web interface is already being adapted to support CriticalMAAS tasks, with new pages for map development, an index of available maps, and pages for individual map datasets (e.g., Huntington, UT). We have produced prototype interfaces for evaluating vector/raster map alignment (Fig. 4), viewing mineral site datasets (Fig. 2), and topological editing of map extractions (Sec. 2.4; Fig. 5). These interfaces will be extended and adapted, but many of the key new technical components (e.g., serving raster datasets; Sec. 2.1) have now undergone initial validation.

4.3 Literature extractions

The xDD system and COSMOS document extraction pipeline are being used to provide literature artifacts and a full-text searchable database ready for TA2 use, both for USGS documents and the broader geologic literature. These endpoints will underpin the development of new capabilities for surfacing document context (Sec. 2.1) and interface development for TA2 (Sec. 2.4).

We are beginning to transition to **GeoKB** as a source of USGS documents. As part of this transition, we switched from using the USGS Zotero instance as the primary metadata source for target documents to the GeoKB SPARQL instance, under the guidance of Sky Bristol. This aligns us with the storage and knowledge plans of Sky's group at USGS. This includes storing the `w3id` stable URLs, which will allow us to link directly to the original source for each USGS PDF.

We have created a document set (`criticalmaas`) defined as the union of these documents with the USGS series publications (doi prefix of `10.3133`). This set is available within the xDD system and queryable using its API. For instance, snippets of documents mentioning the *Bonneterre Dolomite*, a key unit in the Viburnum Trend and type locality of Mississippi Valley-type ore deposits, can be retrieved. Additionally, we are in the process of running COSMOS, word2vec, and doc2vec pipelines for the entire set (these are running in CHTC infrastructure and done to varying degrees of completion). None have live endpoints yet for the entire `criticalmaas` set (though we have them complete for the GeoKB-based articles). *These endpoints will be in place by the Month 3 hackathon.*

5 Index of milestone progress

5.1 Milestones 1 and 2

We are making progress on all proposed milestones. All but one deliverable proposed for execution by Month 4 has crossed key thresholds in readiness and is approaching completion, except for a single deliverable in Task 3B. The early establishment of key capabilities allows us to focus on building integrations with other performers (in all TAs) during and after the Month 3 hackathon.

Task 1: Supply geological data and literature artifacts to CriticalMAAS TAs 1-3

Augment and extend Macrostrat and xDD systems to deliver data and artifacts to TAs 1-3

1A: Extend Macrostrat for TAs 1-3 *Augment Macrostrat capabilities and datasets with functionality for AI-assisted critical mineral assessment.*

1. **Milestone 2** *(Month 4)*: A containerized instance of Macrostrat: *A containerized version of Macrostrat is running (though unstable), and is being used as a base for all development activities*
2. **Milestone 2** *(Month 4)*: Database and software capabilities to ingest and serve raster datasets: *Initial validation complete (Sec. 2.1)*
3. **Milestone 2** *(Month 4)*: User management and authentication: **In initial stages of development, planned by Month 3 hackathon**
4. **Milestone 2** *(Month 4)*: APIs to deliver geologic map and column data to TAs1-3: APIs based around existing map and tileserver APIs have been partially implemented, and deficiencies in data structure and queryability are being identified and evaluated.

1B: Extract literature artifacts using xDD-COSMOS and deliver to TA1-2 *Provide literature artifacts (maps and tables) to TA1-2*

1. **Milestone 1** *(Month 2)*: A vetted corpus of geological literature pertinent to mineral assessment: *The CriticalMAAS corpus is available (Sec. 4.3)*
2. **Milestone 2** *(Month 4)*: Pipeline for delivering contextualized literature artifacts to TA 1 and 2: *COSMOS outputs for maps, table extractions, etc. are available for the CriticalMAAS corpus (Sec. 4.3)*

Task 2: Ingest geological data from TAs 1-3

Incorporate data products produced by TAs 1-3 into Macrostrat

2A: Ingest geologic maps from TA1 and link entities *Incorporate TA1 map data products into harmonized Macrostrat map system*

1. **Milestone 1** *(Month 1)*: Schemas for map data to be accepted by Macrostrat system: *Done as part of TA4 deliverable (Sec. 4.1)*
2. **Milestone 2** *(Month 4)*: Documented ingestion APIs for maps from TA1: *Beginning to produce ingestion CLI and API for TA1 use (Sec. 2.2)*

2B: Ingest geological data from TA2 and link entities *Augment and extend Macrostrat map and column unit data to include mineral assessment-specific criteria*

1. **Milestone 1** *(Month 1)*: Schemas for point-based geological data to be accepted by Macrostrat system: *Done as part of TA4 deliverable (Sec. 4.1)*
2. **Milestone 2** *(Month 4)*: Documented APIs for point-based data ingested from TA2 (and TA1 as applicable): *Started in Weaver repository (Fig. 2)*

Task 3: Build HITL interfaces for model and extraction improvement

Build and deploy interfaces to annotate existing and TA- generated data with expert feedback

Subtask 3A: Annotate and edit geologic maps *Enable dynamic editing and annotation of geologic maps*

1. **Milestone 2** *(Month 4)*: Add widgets for collecting map candidate feedback to Macrostrat's web map interface: *In early development, but not publicly available (Sec. 2.4)*

Subtask 3B: Annotate geological data extractions and linked geological entities *Enable annotation of geological data extracted from descriptive documents*

1. **Milestone 2** *(Month 4)*: Add widgets for collecting linked entity feedback in Macrostrat web interfaces: ***Not yet addressed*** *(Sec. 2.3)*

5.2 Later milestones

We have made some progress to later Phase 1 milestones, as well:

- Subtask 1B **Milestone 4** *(Month 7)*: Pipeline for locating and extracting entities and augmenting Macrostrat database: *In early exploratory phases with CS graduate and undergraduate students supervised by co-PI Venkataraman (Sec. 2.3)*.
- Subtask 3A **Milestone 4** *(Month 7)*: Adapt Mapboard GIS topological editing for map geospatial/topology correction: *Key demonstration/validation has been accomplished (Sec. 2.4)*

6 Index of software repositories

- `DARPA-CRITICALMAAS/schemas`: Schemas for TA1-3 integrations (Sec. 4.1)
- `DigitalCrust/weaver`: Curation and ingestion of geological site datasets (Sec. **??**)
- `Mapboard/topology-manager`: Topological map editing (Sec. 2.4)
- `UW-COSMOS/cosmos-visualizer`: Page annotation and feedback (Sec. 2.4)
- `UW-COSMOS/COSMOS`: PDF entity extraction pipeline (Sec. 2.1)
- `UW-Macrostrat/cli`: Macrostrat system management tools (*currently private*; Sec. 2.1)
- `UW-Macrostrat/criticalmaas`: Index repository for the CriticalMAAS project
- `UW-Macrostrat/macrostrat-xdd`: xDD integration for geologic entity characterization (Sec. 2.3)
- `UW-Macrostrat/map-integration`: Geologic map ingestion and harmonization (Sec. 2.2)
- `UW-Macrostrat/python-libraries`: Python libraries used in multiple applications (Sec. 2.2)
- `UW-Macrostrat/raster-cli`: Prototype tool for ingesting raster datasets (Sec. 2.1)
- `UW-Macrostrat/tiger-macrostrat-config`: Configuration for CHTC infrastructure (*private*; Sec. 2.1)
- `UW-Macrostrat/tileserver`: Server for vector and raster tiles to GIS software (Sec. 2.1)
- `UW-Macrostrat/web-components`: Shared user interface components for Macrostrat user interfaces (Sec. 2.4)
- `UW-Macrostrat/web`: Macrostrat's main web interface (Sec. 2.4)