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

1.1 Success criteria for Macrostrat TA4 system

- Can we provide TA1 and TA2 data to TA3 (and other TA4 performers), en masse?
- Can we provide augment and standardize these datasets to provide relevant lithologic info?

1.2 System goals

Feed TA2 data to TA3

- Make mine report data sub-settable by geologic formation, lithology, etc. through linking to TA1 output (and geologic mapping from other sources)
- Provide these point data objects using standardized APIs that can be queried and assimilated in TA3 modeling

1.3 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. Key new things we would like to emphasize:

We will plan to integrate with tools produced by other TA4 performers:

- Jataware will produce some end-to-end solutions particularly for georeferencing and page-level evaluation of map data objects.
- MTRI will work with TA4 to ensure that geologic mapping data can be filtered and subset

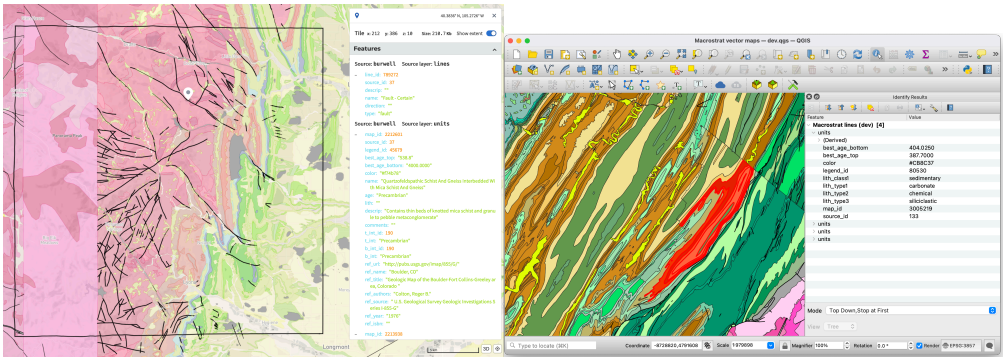
Based on descriptions of bottlenecks in CMA process workflows described by Lawley, and the expected structure of TA1 and TA2 outputs, we forecast that a major problem will be assembing a geologic dataset well-characterized and standardized enough to be useful across scales and study areas to extract 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.

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.

Our software development plan will be organized around three general 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



(a) Tiled Macrostrat output for provision to TA3

(b) Attributed Macrostrat map in QGIS

Figure 1: Macrostrat's tiled output API, for provision to TA3

This is broadly similar to the Tasks 1-3 proposed 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 contextx of this project.

2.1 Providing geologic datasets for CriticalMAAS performers

Our key task for supporting CriticalMAAS is to provide harmonized geologic datasets over stable APIs to other CriticalMAAS performers. They 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.

2.1.1 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. We will continue to refine this output and provide it to TA3. We will work on

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

The key codebase for this work, as well as for raster data provision, is the UW-Macrostrat/tileserver repository.

2.1.2 Mineral site data

We have successfully validated serving and filtering of point datasets relevant to CMA, including the MRDS dataset. Approaches to curating point data and linking it closely to geologic context (enabling spatial and geological time/unit filtering)

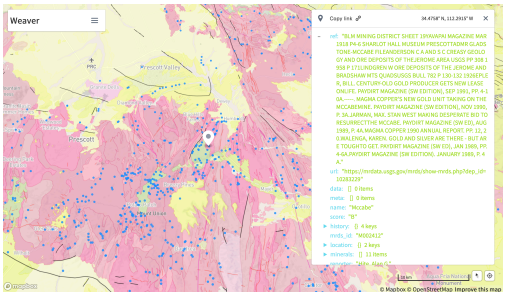


Figure 2: Macrostrat MRDS data layer, showing basic capabilities for point data provision

Mine sites synthesized from TA2 will be linked to geologic context and forwarded to TA3 and TA4 HITL interfaces (ex., MTRI and/or EIS QGIS plugins).

Macrostrat maintains links to other point datasets that may be useful to forward to TA3, such as USGS geochemical databases. If more datasets are needed, we will work with TA3 to identify and integrate them.

2.1.3 Raster datasets

Raster datasets can usually be easily accessed directly by TA3. However, compositing raster datasets across scales, etc. presents demanding workflows that can be supported by TA4 potentially. We have validated and will maintain key capabilities to store and serve raster datasets, to support map feedback and CMA workflows.

- Macrostrat Raster CLI

This can also be used to help support TA1 feedback

2.1.4 A system to orchestrate geological data

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

- Macrostrat CLI
- Macrostrat infrastructure configuration

2.1.4.1 Hackathon targets

- *Month 3 hackathon*: Containerized Macrostrat system that supports basic capabilities
- *Month 6 hackathon*: Data model adjustments and pipelines for storing new data and annotations
- *Base evaluation*: End-to-end system for storing and distributing geological data and literature artifacts

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're going to try to improve the speed and interactivity of Macros-

trat's vector data ingestion pipeline. This system moves from heterogeneous inputs, like Geodatabases, Shapefiles, or the old ArcInfo files you referenced, to the standardized layers that drive Macrostrat APIs.

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

We will build a web-based interface to allow geologists to interactively manipulate map data, and to provide feedback on the quality of the map data. This will operate over:

- Vector datasets in general (for assimilating already digitally-published mapping and TA1 outputs for representation in Macrostrat)
- paired vector/raster map datasets (to facilitate training by TA1)

2.2.1 Relevant software repositories

- Macrostrat CLI holds processing interfaces
- Map integration system will hold map ingestion/harmonization web app
- Python libraries: monorepo for Python libraries used across projects

2.2.2 Milestones

- Initial demo: Month 3 hackathon

2.3 Geologic entity characterization pipeline

- Macrostrat maintains a database of geologic entities, which can be used to further characterize geologic maps

However, this is not of adequate quality to support CMA workflows. This will be rectified by a combination of leveraging outputs from TA1-2, augmenting it with our own AI-assisted literature synthesis, and building new HITL interfaces for characterizing rock units.

Two separate approaches:

1. Find new descriptors of existing entities
2. Find new entities not currently tracked in the database

This will help correct several deficiencies of Macrostrat's current representation of geologic units:

- Lack of information about non-sedimentary units
- Lack of specificity about unit properties

Geologic units in Macrostrat have curated properties, but these are often not rigorous or descriptive enough to provide the level of detail needed for CMA workflows.

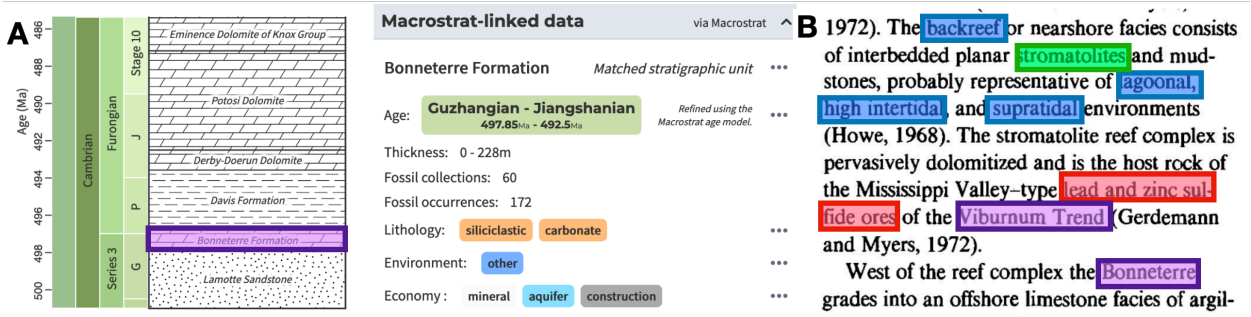


Figure 3: *(a)* Starting user interface and *(b)* potential additional extractions for CMA-focused entity canonicalization tasks

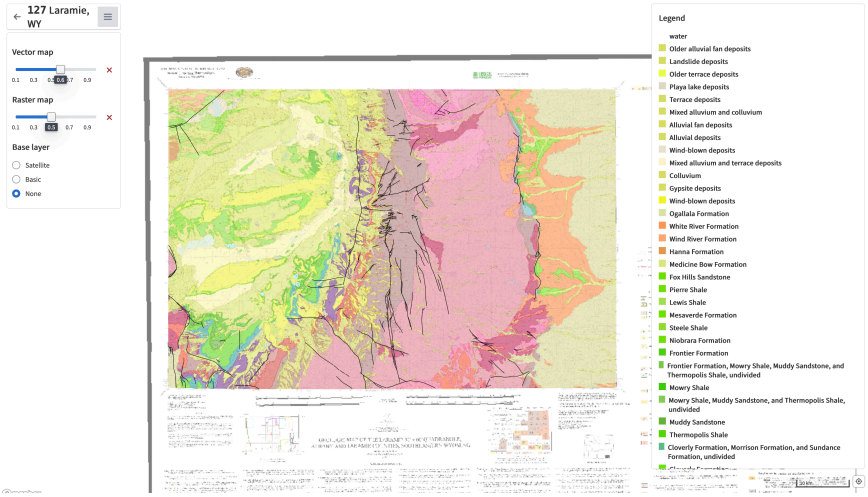


Figure 4: Map interface showing both vector and raster datasets for the same geologic map, in an interface with synthesized legend information

2.4 HITL interfaces for TA1-2 pipeline support

Our plan is to produce key HITL interfaces, especially to support TA1 and TA2.

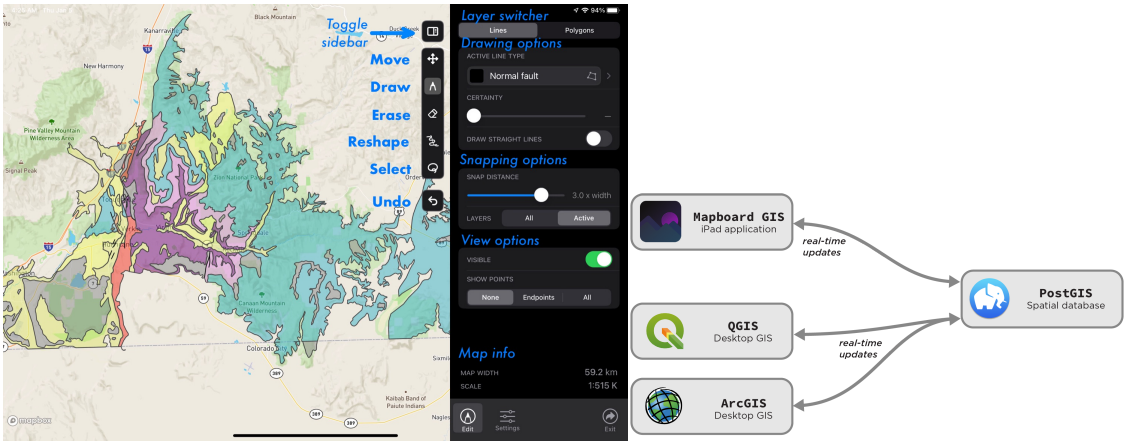
We will also account integrate with feedback interfaces produced by Jataware and MTRI. In particular, Jataware's map-projection system will be a critical precursor to our map ingestion system, and MTRIs QGIS plugin will be a key interface through which TA3 manipulates our vector-tile geologic mapping outputs.

2.4.1 Map feedback interfaces

- Macrostrat web
- Macrostrat web components

2.4.2 Map editing

- Mapboard topology



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

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

Figure 5: Elements of the Mapboard GIS system for TA1 map correction and feedback.

2.4.3 Document-based interfaces

- Coordinate with Jataware for TA2-supporting interfaces
- COSMOS visualizer page-level annotation interface may be adapted, or a Jataware-created tool may be used

2.5 Targets for hackathon events

2.5.1 Month 3 hackathon

- Import pipeline for geologic maps (TA1 outputs)
  - Feedback mechanism for map legend extractions
- Synthesis of these outputs into TA3-ready products
  - First attempts, showing maps in the right structure but not properly attributed

2.5.2 Month 6 hackathon

- Demonstrated pipeline to accrue descriptive characteristics of rock units from literature synthesis
  - Pathway to involve TA2 in providing data to this pipeline
  - Pathway to involve TA3+USGS in providing feedback and HITL effort towards synthesizing geologic entities
- Goal: produce “clean” and highly specific lithologic breakdowns of rock formations amenable to querying by TA3

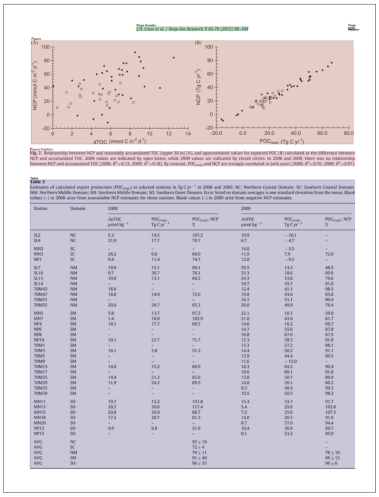


Figure 6: COSMOS image tagger user interface, which is an option for adaptation into HITL systems

2.5.3 Month 9 hackathon

2.5.4 Phase 1 Base evaluation

3 Current status

3.1 System and interaction design

- Macrostrat team has spearheaded the production of schemas for data interchange with TA1-3
- Key additions to the schema: geologic data objects

3.2 Providing literature extractions to TA1 and TA2

The xDD system and COSMOS document extraction pipeline are being used to provide literature artifacts ready for TA2 extractions, both over USGS documents and the broader geologic literature.

We are beginning to transition to **GeoKB** as a source of USGS documents. As part of this transition, 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 entire 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.**

We will continue to integrate with TA2 to build capabilities around accessing and manipulating the literature corpus of documents from USGS and other sources. Since USGS documents are broadly in the public domain, TA2 performers 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.

3.3 Progress to initial milestones

We are making progress on all proposed milestones. All but one deliverable proposed for execution by Month 4 have crossed key thresholds in readiness and are near 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.

3.3.1 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

3.3.1.1 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 but not stable, 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
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.

3.3.1.2 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
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

3.3.2 Task 2: Ingest geological data from TAs 1-3

Incorporate data products produced by TAs 1-3 into Macrostrat

3.3.2.1 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
2. **Milestone 2** *(Month 4)*: Documented ingestion APIs for maps from TA1: Beginning to produce ingestion CLI and API for TA1 use

3.3.2.2 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
2. **Milestone 2** *(Month 4)*: Documented APIs for point-based data ingested from TA2 (and TA1 as applicable): Started in Weaver repository

3.3.3 Task 3: Build HITL interfaces for model and extraction improvement

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

3.3.3.1 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

3.3.3.2 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**

3.4 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.
- Subtask 3A **Milestone 4** *(Month 7)*: Adapt Mapboard GIS topological editing for map geospatial/topology correction: Key demonstration/validation has been accomplished