DATA MANAGEMENT PLAN
The Alaska Soil Data Bank (ASDB): A database for delivering non-NRCS legacy data for digital soil mapping initiatives in Alaska

Expected Data Type. The data generated by this project will be digital soil data generated by acquiring existing non-NRCS soil datasets from academic and agency partners, peer-reviewed literature, and grey literature (i.e. government publications and reports). It is expected that this project will include the acquisition of over 200 datasets and add over 14,000 new observations to the NASIS database for Alaska, particularly in underrepresented areas in the NASIS datasets that are critical areas to gap fill to meet Soil2026 goals.

Data Format. This original data will be acquired and ingested "as is" – if the original data is digital, it will remain in its original format. If the data is held as hard copy records, those will be scanned, digitized, and formatted for NASIS ingestion prior to input into GEMS. The output will be a harmonized dataset in machine-readable format prepared for automated input into NASIS and following all NASIS formatting standards.

Data Storage and Preservation. Data generated by activities outlined in this proposal will be initially stored and maintained in the GEMS platform hosted in the data center of the Minnesota Supercomputing Institute (MSI). MSI performs full off-site backups as part of their disaster recovery plan. GEMS has recurrent Minnesota State Government funding to maintain its core operations, currently with no expiry date. All registered data in GEMS is persistent, given a DataCite DOI minted by our University of Minnesota Libraries partners.

In addition to storage in GEMS and linkage in metadata to original data sources whenever possible, harmonized data will be imported into NASIS in compatible, machine-readable format and therefore all data produced by the project will also be imported into NASIS prior to project end. Per the project narrative, the power of the GEMS platform is that datasets are input "as is", maintaining the data structure of the original authors. GEMS metadata for each dataset includes full reference information. If the dataset is publicly available, the GEMS metadata for that dataset will include information on dataset access, storage platform, and author contact information. This metadata is in JSON format and can be reconfigured and exported for any metadata catalog (in this case specifically targeting the NOTCOM catalog maintained by the NRCS Initial Mapping Team) that may be maintained by NRCS. The GEMS platform excels at utilizing field metadata and fuzzy matching to enable post-input data harmonization. During the dataset input process, in addition to constructing dataset metadata, metadata tags are assigned to each dataset field from a dynamic data dictionary which can be expanded as necessary as more datasets are added. These field metadata tags then become the primary vehicle through which scripts are written to subsequently harmonize the data for export and usage in DSM. There are many advantages to this approach. The primary advantage is that datasets are imported "as is", meaning that data will match the original data source and author format. This increases transparency because there is no prior "re-working" of data before it is added to the database. As scripts are written to harmonize datasets to a defined format using field metadata tags, these scripts will also be available on GitHub to enable any potential user to recreate the eventual harmonized dataset that is utilized for DSM. Finally, because scripting is used for harmonization, output format is flexible. This is critical, because scripts will be developed to harmonize data and output in a format which can be ingested into NASIS.

Data Sharing, Protection, and Public Access. The GEMS agroinformatics platform developed by the University of Minnesota supports the data interoperability and sharing needs of this project. GEMS adheres to and, via its extensive (and still expanding) data handling toolkit, makes practically possible, FAIR(ER) data sharing principles; findable, accessible, interoperable, reusable, ethical—i.e., respects data sensitivities (e.g., partner sourced) or any proprietary (e.g., privately-sourced) data—and reproducible. Non-private data will be made open to the public via GEMS-Open, an open database of functionally interoperable and documented data that adheres to standard and harmonized meta-data protocols. When viewing and analyzing datasets, GEMS requires users to click licenses of each primary data source prior to receiving access to the data. Similarly, all data product derivatives are electronically tagged with links to the product sources in their metadata, ensuring proper credit attributions, as often described by their license agreements. The ASDB will adopt a data policy that accounts for data ownership, allows reuse of data and creation of data derivatives (e.g., pooling of soil data into new larger continental soil applications). We will adopt the *Creative Commons Licensing* scheme for data sharing and attributions. Importantly, the spectrum of private-public data sharing is accounted for in the ASDB. This is an important aspect because some soil data have inherent restrictions of data usage (e.g., soil projects funded by private companies or corporations, soil data with privacy/confidentiality protection), while other soil projects funded by federal resources are considered public goods without data restrictions. Geographic coordinates are in some cases private data that need to be protected for confidentiality reasons and may be released only for public use in aggregated/masked format.

Roles and Responsibilities. PI Jelinski will be responsible ensuring the data management plan is followed by all project participants and will also be responsible for reporting compliance with this data management plan to NRCS. The annual and final reports to NRCS will have an expanded data management plan as an integral component, and will include progress in data sharing (e.g., statistics on shared and accessed data sets, publications, DOIs, outreach materials). The final report will describe the data that was produced during the award period and the components that will be stored and preserved (including the expected duration) after the award ends. The GEMS platform data steward will ensure that the data management plan will be compliant with the Research Terms and Conditions that govern NRCS-funded projects.