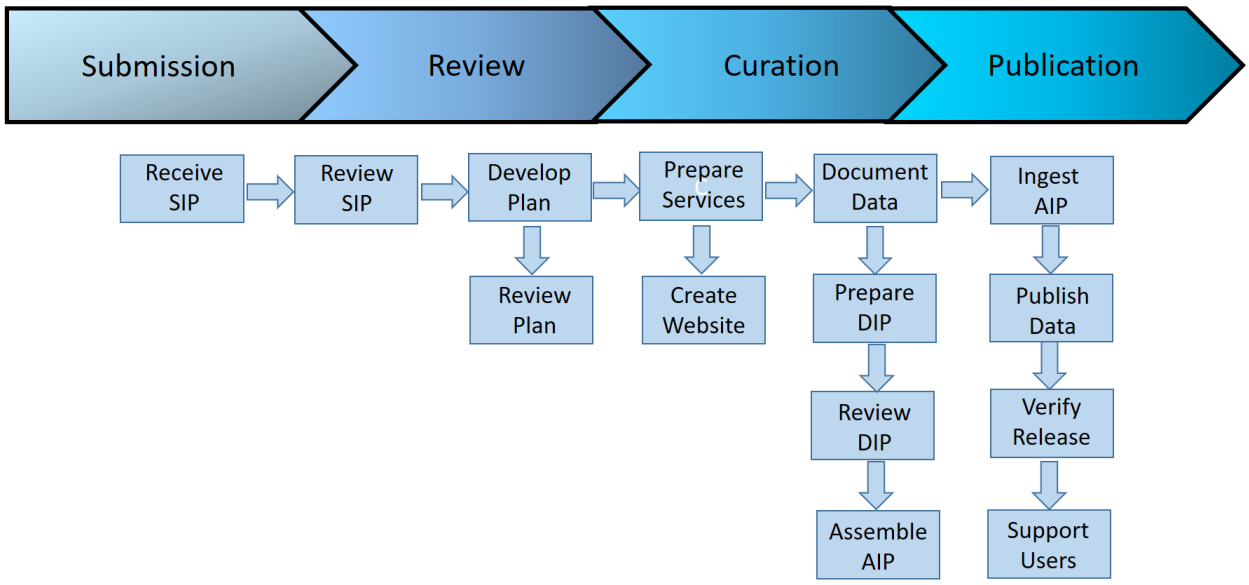
Overview of the SEDAC Data Curation Workflow



The workflow for data that has been accepted for dissemination by SEDAC is described sequentially, below. Some steps are completed simultaneously to release the data in an efficient manner.

Submission

The data producer submits the online form, SEDAC Data Nomination – Request for Acceptance of Data Submission.

Upon completing an initial review, if accepted, SEDAC receives the data and associated information from the data producer as a Submission Information Package (SIP) to be reviewed for potential dissemination to the designated community.

Review

The SEDAC science team conducts an extensive review of the SIP and, if determined to be valuable for the user community, develops the Data Dissemination Plan, which is presented to the SEDAC User Working group (UWG) and NASA representatives for review. If approved, the data producer is notified and data curation begins.

Curation

The data landing page and services are developed on the SEDAC website.

The data are curated and described in metadata and documentation in accordance with the data dissemination plan.

Data producers are invited to contribute to the metadata and the data documentation.

The Dissemination Information Package (DIP) is prepared from the data and is reviewed and verified. Internal alpha and external beta reviews of the data product are conducted.

Publication

The SIP, provenance information and DIP are assembled into the Archival Information Package (AIP). The AIP is verified and ingested into the archive and multiple copies are distributed to secure storage locations.

The data are published on the production platform. A list of frequently asked questions (FAQ) is created and requests for help are addressed by user services staff.

More detailed descriptions of the processes that SEDAC follows are provided, below.

SEDAC Data Management Processes

SEDAC has established data management processes to meet the needs of the community. These processes are described as discrete items, below, for clarity. In practice, many of the processes are conducted simultaneously or in parallel to reduce the potential for delays.

Data Nomination

Data producers who are interested in having SEDAC disseminate their data as open data visit the SEDAC data submission webpage and submit the form, SEDAC Data Nomination – Request for Acceptance of Data Submission. The content of the form is received and reviewed by the Deputy SEDAC Manager for potential relevance to current SEDAC collections and priorities. The SEDAC Lead Project scientist and others are consulted and, if the nomination request is determined to be of potential value to the community, it is assigned to the member of the science team who specializes in the topic of the nominated data to serve as the lead for assessing the data nomination.

Authority

The authority of the submitter is verified by reviewing the referenced literature and cross-referencing the contact information of the submitter with relevant publications and organizational affiliations. The science team and members of the community are consulted to determine whether the submitter is a known member of the community or, in the case of a student, whether their advisor is known and should be consulted to verify the authority of the submitter as the data producer or the designated representative.

Value Assessment

The study description is assessed by the science team to determine how the proposed data product could provide value to the SEDAC user community. Communications are established with the data producer to inform the data producer about relevant policies and procedures and to ask questions. During the assessment of the potential value of the nominated data to the SEDAC user community, access is requested to a review copy of the data or, if applicable, to a sample of the data.

Preliminary Data Review

The process by which access to a review copy of the data was received is verified to ensure that the data are obtained from the data producer or the designated representative. SEDAC staff review the data to determine the potential value to the SEDAC user community. Relevant publications are reviewed, focusing on methods and potential use of the data. Data and software formats, rights, documentation, viability, projected costs, schedules, and other considerations for dissemination are also reviewed. If a decision is made to nominate the data for dissemination, a data dissemination plan is developed and the SEDAC Deputy Manager assigns the appropriate expert among SEDAC staff to collaborate with the data producer and draft the data dissemination plan.

Data Dissemination Plan

The dissemination plan for the data is developed in consultation with the data producer, the SEDAC Deputy Manager and the SEDAC Lead Project Scientist, who also may recommend others whose expertise is needed to develop the plan for disseminating the data. The data dissemination plan is reviewed by the SEDAC Deputy Manager and the SEDAC Lead Project Scientist, who request modifications to the plan, as needed, in preparation for presentation to the SEDAC User Working Group and NASA representatives.

Data Dissemination Plan Review

The SEDAC Deputy Manager or the SEDAC Lead Project Scientist presents the data dissemination plan for the nominated data product to the SEDAC UWG and NASA representatives, either during a scheduled SEDAC UWG Meeting or via email correspondence. Questions about the data are answered or investigated to provide the SEDAC UWG and NASA representative with the information that is needed to make a determination regarding the data dissemination plan. A decision is made whether to approve the data dissemination plan and, if approved, data curation planning begins.

Data Curation and Data Dissemination Planning

Upon approval of the data dissemination plan, comments received during the approval process are assembled for inclusion in the data curation plan. The data dissemination plan is reviewed to identify any conversions and processing necessary for dissemination to current and future users. Tasks to be completed are described in the data curation plan and assigned to a designated internal data product champion and the production team. The data product champion notifies the data producer and requests the data and associated information,

including the open data and information agreement, documentation, and related information and they agree upon the deliverables and transfer method. They decide on the dataset title, version number, rights, collection assignment, dissemination formats, abstract, metadata, documentation, and associated products and services.

The data product champion is responsible for establishing the schedule for each step in the workflow and for shepherding the data through each step. The data product champion also is responsible for identifying and resolving any potential delays or impediments for completing the steps to transition the data throughout the workflow. If the data product champion identifies challenges to completing any steps of the workflow, the SEDAC Deputy Manager is notified with a request to mitigate the challenge.

Transfer of the Submission Information Package (SIP)

The transfer of the data is completed and verified by authenticating the transmitter and transfer method to determine whether the transfer was completed as agreed. Discrepancies are resolved by communicating with the data producer or the designated representative. The data and associated information are assembled into a Submission Information Package (SIP), which is inspected and inventoried to compare it with the items expected. Any missing data and information is requested when communicating with the data producer or designated representative.

Inspection of SIP

Inspections of the SIP are conducted for viruses or security threats, sensitive or confidential information, and intellectual property rights for open data, integrity verification, and completeness of inventory, If the results of any of the inspections are unsatisfactory, the data producer or designated representative is contacted and the issue is resolved before accepting the SIP.

Validate Data Integrity and Generate Fixity Values

The contents of the SIP are inventoried using Digital Record Object Identification (DROID) Software. Fixity values for all items in the SIP are generated to verify and validate the integrity of the data. If the SIP is received with fixity values, such as SHA-1 or MD-5 digests, the received fixity values are compared with the generated fixity values to ensure the integrity of the data received. The data is assigned a unique identifier for internal use, described in the roster of UWG-approved data, and recorded on the data holdings list. Upon receiving the SIP and verifying that it has been received as agreed in good order, the data producer or designated

representative is notified to acknowledge receipt when SEDAC has taken possession and accepted responsibility for the preservation and dissemination of the data.

Create and Review Data Landing Page, Metadata, and Data Documentation

The dissemination plan for the dataset guides the data publication effort. The items to be published as open data artifacts, including the data landing page, metadata and data documentation, are prepared in collaboration with data producers and reviewed for completeness and consistency. The data and related items are assembled on the data landing page within the alpha environment and reviewed by the data production team and the data product champion. A list of frequently asked questions (FAQ) is produced. The completed elements are subsequently reviewed by the SEDAC Configuration Management Board (CMB) and revisions requested by the CMB are described in the data curation plan. Upon completing the requested revisions in alpha, the data landing page is copied to the beta environment for external review by representatives of the SEDAC User Working Group, NASA, and the community. Revisions requested during the beta review are added to the data curation plan and completed. The CMB reviews the data curation plan and the revised beta web pages and, if necessary, requests additional refinements, which are recorded in the data curation plan and completed prior to dissemination.

Assemble and Store the Archival Information Package

The Dissemination Information Package (DIP), SIP, and supplemental information, including an open data and information agreement, dissemination plan, curation plan, digital signatures, and inventory information are assembled into an Archival Information Package (AIP) using the Bagit library of software. Identical preservation, security, and circulation master copies are created. The integrity of each copy is verified and the copies are stored redundantly in distributed locations.

Publish and Support Use of Data

Upon successful completion of all reviews and archival processes, the data are released within the production environment. Reviews are conducted to verify that the production environment is correct and complete. The user services team supports requests for information about the data and forwards such requests to the appropriate expert.

Continuous Quality Improvement

As part of its commitment to serving the community, SEDAC engages in continuous quality improvement (CQI) to enhance its data stewardship capabilities on an ongoing basis. CQI at SEDAC focuses on improving the effectiveness and efficiency of activities performed during the entire data curation workflow. During each step of the workflow, SEDAC managers and staff explore opportunities to correct potential inefficiencies in the process that could be improved as well as examining potential risks to safety and security that could be avoided. Likewise, sources of errors are investigated to determine how they can be avoided. The SEDAC CMB regularly contributes to CQI by identifying ways to reduce the need for redundant revisions to data products and by identifying opportunities to adopt new techniques for improving the submission, development, dissemination, and support of data products and services. In addition, SEDAC staff members regularly engage in professional development activities and routinely discuss new developments in available resources, processes, and technologies with other data stewards, producers, and users to identify opportunities in which activities conducted during any of the steps of the workflow can be improved. When approaches that have the potential for improving processes are identified, they are proposed, discussed, tested extensively prior to adoption to avoid any unintended consequences. Upon adoption of proposed revisions, written procedures and related documents are modified and staff are trained to follow the adopted changes. As part of its CQI efforts, SEDAC regularly completes a variety of self-assessments, internal reviews, and external audits of its infrastructure, capabilities, and processes to identify opportunities for improving upon the ways that data, related information resources, and services are provided to the community. NASA also conducts routine audits of SEDAC to ensure that existing and new requirements are being met.