CDI Science Support Framework

The Community for Data Integration (CDI) represents a dynamic community of practice focused on advancing scientific data and information management and integration capabilities within the United States Geological Survey (USGS).

In 2012, the CDI Coordinators developed a Science Support Framework (SSF) (Figure 2) that categorizes and relates the activities and processes through which research data flows and within and upon which the CDI operates. It is these categories that provide the focus and a framework for coordination and integration of current and future CDI-funded projects. A more detailed explanation of the CDI SSF categories and the direction of data flow through the framework are included towards the end of this page.

CDI Overview: of the CDI Operational Context

Since 2009, CDI has funded a variety of projects that support the overarching goal of data integration (Proposals). USGS and other researchers conduct monitoring, assessment, and research activities that generate data assets, which through the application of business, computational, and analytic processes and technologies are converted into information that contributes to our understanding of the Earth's physical and biological systems. It is within this context that data management and integration occurs and where the CDI operates (Figure 1). The CDI has provided funding support for projects that promote data integration and

- Focus on short-term benefits to science
- Leverage existing capabilities
- Apply solution/methodology that can be replicated
- Ensure sustainability
- Seek substantial return on investment
- Expose corporate data
- Organize science models and outputs
- Preserve and access project data

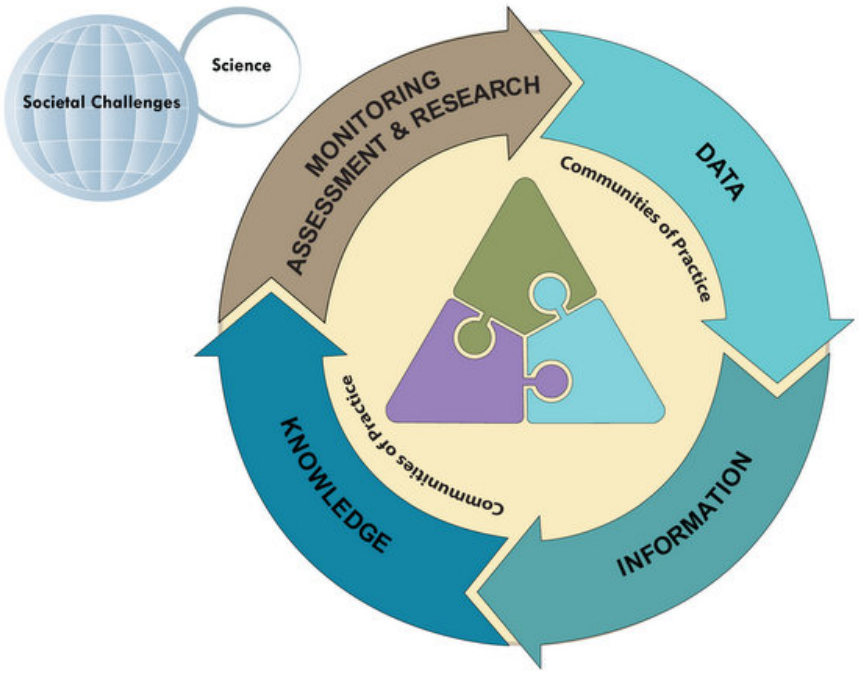
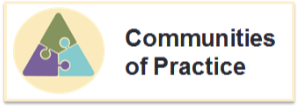


Figure 1: Overview of CDI Operational Context

Breakdown of the Components of the CDI Overview (Figure 1)



Communities of practice include scientists, the CDI as a whole, CDI Working Groups, external partners, and the human network of scientific domain collaborators.

Computational tools and services include applications, Web services, data discovery tools, models, semantic services and tools, infrastructure, data brokers, and visualization tools.

Management, policy, and standards include data stewardship, the implementation of the Science Data Lifecycle, knowledge management, data standards, governance, and policy.

Data and information assets include persistent archives, data registries, catalogs, data, metadata, derived information products, knowledge bases, and vocabularies/ontologies.

CDI Science Support Framework (SSF)

The CDI SSF provides a conceptual architecture that: illustrates how the CDI contributes to Bureau-level data integration efforts; and defines how current and future CDI projects fit within the framework.

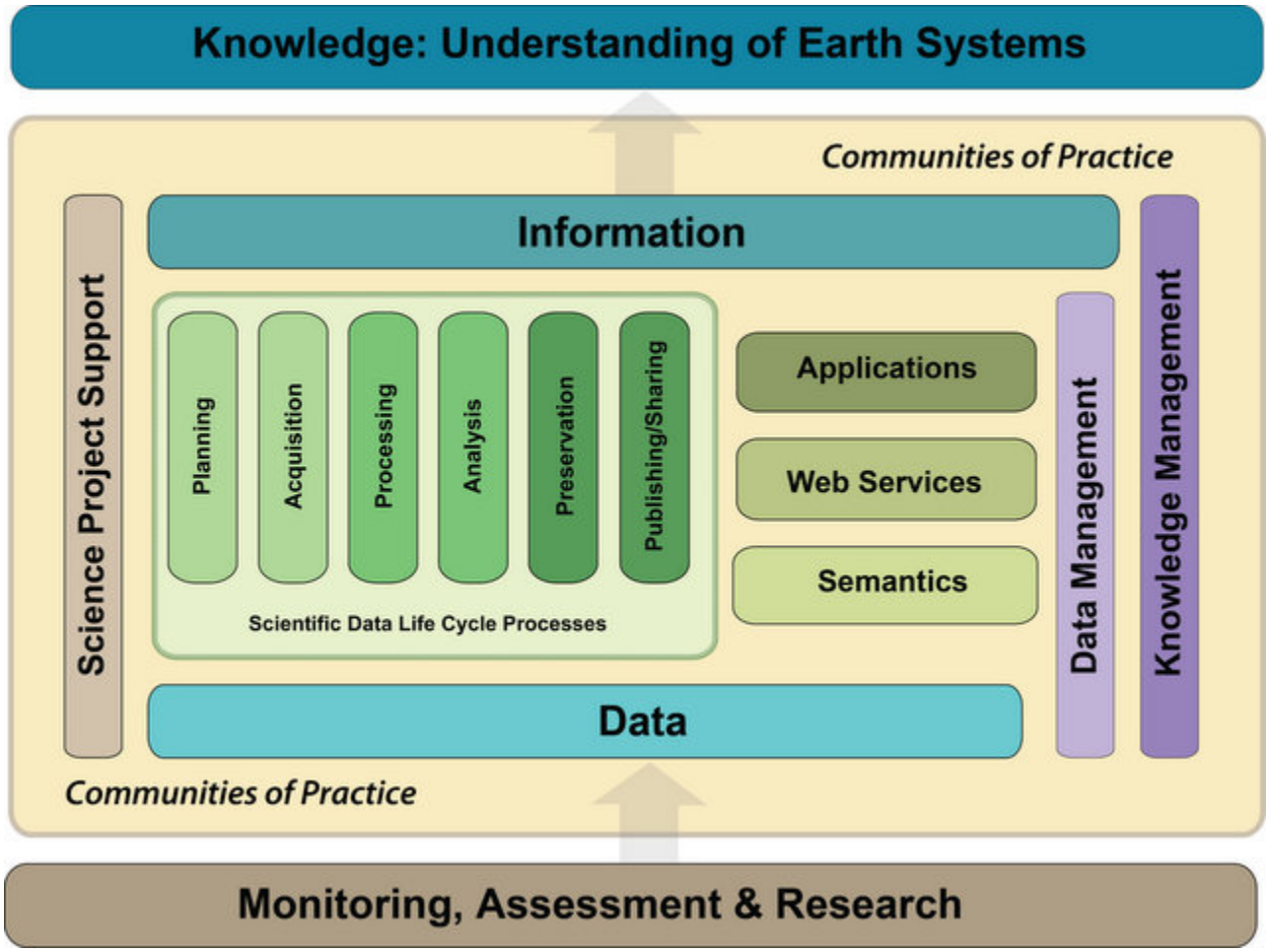


Figure 2: CDI Science Support Framework (SSF)

* Note that the color of the Framework elements (Figure 2) match those of the Overview elements in Figure 1.

USGS SCIENTISTS conduct **MONITORING, ASSESSMENT, AND RESEARCH** that generates **DATA ASSETS.** Through the application of business, computational, and analytical processes and technologies, these USGS **DATA ASSETS** flow vertically through the SSF from a base of **MONITORING, ASSESSMENT, AND RESEARCH** through the Science Data Lifecycle Model (SDLM) processes, applications, Web services, and semantics. The **DATA ASSETS**are transformed into **INFORMATION** products that benefit from data and knowledge management and also increase **KNOWLEDGE** and understanding of the Earth's physical and biological systems. Data assets also flow horizontally through the SSF from and through science projects to data and knowledge management.

The horizontal elements in the SSF represent the "what" of the CDI: products and tools, the things that contribute to the advancement of scientific data and that lead to the development of knowledge and understanding of the Earth's systems.

The vertical elements in the SSF represent the "how" of the CDI: the processes, the implementation of standards and best practices, and the interactions among people, data, and technology used to achieve data integration.

Individual Framework element descriptions:

* Note that the color of the Framework elements match those of the Overview elements

Science Inputs (the brown elements)

Monitoring, Assessment, & Research: USGS scientists conduct monitoring, assessment, and research that generates data assets. Through the application of business, computational, and analytical processes and technologies, these assets are converted into information products that can be shared with other researchers, stakeholders, and citizens to increase our knowledge and understanding of the Earth's physical and biological systems.

Science Project Support: Successful science projects encompass a range of activities represented in the SDLM. At each step in the cycle, researchers and data stewards rely on an array of sophisticated tools and services for data, information and knowledge discovery, acquisition, integration, management, and sharing.

Communities of Practice (the tan element)

Communities of practice are the foundation for CDI and all its products – the communities of people working towards the goal of advancing scientific data and information management and data integration across the USGS.

Data & Information Assets (the blue elements)

USGS assets include **Data** (e.g., raw data, databases, and linked open data (RDF[1])); **Information** or derived/interpreted information products in the broad sense (e.g., published or shared maps, reports, datasets); and **Knowledge** of all types and in all forms — recorded, organized, and preserved in the form of various artifacts. Knowledge can then be improved; shared across groups, organizations, and domains; and reused to support individual and group learning and research.

Computational Tools & Services (the green elements)

Science Data Lifecycle processes include tools and services that move data through the SDLC, human and machine interactions, and interactions with data through technology.

Detailed descriptions of SDLC Processes:

- **Planning** – A documented sequence of intended actions to identify and secure resources and gather, maintain, secure, and utilize data assets;
- **Acquisition** – The series of actions for collecting or adding to data assets;
- **Processing** – A series of actions or steps performed on data to verify, organize, transform, integrate, and extract data in an appropriate output form for subsequent use;
- **Analysis** – A series of actions and methods performed on data that help describe facts, detect patterns, develop explanations, and test hypotheses;
- **Preservation** – Actions and procedures to keep data for some period of time; to set data aside for future use; and
- **Publishing/Sharing** – To prepare and issue, or to disseminate data or information products.

Semantics convert raw data into data that can be interpreted by machines: Machine Readable Metadata, Semantic Mediation for Data Integration & Discovery, Ontologies/Vocabularies, and World Wide Web Consortium Standards.

Web Services include machine to machine data exchange, SOAP,[2] REST,[3] SPARQL[4] EndPoints, and other protocols and services.

Applications include human readable data services and user interfaces to data driven applications.

Management, Policy, & Standards (the purple elements)

Data Management includes data and metadata standards and policies and occurs in all phases of the Science Data Lifecycle from scientific research to finished information products.

Knowledge Management involves the creation, standardized documentation, and organization of knowledge using tools such as SKOS[5] Vocabularies and information modeling, resulting in the formation of knowledge bases.

[1] Resource Description Framework

[2] Simple Object Access Protocol

[3] REpresentational State Transfer

[4] SPARQL Protocol and RDF Query Language

[5] Simple Knowledge Organization Systems