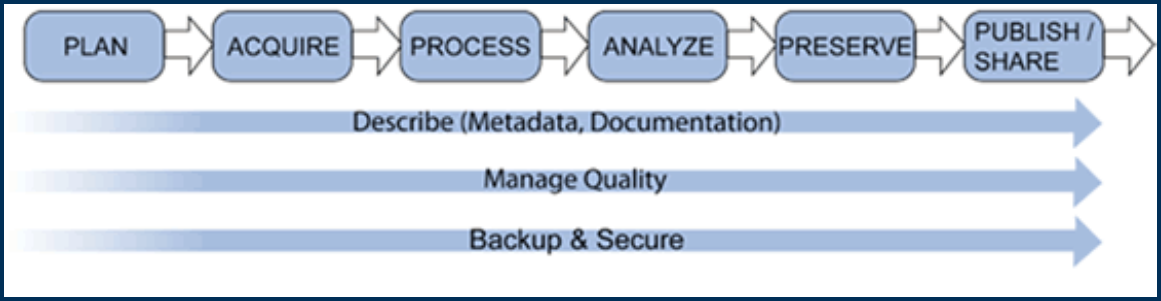
USGS
science for a changing world

USGS Data Management Training Modules:

USGS Science Data Lifecycle



Slides and slide notes are available for download from the Data Management website.

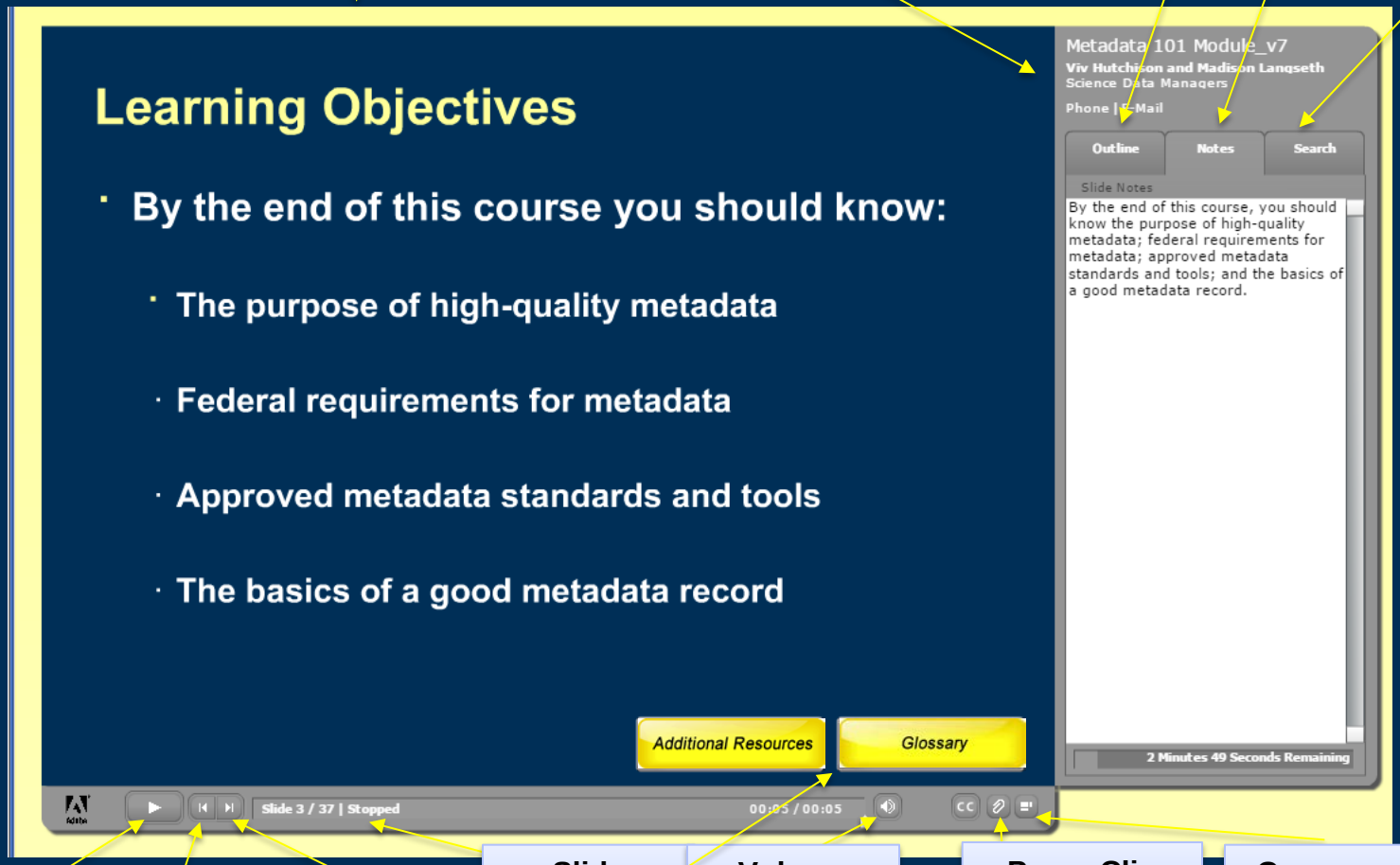
U.S. Department of the Interior
U.S. Geological Survey

Version 1, April 2016

Welcome to the USGS Data Management Training Modules, "USGS Science Data Lifecycle."

Slides and slide notes are available for download from the Data Management website: https://www2.usgs.gov/datamanagement/training/modules.php.

Course Navigation 101



Presenter Window

Instructor's Contact Info

Notes Tab

Outline Tab

Search Tab

Navigation Side Bar

Play

Backward

Forward

Slide Number

Volume Control

Paper Clip Attachments

Compress Nav. Bar

Additional Resources and Glossary buttons

USGS

Course Navigation 101

Each lesson starts with a title screen followed by lecture slides. The screen consists of a presentation window, navigation side-bar, and navigation base bar. The **presentation window** is used to view the slides that summarize the lecture material. The **gray navigation sidebar** is to the right of the presentation window. **At the top** you will find contact information for the course coordinators. The rest of **the sidebar is divided into three panes**:

- The **Outline** pane lists the slides in the lesson. You can use this pane to go to any slide within the lesson.
- The **Notes** pane shows the **lecture material** for each slide. **The principal content of each lesson is contained in the lecture material and should not be skipped**. However, we recommend reading the lecture material from a *downloaded file* because the scientific formatting in the downloaded version may be more accurate than in the version shown in the notes pane.
- The **Search** pane allows you to search for any word in the lesson and will show all slides containing that word.
- The **gray navigation bar** below the presentation window contains:
    - "Forward" and "Backward" buttons to advance and return to previously viewed slides.
    - The "Fast Forward" button.
    - The status bar displaying the current slide number.
    - Controls for sound and volume.
    - A toggle button (lower right) changes the view of the Navigation Side Bar to full screen or to a compressed icon.

YELLOW BUTTONS: Additional Resources and Glossary

- **Additional Resources** – Provides additional references, suggested training, or other information.
- **Glossary** – Click here to find the definitions of terms used in the course.

Learning Objectives

- By the end of this course you should know:
  - What is a science data lifecycle?
  - Why a science data lifecycle is both important and useful.
  - What are the elements of the USGS science data lifecycle and how they are connected.
  - Roles and responsibilities.
  - Where to go for more information.

Learning Objectives

By the end of this module, you should be able to answer the following questions… What is a science data lifecycle? Why is a science data lifecycle important and useful? What are the elements of the USGS science data lifecycle, and how are they connected? What are the difference roles and responsibilities? Where do you go if you need more information?

Suggested Citation

- Henkel, H.S., Hutchison, V.B., Langseth, M.L., Thibodeaux, C.J., Zolly, L., 2015, USGS data management training modules—USGS science data lifecycle: U.S. Geological Survey, http://dx.doi.org/10.5066/F7RJ4GGJ.

Suggested Citation

Henkel, H.S., Hutchison, V.B. Langseth, M.L., Thibodeaux, C.J., Zolly, L., 2015, USGS data management training modules—USGS science data lifecycle: U.S. Geological Survey, http://dx.doi.org/10.5066/F7RJ4GGJ.

A Warm-up Question

Before we get started, a quick question:

What is a data lifecycle (which apply)?

a) A high-level overview of actions, operations, or processes, started at different stages.

b) A record of data—from conception through preservation and sharing.

c) Something that clearly connects data management activities with research project plans.

A Warm-up Question

Before we get started, here's a quick warm-up question – what is a data lifecycle? Is it

a) A high-level overview of actions, operations, or processes, started at different stages,

b) A record of data - from conception through preservation and sharing,

c) Something that clearly connects data management activities with research project plans?

Answer to the Warm-up Question

What is a data lifecycle?

a) A high-level overview of actions, operations, or processes, started at different stages

b) A record of data - from conception through preservation and sharing

c) Something that clearly connects data management activities with research project plans

The answer – all of them! **The data lifecycle is a logical series of connected steps or processes that illustrates the connection between data management and research projects.**

USGS

Additional Resources

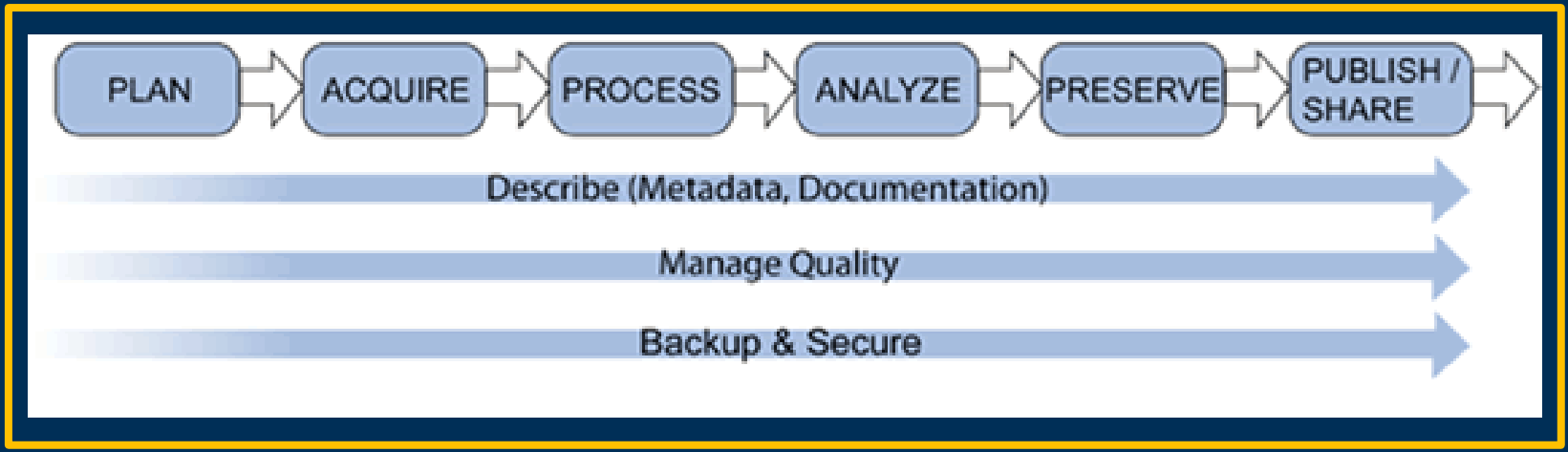
Glossary

A Warm-up Question (Cont.)

The answer is – all of them! The data lifecycle is a high-level, logical series of connected steps or processes from conception through preservation and sharing, that illustrates the connection between data management and research projects.

Science Data Lifecycle

Reference point: where is this module in the lifecycle?



Everywhere!

(Very exciting!)

USGS

Additional Resources

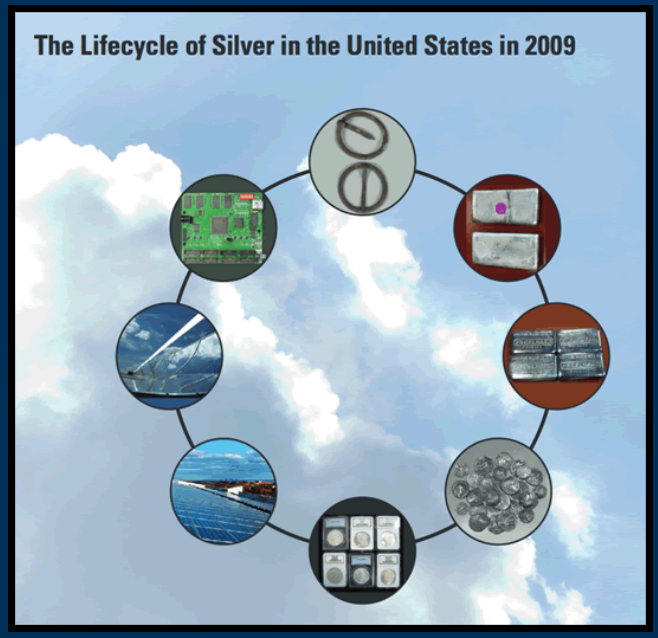
Glossary

Science Data Lifecycle

Let’s check to see where this module fits within the science data lifecycle. Since this lesson deals with the entire science data lifecycle, it fits everywhere! Very exciting!

What is a Lifecycle?

- A lifecycle is a graphical representation of a series of related processes, grouped into topics.



Goonan, T.G., 2014, The lifecycle of silver in the United States in 2009: U.S. Geological Survey Scientific Investigations Report 2013–5178, 17 p., *http://dx.doi.org/10.3133/sir20135178*.

USGS

Additional Resources

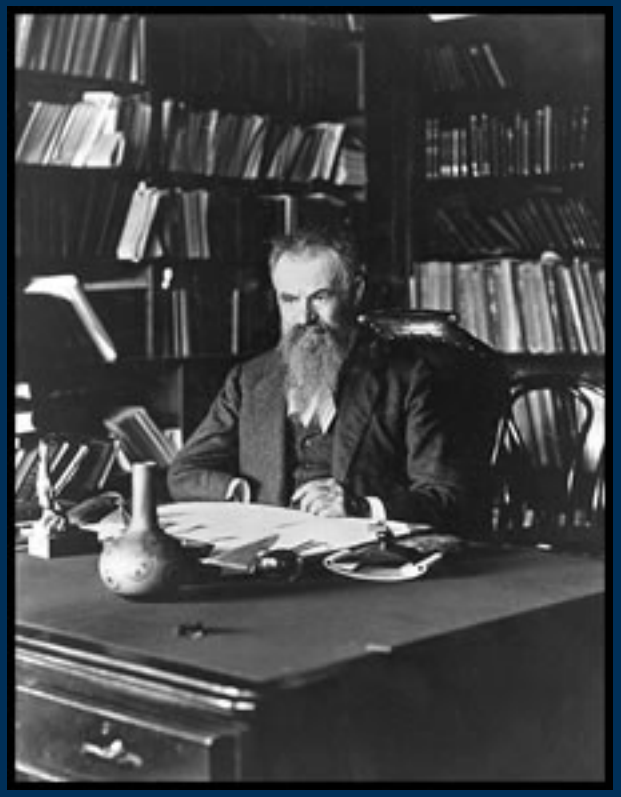
Glossary

What is a Lifecycle?

So what exactly is a lifecycle? A lifecycle is simply a graphical representation of a series of related processes, grouped into topics.

What is a Science Data Lifecycle?

- A science data lifecycle describes the various stages of data management.
  - From start to end – planning
    and acquiring through preserving and sharing.
  - Includes cross-cutting, supporting items like metadata, managing quality, and securing data.



John Wesley Powell

```
▼<metadata>
  ▼<idinfo>
    ▼<citation>
      ▼<citeinfo>
          <origin>Eduardo Patino</origin>
          <origin>Jeff Woods</origin>
          <origin>Mark Zucker</origin>
          <origin>Gordon Anderson</origin>
          <pubdate>2013</pubdate>
          <title>South Florida Hydrology Database</title>
```

USGS

Additional Resources

Glossary

What is a Science Data Lifecycle?

So what is a science data lifecycle? A science data lifecycle describes the various stages of data management from start to end, from planning and acquiring through preserving and sharing.

It also includes cross-cutting, supporting items like metadata, managing quality, and securing data.

Why is a Science Data Lifecycle Useful?

- A Science Data Lifecycle offers a high-level overview of actions, operations, or processes.
  - Individual and Survey-wide.
  - Necessary in handling, documenting, preserving, and providing access to the Bureau's science data.

→ The resulting well-curated data resources, which researchers can re-use, are critical to integrated science and extend the value of the data.

Only some of these vials have labels. What are in the rest?

USGS

Additional Resources

Glossary

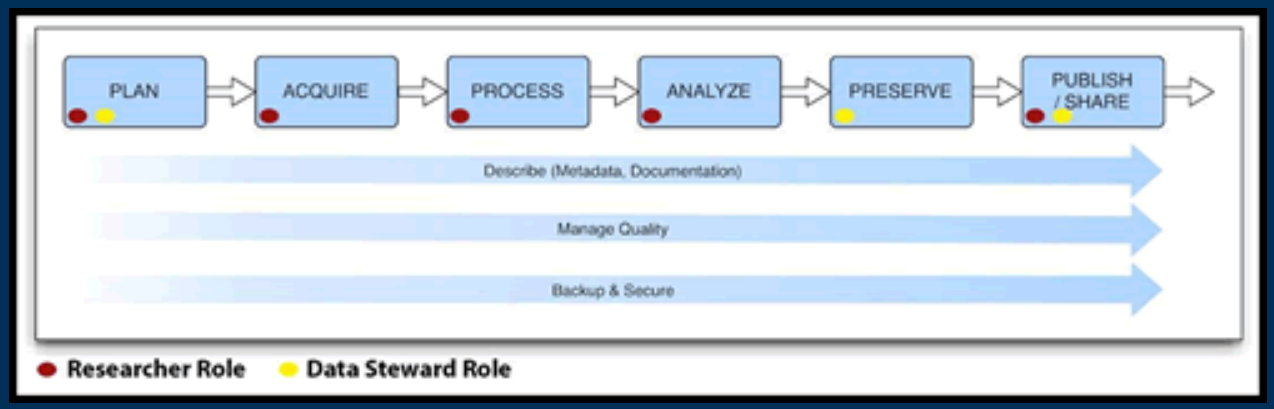
Why is a Science Data Lifecycle Useful?

Why is a science data lifecycle useful? A science data lifecycle offers a high-level overview of actions, operations, or processes for both individuals and Survey-wide. The lifecycle is necessary in handling, documenting, preserving, and providing access to the Bureau's science data.

The resulting well-curated data resources, which researchers can re-use, are critical to integrated science and extend the value of the data.

The picture on the right contains several vials full of something. But only a few of the vials have labels, and many do not. What is in the vials without labels? And are they worth saving?

Why is a Science Data Lifecycle Useful?



- Provides a way for researchers data managers/stewards to see roles/responsibilities.
  - These are expected paths; individual projects may adjust as needed.
- Lead Researcher/PI is not always expected to conduct each role, but are expect to ensure each role is completely accomplished.
- Left-to-right flow of its narrative and a clearly defined starting point that aligns with the inception of the research project.
- Note the placement of an arrow after Publish/Share; indicates the output of the model is used as input for another project.

USGS

Additional Resources

Glossary

Why is a Science Data Lifecycle Useful?

A science data lifecycle also provides a way for researchers and data managers or stewards to easily see roles and responsibilities. By combining the graphic with the identification of roles, it helps clarify responsibilities and help show the researcher where and when to go for help. Note that these are the expected pathways, but individual projects may adjust the roles and responsibilities, even revisiting steps, as needed.

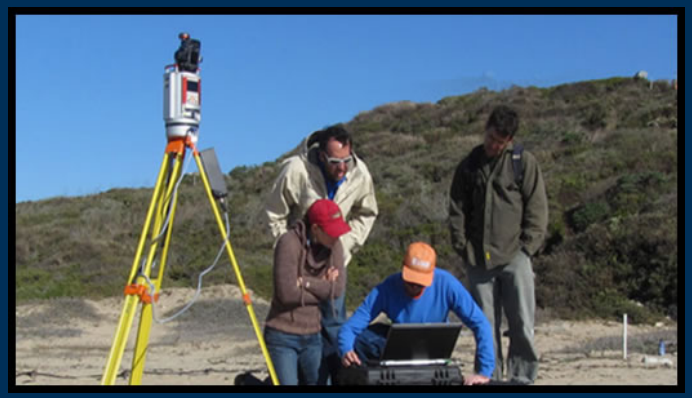
In the graphic, certain steps within the lifecycle are handled by the researcher (or scientist), other steps are handled by the data steward (or data manager), and for other steps both the researcher and steward or manager work together. Also, while the Lead Researcher or PI is not always expected to conduct each role herself/himself, they are expect to ensure each role is completely accomplished.

The left-to-right flow of the graphic's narrative communicates directionality, and a clearly defined starting point aligns with the inception of the research project. Note the placement of an arrow after publish and share; this is to indicate that the output of the model is used as input for another project.

Why is a Science Data Lifecycle Useful?

- Way to facilitate shared recognition and understanding of the necessary steps to:
  - → document,
  - → protect, and
  - → make available

  the Bureau's valued data assets.
- Also serves as a structure to help the USGS evaluate and improve policies and practices for managing scientific data, and to identify areas in which new tools and standards are needed.



USGS

Additional Resources

Glossary

Why is a Science Data Lifecycle Useful? (Cont.)

Another way a science data lifecycle is useful is that it is a way to facilitate shared recognition and understanding of the necessary steps to document, protect, and make available the Bureau's valued data assets. It also serves as a structure to help the USGS evaluate and improve policies and practices for managing scientific data, and to identify areas in which new tools and standards are needed.

How can a SDL be Useful to you? Some Examples…

- Ensures data is available to you—and others—after the project has ended.
- Helps manage data.
- Makes accessing and using the data more efficient.
- Helps document data.

Date	Water Level (feet NAVD88)
2013-03-10	9.02
2013-03-11	9.01
2013-03-12	9
2013-03-13	8.99
2013-03-14	8.97
2013-03-15	8.95
2013-03-16	8.93
2013-03-17	8.91
2013-03-18	8.9
2013-03-19	8.99

- Provides for a better understanding of the data – limitations, appropriateness.
- Helps reduce duplication of data collection.

How can a SDL be Useful to you?

Here are some examples of how the science data lifecycle can be useful to you—you may have other examples in addition to these:

- Ensures data is available to you—and others—after the project has ended.

- Helps manage data from the start of the project through its conclusion.

- Makes accessing and using the data more efficient.

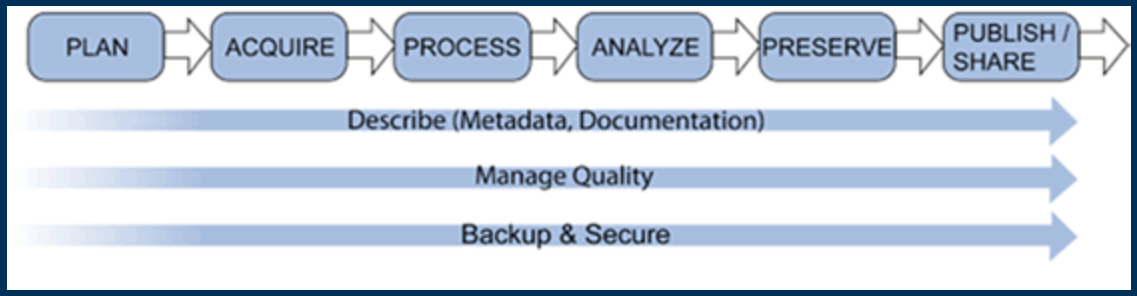
- Helps document data completely.

- Provides for a better understanding of the data – limitations, appropriateness.

- Helps reduce duplication of data collection.

What are the Elements of the USGS SDL?

- Plan
- Acquire
- Process
- Analyze
- Preserve
- Publish/Share



Including the cross-cutting elements of:

- Describe - Manage Quality - Backup & Secure

USGS

Additional Resources

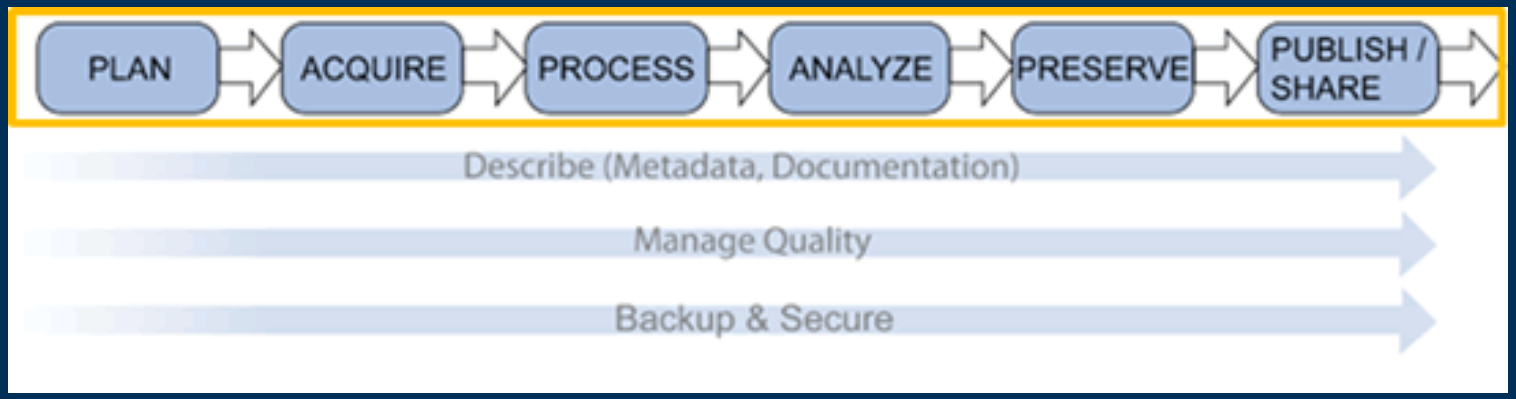
Glossary

What are the Elements of the USGS SDL?

Let's talk about the individual elements of the USGS science data lifecycle. Across the top of the graphic is a series of elements that follow one another: Plan, Acquire, Process, Analyze, Preserve, and Publish/Share. On the bottom, there are three cross-cutting elements: Describe (including metadata and documentation), Manage Quality, and Backup and Secure. Why are there two sets of elements? Let's find out….

Why are Elements Displayed Differently?

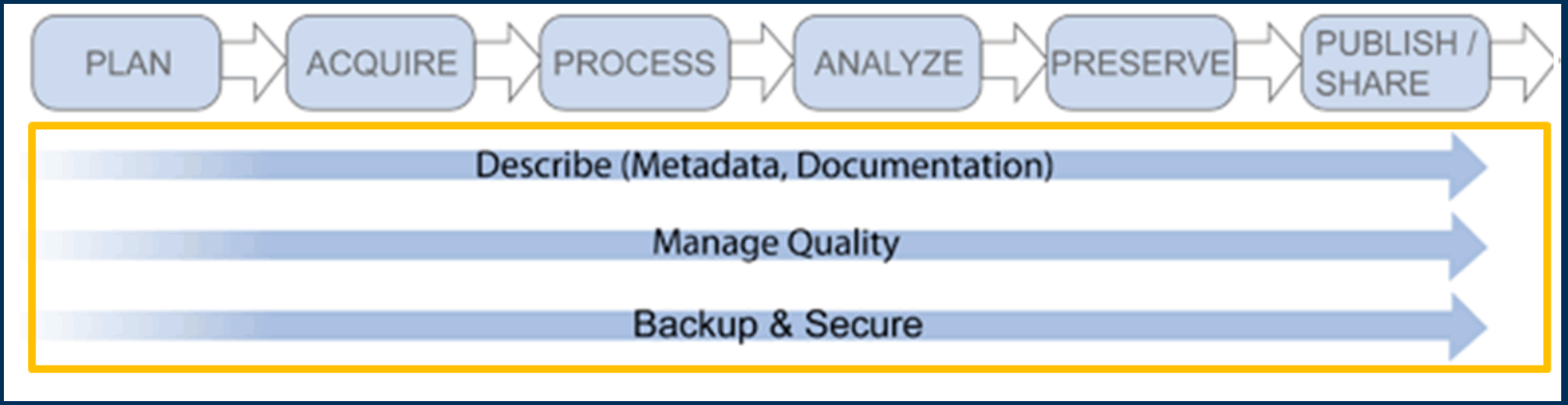
Each of the primary elements of the Model addresses discrete activities and outputs unique to that stage.



Why are Elements Displayed Differently?

Why are the science data lifecycle elements displayed differently? Each of the primary elements, the ones that are listed at the top of the graphic, addresses discrete activities and outputs unique to that stage. Using the elements listed helps USGS scientists understand and manage for the lifecycle of data and information products and preserved for access and use beyond the life of research projects.

Why are Elements Displayed Differently?



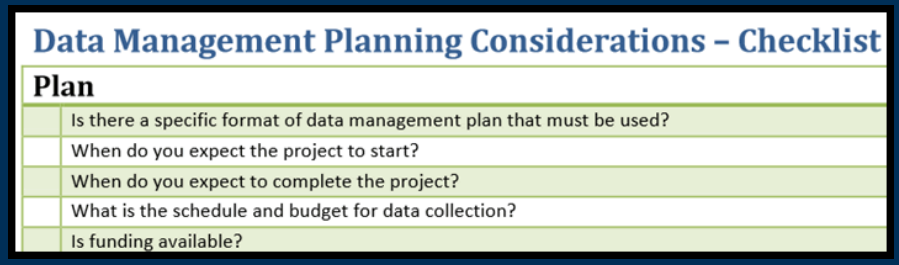
Other critical activities must be performed continually across all stages of the lifecycle to help support effective data management.

Why are Elements Displayed Differently? (Cont.)

What about the elements displayed on the bottom? These cross-cutting items are critical activities that must be performed continually across all stages of the science data lifecycle. Many of these activities are left until the end of the project, if performed at all. When that happens, the documentation is incomplete and not as accurate, mistakes can be found too late, and data permanently lost because proper backups were not performed along the way. Effective data management requires documentation, quality control, and ensuring your data is backed up and secured to limit data loss. It also means a better data product in the end.

Plan: 1st Model Element

- Help scientists consider all activities related to the project's data assets
- During this stage:
  - All elements of the Model should be evaluated, addressed, and documented.
  - Consider approaches, needed resources (including funding and personnel), and intended outputs.



Data Management Planning Considerations – Checklist	
Plan	
	Is there a specific format of data management plan that must be used?
	When do you expect the project to start?
	When do you expect to complete the project?
	What is the schedule and budget for data collection?
	Is funding available?

→ *A data management plan is the recommended output of this element of the Model.*

USGS

Additional Resources

Glossary

Plan: 1st Model Element

Plan, the first Model element, is intended to assist scientists in assuring consideration of all activities related to the handling of the project's data assets, from project inception to publication and archiving. During this stage, all elements of the Model should be evaluated, addressed, and documented. The project team should consider approaches, needed resources (including funding and personnel), and intended outputs for each stage of the data lifecycle. A data management plan is the recommended output of this element of the Model.

Acquire: 2nd Model Element

- Activities through which new or existing data are collected, generated, or considered and evaluated for re-use.
- Emphasizes the importance of considering relevant USGS policies and best practices that maintain the provenance and integrity of the data as a USGS information product:
  - Provenance is defined as the complete, chronological history of ownership and modifications.

→ The outputs of this element are the project's data inputs.

USGS

Additional Resources

Glossary

Acquire: 2nd Model Element

Acquire, the second Model element, represents the activities through which new or existing data are collected, generated, or considered and evaluated for re-use. Stream gage data, historical maps, seismology motion sensor outputs, biological records, and satellite observations are examples of acquired data and information that represent the diverse and robust variety of science data inputs to USGS research. Scientists are skillful in designing data acquisition techniques to address research questions; in the USGS context, this element emphasizes the importance of considering relevant USGS policies and best practices that maintain the provenance and integrity of the data as a USGS information product. Provenance is defined as the complete, chronological history of ownership and modifications

The outputs of this element are the project's data inputs.

Process: 3rd Model Element

- Activities associated with preparation of new or previously collected data inputs.
- Reminds scientists that USGS standards and tools are available that can meet project requirements while also building a Bureau-wide foundation of data for integrated science.

→ *The outputs of this element are datasets that are ready for integration and analysis.*

USGS

Additional Resources

Glossary

Process: 3rd Model Element

Process, the third Model element, represents various activities associated with preparation of new or previously collected data inputs. Processing of input data may entail definition of data elements; integration of disparate datasets; extraction, transformation, and load operations; and application of calibrations to prepare the data for analysis. The Process element in the Model reminds scientists that USGS standards and tools are available that can meet project requirements while also building a Bureau-wide foundation of data for integrated science. The outputs of this element are datasets that are ready for integration and analysis.

Analyze: 4th Model Element

- Exploration and interpretation of processed data, where hypotheses are tested, discoveries made, conclusions drawn
- New data are generated, versions are tracked, and processes are documented
- Data management during analysis improves efficiency of data analysis activities, preserves documentation that is critical for scientific integrity, and creates a foundation for future research

→ *The outputs of this element are interpretations or new datasets, which often are published in written reports or machine-readable formats such as map layers or numerical modeling results.*

USGS

Additional Resources

Glossary

Analyze: 4th Model Element

Analyze, the fourth Model element, represents the activities associated with the exploration and interpretation of processed data, where hypotheses are tested, discoveries are made, and conclusions are drawn. Analytical activities include summarization, graphing, statistical analysis, spatial analysis, and modeling, and are used to produce scientific results and information. In this element, new data are generated, versions are tracked, and processes are documented. Data management during analysis improves the efficiency of data analysis activities, preserves documentation that is critical for scientific integrity, and creates a foundation for future research. The outputs of this element are interpretations or new datasets, which often are published in written reports or machine-readable formats such as map layers or numerical modeling results.

Preserve: 5th Model Element

- Activities associated with storing data for long-term use and accessibility:
  - Long term: (beyond the life of the project).
- Preservation often is not considered until the end stage of a project, when it might be neglected because of the pressure of project budgets, timetables.
  - Data preservation agreements to ensure availability of data.

USGS

Additional Resources

Glossary

Preserve: 5th Model Element

Preserve, the fifth Model element, represents the activities associated with storing data for long-term- well beyond the life of the project - use and accessibility. Preservation often is not considered until the end stage of a project, when it might be neglected because of the pressure of project budgets and timetables. It is recommended that during this step, agreements are made with an operational unit of the USGS to preserve the data beyond the life of the project.

Preserve: 5th Model Element

- Deliberately placed ahead of Publish/Share in the Model; reminder that Federally funded scientists must plan for:
  - long-term preservation of data, metadata, ancillary products, application-neutral storage formats, and any additional documentation.
- USGS policy requires preserving scientific data and information developed by the Bureau's information and research programs.

- All scientific data produced as a result of USGS funding must be preserved.

→ *Preservation ensures availability and re-use.*

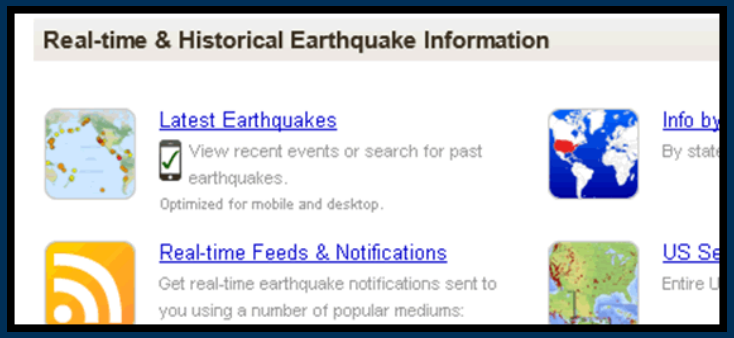
Preserve: 5th Model Element (Cont.)

The deliberate placement of this element ahead of Publish/Share in the Model is a reminder that federally funded scientists must plan for the long-term preservation of data, metadata, ancillary products, application-neutral storage formats, and any additional documentation, to ensure availability and re-use.

USGS policy requires preserving scientific data and information developed by the Bureau's information and research programs. All scientific data produced as a result of USGS funding must be preserved.

Publish/Share: 6th Model Element

- Combines the Bureau's concepts of traditional peer-reviewed publication with the distribution of data through Web sites, data catalogs, social media, and other venues.
  - Publication and dissemination of USGS data and information are critical components of the USGS Mission.
- Focal point of recent Federal directives is to **increase access** to the results of federally funded research.



→ *Reminds scientists that data, as well as traditional publications, are research products.*

USGS

Additional Resources

Glossary

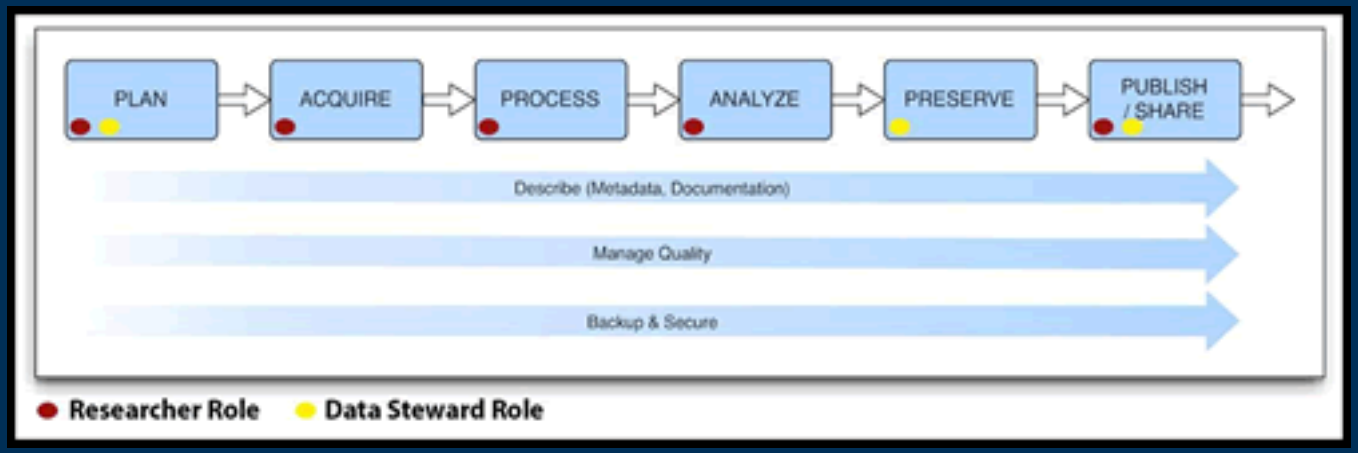
Publish/Share: 6th Model Element

Publish/Share, the sixth element in the Model, combines the Bureau's concepts of traditional peer-reviewed publication with the distribution of data through Web sites, data catalogs, social media, and other venues. Publication and dissemination of USGS data and information are critical components of the USGS Mission, and are a focal point of recent Federal directives to increase access to the results of federally funded research. This element reminds scientists that data, as well as traditional publications, are research products.

Roles and Responsibilities

- Data management activities require specialized knowledge and skills, including methods and standards.
- The Model recognizes two different roles:

Researchers and Data Stewards



USGS

Additional Resources

Glossary

Roles and Responsibilities

The USGS Science Data Lifecycle Model includes data management activities that require specialized knowledge and skills, as well as ongoing education about methods and standards. The Model encourages researchers to plan project teams that recognize two different roles: researchers and data stewards.

Roles and Responsibilities

- Researchers acquire, process, and analyze data and information and are responsible for publishing and sharing it.
- Data stewards work alongside and support researchers/scientists.
  - Manage another's data or information to ensure that they can be reused and accessed.
  - Ensure requirements are met, and data documentation is developed and maintained.
  - Background in data management.

Roles and Responsibilities (Cont.)

Researchers collect, process, and analyze data and information and are responsible for publishing and sharing it.

Data stewards work along side and support the researchers and scientists. They manage another's data or information to ensure that they can be used to draw conclusions or make decisions. They also ensure official agency records requirements are met, and data documentation is developed and maintained. Data stewards often have a background in data management.

Roles and Responsibilities

- Researcher and data steward roles may fall to one person for one or more lifecycle stages or may be divided among multiple individuals with discrete responsibilities.
  - Multiple personnel may oversee the various data lifecycle elements.
- Reminder: Lead Scientist/PI is responsible for ensuring that each element is addressed throughout the project.

USGS

Additional Resources

Glossary

Roles and Responsibilities (Cont.)

Depending on staff expertise, the researcher and data steward roles may fall to one person for one or more lifecycle stages or may be divided among multiple individuals with discrete responsibilities. Although multiple personnel may oversee the various data lifecycle elements, the lead scientist or PI is responsible for ensuring that each element is addressed throughout the life of the project.

Summary/Recap

- The science data lifecycle is a high-level view of data.
  - Guides data management from conception through preservation and sharing.
- Illustrates how data management activities relate to project workflows.
- Assists with understanding the expectations of proper data management.

USGS

Additional Resources

Glossary

Summary/Recap

Let’s recap what we have learned in this module:

The science data lifecycle gives a high-level view of data, and guides data management from conception through preservation and sharing. It also illustrates how data management activities are related to project workflows. Lastly, it assists with understanding the roles and expectations of proper data management.

Summary/Recap

- There are six elements to the USGS Science Data Lifecycle: Plan, Acquire, Process, Analyze, Preserve, and Publish/Share.
  - They addresses discrete activities and outputs unique to that stage.
- There are three cross-cutting activities: Document, Manage Quality, Backup and Secure.
  - These steps are performed continually across all stages of the lifecycle to help support effective data management.

Summary/Recap (Cont.)

There are six elements to the USGS Science Data Lifecycle: Plan, Acquire, Process, Analyze, Preserve, and Publish/Share. They addresses discrete activities and outputs unique to that stage.

There are three cross-cutting activities: Document, Manage Quality, Backup and Secure. These steps are performed continually across all stages of the lifecycle to help support effective data management.

Summary/Recap

- By applying the model to research, USGS scientists can ensure data products will be:
  - Well-described,
  - Preserved,
  - Accessible, and
  - Fit for re-use.

USGS

Additional Resources

Glossary

Summary/Recap (Cont.)

In applying the Model to research activities, USGS scientists can ensure that data products will be well-described, preserved, accessible, and fit for re-use.