USGS
science for a changing world

USGS Data Management Training Modules:

Value of Data Management

"*Data is a precious thing and will last longer than the systems themselves.*"
-- Tim Berners-Lee

…. But only if that data is properly managed!

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

Version 1, October 2013

Narration: Introduction Slide 1

USGS Data Management Training Modules – the Value of Data Management

Welcome to the USGS Data Management Training Modules, a three part training series that will guide you in understanding and practicing good data management. Today we will discuss the Value of Data Management.

As Tim Berners-Lee said, “Data is a precious thing and will last longer than the systems themselves.” ……We will illustrate here that the validity of that statement depends on proper management of that data!

Module Objectives

- Describe the various roles and responsibilities of data management.
- Explain how data management relates to everyday work.
- Emphasize the value of data management.

From Flickr by cybrarian77

USGS

Additional Resources

Glossary

Narration: Slide 2

Module Objectives

In this module, you will learn how to:

1. Describe the various roles and responsibilities of data management.
2. Explain how data management relates to everyday work and the greater good.
3. Motivate (with examples) why data management is valuable.

These basic lessons will provide the foundation for understanding why good data management is worth pursuing.

Terms and Definitions

- **Data Management (DM)** – the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.[1]
- **Data Stewardship** – taking responsibility for a set of data for the well being of the larger organization, and operating in service to, rather than in control of, those around us.[2]
- **Metadata** – information that describes a dataset, such that a dataset can be understood, re-used, and integrated with other datasets.[2]

[1] DAMA Data Management Body of Knowledge (DAMA-DMBOK)
[2] USGS Data Management Website

Terms and Definitions

Before we get started, we want to define a few key terms.

Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets. Data management involves the entire lifecycle of a project or data set – including data management plans, preservation, metadata and documentation, securing, and sharing.

Data stewardship involves taking responsibility for a set of data for the well being of the larger organization, and operating in service to, rather than in control of, those around us. In some organizations, there are specific roles and responsibilities assigned to staff acting as stewards who manage data or information to ensure that the data can be used to draw conclusions or make decisions.

Lastly, metadata is simply data about data. It is information that describes a dataset, such that a dataset can be understood, reused, and integrated with other datasets. Metadata describes the who, what, where and when.

For more information on these terms, and others found in this presentation, visit the USGS Data Management website at http://www2.usgs.gov/datamanagement.

Data Realities

Photos from USGS

We create massive amounts of data

- Data is currently collected in many ways: from sensors, sensor networks, remote sensing, observations, and more - calls for increased attention to DM and stewardship.

Data loss happens frequently and without warning.

- Natural disasters, hardware failures, human error…

Poor data management affects everyone!

USGS

Additional Resources

Glossary

Data Realities

Data is being generated in massive quantities daily. Improvements in technology enable higher precision and coverage in data acquisition, requiring higher capacity systems to store and migrate the data. This increases the importance of managing, integrating, and re-using data.

Creating more data comes with a harsh reality – there is more data that can be lost. Data loss happens frequently, often without warning, and can happen due to a variety of factors, such as natural disasters (like floods), hardware failures (the death of a disk), or human error (accidentally deleting the wrong files).

Poor data management doesn't just affect the person responsible for the managing the data – it affects everyone. For example, an office may have a collection of sediment cores, but they have been stored without any documentation – which is an example of bad data management. The cores are almost useless without the associated documentation which would be needed by scientists in order to use the cores for research. In this case, bad data management affects the person who collected the cores – without being able to use them, the time, effort, and funding that went into the collection of the cores is essentially lost. Another scientist who might have been able to use the cores in another project will now have to go out and recollect the samples - wasting time and money on the duplicative effort. And the manager who funds the research project may not be inclined to provide additional funding in the future because of the loss of efficiency and money.

Data Management Roles

Who is involved in data management, what are their concerns, and what roles do they have?

- Scientists
- Data Stewards
- Managers

Photo from USGS

USGS

Additional Resources

Glossary

Narration: Slide 5

Data Management Roles

So who is responsible for doing data management? Why should they care and what roles do they have?

While IT and others do play very important support roles, for this lesson, we will focus on three main groups that affect and are affected by data management: the scientists, data stewards, and managers.

Scientists

Photo from USGS

- Often busy with their own research.
- Feel they don't have the time or background to do DM.
- "I'll get to it later" – but later never happens.
- Have concerns about releasing data and making it available for access/discovery:
  - Misuse of data ("Someone will misunderstand my results or use them incorrectly..").
  - Getting scooped ("Someone will scoop my research … I want to publish it first …").

USGS

Additional Resources

Glossary

Narration: Slide 6

Scientists

Our first group is the scientists.

- Often they are busy with their own research and running science projects.
- They may feel that they don't have the time or proper background to do data management.
- They may intend on getting to things like metadata and documentation towards the end of a project, but frequently it doesn't get done.
- They may have concerns about releasing data and making it available.
- They may not want to provide access to their data for fear of someone misusing their data or another scientist publishing results and findings before they have had a chance.

Scientists (cont.)

What role do they play?

- Know the data best, so they are the best people to document it:
  - Metadata, DM plans, file names, database fields.
- Want to know what data is available and would like it documented:
  - If you keep your data managed, you and others will be able to find it.

Photo from USGS

USGS

Additional Resources

Glossary

Narration: Slide 7

Scientists (cont.)

Since scientists generate data, and because data needs to be managed properly, what role do scientists play in data management? First, because they have either created the data or work with the data, they would know the data best – and so they are the best people to provide accurate documentation for the data (creating the metadata). They also will have the best idea of the lifecycle of the data, and are the logical choice for outlining a data management plan. And because they are working with the data on a regular basis, they more so than anyone else would know what would be logical file names, database field names, etc.

When scientists are doing research, frequently they assess what data is already out there that they might be able to use. If the data has been well-managed, and has been documented and made available to others, scientists can reuse the data. In addition, if you keep your own data well-managed, not only will others be able to find it, but so will you!

Data Stewards

Data Stewards are those business subject matter experts, who manage another's facts or information to ensure that they can be used to draw conclusions or make decisions[1]**:**

- May have some knowledge of DM, but maybe have learned bad habits, have outdated knowledge, or a limited background.
- Need more training, but no funds or no one to ask for help.

Data Stewards help support scientists and practice good data management.

[1] USGS Data Management Website

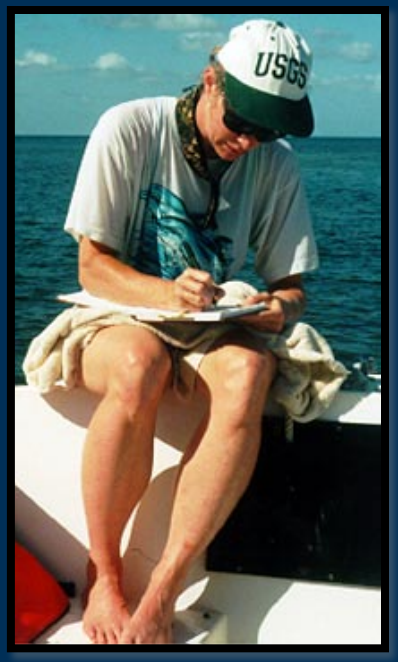


Photo from USGS

USGS

Additional Resources

Glossary

Data Stewards

Data Stewards are those business subject matter experts, who manage another's facts or information to ensure that they can be used to draw conclusions or make decisions.

They generally have some knowledge of data management, but maybe have learned bad habits, have outdated knowledge, or a limited background.

Data stewards may need more training, but there may not be any funds or resources to ask for help.

Still Data stewards have an important role to play to help support scientists and practice good data management.

Managers

- Responsible for projects/programs.
- Need to answer data calls, respond to compliance policies
  (i.e. OMB/OSTP memos).
- Find ways to save money and create efficiencies.
- Primary objective is supporting/funding science; DM comes later, if at all.

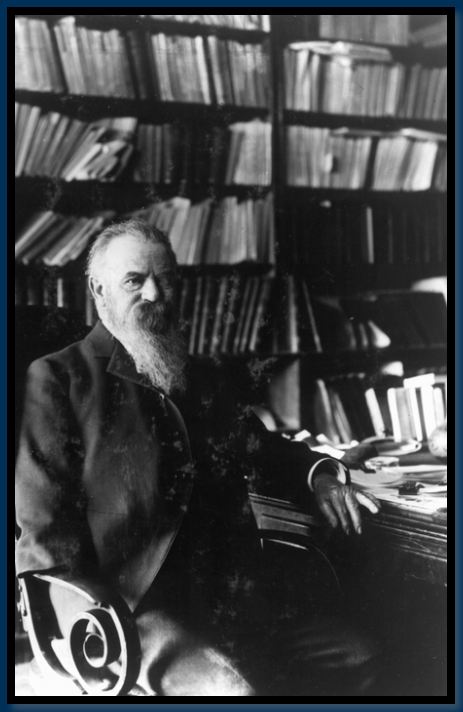


Photo from USGS

Role to ensure DM practiced at all levels of the organization, right people available for support.

USGS

Additional Resources

Glossary

Managers

Lastly, managers also need to understand data management. They may not perform data management in the same way that scientists or data stewards do, but they too can benefit from good practices. Descriptive file names, a best practice of data management, will help managers quickly find the file they need. Ensuring data is backed up, another data management best practice, means limiting or even eliminating data loss.

But data management isn't just good practice – it's often required. Recently, several new policies and memos, including those from the Office of Science Technology and Policy (OSTP) and the Office of Management and Budget (OMB), have begun to require government agencies to provide open access to their data. Open access can only be accomplished through good data management.

A major function of a manager's job is to find ways to save money and create efficiencies. As noted in our previous example of the sediment cores, bad data management can cause loss of time, money, and effort.

A manager's main focus is supporting and funding science, but finding funding for and supporting data management is often something that may come later – if at all.

So what role can managers play in data management, apart from good data management habits? They can ensure that data management is practiced by all levels in their organization and also provide data management support where needed. Often there is little or no funding available in a project's budget for data management, but the USGS has resources out there that can help – groups like the Data Management Working Group and the Science Data Coordinator Network, in addition to the Data Management website.

Why Should you Care?

- **Data are valuable assets:**
  - If DM done right in the beginning, more time and money become available for future activities.
  - Data are expensive and time consuming to collect, and some results can't be reproduced.
  - DM saves time and therefore MONEY!
- **Research needs to be transparent:**
  - Transparent science to enable your data to be reproduced to confirm your findings.
  - Data can be called into question in court or by other scientists. Must be able to defend and explain the data and the methods you used.

"Eighty percent of a scientist's effort is spent discovering, acquiring, documenting, transforming, and integrating data, whereas only 20 percent of the effort is devoted to more intellectually stimulating pursuits such as analysis, visualization, and making new discoveries." - Bill Mitchner of DataONE

USGS

Additional Resources

Glossary

Why Should you Care?

Still need to be convinced about why you should care?

Data are valuable assets. Scientists can do more science and less DM if it is done properly in the beginning, both for finding/working with their own data and data from others. Data are often expensive and time-consuming to collect, and some results can't be reproduced (such as biological observations). Good data management saves time and therefore money.

Research also needs to be transparent. Metadata and documentation may not be exciting, but research needs to be transparent for a variety of reasons. To be published and peer reviewed, data must be transparent to enable your data to be reproduced to confirm your findings. In some cases, the results of scientists have been called into question, either in court or by other scientists, and they must be able to defend and explain the data and methods used. This is especially important for federal agency scientists who may have to testify during litigation procedures.

A quote from Bill Michener of DataONE states, "Eighty percent of a scientist's effort is spent discovering, acquiring, documenting, transforming, and integrating data, whereas only 20 percent of the effort is devoted to more intellectually stimulating pursuits such as analysis, visualization, and making new discoveries."

We need to provide that 80% of our time with good data management practices.

Why Should you Care? (cont.)

- **It's required:**
  - Federal employees are subject to open data policies (OSTP, OMB, Information Quality Act, FOIA) and security requirements (Deepwater Horizon Oil Spill in the Gulf of Mexico).
  - Many funding agencies now require some level of data management (USGS Powell Center, NSF, publishers).
- **You're part of a bigger picture:**
  - There is a bigger, more global perspective – requires big data.
  - Data-calls require good DM to find, understand, and integrate data.
  - Data can be re-used providing far more value and may lead to more co-authorship and credit.
  - Regardless of the archaic "Publish or Perish" mentality, your data are worth as much (or more) than the resulting publication.

Why Should you Care? (cont.)

If that isn't enough reason, there might be requirements to manage the data. Federal employees and those receiving federal funding are subject to the "Open Data" policies and laws such as the OSTP and OMB memos mentioned previously, and the Information Quality Act and FOIA. Disasters such as the Deepwater Horizon oil spill require that we can quickly access data from many sources so that we can respond in a timely manner. We are also required to have the data and information related to the oil spill be properly archived and documented – key components of data management. In addition, many funding sources, such as the USGS Powell Center and the National Science Foundation, now require some level of data management planning. Peer-reviewed journals are also moving toward requirements for the raw data to be submitted alongside the article.

Finally, data management is important because you and your data may be part of something bigger than the research project that collected the data. Regional, national, and worldwide scientific challenges, like climate change, require massive amounts of data to be integrated. This requires that the data be managed correctly and common standards to be in place to integrate the data seamlessly.

Data can also be reused in ways the original data creator never imagined. For example, George Washington was not known for his survey data but it is still being used and built upon today. And each time your data is reused, you receive credit and citations.

Regardless of the archaic "Publish or Perish" mentality, your data are worth as much (or more) than the resulting publication. With good data management, the data can be re-used for many publications to come, long after the scientist retires.

Use Case: Whirling Disease

- Whirling Disease
- Rock Creek Project

Photo by Dr. Thomas L. Wellborn

Photo by M.E. Markiw

Photo by USGS Western Fisheries Science Center

USGS

Additional Resources

Glossary

Narration: Slide 12 -- Contains video of Janice Gordon talking

Use Case: Whirling Disease

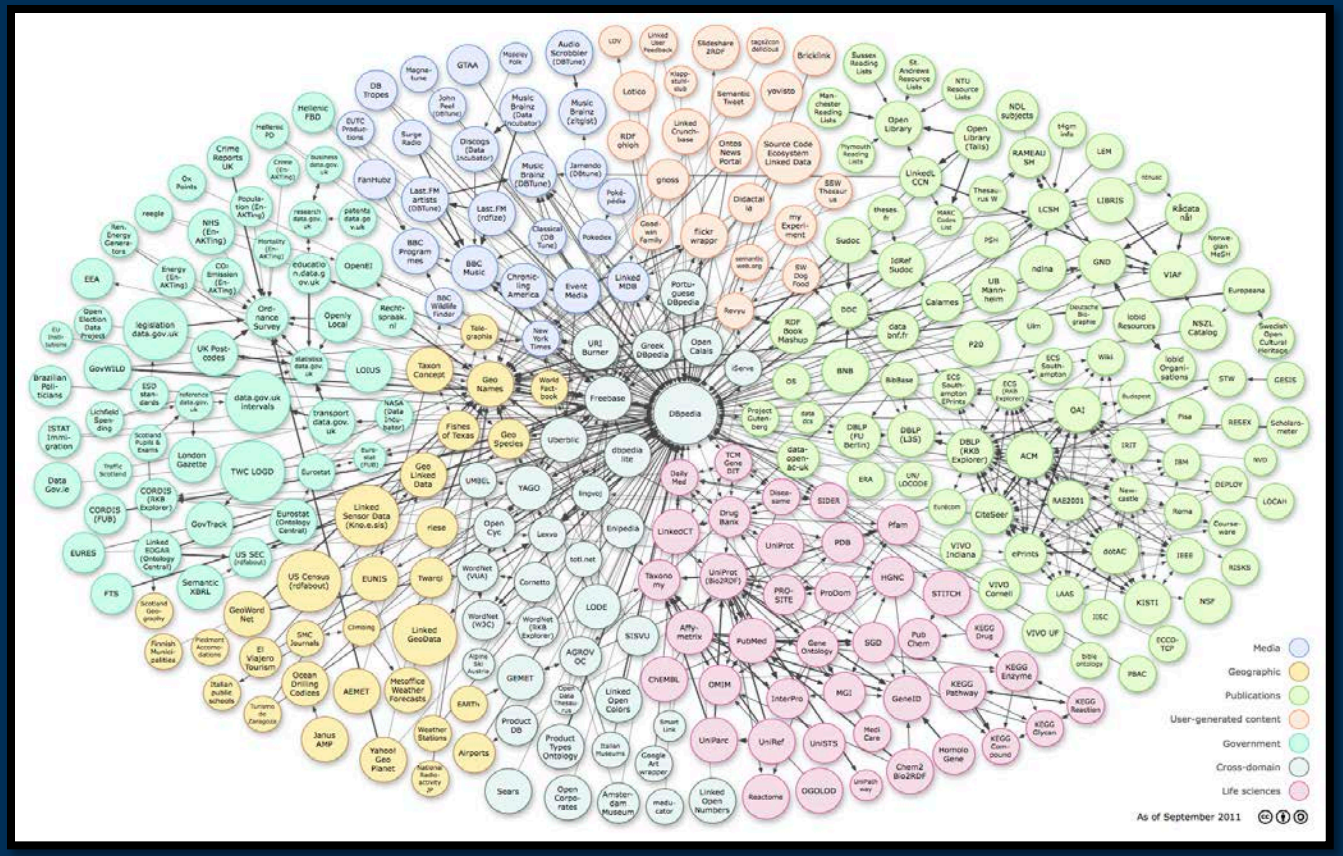
Hi, my name is Janice Gordon. I work for the USGS Core Science and Analytics and Synthesis Program.

Whirling disease is a parasite that inflicts salmonoid fish and has contributed to declines in trout populations across the country. The rock creek project conducted a study of the disease organism in several tributaries of Rock Creek from 1998 through 2003 with funding provided through a partnership with USGS. The data from the study were housed in a data repository hosted at a university.

I wanted to republish the data from the study before it was taken offline due to funding cuts.

Use Case: Whirling Disease

- **Republish in Linked Open Data (LOD)**



"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch"

USGS

Additional Resources

Glossary

Narration: Slide 13 -- Contains video of Janice Gordon talking

I wanted to make the dataset available online as Linked Open Data as part of a semantic web technology use case.

This involved converting data within excel spreadsheets into the standardized Resource Description Framework or RDF data model.

Challenges…

- **Data/Metadata Unavailable**

From Flickr by Sybren A. Stuvel

Narration: Slide 14 -- Contains video of Janice Gordon talking

The university shut down the data repository portal, no longer making the data available. Luckily, I was able to obtain a copy of the archived data and metadata from the university directly.

Unclear Site Names

Year	Lower Middle Fork	Upper East Fork (at bridge)	Bohrnsens's Bridge
1998	--	E. Fk Rock Cr.- RCS4	--
1999	--	RCS-4-99 (EAST FK. ROCK CR.)	--
2000	RCT-3-00(Mid FK. Rock Cr. Lower)	RCT-2-00(E.Fk. Rock Cr)	RCR-6-00(Boreson Ranch)
2001	RCT-3-01(Middle Fk Rock Cr.)	RCT-2-01(E. Fk Rock Cr.)	RCS-6-01(Boreson's Bridge)
2002	?	EFK02-2(FS rd 5106 x-ing)	RCS02-3(Bohrnson's brdg)
2003	RCT03-11 M. Fk. 1(org. lower), FS rd. #70 brdg x-ing	RCT03-13 E. Fk. 3, org. site,@ USFS rd. 5106 x-ing	RCS03-8 Bohrnsen's Ranch

USGS

Additional Resources

Glossary

Narration: Slide 15 -- Contains video of Janice Gordon talking

When I began looking at data from the project, I had difficulty determining the site location names where the samples were taken.

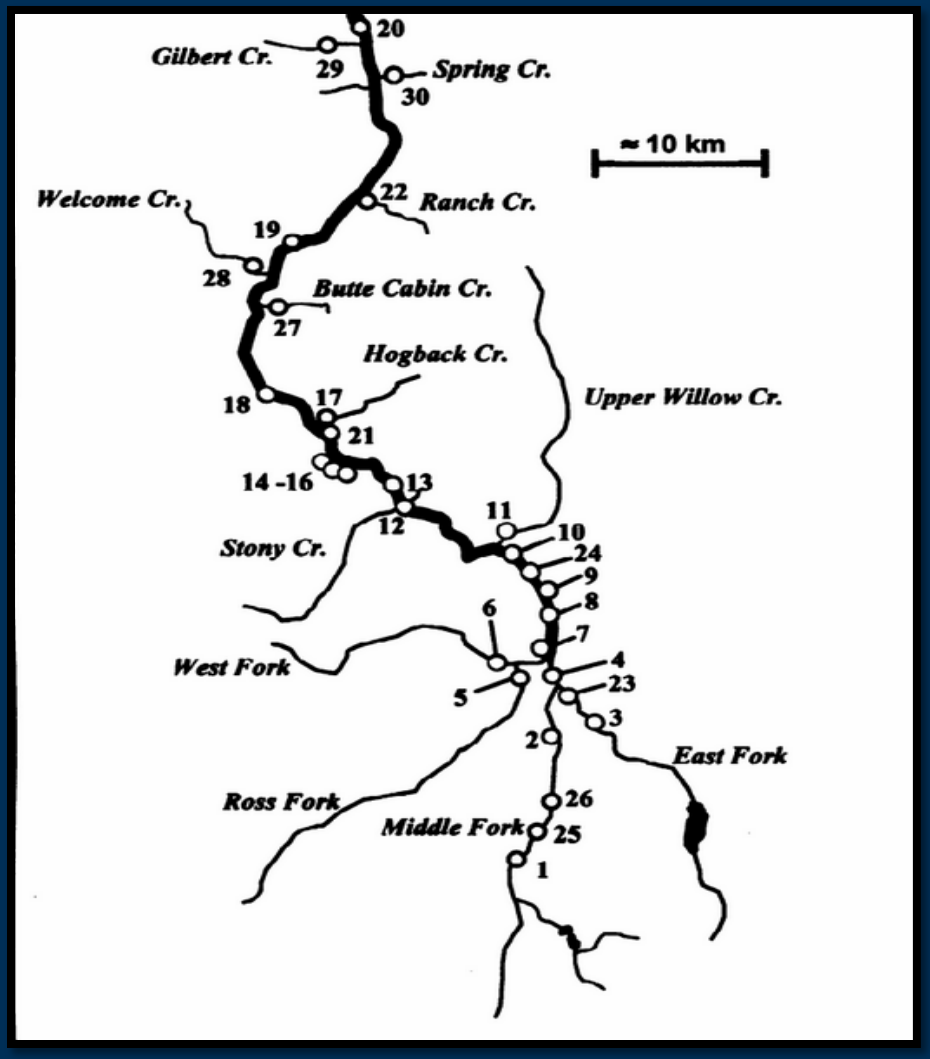
There were 30 sampling sites defined in the publication associated with the project data, however the site names in the dataset were not consistent.

As you can see from this example from 3 sites, the labeling changed from year to year. In the Upper East Forks site, the labels changed from abbreviations of East Fork to RCS, RDT back to EFK. Another example had misspelled names such as Boreson Ranch, then Borhnson's bridge.

The site names seemed to have codes, however there was no data dictionary available to explain the meaning. There were also no notes in the spreadsheet explaining which sites had and had not been sampled each year. It was nearly impossible to resolve which creek they were referring to from the spreadsheets themselves.

This chart shows my assumptions based on the information in the spreadsheets and the publication, but I can't be sure these names are correct.

No Exact GPS Locations



Drawing from Rock Creek Project

USGS

Additional Resources

Glossary

Narration: Slide 16 -- Contains video of Janice Gordon talking

The dataset also did not include GPS coordinates for each site. The publication only had a vague black and white drawing of the site locations.

I tried to find the authors of the paper but the contact person had left his position. I did get in touch with a research technician who was involved with the project, but she didn't have any additional details she could provide me.

I tried to make rational assumptions of which sites the data referred to, by process of elimination and comparisons with the paper. The FGDC metadata record didn't have detailed site information so that didn't help me either.

I documented my assumptions but in the end I gave up using the dataset for my use case because there were too many unknowns in order to validly re-use and republish these study data.

Data Should be Managed to:

- Maximize the effective use and value of data and information assets.
- Continually improve data quality including: data accuracy, integrity, integration, timeliness of data capture and presentation, relevance and usefulness.
- Ensure appropriate use of data. and information
- Facilitate data sharing.
- Ensure sustainability and accessibility in long-term for re-use in science.

Photo from USGS

USGS

Additional Resources

Glossary

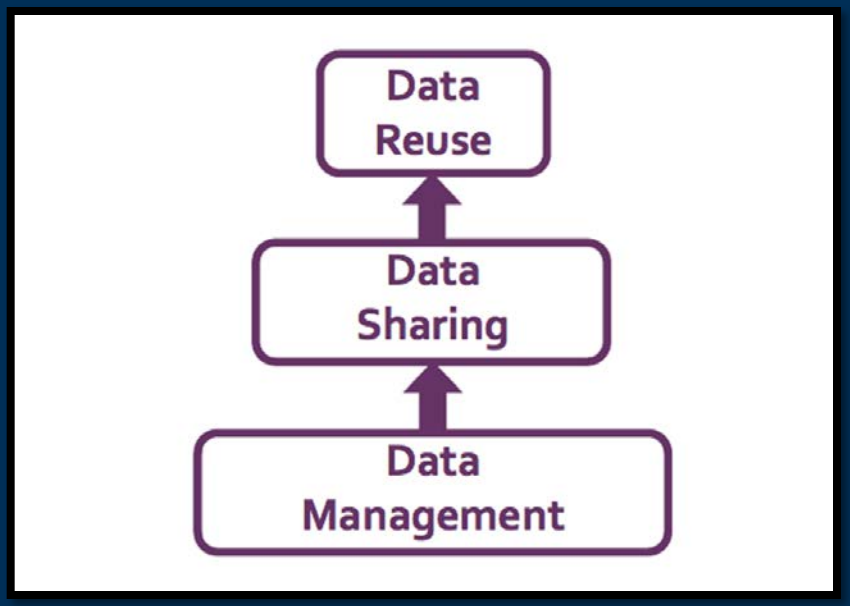
For everyone involved in the data's lifecycle, data should be managed to:

- Maximize the effective use and value of data and information assets.
- Continually improve the quality including: data accuracy, integrity, integration, timeliness of data capture and presentation, relevance and usefulness.
- Ensure appropriate use of data and information.
- Facilitate data sharing.
- Ensure sustainability and accessibility in long term for re-use in science.

These principles enable the data to continue to be useful and relevant for the long term.

The Bottom Line

- Data Management makes good sense:
  - Saves time, money, effort by enabling transparency, reproducibility, knowledge transfer, preservation, and access.
  - Facilitates data sharing, archiving, and publishing data.
  - Enables the ability to reuse, reproduce, and integrate data.



Narration: Slide 18

The Bottom Line

The bottom line is that data management makes good sense.

It saves time, money, and effort by enabling transparency, reproducibility, knowledge transfer, preservation, and access.

It facilitates data sharing, archiving, and publishing data.

These activities ultimately enable the ability to reuse, reproduce, and integrate data which provides for the potential for new discoveries.

Key Points

- The increasing data deluge around us and the threat of data loss can be addressed with good data management.
- Data management relates directly to your roles and responsibilities.
- Data is valuable and the cost of not performing data management can be very high.
- Data management enables transparency for reproducibility of findings and for defense in litigation.
- Policies and requirements are addressing data management in Federal and non-Federal organizations.
- Big data and data integration efforts require consistent management of data.

Key Points

In this module, we have covered the value of data management.

As you have seen, the increasing data deluge around us and the threat of data loss can be addressed with good data management. Good data management can help us better integrate data and reduce the risk of losing valuable data.

Data management relates directly to your roles and responsibilities, whether you are a scientist, data steward, or manager. Ensuring good data management is part of everyone's jobs.

Data is valuable and the cost of not performing data management can be very high. Making up for the data loss or loss of knowledge transfer when someone leaves or retires can be time consuming and costly.

Data management enables transparency for reproducibility of findings and for defense in litigation.

Policies and requirements are addressing data management in Federal and non-Federal organizations. You need to ensure that you understand what is required of you when you conduct your science, apply for grants, and submit to publishers.

Big data and data integration efforts require consistent management of data. Considerable thought must be given to creating common standards, planning, processing, and preserving.