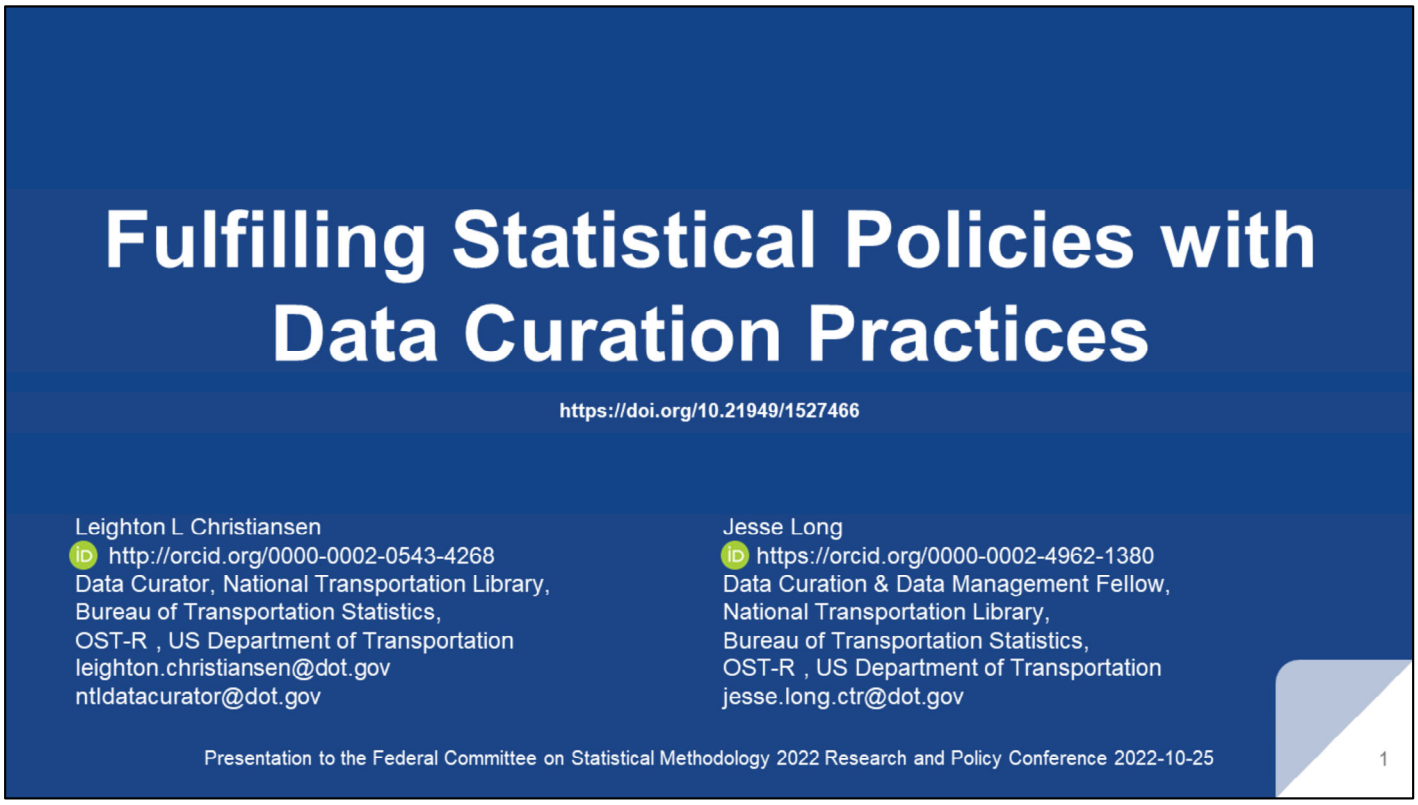


Leighton:
Slide Title: Fulfilling Statistical Policies with Data Curation Practices

[FCSM Scripted Text]
Thank you for that introduction. My co-author, Jesse Long, the Data Curation & Data Management Fellow at NTL, is away at another conference and sends her regrets. Over the course of the presentation, I will go into more depth on what data curation is, and how we believe the practices of data curation can help you to make federal statistics more transparent, as well as accessible, interoperable, and preservable.

[Slide]
[Next speaker: Leighton]
[Time: 25 seconds]

[Slide Text Not Read]
Fulfilling Statistical Policies with Data Curation Practices

Leighton L Christiansen http://orcid.org/0000-0002-0543-4268
Data Curator, National Transportation Library,
Bureau of Transportation Statistics,
OST-R , US Department of Transportation
leighton.christiansen@dot.gov
ntldatacurator@dot.gov

Jesse Long https://orcid.org/0000-0002-4962-1380
Data Curation & Data Management Fellow,
National Transportation Library,
Bureau of Transportation Statistics,
OST-R , US Department of Transportation
jesse.long.ctr@dot.gov

Leighton:
Slide Title: Overview

[FCSM Scripted Text]
A brief overview of what we will discuss in the next 20 minutes.

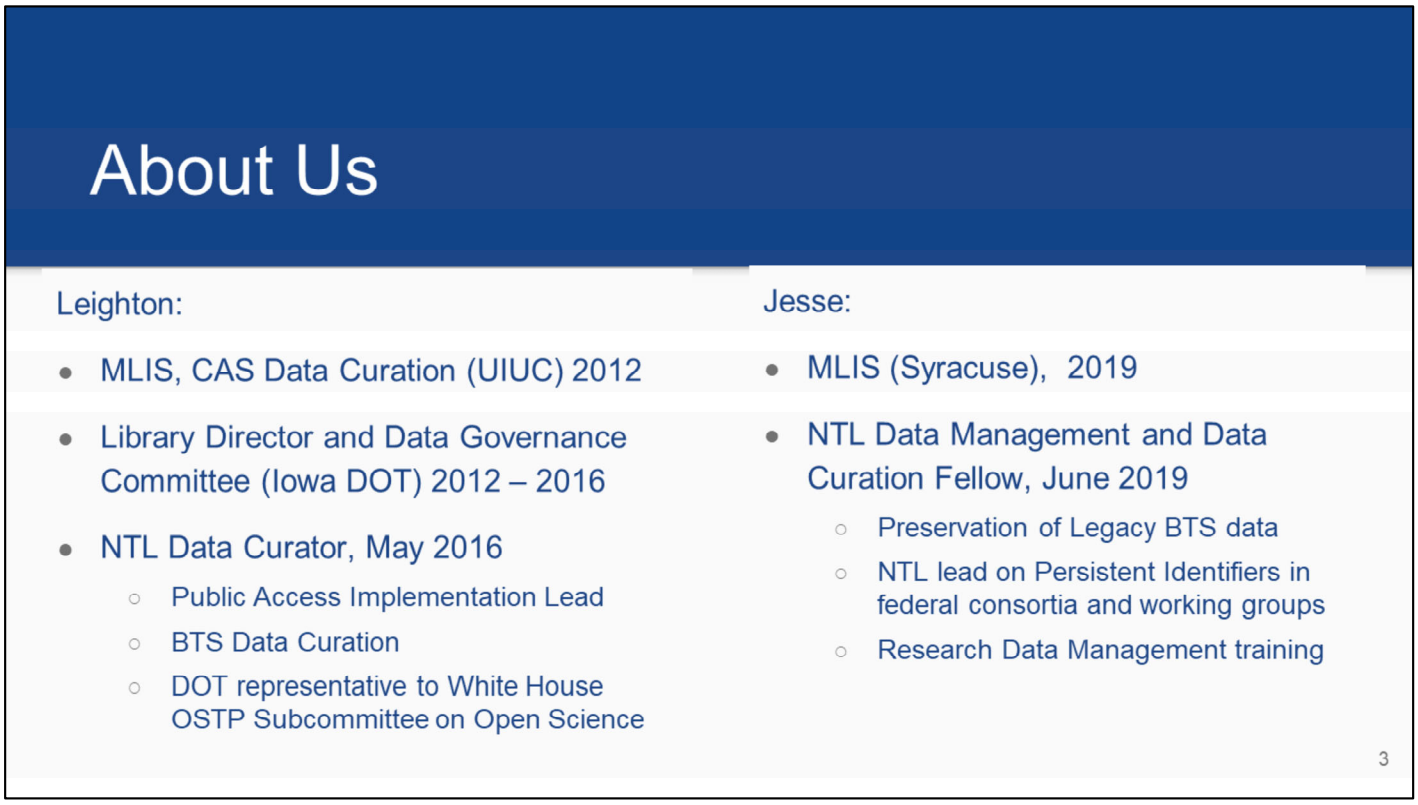
- Statistical Laws & Practices
- About Data Curation
- Data Curation for Transparent Statistics: Suggestions
- Conclusions

[Slide]
[Next speaker: Leighton]
[Time: 20 seconds]
[Total time: 0:45]

[Complete Slide Text]
Overview

- About Us
- Statistical Laws & Practices
- About Data Curation

- Data Curation for Transparent Statistics: Suggestions
- Conclusions
- Questions
- Supplemental Slides



Leighton & Jesse:
Slide Title: About Us
[FCSM Scripted Text]: Slide skipped in the interest of time

[Next Slide]

[Extended Text, not presented]
As we move into the section on data curation, just a little bit more about us and our roles.

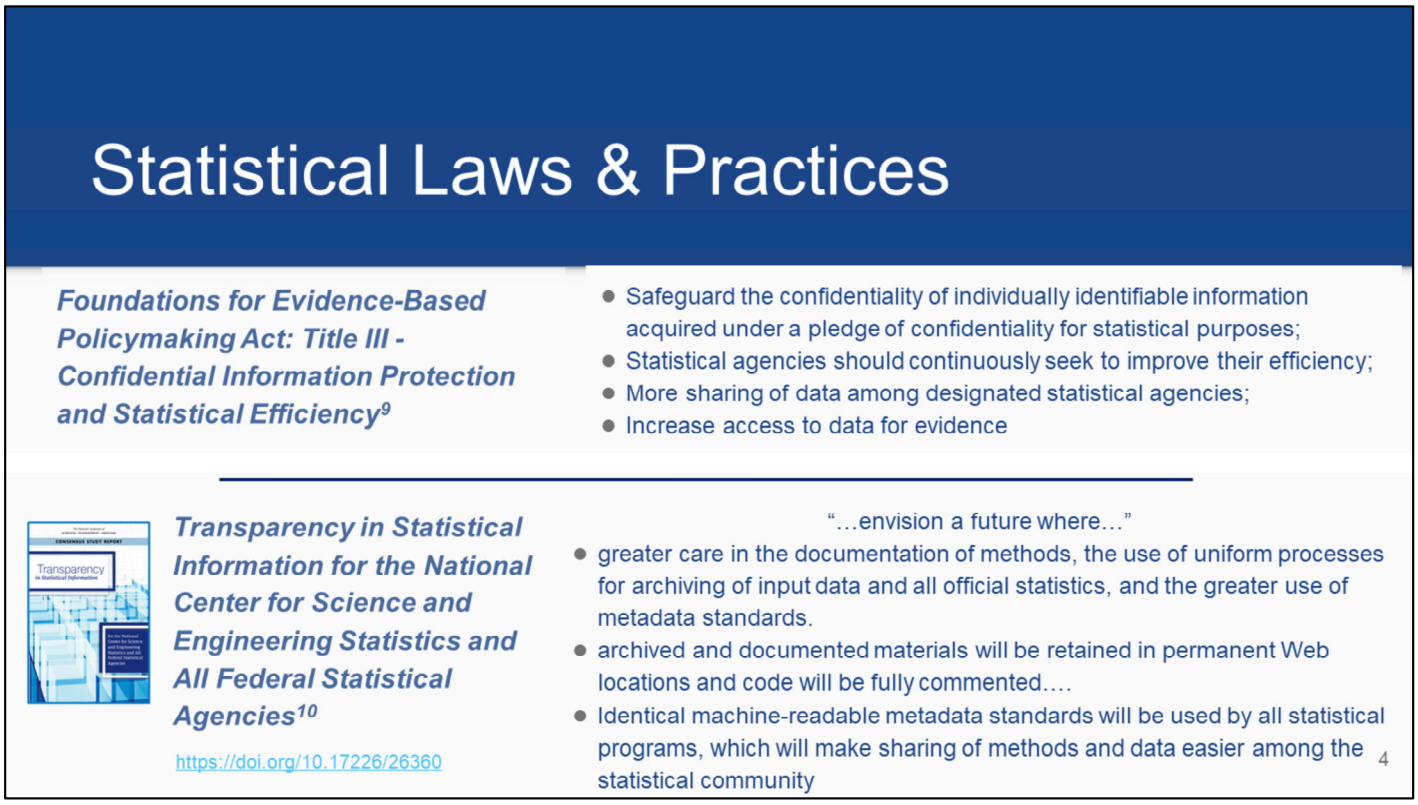
Leighton: I graduated from the University of Illinois at Urbana-Champaign, with a Masters of Library and Information Science, and Certificate of Advanced Study in Data Curation, in 2012. From their I became the director of the Iowa Department of Transportation Library. While the bulk of my duties centered around research librarianship, information preservation, and digitization, I was an early advocate for data management and public access, and served on the Iowa DOT's data governance committee. In 2016 I came to NTL to serve as the first Data Curator in the department. The major foci of my position are implementing the DOT public access plan for research data, and building a culture of data management, curation, preservation, and sharing for BTS-created statistical data. I also ensure BTS data is represented in the DOT data catalog: data.transportation.gov, with complete metadata and documentation.

These records are then shared with the federated search engine, data.gov, so that the public can find and make use of BTS statistical data. Further, I serve as DOT representative to White House OSTP Subcommittee on Open Science

Jesse: I recently graduated from Syracuse University, with a Masters of Library and Information Science. My studies and interests revolved digital data, digitization, and preservation. This led me to the National Transportation Library and my position as a Data Management and Data Curation Fellow. Already I have been able to further explore and build skills around these interests, by working with legacy datasets to ensure they are preserved for future use. I also lead our effort around the use of persistent identifiers for research outputs, people, and entities, and I provide research data management training for tools and practices.

And because BTS is unique among federal statistical agencies in having its own library and data curation team, we hope that BTS can offer some new outlooks on making federal statistics transparent. We believe there are practices within data curation that you all will find helpful

[Slide]
[Next speaker: **Leighton**]
[Time: 0:00 seconds]
[Total time: 0:45]



Leighton:
Slide Title: Statistical Laws & Practices

[FCSM Scripted Text]
So why do statistical agencies want to make their statistics more transparent?
First, there is law.
The 2019 Foundations for Evidence-Based Policymaking Act: Title III - Confidential Information Protection and Statistical Efficiency calls on statistical agencies to:

- Safeguard the confidentiality of individually identifiable information acquired under a pledge of confidentiality for statistical purposes;
- Statistical agencies should continuously seek to improve their efficiency;
- designated statistical agencies should share more data; and,
- Increase access to data for evidence.

In the evidence act we find calls to make statistics more transparent both among statistical agencies and to the public.

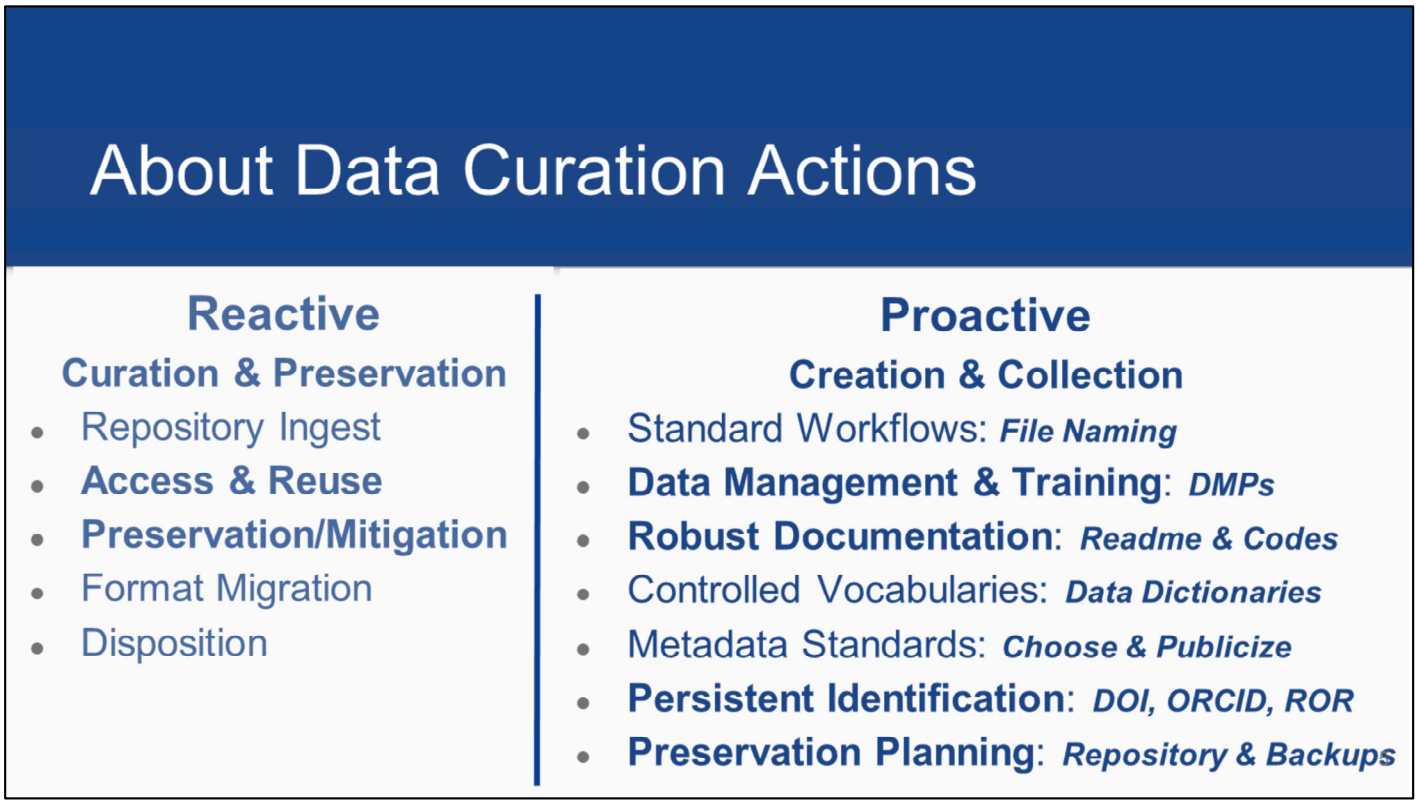
Second, there is creating a better future for statistics users.
In the 2022 National Academies publication "Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical

Agencies" the research panel envisions a future where:

- There is greater care in the documentation of methods, and the greater use of metadata standards;
- Archived and documented materials will be retained in permanent Web locations;
- Identical machine-readable metadata standards will be used by all statistical programs, which will make sharing of methods and data easier among the statistical community.

You might be asking yourself: "How do we fulfill the law and get to that great sounding future?"
By harnessing the practices, tools, and expertise of data curation, I would argue.

[Slide]
[Next speaker: Leighton]
[Time: 1:40 minutes]
[Total time: 3:00 minutes]



Leighton:
Slide Title: About Data Curation Actions

[FCSM Scripted Text]
What is data curation? To help you picture data curation, let's think of a more familiar use of the words "curation" or "curator".

Reactive Actions:
If you think about the common usage of the words "curation" or "curator," you most likely think of a museum curator, who focuses on the preservation of objects that have already been created. And data curators certainly can and do perform many of those same "reactive" curation and preservation tasks.
These can include:

- **Access and Reuse:** Making objects finable and accessible to researchers and the public. For physical objects, this might mean putting an object on display. For digital objects, this could mean exposing the object metadata to search engines.
- **Preservation/Mitigation:** For a physical object preservation and mitigation may mean housing the object in an environmentally controlled lab space to slow the effects of aging. For digital objects, this can mean holding master copies in dark storage, performing regular fixity checks looking for bit rot, and migrating master

and use copies from drives to new drives every few years to avoid data loss from media degradation.

You can see that data curators perform many of the same ***reactive*** actions as physical object curators. However, digital data curators also want to be in a position to take ***proactive*** steps as well.

Proactive Actions

In order to be more proactive, data curators want to be embedded in data collection projects from the very beginning. By implementing data management strategies at the time of data creation we can improve data preservation outcomes for years or decades in the future. Approaching data curation and preservation for legacy or already existing datasets, is often harder, and suffers from incomplete knowledge or information due to limited documentation.

Outcomes can be improved by taking ***proactive actions*** and planning for long-term data preservation and sharing from the beginning of a project.

The proactive actions that curators want to help data collection teams implement include:

- **Data Management & Training:** The most crucial step to take before any data is collected is writing a data management plan (or DMP). You might ask "Does every data collection activity need a DMP?" My response is that every data collection action ***deserves*** a robust data management plan. DMPs can go a long way to making data preservable, interoperable, and transparent. Data curators can help a data collection team draft, revise, implement, and update their DMP.
- **Robust Documentation**: An embedded data curator can assist the team with creating robust documentation. This can include:
    - writing up readme files and data dictionaries;
    - checking for the presence of code tables, data dictionaries, and supplementary files;
    - researching, suggesting, and implementing domain appropriate metadata schema; and,
    - Building a complete data package to improve preservation and transparency for the dataset.
- **Persistent identification:** Persistent identifiers (or PIDs) eliminate ambiguity and confusion with published research, because they provide unique identification. There are PIDs for objects, people, and organizations.
- **The final proactive action step I will mention is Preservation Planning**: Data curators can help data collection teams identify, ahead of data collection, likely target

repositories for the data types generated from the project, storage size needs based on the planned data collection, local backup strategies, and vet repositories based on those strategies and backup server locations.

Now, in the descriptions of reactive and proactive curation actions, you may recognize some actions your data collection teams already take.

However, unless each of these steps is included in your data collection, your data and your statistics cannot reach maximum transparency. Transparency that is planned for and included at the beginning of data collection is more efficient and impactful than transparency attempted after the fact.

[Slide]
[Next speaker: Leighton]
[Time: 4:00 minutes]
[Total time: 7:00 minutes]

[Extended Workshop Text, not presented at FCSM]
What is data curation? To help you picture data curation, let's think of a more familiar us of the words "curation" or "curator".

Reactive Actions:
If you think about the common usage of the words "curation" or "curator," you most likely think of a museum curator, who focuses on the preservation of objects that have already been created. And data curators certainly can and do perform many of those same "reactive" curation and preservation tasks.
These can include:

- Repository Ingest: Seeking and accepting objects to be added to a museum special collection, if a physical object, or adding a local copy of a digital object to digital repository, based on you collection development policy.
- **Access and Reuse:** Making objects finable and accessible to researchers and the public. For physical objects, this might mean putting an object on display. For digital objects, this could mean exposing the object metadata to search engines.
- **Preservation/Mitigation:** For a physical object preservation and mitigation may mean housing the object in an environmentally controlled labs space to slow the effects of aging. For digital objects, this can mean holding master copies in dark storage, performing regular fixity checks looking for bit rot, and migrating master and use copies from drives to new drives every few years to avoid data loss from media degradation.
- Format Migration: Now for format migration, I want to start with the digital objects.

When we talk about format migration, we usually mean taking the data recorded by people, sensors, or machines, and converting that data file from its original, proprietary format, into a more universally accessible or open format. For example, say you hired a contractor to take hand-held tablets out on to street corners to survey citizens on their opinions on a specific topic. The contractor would likely write an app that would store the data in a form that was easily stored and rendered on the tablet. But would that data be easy to read on your desktop machine using SAS or a spreadsheet program? Maybe not, and you would need to migrate that data from the tablet format to something ubiquitous such as comma separated value, or CSV, so that you would us it on any machine, far into the future. What you have done is preserved the intellectual content of the data object, even if you have not preserved it in its original form. We do the same thing when we digitize books or printed reports. We often unbind them, destroying the original container, in order to rapidly bulk scan the pages, preserving the text as PDFs or plain text documents. Format migration for museum pieces might include making a mold and a plaster copy of a famous sculpture, or making a scale model of object too large to preserve, such as the Colosseum in Rome: We have preserved some aspect of the original item by migrating that information to a new format or media, even if we cannot save the original experience or materials of the item.

- Disposition: There may come a time in an objects life when we decide to dispose of it. This can take a number of different forms. Museums and galleries may sell a particular work because they want to raise money for capital projects or to make other purchase. A gallery may decide that a physical, sacred object should be returned to the culture that created it. For data curators, we may decide to delete a large dataset is no longer of interest to the scientific community, as no one has requested access in a number of decades, or the dataset has been largely rebuffed by the research community because of questionable collection methods or erroneous data. We might even be forced into such a situation due to resources constraints: we have run out of server space and need to make room for new, cutting edge data, therefore, some legacy data has to go.

You can see that data curators perform many of the same ***reactive*** actions as physical object curators. However, digital data curators also want to be in a position to take ***proactive*** steps as well, and Jesse will describe these.

Proactive Actions

In order to be more proactive data curators want to be embedded in data collection projects from the very beginning. By implementing data management strategies at the time of data creation we can improve data preservation outcomes for years or decades in the future. Approaching data curation and preservation for legacy or already existing datasets, is often harder, and suffer from incomplete knowledge or information due to limited documentation.

Outcomes can be improved by taking ***proactive actions*** and planning for long-term data preservation and sharing from the beginning of a project.

The proactive actions that curators want to help data collection teams implement are:

- Standard Workflows: Data curators can help data collectors document and standardize work flows and data stewardship practices. A very simple practice that many teams overlook is a standard and documented file naming convention. Files names should be human readable, contain some project intelligence, and include a date and timestamp for version control. There is nothing worse than having 16 files in your folder called "data" or "full text". Determining a file naming structure is an important step to take before any data is ever collected.
- **Data Management & Training:** Another crucial step to take before any data is collected is writing a data management plan (or DMP). You might be ready to ask "Does every data collection activity need a DMP?" My response is that every data collection action ***deserves*** a robust data management plan. DMPs can go a long way to making data preservable, interoperable, and transparent. Data curators can help a data collection team draft, revise, implement, and update their DMP.
- **Robust Documentation**: An embedded data curator can assist the team with creating robust documentation. This can include:
    - writing up readme files and data dictionaries;
    - checking for the presence of code tables, data dictionaries, and supplementary files;
    - researching, suggesting, and implementing domain appropriate metadata schema; and,
    - Building a complete data package to improve preservation and transparency for the dataset.
- Controlled Vocabularies: A data curator can research and suggest implementation of an existing controlled vocabulary to make variable name, meanings, and specifications standard and interoperable. Using existing controlled vocabularies makes writing a data dictionary much easier. Additionally, the data curator can help crosswalk controlled vocabularies to make translating between vocabularies easier.
- Metadata Standards: In addition to controlled vocabularies, data curators can suggest appropriate metadata standards, or help crosswalk between existing and new standards. Either way, the data curator will help the data collection team choose the necessary standards, as well as document and publicize the metadata standards in use. Publicizes chosen metadata standards is a great step towards increasing transparency, and helps data users read and use your data.
- **Persistent identification:** Persistent identifiers (or PIDs) eliminate ambiguity and confusion with published research, because they provide unique identification. There are PIDs for objects, people, and organizations.
    - For Objects: there are Digital Object Identifiers (DOIs). DOIs are typically used

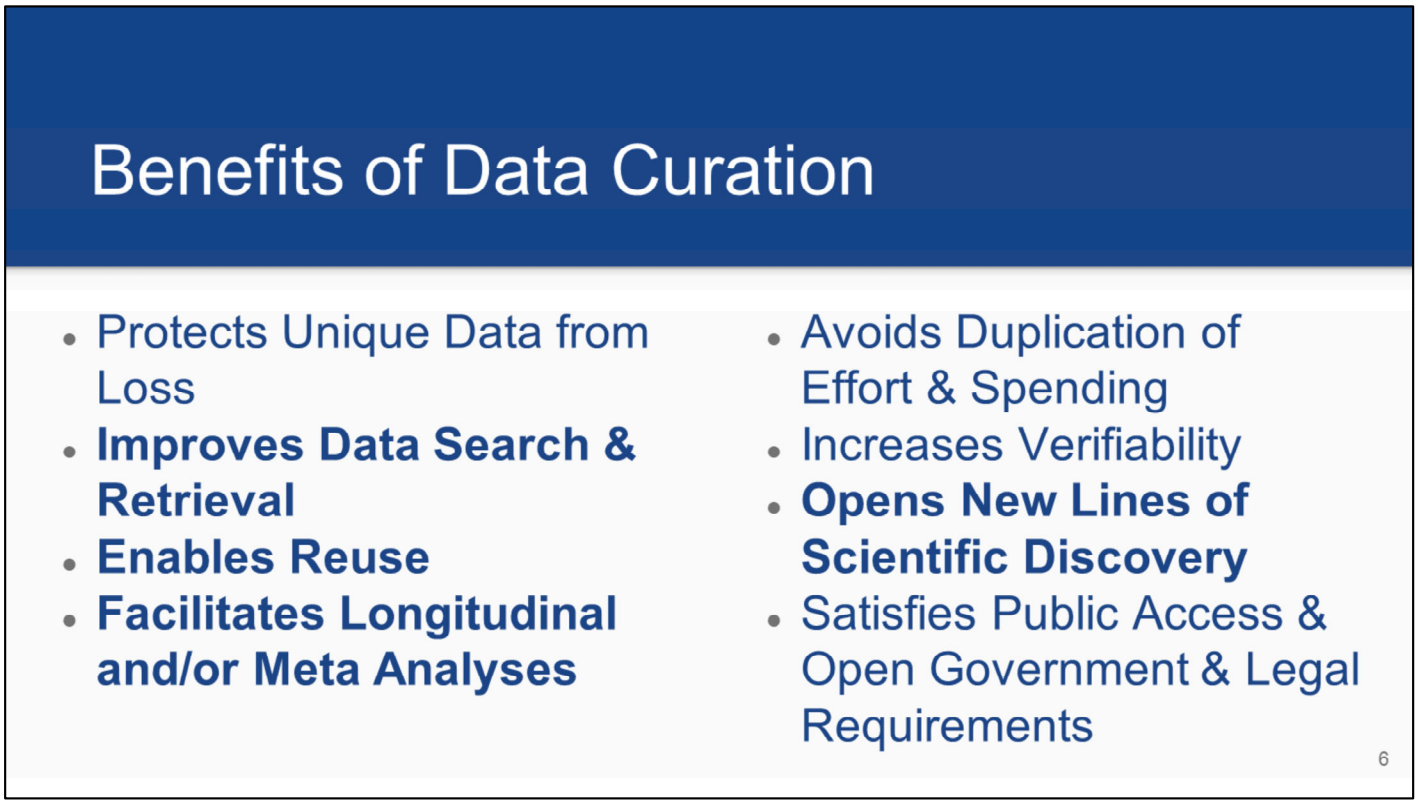
for publications, images, audio files, measurement instruments, or any other THING that can either have a digital presence in a networked environment, or can be described by a web page or digital metadata file. The DOI may point either directly to the object or its digital "landing page." There are many "brands" of PIDs for things, I only mention DOIs here because they are most common.

- Open Researcher and Contributor Identifiers (or ORCIDs) are used to uniquely identify people. On the title slide, you saw both of our ORCIDs. My ORCID iD which is a https link, containing the protocol and 16 digits, that leads to a web page where I have a profile that records my works and uniquely disambiguates me from all other humans named Jesse Long.
- Finally, for organization there is the Research Organization Registry (or ROR). It is a fairly new initiative to build an open, collaborative research organization identification schema and registration service. The ROR identifier system is controlled by the research organizations themselves, and seeks to be interoperable among systems. This is a different model than identifiers such as FundRef, which are controlled by scholarly journal publishers, and are often used only within the family of journals put out by that publisher.

- **The final proactive action step I will mention is Preservation Planning**: Data curators can help data collection teams identify, ahead of data collection, likely target repositories for the data types generated from the project, storage size needs based on the planned data collection, plan for local backup strategies, and vet repositories based on those strategies and backup server locations.

Now, in the descriptions of reactive and proactive curation actions, you probably recognize some actions your data collection teams already take.

However, one of our suggestions to you is that unless each of these steps is included in your data collection, your data and your statistics cannot reach maximum transparency. Transparency that is planned for and included at the beginning of data collection is more efficient and impactful than transparency created after the fact.



Leighton:
Slide Title: Benefits of Data Curation

[FCSM Scripted Text]
Data curation provides data creators and data consumers a great number of benefits. Some of these benefits include:

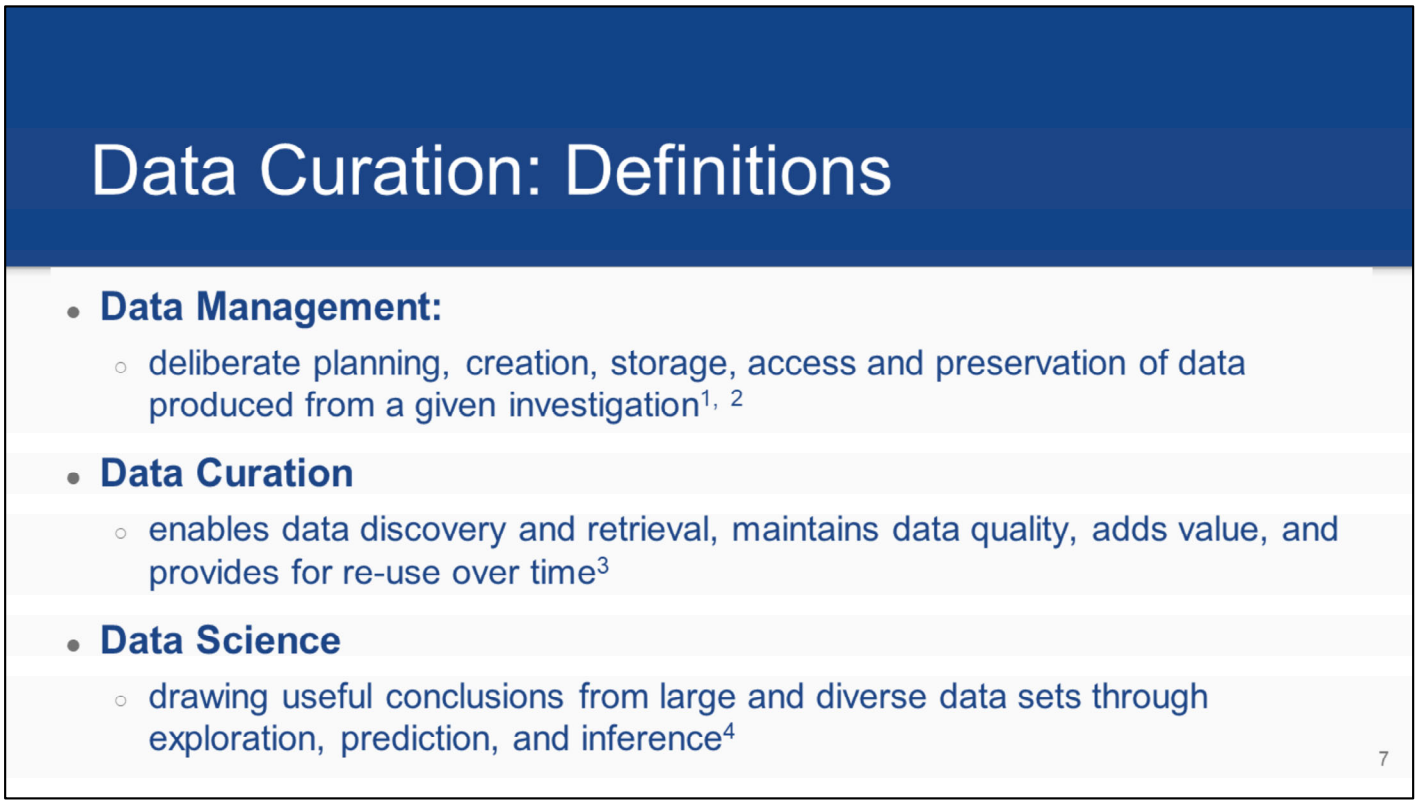
- allowing for new research in the future by freeing up research funds or by enabling meta-analyses.
- allowing for data reuse in ways not intended by the original researchers.
- ensuring that data creators and data consumers can locate the data that fits their needs.

By employing data management best practices, metadata standards, and open data formats, we can improve interoperability among datasets, not just inside our organizations, but for external users as well. And not just improved interoperability among datasets we or you create, but we can help improve interoperability with weather data, census data, public health data, space data, etc., to open up new avenues of discovery among researchers as well as citizen scientists. This is perhaps the most promising outcome of curating data for transparency.

[Slide]
[Next speaker: **Leighton**]
[Time: 1:00 minutes]
[Total time: 8:00 minutes]

[Complete Slide Text]
Slide Title: Benefits of Data Curation

- Protects Unique Data from Loss
- Improves Data Search & Retrieval
- Enables Reuse
- Facilitates Longitudinal and/or Meta Analyses
- Avoids Duplication of Effort & Spending
- Increases Verifiability
- Opens New Lines of Scientific Discovery
- Satisfies Public Access & Open Government & Legal Requirements



Leighton:
Slide Title: Data Curation: Definitions

[FCSM Scripted Text]
We have talked about data curation actions, and the benefits of data curation. I would like to go back to some key definitions.

Data Management: “In the context of research and scholarship, "Data Management" refers to the storage, access and preservation of data produced from a given investigation. Data management practices cover the entire lifecycle of the data, from planning the investigation to conducting it, and from backing up data as it is created and used; to long term preservation of data deliverables after the research investigation has concluded.”

“Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation.”

“Data Science is about drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference, using the skills and practices of statistics, information science, and computer programming.”

I hope you can see that each of these terms is distinct. As we talk about data management, data curation, and data science, you will likely recognize your own work touches on some of the actions described in these definitions.
What may be missing from your current practice is seeing these individual actions as part of a holistic strategy for sharing and preserving data.
Let us look at linking these actions next.
[Slide]
[Next speaker: Leighton]
[Time: 1:35 minutes]
[Total time: 9:35 minutes]

[Extended Workshop Text, not presented at FCSM]
We have talked about data curation actions, and the benefits of data curation. I would like to go back to some key definitions.

“In the context of research and scholarship, "Data Management" refers to the storage, access and preservation of data produced from a given investigation. Data management practices cover the entire lifecycle of the data, from planning the investigation to conducting it, and from backing up data as it is created and used; to long term preservation of data deliverables after the research investigation has concluded.”

> 1: Source: University Library, Texas A&M University. “Data Management Defined - Research Data Management - Guides at Texas A&M University.” Research Data Management, October 1, 2013. http://guides.library.tamu.edu/DataManagement

Or to borrow a plain language definition from Kristin Briney, (page 7) “Data management is the compilation of many small practices that make your data easier to find, easier to understand, less likely to be lost, and more likely to be usable during a project or ten years later.”

> 2: Source: Briney, Kristin. 2015. *Data management for researchers: organize, maintain and share your data for research success*. (6)

“Data curation is the active and ongoing management of data through its lifecycle of

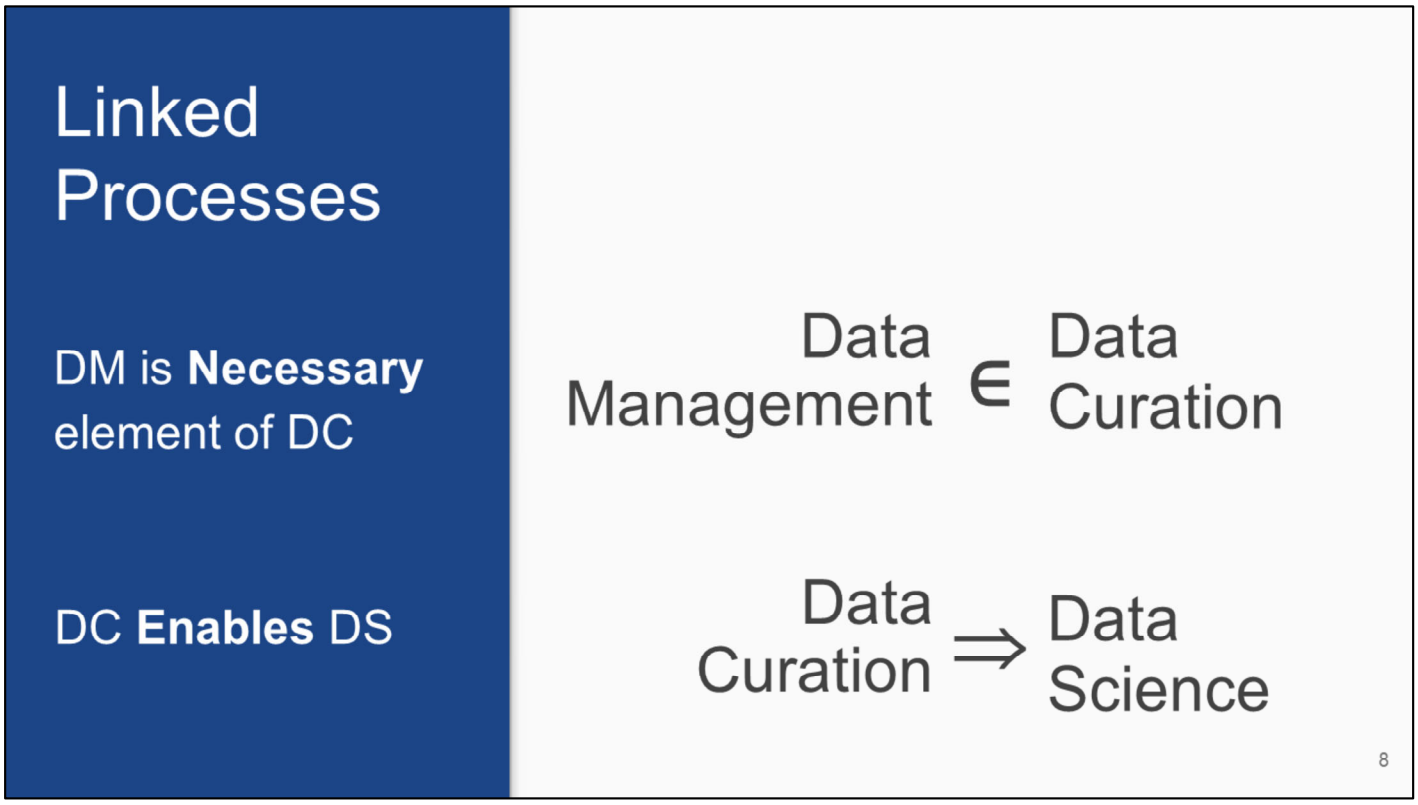
interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation."

> 3: Source: Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. "Specialization in Data Curation," 2013. http://www.lis.illinois.edu/academics/programs/specializations/data_curation.

"Data Science is about drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference, using the skills and practices of statistics, information science, and computer programming."

> 4: Based on: http://www.inferentialthinking.com/chapter1/what-is-data-science.html

As we talk about data management and data curation, you will likely recognize your own work touches on many of the actions described in these definitions.
What may be missing from your current practice is seeing these individual actions as part of a holistic strategy for sharing and preserving data.
Lets look at linking these actions next.
[Slide]
[Next speaker: Leighton]
[Time: 2:40 minutes]
[Total time: 11:00 minutes]



Leighton:
Slide Title: Linked Processes

[FCSM Scripted remarks]
Here I attempt to illustrates the interconnectedness of skills that we hope to harness to improve the transparency of federal data and statistics.

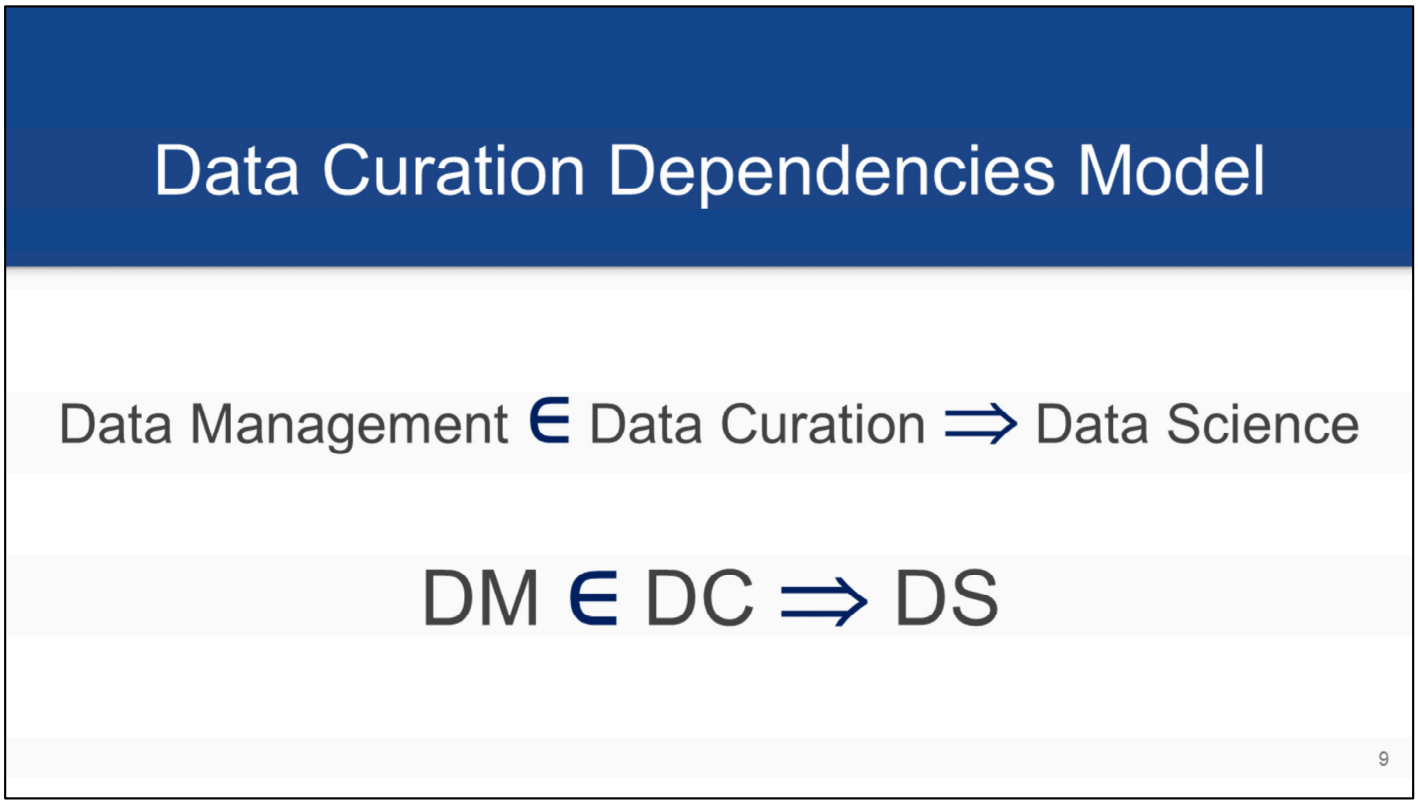
- Data Management is a ***necessary element*** of Data Curation. And to enable good Data Curation, it often means that we have to encourage researchers and data collections to think beyond a specific investigation or survey, and adapt good data management practices to meet future needs.

- Good Data Curation, in turn, ***enables*** broader, longitudinal Data Science. By preserving and adding value to data, Data Curation makes the task of Data Science more efficient and effective, as well as opening new output possibilities.

Let us abstract a step further.

[Slide]
[Next speaker: Leighton]
[Time: 1:00 minutes]

[Total time: 10:35 minutes]



Leighton:
Slide Title: Data Curation Dependencies Model

[FCSM Scripted Remarks]
To visualize these dependencies altogether, I have created this pseudo-equation: Data Management (DM) is a ***necessary element*** of Data Curation (DC) which ***enables*** Data Science (DS).

Data Management ∈ Data Curation ⇒ Data Science

DM ∈ DC ⇒ DS

Where:

- ∈ stands for "necessary element", and,
- ⇒ stands for "enables"

As far as I know, the above dependency model is original to me, beginning in July 2016.

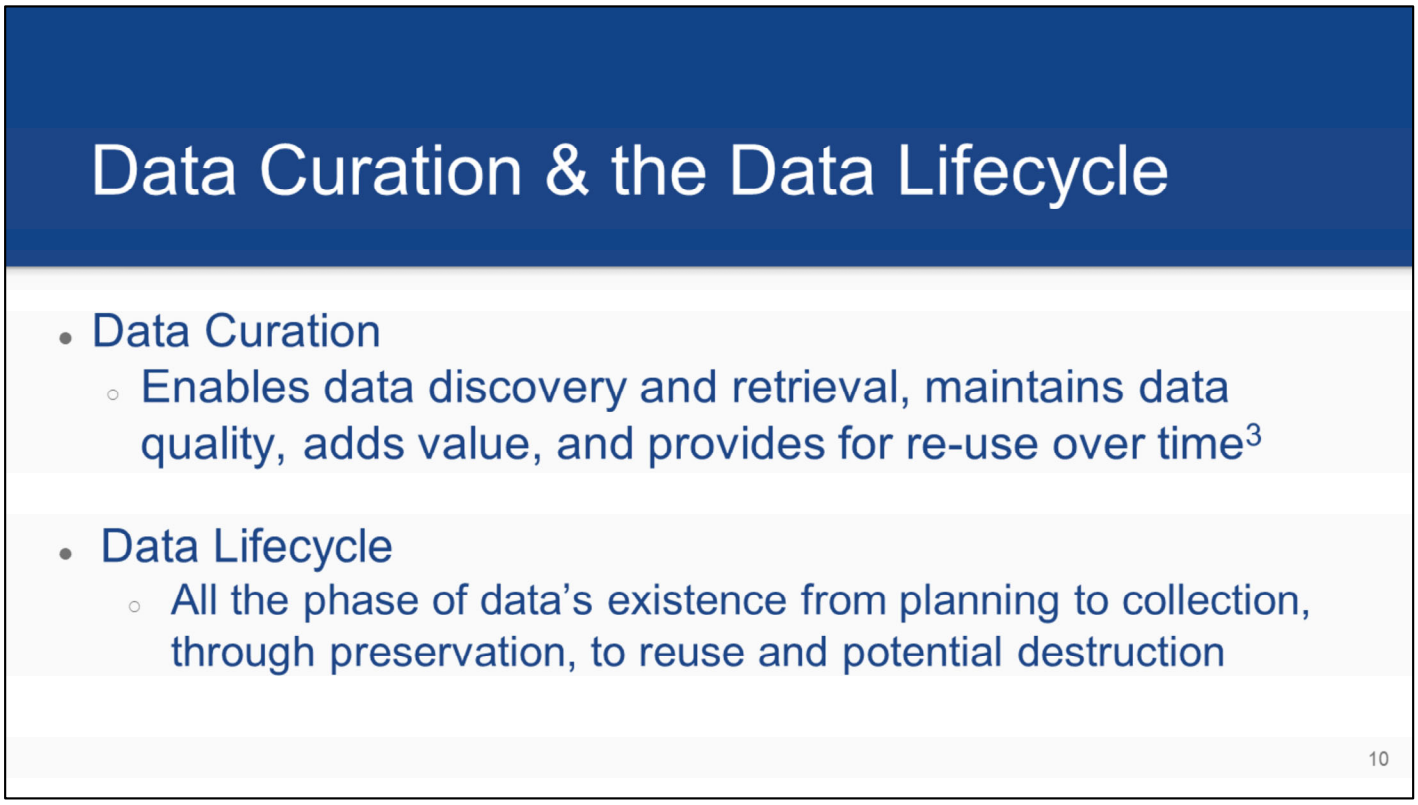
Soapboxing time: The reason I think that talking about these linked processes is important is that it helps to provide context for public access and statistical transparency policies and laws for federal data and statistics: there is a hope that by opening these datasets more broadly to the research, business, scientific, policy, and public communities, that new discoveries can be made by current and future data scientists. So, we as data creators and collectors of today have a responsibility to the data users

who follow us. And those data users may even be us!!

That responsibility can be summed up as: We should use the best resources and practices at our disposal to steward those data and statistics into the future, for as long as they will be of interest. And given the nature of many federal statistical surveys, the period of interest may extend decades or longer, as folks at the Census well know.

So now let us get back to the definition of data curation.

[Slide]
[Next speaker: **Leighton**]
[Time: 1:40 minutes]
[Total time: 12:10 minutes]



Leighton:
Slide Title: Data Curation & the Data Lifecycle

[FCSM Scripted Text]
To review, Data Curation, "is the active and ongoing management of data through its ***lifecycle of interest and usefulness*** to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use ***over time*** through activities including authentication, archiving, management, preservation, and representation."

Important phrases in this definition are "lifecycle of interest and usefulness" and "over time." This means that data curators see data and statistical data as having what we refer to as a "lifecycle."

There are a number of definitions of the data lifecycle. To paraphrase many of them you could say the data lifecycle is "All the phases of data's existence from planning to collection, through preservation, to reuse and potential destruction." There are also a number of graphic models of the data lifecycle, but we will only look at one today.

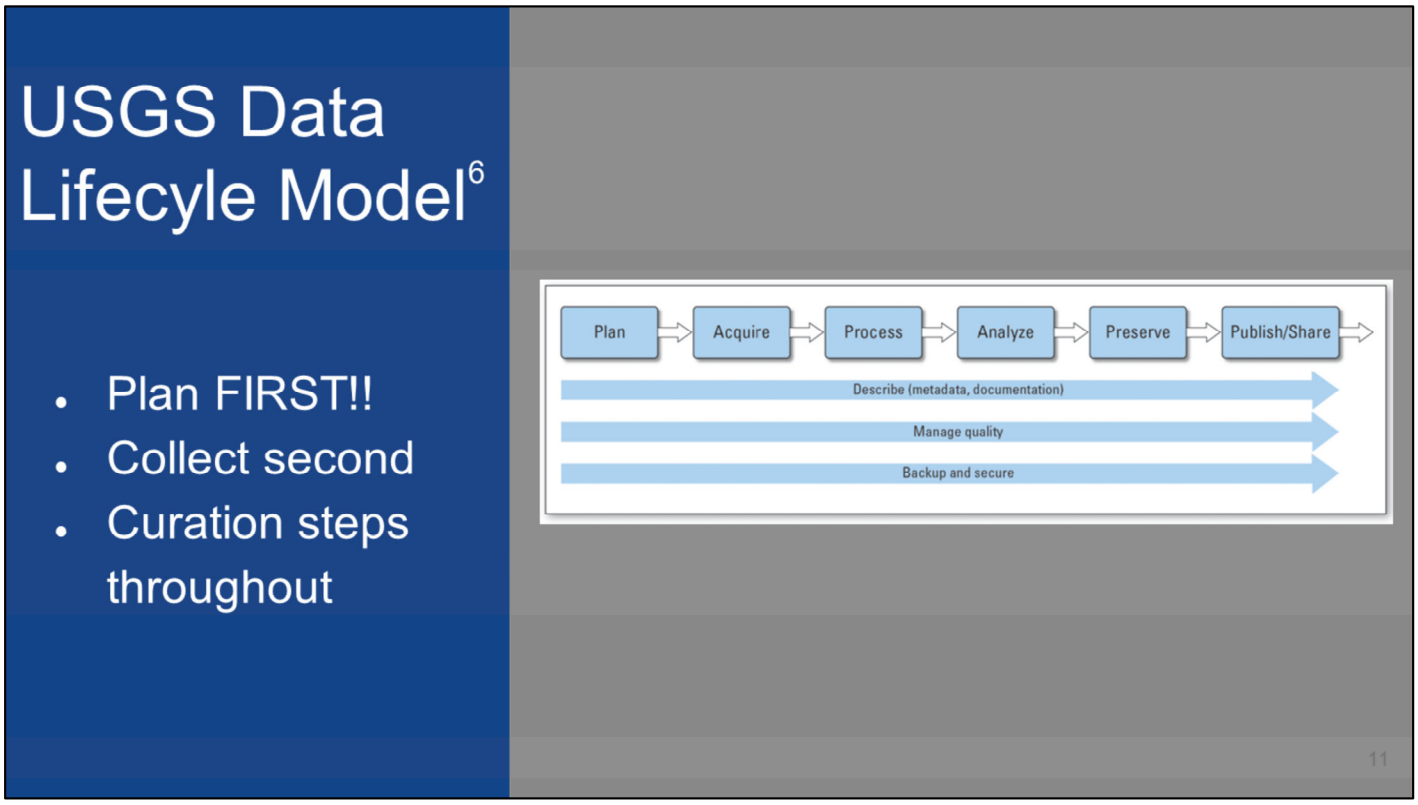
[Slide]
[Next speaker: Leighton]

[Time: 1:00 minutes]
[Total time: 13:10 minutes]

[Slide Text]

- Data Curation
  - Enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time[3]

- Data Lifecycle
  - All the phase of data's existence from planning to collection, through preservation, to reuse and potential destruction



Leighton:
Slide Title: USGS Data Lifecycle Model[6]

[FCSM Scripted Text]
This is the U.S. Geological Survey (USGS) Data Lifecycle model, it is one of the simpler versions of the data lifecycle and one we view as more beneficial, because the very first action is planning. Planning, before data collection, is the most important step.

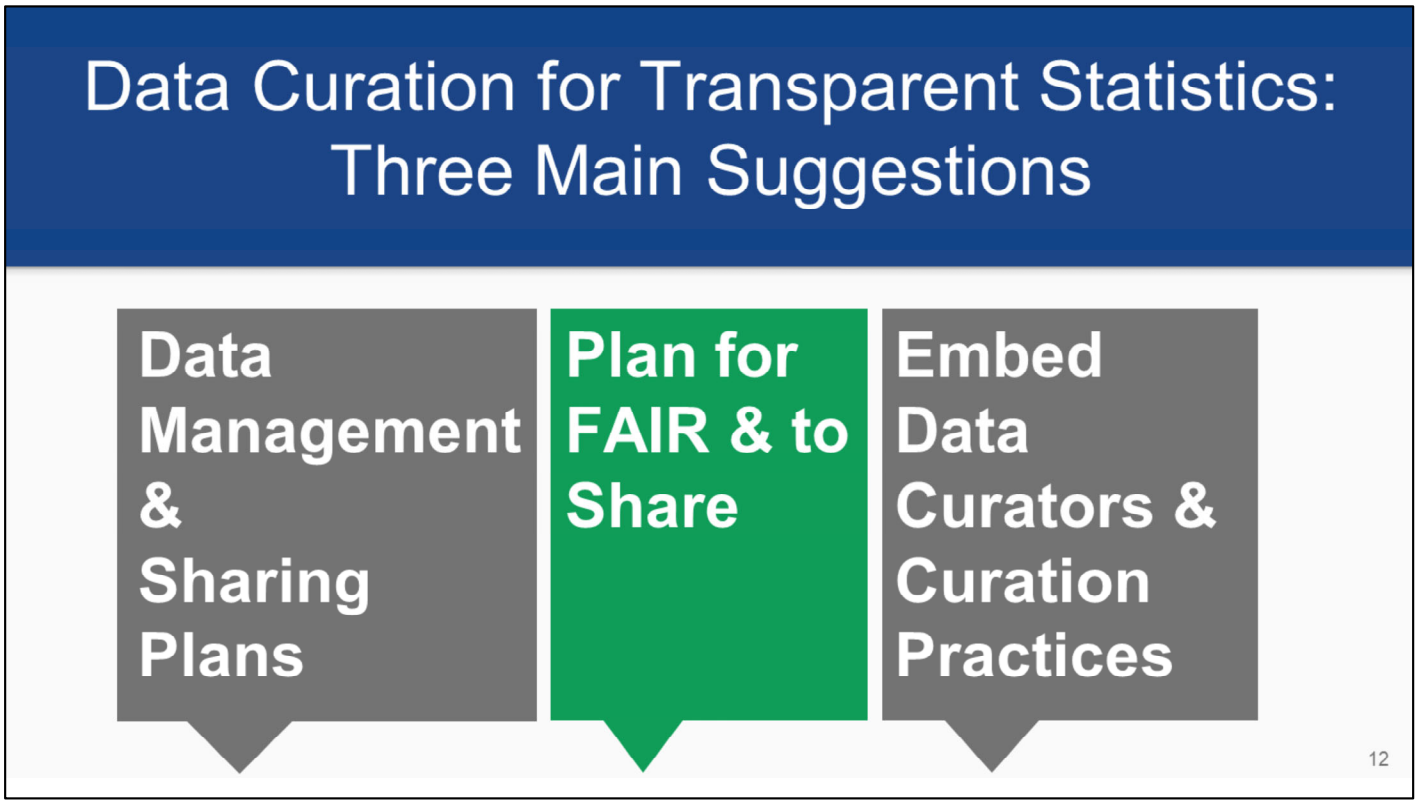
During the Planning step a number of things needed to be determined, such as:

- What data is going to be collected?
- How data will be collected?
- How will the data be organized?
- Who will be responsible for data?
- When will backups occur?
- Will there be sensitive data collected and if so how will it be handled?
- Whether and how much data will be shared, in the end?

Only after all of these questions are answered, should data be collected. If you plan preservation, sharing, and transparency at the beginning, it makes preservation, sharing, and transparency goals much easier to achieve. Now that we have talked about data curation it is time to take a look at specific suggestions.

[Slide]
[Time: 0:45 seconds]
[Total time: 14:00 minutes]

[Slide Text]
Plan FIRST!!
Collect second
Curation steps throughout
Central portion of slide shows image of the USGS data lifecycle model, described with the following Alt Text:
USGS Data Lifecycle Model. Steps are in rectangles in this order: Plan; Acquire; Process; Analyze; Preserve; Publish/Share. Actions that happen through each step include: Describe with metadata and documentation; Manage quality; and, Backup and secure.

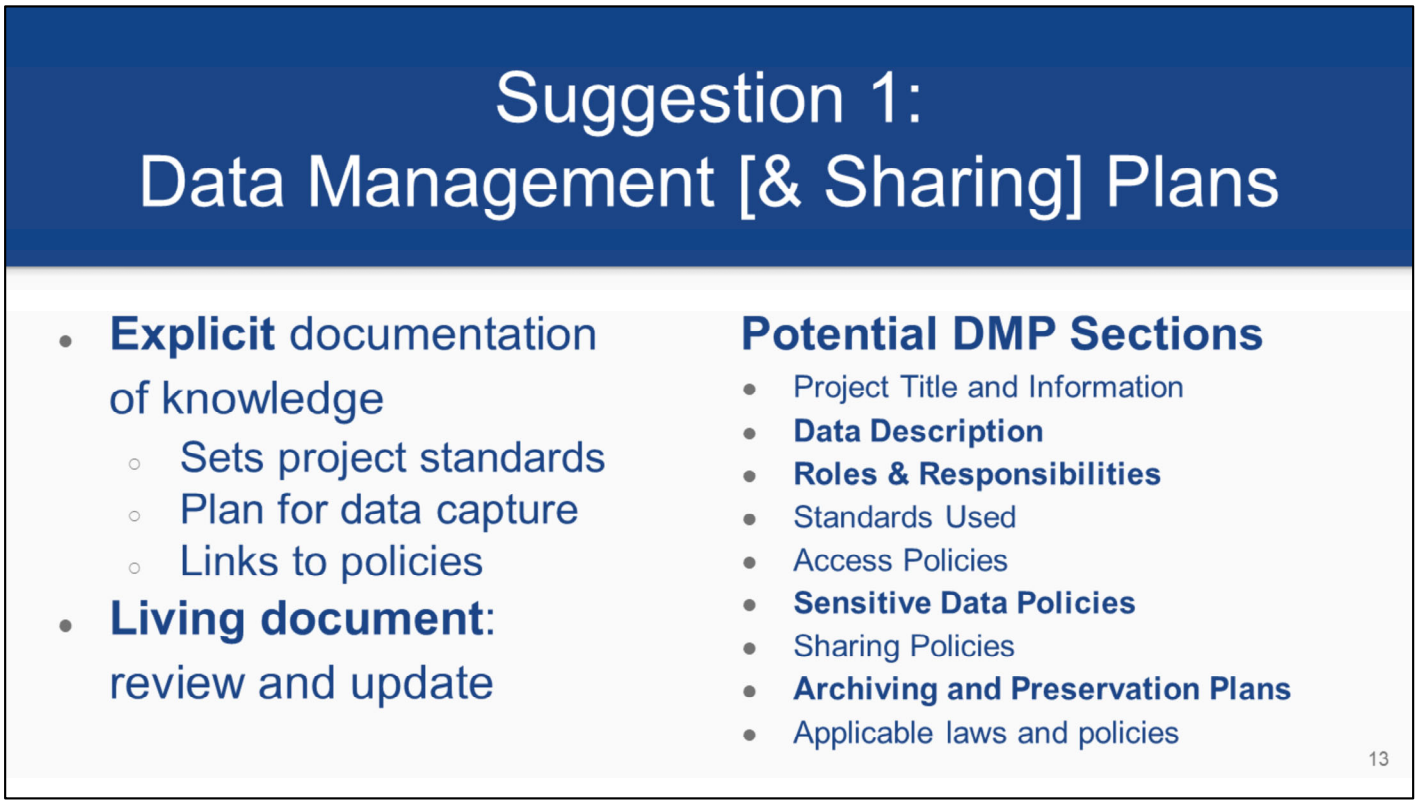


Leighton:
Slide Title: Data Curation for Transparent Statistics: Three Main Suggestions

[FCSM Scripted Text]
Our three major suggestions for making federal statistics more transparent are:

1. Creating Data Management & Sharing Plans;
2. Plan for FAIR & to Share; and,
3. Embed Data Curators & Curation Practices.

Let us look at these suggestions in a bit more detail.
[Slide]
[Next speaker: **Leighton**]
[Time: 0:15 minutes]
[Total time: 14:15 minutes]



Leighton:
Slide Title: Suggestion 1: Data Management [& Sharing] Plans

[FCSM Scripted Text]
The first suggestion is Data Management [and Sharing] Plans. The single most important step you can take to make your statistical data, or any other dataset, more transparent, is to start by making your data collection and storage needs as transparent as possible to yourself and your team.

A Data Management Plan, when created during the project planning phase, can help you think through all of your data externalities and dependencies, as well as plan for access, storage, sharing, and preservation. A DMP becomes a knowledge management document for your team.

A DMP document should:
- Makes all knowledge and information explicit. This includes:
- Types of data to be collected
- How data will be collected
- File types generated:
- Standardized file naming conventions
- How data will be anonymized to protect sensitive information, if needed

- Chosen repository, etc.

The types of information you should make explicit can go on. The point is to record everything you can think of, documenting information project staff need to know, and looking for weak spots that would put the data, the project, or subjects at risk. Then work to eliminate risks prior to data collection.

The most up-to-date version of the DMP should be accessible to all project staff during the entire life of the project.

Also, it is important to consider a DMP as a living document. DMPs should be reviewed as frequently as necessary, and should be updated to capture every project change.

The sections of a DMP can vary and you will be best served by creating an organizational template for consistency and efficiency, making changes as needed. The 2023 NIH Data Management and Sharing Policy[11] https://www.oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy offers a nice template that many federal agencies are moving towards through our work together in the OSTP Subcommittee on Open Science.

You may have noticed that many potential sections of the DMP relate directly to curation actions shown in the data lifecycle. A DMP can serve as a guide to the curation of data and statistics.

A final, public version of your DMP should be included as part of your data package when the data is shared or made public. This public version can be redacted a bit, if needed, for public consumption.

[Slide]
[Next speaker: Leighton]
[Time: 2:00 minutes]
[Total time: 16:15 minutes]

[Extended Workshop Text, not presented at FCSM]
The first suggestion is Data Management and Sharing Plans. The single most important step you can take to make your statistical data, or any other dataset, more transparent, is to start by making your data collection and storage needs as transparent as possible to yourself and your team.

A Data Management Plan, or Data Management & Sharing Plan, when created during the project planning phase, can help you think through all of your data externalities and

dependencies, as well as plan for access, storage, sharing, and preservation.

A DMP document should:

- Makes all knowledge and information ***explicit***. This includes:
  - Project Lead
    - Who takes over, owns intellectual property
  - Staff
  - **Types of data to be collected**
  - When data will be collected
  - Who will collect data
  - **How data will be collected**
    - by humans or machines?
  - **File types generated**:
    - Proprietary file formats or open?
  - **Standardized file naming conventions**
  - File sizes expected or estimated
  - Data access levels
  - **How data will be anonymized to protect sensitive information, if needed**
  - **Chosen repository**
  - Link to repository contacts and policies
  - Organizational IT contacts, and
  - Policies, laws, and Institutional Review Board (IRB) rules that affect data collection

The types of information you should make explicit can go on. The point is to record everything you can think of, documenting information project staff need to know, and looking for weak spots that would put the data, the project, or subjects at risk. Then fix the plan to eliminate that risk.

The most up-to-date version of the DMP should be accessible to all project staff during the life of the project.

Also, it is important to consider a DMP as a living document. DMPs should be reviewed as frequently as necessary, and should be updated to capture every project change.

- During the planning phase, this might include a review at every team meeting: questioning did we change anything that affects the DMP? And noting those changes
- During data collection, that might be monthly, or at other key milestones.
- During data analysis, that might be quarterly.
- After publication, the review might be annual, to catch IT infrastructure changes.
- Then while the data is archived, it might be every few years.

The sections of a DMP can vary and you will be best served by creating an organizational template for consistency and efficiency, making changes as needed.

Your DMP template may have the following sections:

- Project Title and Information
- Data Description
- Roles & Responsibilities
- Standards Used
- Access Policies
- Sensitive Data Policies
- Sharing Policies
- Archiving and Preservation Plans, and
- Applicable laws and policies

You may have noticed that many potential sections of the DMP relate directly to curation actions shown in the data lifecycle. A DMP can serve as a guide to the curation of data and statistics.

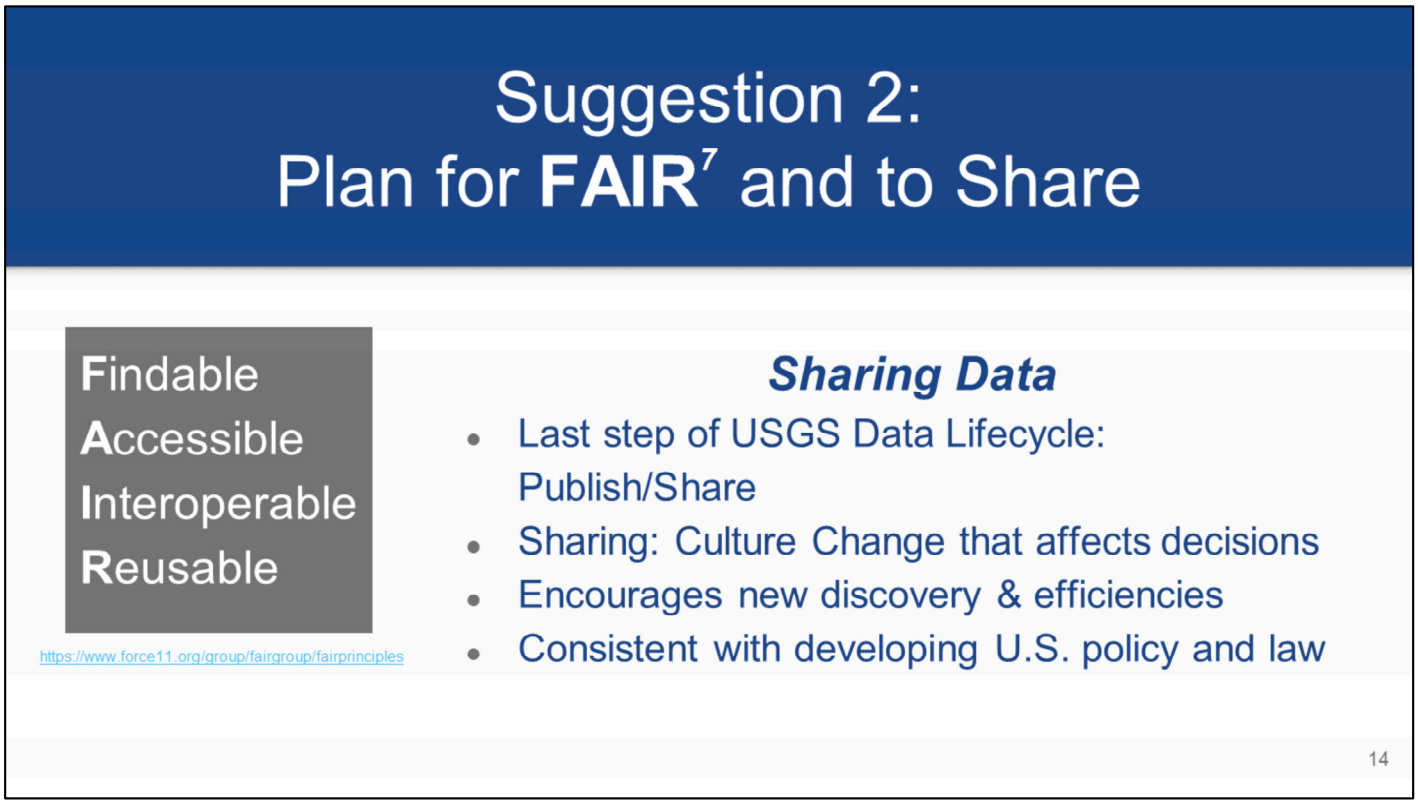
A final, public version of your DMP should be included as part of your data package when the data is shared or made public. This public version can be redacted a bit, if needed, for public consumption.

If you have planned to make your data collection and project actions as transparent as possible to your own team, it will be easier to make your statistics transparent to external audiences.

When we think back to the charge to the panel, DMPs fulfill all 4 charges:

1. DMPs are a best practice
2. DMPs serve as guidance, record standards, and describe and document tools, as well as layout archiving plans
3. Early planning helps to minimize cost.
    1. You can go back and document data after your create it, and migrate it to open formats after you create it, but that is an expense in both time and resources.
4. DMPs can be implemented today, or at any stage, to improve later performance. They are the lowest of the low hanging fruit.

[Slide]
[Next speaker: **Leighton**]
[Time: 2:00 minutes]
[Total time: 16:30 minutes]



Leighton:
Slide Title: Suggestion 2: Plan for FAIR and to Share

[FCSM Scripted Text]
Data professionals across the globe are working at these same issues. Groups of data professional have come up with various principles for making data more shareable and to improve preservation. One of these sets of principles, created by FORCE11 in 2014, is the FAIR principles[7]. The goals of FAIR are to make data and metadata more findable, accessible, interoperable, and reusable.
There are 15 steps or practices that researchers can apply to data and metadata to make them more FAIR.
I also believe the FAIR principles could easily be extended to paradata.
Adoption of some, or most, of these principles would go a long way to making federal statistics more transparent.
Here are a few examples from the FAIR Principles
To Be Findable:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.

To Be Accessible:

- A2 metadata are accessible, even when the data are no longer available.

To Be Interoperable:

- I2. (meta)data use controlled vocabularies that follow FAIR principles.

To Be Re-Usable:

- R1.1. (meta)data are released with a clear and accessible data usage license.

Looking at Data Sharing you may question, Why should we plan for the sharing of federal data and statistics?

As we see in the USGS data lifecycle model, sharing is an assumed part of the lifecycle. This is where scientific practice around digital data is heading, a fact embraced by evolving federal laws and policies. Globally, these actions are often referred to as "open science." For a brief overview of open science, please see my presentation "U.S. Open Science Policy Perspectives & Transportation: Open Science in Transportation: Challenges and Opportunities in a COVID-19 Era"`[12] at https://doi.org/10.21949/1520725

Moving towards sharing is a culture change, it is not simply a technological fix. That culture change should affect every decision about how we collect and analyze our data, and share our statistics. Sharing has to be acknowledge at the front, as it creates dependencies downstream in data collection projects.

[Slide]
[Time: 2:00 minutes]
[Total time: 18:15 minutes]

[Extended Workshop Text not presented at FCSM]
Leighton:
Data professionals across the globe are working at these same problems and issues. Groups of data professional have come up with various principles for making data more shareable and improve preservation. One of these sets of principles, created by FORCE11 in 2014, is the FAIR principles. The goals of FAIR are to make data more findable, accessible, interoperable, and reusable.
There are 15 steps or practices that researchers can apply to data and metadata to make them more FAIR.
I also believe the FAIR principles could easily be extended to paradata.
Adoption of some, or most, of these principles would go a long way to making federal statistics more transparent.

[Highlight a few from below.]

FAIR Data Principles:

1. Are an evolving best practice

2. Are principles, therefor Provide some guidance, encourage use of standards, but are tool agnostic.
3. Making data interoperable helps to minimize cost of future data collections and analysis.
4. Some FAIR Data principles can be implemented today, others will take time.

[Do not read this long section: For reference only]
For more on the FAIR Data Principles go to https://www.force11.org/group/fairgroup/fairprinciples

- **To be Findable:**
- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
  F2. data are described with rich metadata.
  F3. (meta)data are registered or indexed in a searchable resource.
  F4. metadata specify the data identifier.
- **TO BE ACCESSIBLE:**
- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
  A1.1 the protocol is open, free, and universally implementable.
  A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
  A2 metadata are accessible, even when the data are no longer available.
- **TO BE INTEROPERABLE:**
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  I2. (meta)data use vocabularies that follow FAIR principles.
  I3. (meta)data include qualified references to other (meta)data.
- **TO BE RE-USABLE:**
- R1. meta(data) have a plurality of accurate and relevant attributes.
  R1.1. (meta)data are released with a clear and accessible data usage license.
  R1.2. (meta)data are associated with their provenance.
  R1.3. (meta)data meet domain-relevant community standards.

Leighton:

Looking at Data Sharing you may question, Why should we plan for the sharing of federal data and statistics?

As we see in the USGS data lifecycle model, and most others, sharing is an assumed part of the lifecycle. This is where scientific practice around digital data is heading, a fact embraced by evolving federal laws and policies.

Moving towards sharing is a culture change, it is not simply a technological fix. That culture change should affect every decision about how we collect and analyze our data, and share our statistics. Sharing has to be acknowledge at the front, as it creates dependencies down stream in data collection projects.

Some key benefits of data sharing are

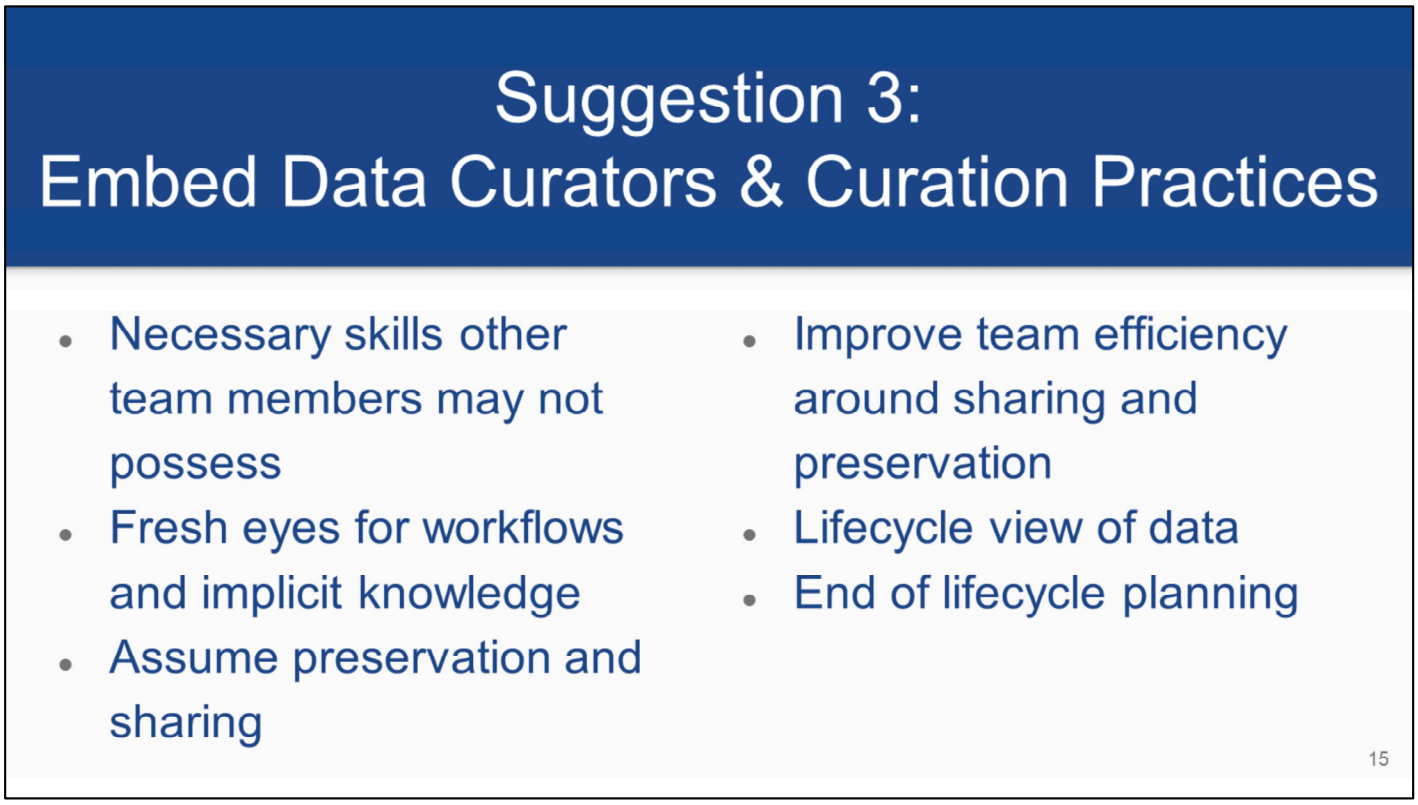
- It encourages new discovery.
- It enables re-use, by the original data collectors, their federal partner agencies, or other researchers. This means a potential greater return on investment., and again

- Data and statistical sharing is consistent with both policy and law.

Keeping the charge to the panel in mind, data sharing:

1. Is an evolving best practice
2. Planning for sharing helps to minimize cost of future data collections and analysis.
3. Sharing can be implement right away, and can be graduated, based on the role of the user in relationship to the data: some are allowed to see more sensitive data, others only the public release.

[Slide]
[Next speaker: Jesse]
[Time: 1:00 minutes]
[Total time: 17:30 minutes]



Leighton:
Slide Title: Suggestion 3: Embed Data Curators & Curation Practices

[FCSM Scripted Text]
The final suggestion is, Embed Data Curators.

This suggestion may seem a little self-serving at first glance. However, if your goal is to gather a team of professionals best able to carry out a data collection project, analyze data into statistics, and document, preserve, and share the statistics, your team deserves a trained, professional data curator.

- Data curators possess technical and research skills other team members won't have, but will contribute directly to data and statistical transparency.
- Data curators, can serve as fresh eyes on repeated data collection projects and make explicit knowledge that is implicit and "obvious" to the team.
- Data curators work under the assumption that data should be shared, while remaining aware of data sensitivity.
- Data curation practices will improve team efficiency around sharing, preservation, and transparency, by default.
- Curators take a lifecycle view of the data, and can relieve other team members of that duty.

- Data curators can also plan for end of data lifecycle events and disposition, in ways consistent with established best practices.

In BTS, both Jesse and I have experience coming in after fact to try to package and share data. One example is the Omnibus Household Survey. The datasets were 15 to 20 years old, and much documentation was trapped in HTML or missing altogether.

Jesse and are now working with our BTS and Census colleagues to plan for the preservation and sharing of the Vehicle Inventory and Use Survey (VIUS) data as that survey gets resurrected. We expect that by being engaged from the beginning we will be able to help BTS, Census, and researchers get more value from VIUS going forward and backward in time.

[Slide]
[Next speaker: Leighton]
[Time: 1:30 minutes]
[Total time: 20:00 minutes]

[Extended Workshop Text, not presented at FCSM]
The final suggestion is, Embed Data Curators.

This suggestion may seem a little self-serving at first glance. However, if your goal is to gather a team of professionals best able to carry out a data collection project, analyze data into statistics, and document, preserve, and share the statistics, your team deserves a trained, professional data curator.

- Data curators possess technical and research skills other team members won't have, but will contribute directly to data and statistical transparency.
- Data curators, can serve as fresh eyes on repeated data collection projects and make explicit knowledge that is implicit and "obvious" to the team.
- Data curators work under the assumption that data should be shared, while remaining aware of data sensitivity.
- Data curation practices will improve team efficiency around sharing, preservation, and transparency, by default.
- Curators take a lifecycle view of the data, and can relieve other team members of that duty.
- Data curators can also plan for end of data lifecycle events and disposition, in ways consistent with established best practices.

Over the past few months I have been working with a legacy dataset, the Omnibus Household Surveys. The data collected from these surveys is now roughly 15 to 20 years

old and as I have worked to create complete data packages for each dataset. Over the course of this work I have run into many issues that we mentioned earlier in the presentation, such as inconsistent naming structure, unknown locations for the data, and limited documentation.

- **An example (if time permits), is a file that I came across was simply named "disposition." Within the documentation I had I was unable to understand the purpose of the data within this file, and was further confused since the work "disposition" was used in the documentation I did have to reference various variables. The lack of documentation around this file resulted in me spending wasted time with this file, the data dictionary, and the main dataset, until I was finally able to piece together its purpose. This example. . .**

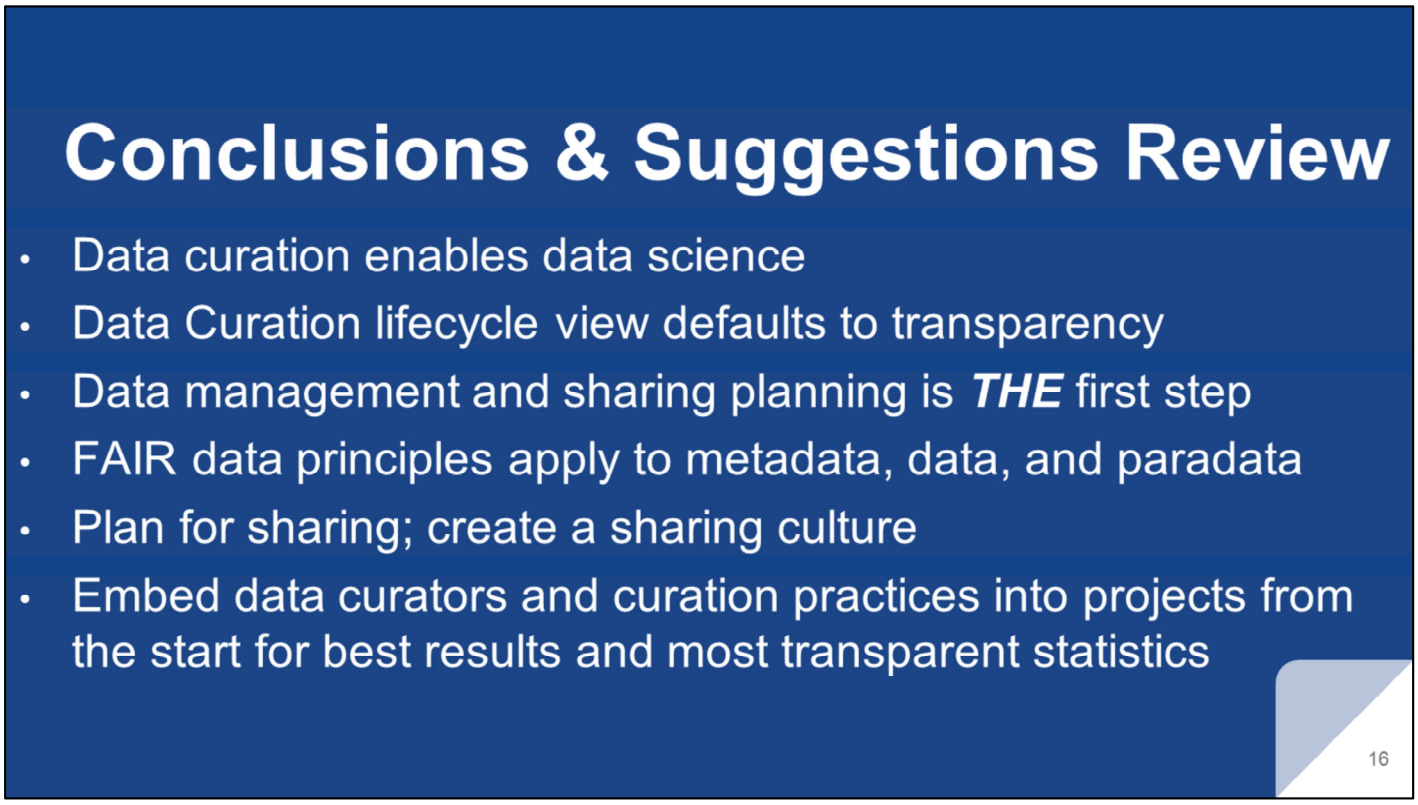
Further demonstrating the need for embedded data curator throughout the lifecycle and confirming the difficulties and limited abilities when the actions are reactive.

In the future, I will serve as an embedded data curator for Office of Airline Information (OAI), to prevent such issues from occurring. I will implement the suggestions and strategies we have outlined in this presentation to achieve greater transparency when it comes to BTS data.

Keeping the charge to the panel in mind, data curators and curation actions:

1. Are an evolving best practice
2. Are resources for guidance, standards, and tool.
3. Curation practices help to minimize cost of future data collections and analysis.
4. Some curation practice can be implemented today, others will take to mature.

[Slide]
[Next speaker: **Leighton**]
[Time: 2:00 minutes]
[Total time: 19:00 minutes]



Leighton:
Slide Title: Conclusions & Suggestions Review

[FCSM Scripted Text]
Conclusions & Suggestions Review

- Data curation enables data science
- Data Curation lifecycle view defaults to transparency
- Data management and sharing planning is ***THE*** first step
- FAIR data principles apply to metadata, data, and paradata
- Plan for sharing; create a sharing culture
- Embed data curators and curation practices into projects from the start for best results and most transparent statistics

[Slide]
[Next speaker: Leighton]
[Time: 0:30 minutes]
[Total time: 20:30 minutes]

Leighton:
Throughout the presentation and the slide notes you will find references to the following materials.

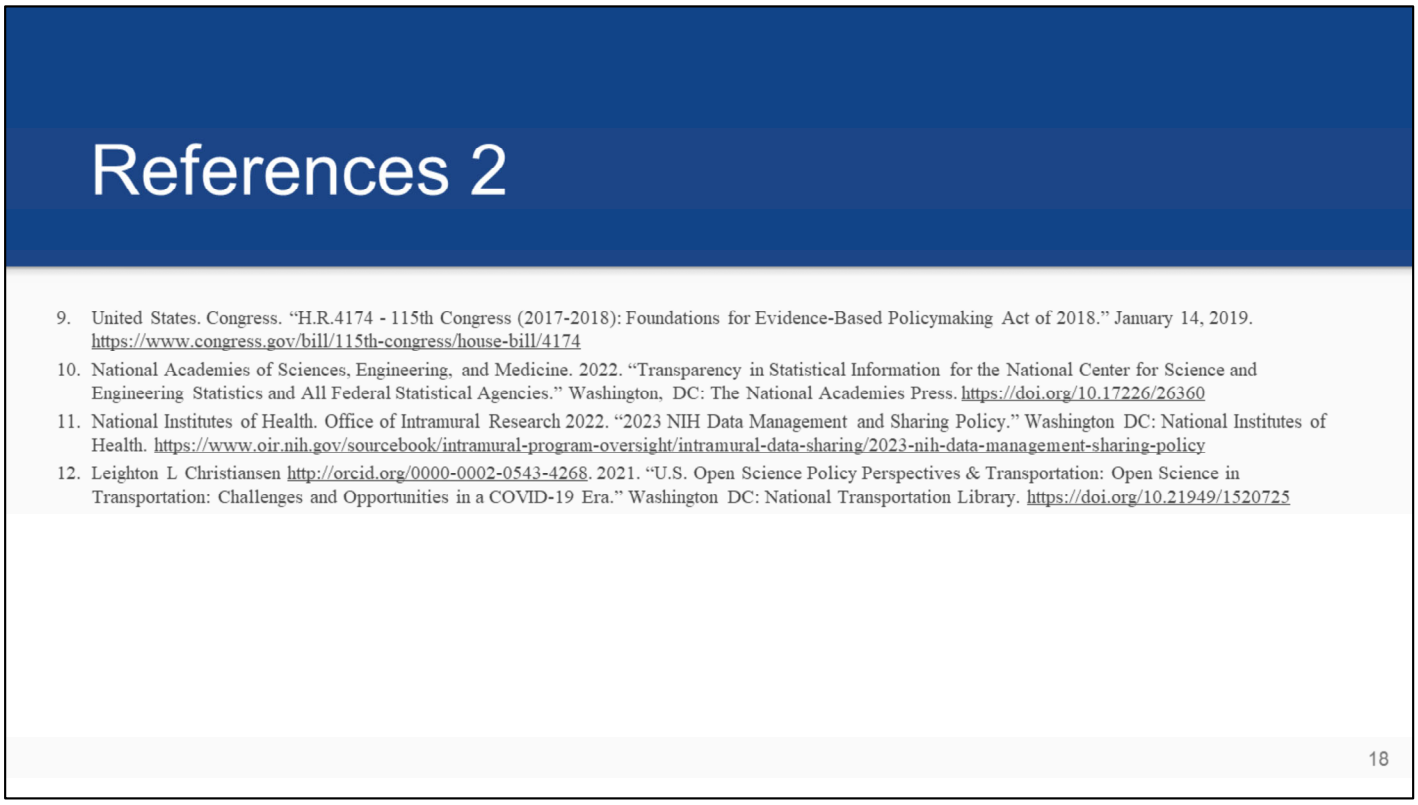
[Slide text]

1. University Library, Texas A&M University. "Data Management Defined - Research Data Management - Guides at Texas A&M University." Research Data Management, October 1, 2013. http://guides.library.tamu.edu/DataManagement
2. Briney, Kristin. 2015. Data management for researchers: organize, maintain and share your data for research success. http://www.pelagicpublishing.com/data-management-for-researchers.html
3. Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. "Specialization in Data Curation," 2013. http://www.lis.illinois.edu/academics/programs/specializations/data_curation
4. Definition based on Ani Adhikari and John DeNero, "The Foundations of Data Science" http://www.inferentialthinking.com/index.html "What is Data Science" http://www.inferentialthinking.com/chapter1/what-is-data-science.html
5. Digital Curation Centre. Data Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model
6. Faundeen, J.L., Burley, T.E., Carlino, J.A., Govoni, D.L., Henkel, H.S., Holl, S.L.,

Hutchison, V.B., Martín, Elizabeth, Montgomery, E.T., Ladino, C.C., Tessler, Steven, and Zolly, L.S., 2013, The United States Geological Survey Science Data Lifecycle Model: U.S. Geological Survey Open-File Report 2013–1265, 4 p., http://dx.doi.org/10.3133/ofr20131265

7. FORCE11. "The FAIR Data Principles." 2016. https://www.force11.org/group/fairgroup/fairprinciples
8. Allen, Robert, & Hartland, David. (2018, May 21). FAIR in practice - Jisc report on the Findable Accessible Interoperable and Reuseable Data Principles (Version 1). Zenodo. http://doi.org/10.5281/zenodo.1245568

[Slide]
[Next speaker: Leighton]
[Time: 0:30 minutes]
[Total time: 38:37 minutes]

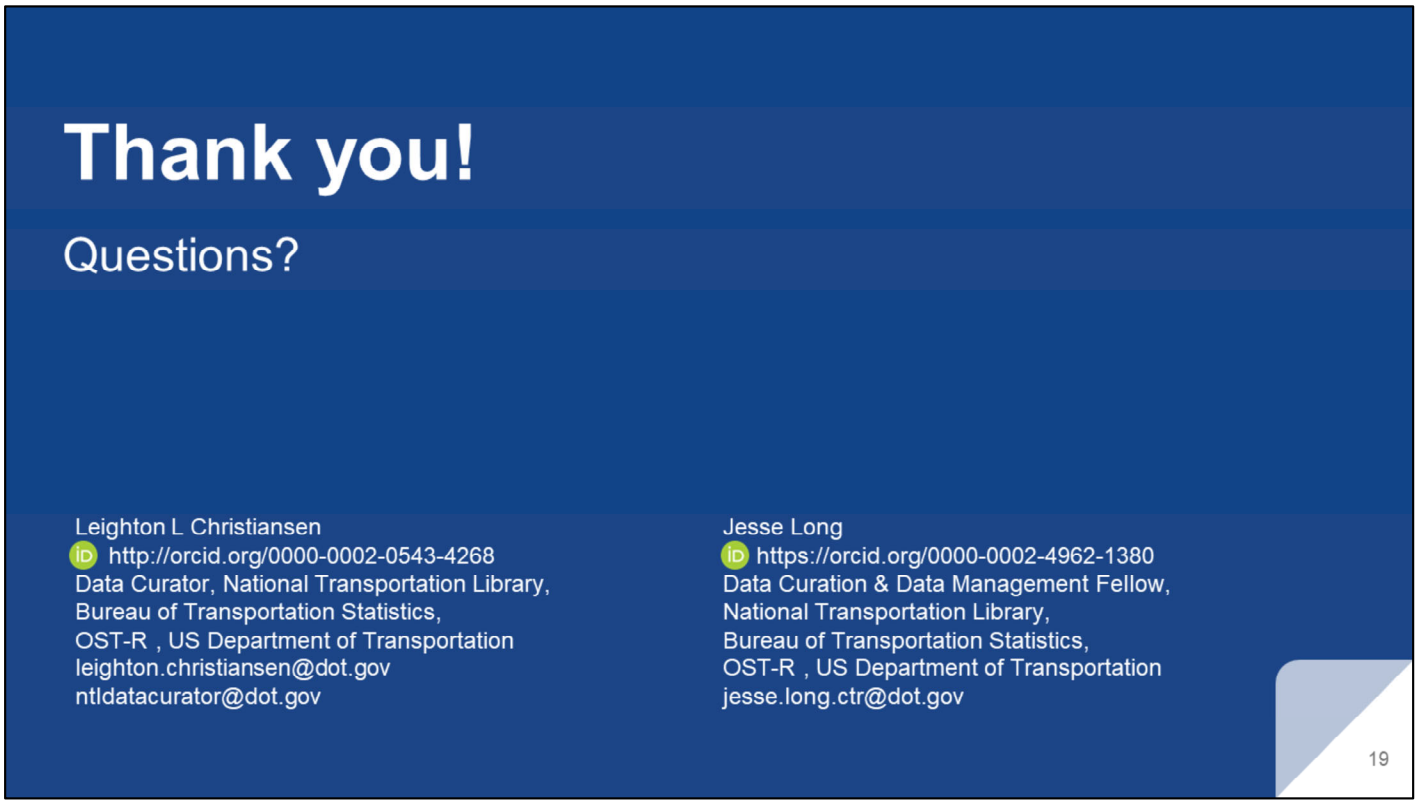


Leighton:
Throughout the presentation and the slide notes you will find references to the following materials.

[Slide Text]

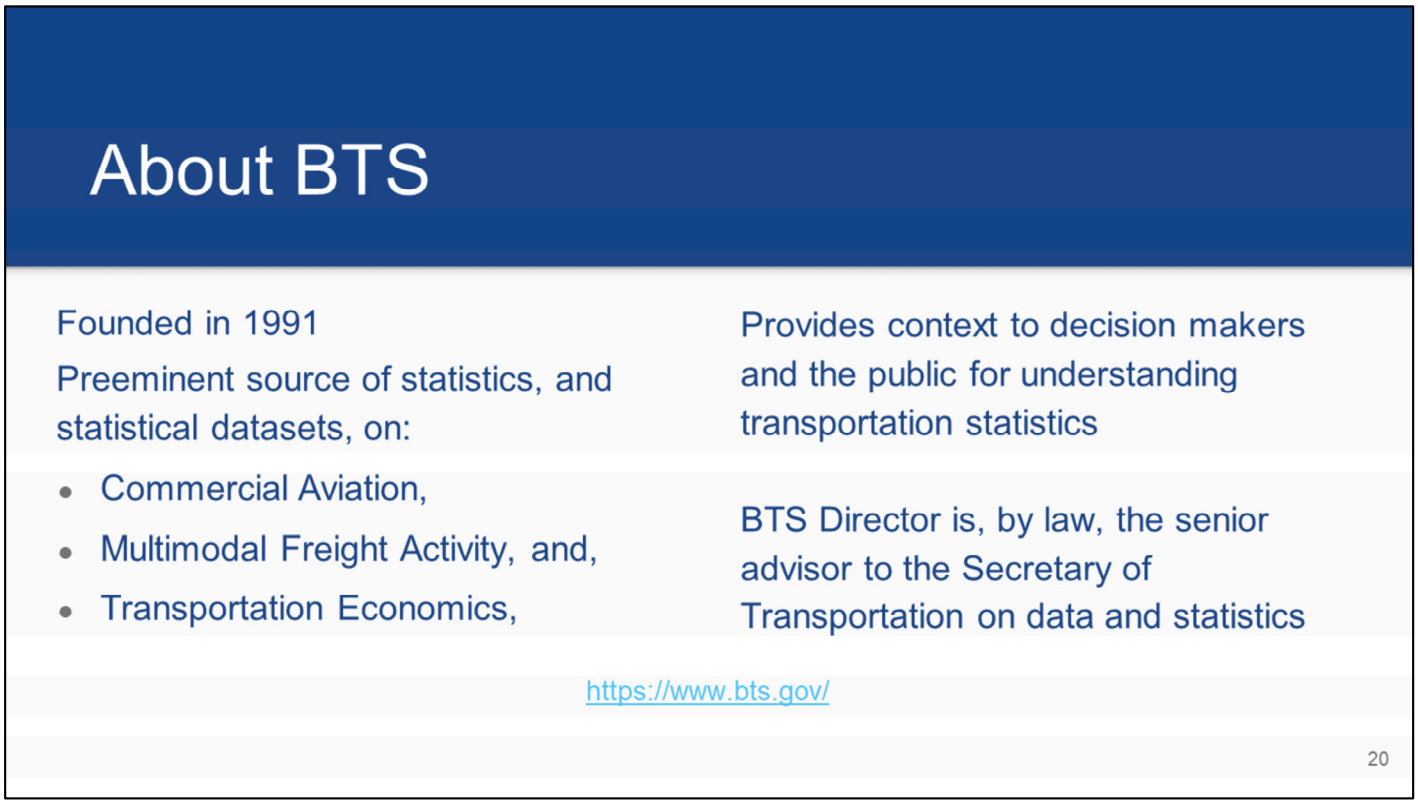
9. United States. Congress. "H.R.4174 - 115th Congress (2017-2018): Foundations for Evidence-Based Policymaking Act of 2018." January 14, 2019. https://www.congress.gov/bill/115th-congress/house-bill/4174
10. National Academies of Sciences, Engineering, and Medicine. 2022. "Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies." Washington, DC: The National Academies Press. https://doi.org/10.17226/26360
11. National Institutes of Health. Office of Intramural Research 2022. "2023 NIH Data Management and Sharing Policy." Washington DC: National Institutes of Health. https://www.oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy
12. Leighton L Christiansen http://orcid.org/0000-0002-0543-4268. 2021. "U.S. Open Science Policy Perspectives & Transportation: Open Science in Transportation: Challenges and Opportunities in a COVID-19 Era." Washington DC: National Transportation Library. https://doi.org/10.21949/1520725

[Slide]
[Next speaker: Leighton]
[Time: 0:30 minutes]
[Total time: 38:37 minutes]



Leighton: We would now be happy to answer any questions.

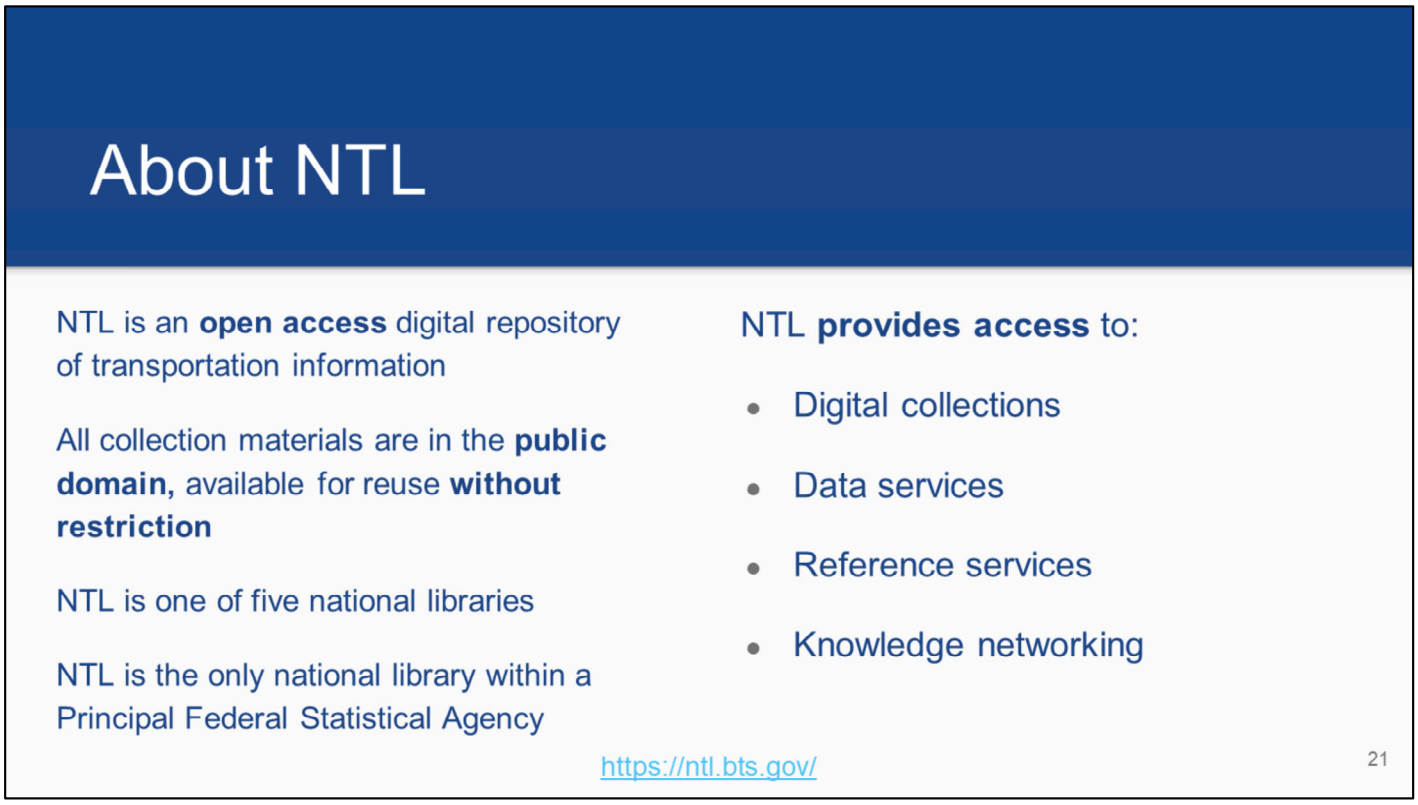
[Slide]
[Next speaker: Leighton]
[Time: 0:23 minutes]
[Total time: 39:00 minutes]



Leighton:
In the interest of time, I am going to keep this intro short. I just want to say that due to a number of laws and policies, BTS has a deep commitment to making our products sharable, and increasingly transparent. You can read more detail about BTS in the slide notes of this presentation, or at bts.gov
[Skip these 2 bullets: for reader reference]

- The Bureau of Transportation Statistics (BTS), part of the Department of Transportation (DOT) is the preeminent source of statistics on commercial aviation, multimodal freight activity, and transportation economics, and provides context to decision makers and the public for understanding statistics on transportation. BTS assures the credibility of its products and services through rigorous analysis, transparent data quality, and independence from political influence. BTS promotes innovative methods of data collection, analysis, visualization, and dissemination to improve operational efficiency, to examine emerging topics, and to create relevant and timely information products that foster understanding of transportation and its transformational role in society. The Bureau's National Transportation Library (NTL) is the permanent, publicly accessible home for research publications from throughout the transportation community; the gateway to all DOT data; and the help line for the Congress, researchers, and the public for information about transportation.
- The BTS Director is by law the senior advisor to the Secretary of Transportation on data and statistics

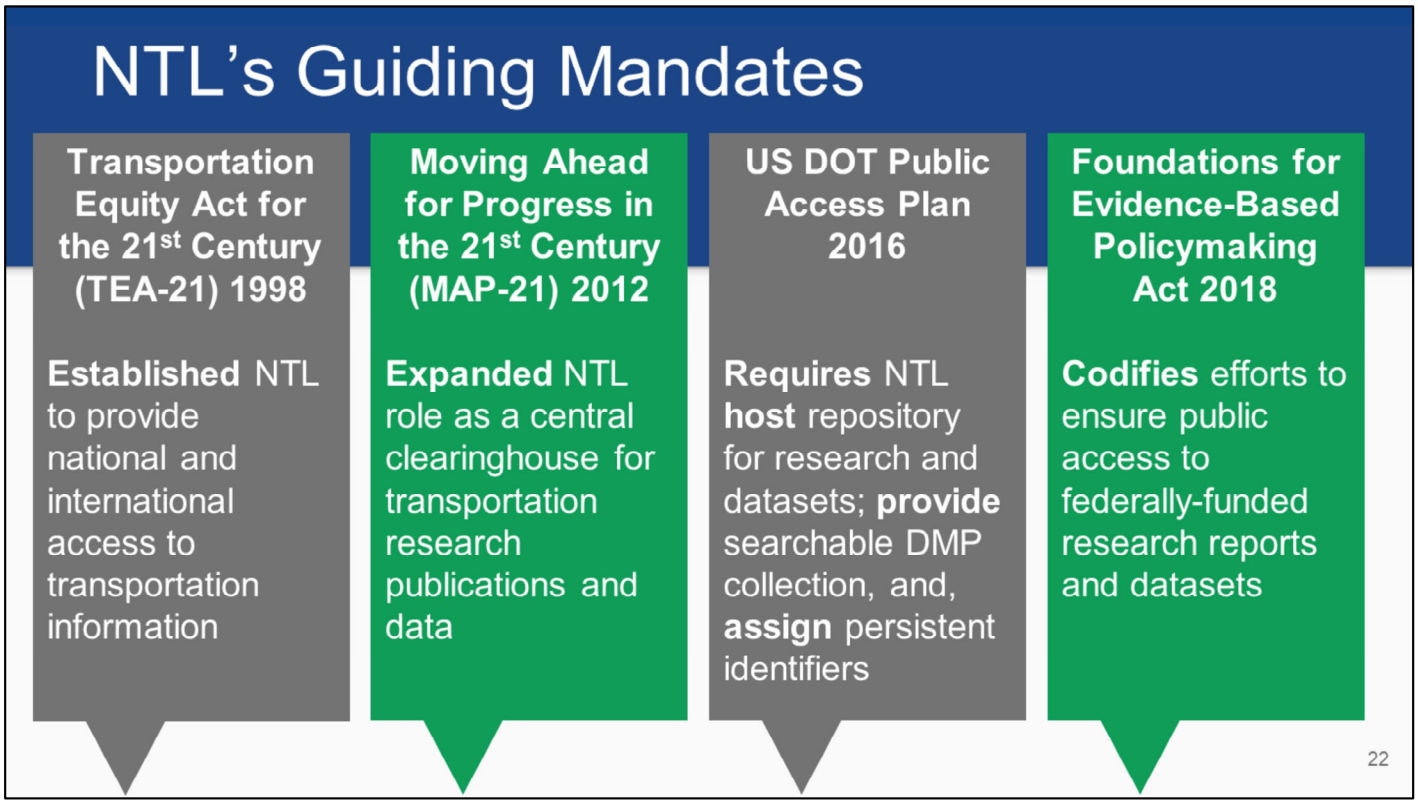
[Slide]
[Next speaker: Leighton]
[Time: 39 seconds]
[Total time: 1:53]



Leighton:
Very quickly: Fulfilling the mandates establishing it, the NTL is an open access digital repository of transportation information, providing access to digital collections, data services, reference services, and facilitating knowledge networking in the transportation research community.
NLT is the only library within a Principal Federal Statistical Agency. And based on this unique outlook, our direct relationship to public access to and preservation of statistical data, we were invited to address the NAS "Committee on Transparency and Reproducibility of Federal Statistics for NCSES". This presentation is based on that work and a similar presentation made to NCSES in October 2019.

All materials in the NTL collection are in the public domain, available for reuse without restriction.
The five national libraries are: Library of Congress, Washington, D.C.; National Library of Education; National Transportation Library; National Library of Medicine, Bethesda; National Agricultural Library

[Slide]
[Next speaker: Leighton]
[Time: 42 seconds]
[Total time: 2:38]



Leighton:

NTL was created in 1998 under the Transportation Equity Act for the 21st Century The act mandated the "establish[ment] and maintain[tenence] of a National Transportation Library" to host "a collection of statistical and other information needed for transportation decision making at the Federal, State, and local levels."

To fulfill this mandate, NTL:

1 Provides national and international access to transportation information by maintaining a digital repository of full-text documents and datasets

1 Coordinates information creation and dissemination

1 Offers reference services to the transportation community

The NTL mandate was extended and expanded in 2012 under the Moving Ahead for Progress in the 21st Century Act (MAP-21) . NTL is tasked to:

1 "Acquire, preserve and manage transportation information and information products and services for use by DOT, other Federal agencies, and the public;

1 To serve as: " a central repository for DOT research results and technical publications;" & as the Central clearinghouse for transportation data information of the Federal Government at the US DOT.

1 Coordinate & lead policy for transportation information access

1. Develop a “comprehensive transportation information and knowledge network”
1. Publicize, facilitate, and promote access to transportation information products and services

In 2013, the White House Office of Science and Technology Policy issued a memorandum, *Increasing Access to the Results of Federally Funded Scientific Research*, requiring all Executive Departments and Agencies spending more than $100 million/year on R&D to ensure public access to peer-reviewed publications and digital datasets arising from federally-funded scientific research.

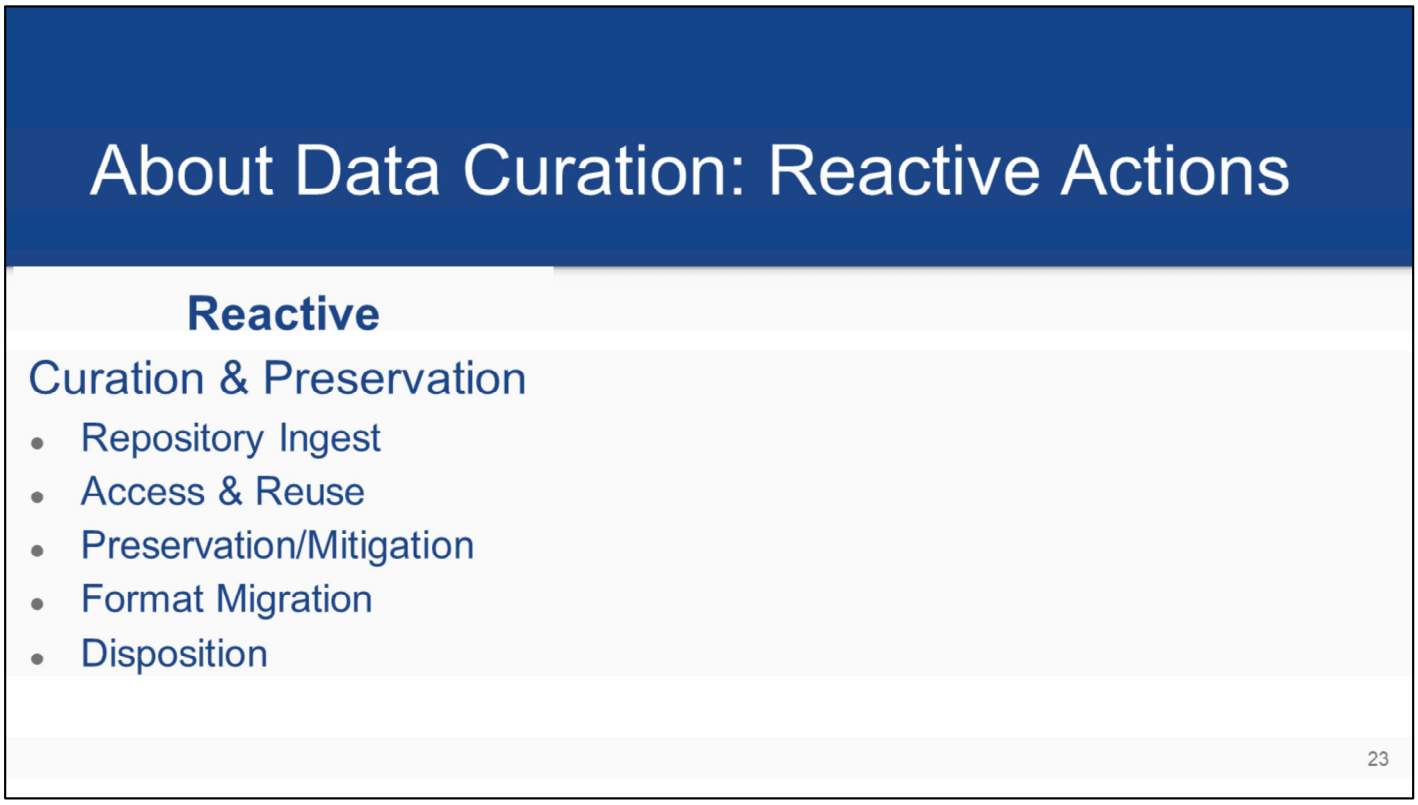
From 2013, NTL has been the centerpiece of US DOT’s response the memo, serving as the public repository and point of access for DOT-funded research.

The US DOT “Plan to Increase Public Access to the Results of Federally-Funded Scientific Research” went into effect on January 1, 2016. Within that plan, NTL is specifically called out to: host repository for federally funded research reports and datasets; provide searchable collection of public access data management plans; and, assign DOIs to resources.

Title II of the Foundations for Evidence-Based Policymaking Act of 2018, signed in February 2019, codifies federal agencies’ and NTL’s work to ensure public access to federally-funded research reports and datasets.

Now, because BTS is unique among federal statistical agencies in having its own library, and data curator and data curation team, we hope that BTS can offer some new outlooks on making federal statistics transparent. I believe there are practices within data curation that you all will find helpful.

[Slide]
[Next speaker: Leighton & Jesse]
[Time: 1:31 seconds]
[Total time: 3:27]



Leighton: (fewer examples or shorter)

So if you think about the common usage of the words "curation" or "curator," you most likely think of a museum curator, who focuses on the preservation of objects that have already been created. And data curators certainly can and do perform many of those same "reactive" curation and preservation tasks.
These can include:

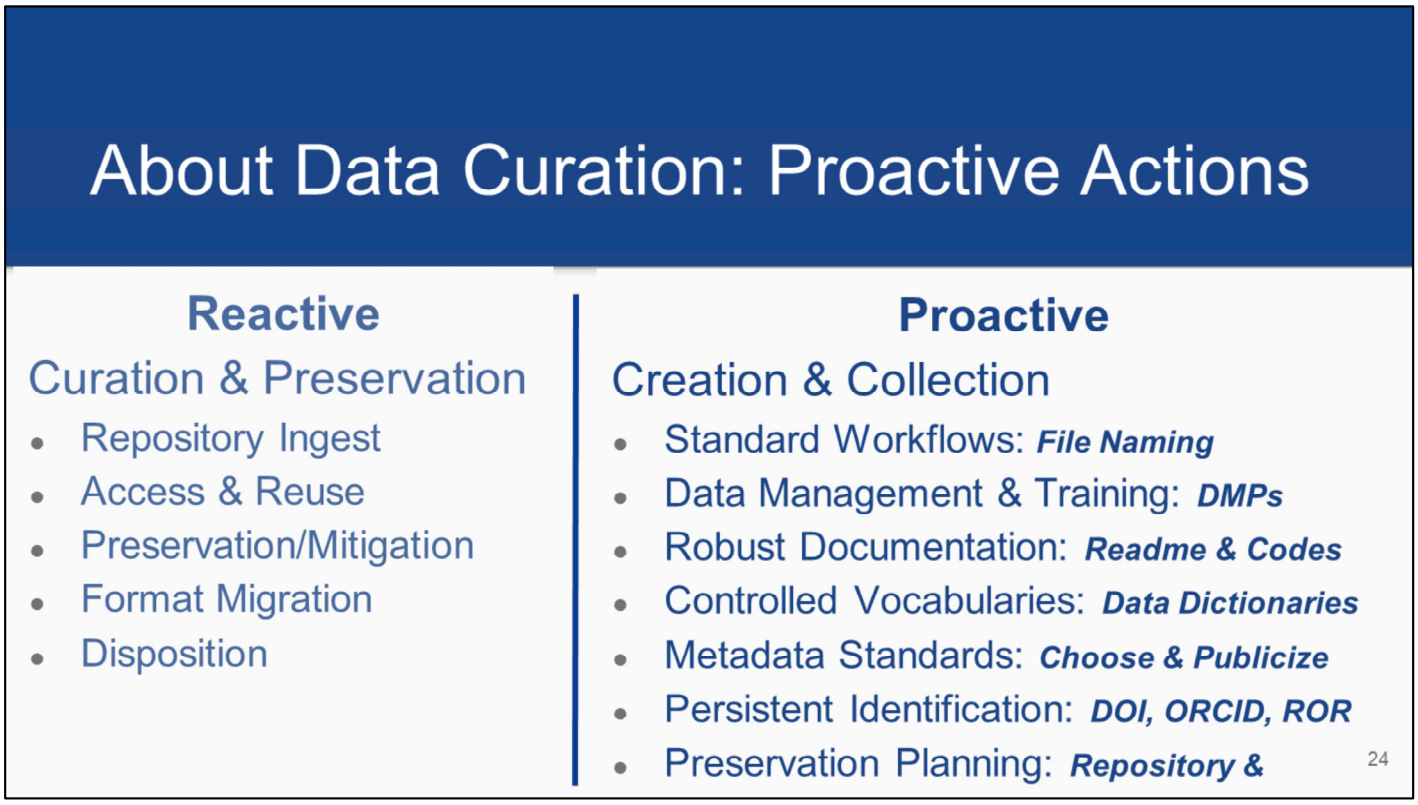
- Repository Ingest: Seeking and accepting objects to be added to a museum special collection, if a physical object, or adding a local copy of a digital object to digital repository, based on you collection development policy.
- Access and Reuse: Making objects finable and accessible to researchers and the public. For physical objects, this might mean putting an object on display. For digital objects, this could mean exposing the object metadata to search engines.
- Preservation/Mitigation: For a physical object preservation and mitigation may mean housing the object in an environmentally controlled labs space to slow the effects of aging. For digital objects, this can mean holding master copies in dark storage, performing regular fixity checks looking for bit rot, and migrating master and use copies from drives to new drives every few years to avoid data loss from media degradation.
- Format Migration: Now for format migration, I want to start with the digital objects. When we talk about format migration, we usually mean taking the data recorded by

people, sensors, or machines, and converting that data file from its original, proprietary format, into a more universally accessible or open format. For example, say you hired a contractor to take hand-held tablets out on to street corners to survey citizens on their opinions on a specific topic. The contractor would likely write an app that would store the data in a form that was easily stored and rendered on the tablet. But would that data be easy to read on your desktop machine using SAS or a spreadsheet program? Maybe not, and you would need to migrate that data from the tablet format to something ubiquitous such as comma separated value, or CSV, so that you would us it on any machine, far into the future. What you have done is preserved the intellectual content of the data object, even if you have not preserved it in its original form. We do the same thing when we digitize books or printed reports. We often unbind them, destroying the original container, in order to rapidly bulk scan the pages, preserving the text as PDFs or plain text documents. Format migration for museum pieces might include making a mold and a plaster copy of a famous sculpture, or making a scale model of object too large to preserve, such as the Colosseum in Rome: We have preserved some aspect of the original item by migrating that information to a new format or media, even if we cannot save the original experience or materials of the item.

- Disposition: There may come a time in an objects life when we decide to dispose of it. This can take a number of different forms. Museums and galleries may sell a particular work because they want to raise money for capital projects or to make other purchase. A gallery may decide that a physical, sacred object should be returned to the culture that created it. For data curators, we may decide to delete a large dataset is no longer of interest to the scientific community, as no one has requested access in a number of decades, or the dataset has been largely rebuffed by the research community because of questionable collection methods or erroneous data. We might even be forced into such a situation due to resources constraints: we have run out of server space and need to make room for new, cutting edge data, therefore, some legacy data has to go.

You can see that data curators perform many of the same ***reactive*** actions as physical object curators. However, digital data curators also want to be in a position to take ***proactive*** steps as well, and Jesse will describe these.

[Slide]
[Next speaker: **Jesse**]
[Time: 4:00 seconds]
[Total time: 10:00]



Jesse: [take your time. Important new information]

In order to be more proactive data curators want to be embedded in data collection projects from the very beginning. By implementing data management strategies at the time of data creation we can improve data preservation outcomes for years or decades in the future. Approaching data curation and preservation for legacy or already existing datasets, is often harder, and suffer from incomplete knowledge or information due to limited documentation.

Outcomes can be improved by taking ***proactive actions*** and planning for long-term data preservation and sharing from the beginning of a project.

The proactive actions that curators want to help data collection teams implement are:

- Standard Workflows: Data curators can help data collectors document and standardize work flows and data stewardship practices. A very simple practice that many teams overlook is a standard and documented file naming convention. Files names should be human readable, contain some project intelligence, and include a date and timestamp for version control. There is nothing worse than having 16 files in your folder called “data” or “full text”. Determining a file naming structure is an important step to take before any data is ever collected.
- Data Management & Training: Another crucial step to take before any data is

collected is writing a data management plan (or DMP). You might be ready to ask "Does every data collection activity need a DMP?" My response is that every data collection action ***deserves*** a robust data management plan. DMPs can go a long way to making data preservable, interoperable, and transparent. Data curators can help a data collection team draft, revise, implement, and update their DMP.

- Robust Documentation: An embedded data curator can assist the team with creating robust documentation. This can include:
  - writing up readme files and data dictionaries;
  - checking for the presence of code tables, data dictionaries, and supplementary files;
  - researching, suggesting, and implementing domain appropriate metadata schema; and,
  - Building a complete data package to improve preservation and transparency for the dataset.
- Controlled Vocabularies: A data curator can research and suggest implementation of an existing controlled vocabulary to make variable name, meanings, and specifications standard and interoperable. Using existing controlled vocabularies makes writing a data dictionary much easier. Additionally, the data curator can help crosswalk controlled vocabularies to make translating between vocabularies easier.
- Metadata Standards: In addition to controlled vocabularies, data curators can suggest appropriate metadata standards, or help crosswalk between existing and new standards. Either way, the data curator will help the data collection team choose the necessary standards, as well as document and publicize the metadata standards in use. Publicizes chosen metadata standards is a great step towards increasing transparency, and helps data users read and use your data.
- Persistent identification: Persistent identifiers (or PIDs) eliminate ambiguity and confusion with published research, because they provide unique identification. There are PIDs for objects, people, and organizations.
  - For Objects: there are Digital Object Identifiers (DOIs). DOIs are typically used for publications, images, audio files, measurement instruments, or any other THING that can either have a digital presence in a networked environment, or can be described by a web page or digital metadata file. The DOI may point either directly to the object or its digital "landing page." There are many "brands" of PIDs for things, I only mention DOIs here because they are most common.
  - Open Researcher and Contributor Identifiers (or ORCIDs) are used to uniquely identify people. On the title slide, you saw both of our ORCIDs. My ORCID iD which is a https link, containing the protocol and 16 digits, that leads to a web page where I have a profile that records my works and uniquely disambiguates me from all other humans named Jesse Long.
  - Finally, for organization there is the Research Organization Registry (or ROR). It is a fairly new initiative to build an open, collaborative research organization

identification schema and registration service. The ROR identifier system is controlled by the research organizations themselves, and seeks to be interoperable among systems. This is a different model than identifiers such as FundRef, which are controlled by scholarly journal publishers, and are often used only within the family of journals put out by that publisher.

- The final proactive action step I will mention is Preservation Planning: Data curators can help data collection teams identify, ahead of data collection, likely target repositories for the data types generated from the project, storage size needs based on the planned data collection, plan for local backup strategies, and vet repositories based on those strategies and backup server locations.

Now, in the descriptions of reactive and proactive curation actions, you probably recognize some actions your data collection teams already take.

However, one of our suggestions to you is that unless each of these steps is included in your data collection, your data and your statistics cannot reach maximum transparency. Transparency that is planned for and included at the beginning of data collection is more efficient and impactful than transparency created after the fact.

[Slide]
[Next speaker: Jesse]
[Time: 5:48 minutes]
[Total time: 16 minutes]

Leighton:

Potential challenge to implementing good statistical transparency is that according to a 2018 JISC report "FAIR in Practice," tools to automate metadata capture and documentation are still lacking. Tool creation needs a great deal more effort and collaboration.

[Slide]
[Next speaker: Leighton]
[Time: 0:30 minutes]
[Total time: 18:30 minutes]