DataONE Webinar Series

Open Science and Data: Towards Optimizing the Research Process

Dr Jean Claude Guédon
Université de Montréal
Enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it.

DataONE network of Member Nodes: www.dataone.org/current-member-nodes

DataONE search tool: https://cn.dataone.org/onemercury/
Lesson 1: Data Management

The world of data around us

The data deluge has created a surge of information that needs to be well-managed, discoverable, and accessible.

The amount of available storage is not keeping pace with the amount of data being produced.

Causes of data loss
- Natural disasters
- Facilities infrastructure failures
- Storage failure
- Server hardware or software failure
- Application software failure
- Human errors
- Malicious attack
- Format obsolescence
- Loss of competencies
- Loss of funding
- Loss of institutional commitment

Costs of not doing data management can be very high!

Why manage data:
- the researcher perspectives
- Keep yourself organized = find your own file!
- Track your processes for reproducibility
- Better version control of data
- More efficient data quality control
- More backups to avoid data loss
- Format your data for reuse by yourself & others
- Document your data for understandability and reuse
- Prepare it to share it & gain credibility and recognition for your scientific efforts

Data management facilitates sharing and reuse.

Data Reuse Example
Researchers reused and aggregated data from several different sources to determine migration routes for specific bird species.

The Case for Data Management
If data are:
- Well-organized
- Documented
- Protected
- Accessible
- Verified as to accuracy & validity

The results are:
- High quality data
- Data that is easy to share and reuse
- Citation & credibility to researcher
- Cost savings to further science

Lesson 2: Data Sharing

Address data sharing throughout the data lifecycle

Description of data to be produced

How data will be acquired

How data will be processed

How data will be acquired

Make it easier to keep your data organized for (and how it turned out!)

Previous what has been done in your lab in the past

Prevents duplication of effort > it will be more obvious

To the research community

To the research sponsor/funder

To the public

Lesson 3: Data Management Planning

The Data Management Plan (DMP)

A DMP outlines what you will do with your data during and after you complete your research project.

It is a formal (but evolving) document in which you lay out a plan to ensure that your data will be safe for during and after you complete your research project.

A DMP outlines what you will do with your data

- Quality assurance and control processes to be
- File formats of data
- How data will be processed

Better access to information leads to better understanding and contributions toward potential collaborations

Improved data quality

An authoritative source

Costs of not doing data management can be very high!

Loss of institutional commitment

Loss of competencies

Format obsolescence

Server hardware or software failure

Storage failure

Facilities infrastructure failures

Natural disasters

Causes of data loss

The Data Lifecycle

The stages through which well-managed data passes from project inception to conclusion.

Plan
- Design
- Data Collection
- Data Processing Desc.
- Abstract

Integrate
- Data Retrieval
- Data Staging
- Data Processing

Analyze
- Data Quality Assurance
- Data Quality Assurance
- Data Quality Assurance

Discoverable, robust metadata.

What metadata are needed?

How will metadata be created and/or captured?

Include unique IDs & citation information.

Publish metadata via a portal or clearinghouse.

Create discoverable, robust metadata.

Concerns about data sharing

Concerns about data sharing

License
- Data Reuse
- Data Sharing
- Data Management

Copyright issues

Ethical issues

Obligations

What data will be preserved?

Lesson 2: Data Sharing

Address data sharing throughout the data lifecycle

Description of data to be produced

How data will be acquired

How data will be processed

How data will be acquired

Make it easier to keep your data organized for (and how it turned out!)

Previous what has been done in your lab in the past

Prevents duplication of effort > it will be more obvious

To the research community

To the research sponsor/funder

To the public

Lesson 3: Data Management Planning

The Data Management Plan (DMP)

A DMP outlines what you will do with your data during and after you complete your research project.

It is a formal (but evolving) document in which you lay out a plan to ensure that your data will be safe for during and after you complete your research project.

A DMP outlines what you will do with your data

- Quality assurance and control processes to be
- File formats of data
- How data will be processed

Better access to information leads to better understanding and contributions toward potential collaborations

Improved data quality

An authoritative source

Costs of not doing data management can be very high!

Loss of institutional commitment

Loss of competencies

Format obsolescence

Server hardware or software failure

Storage failure

Facilities infrastructure failures

Natural disasters

Causes of data loss

The Data Lifecycle

The stages through which well-managed data passes from project inception to conclusion.

Plan
- Design
- Data Collection
- Data Processing Desc.
- Abstract

Integrate
- Data Retrieval
- Data Staging
- Data Processing

Analyze
- Data Quality Assurance
- Data Quality Assurance
- Data Quality Assurance

Discoverable, robust metadata.

What metadata are needed?

How will metadata be created and/or captured?

Include unique IDs & citation information.

Publish metadata via a portal or clearinghouse.

Create discoverable, robust metadata.

Concerns about data sharing

Concerns about data sharing

License
- Data Reuse
- Data Sharing
- Data Management

Copyright issues

Ethical issues

Obligations

What data will be preserved?

Lesson 2: Data Sharing

Address data sharing throughout the data lifecycle

Description of data to be produced

How data will be acquired

How data will be processed

How data will be acquired

Make it easier to keep your data organized for (and how it turned out!)

Previous what has been done in your lab in the past

Prevents duplication of effort > it will be more obvious

To the research community

To the research sponsor/funder

To the public

Lesson 3: Data Management Planning

The Data Management Plan (DMP)

A DMP outlines what you will do with your data during and after you complete your research project.

It is a formal (but evolving) document in which you lay out a plan to ensure that your data will be safe for during and after you complete your research project.

A DMP outlines what you will do with your data

- Quality assurance and control processes to be
- File formats of data
- How data will be processed

Better access to information leads to better understanding and contributions toward potential collaborations

Improved data quality

An authoritative source

Costs of not doing data management can be very high!

Loss of institutional commitment

Loss of competencies

Format obsolescence

Server hardware or software failure

Storage failure

Facilities infrastructure failures

Natural disasters

Causes of data loss

The Data Lifecycle

The stages through which well-managed data passes from project inception to conclusion.

Plan
- Design
- Data Collection
- Data Processing Desc.
- Abstract

Integrate
- Data Retrieval
- Data Staging
- Data Processing

Analyze
- Data Quality Assurance
- Data Quality Assurance
- Data Quality Assurance

Discoverable, robust metadata.

What metadata are needed?

How will metadata be created and/or captured?

Include unique IDs & citation information.

Publish metadata via a portal or clearinghouse.

Create discoverable, robust metadata.

Concerns about data sharing

Concerns about data sharing

License
- Data Reuse
- Data Sharing
- Data Management

Copyright issues

Ethical issues

Obligations

What data will be preserved?

Lesson 2: Data Sharing

Address data sharing throughout the data lifecycle

Description of data to be produced

How data will be acquired

How data will be processed

How data will be acquired

Make it easier to keep your data organized for (and how it turned out!)

Previous what has been done in your lab in the past

Prevents duplication of effort > it will be more obvious

To the research community

To the research sponsor/funder

To the public

Lesson 3: Data Management Planning

The Data Management Plan (DMP)

A DMP outlines what you will do with your data during and after you complete your research project.

It is a formal (but evolving) document in which you lay out a plan to ensure that your data will be safe for during and after you complete your research project.

A DMP outlines what you will do with your data

- Quality assurance and control processes to be
- File formats of data
- How data will be processed

Better access to information leads to better understanding and contributions toward potential collaborations

Improved data quality

An authoritative source

Costs of not doing data management can be very high!

Loss of institutional commitment

Loss of competencies

Format obsolescence

Server hardware or software failure

Storage failure

Facilities infrastructure failures

Natural disasters

Causes of data loss

The Data Lifecycle

The stages through which well-managed data passes from project inception to conclusion.

Plan
- Design
- Data Collection
- Data Processing Desc.
- Abstract

Integrate
- Data Retrieval
- Data Staging
- Data Processing

Analyze
- Data Quality Assurance
- Data Quality Assurance
- Data Quality Assurance

Discoverable, robust metadata.

What metadata are needed?

How will metadata be created and/or captured?

Include unique IDs & citation information.

Publish metadata via a portal or clearinghouse.

Create discoverable, robust metadata.

Concerns about data sharing

Concerns about data sharing

License
- Data Reuse
- Data Sharing
- Data Management

Copyright issues

Ethical issues

Obligations

What data will be preserved?

Lesson 2: Data Sharing

Address data sharing throughout the data lifecycle

Description of data to be produced

How data will be acquired

How data will be processed

How data will be acquired

Make it easier to keep your data organized for (and how it turned out!)

Previous what has been done in your lab in the past

Prevents duplication of effort > it will be more obvious

To the research community

To the research sponsor/funder

To the public

Lesson 3: Data Management Planning

The Data Management Plan (DMP)

A DMP outlines what you will do with your data during and after you complete your research project.

It is a formal (but evolving) document in which you lay out a plan to ensure that your data will be safe for during and after you complete your research project.

A DMP outlines what you will do with your data

- Quality assurance and control processes to be
- File formats of data
- How data will be processed

Better access to information leads to better understanding and contributions toward potential collaborations

Improved data quality

An authoritative source

Costs of not doing data management can be very high!

Loss of institutional commitment

Loss of competencies

Format obsolescence

Server hardware or software failure

Storage failure

Facilities infrastructure failures

Natural disasters

Causes of data loss

The Data Lifecycle

The stages through which well-managed data passes from project inception to conclusion.

Plan
- Design
- Data Collection
- Data Processing Desc.
- Abstract

Integrate
- Data Retrieval
- Data Staging
- Data Processing

Analyze
- Data Quality Assurance
- Data Quality Assurance
- Data Quality Assurance

Discoverable, robust metadata.

What metadata are needed?

How will metadata be created and/or captured?

Include unique IDs & citation information.

Publish metadata via a portal or clearinghouse.

Create discoverable, robust metadata.

Concerns about data sharing

Concerns about data sharing

License
- Data Reuse
- Data Sharing
- Data Management

Copyright issues

Ethical issues

Obligations

What data will be preserved?

Lesson 2: Data Sharing

Address data sharing throughout the data lifecycle

Description of data to be produced

How data will be acquired

How data will be processed

How data will be acquired

Make it easier to keep your data organized for (and how it turned out!)

Previous what has been done in your lab in the past

Prevents duplication of effort > it will be more obvious

To the research community

To the research sponsor/funder

To the public
DataONE Webinar Series
www.dataone.org/webinars

Upcoming Webinar Event
www.dataone.org/upcoming-webinar

Previous Webinar Events (Recording and Forum)
www.dataone.org/previous-webinars

Next Webinar
Tuesday March 10th 0900 PT / 12 noon ET
Dr Cameron Neylon, Public Library of Science

#DWS2015
@DataONEorg
Quick Poll

✓

✗
Open Science and Data: Towards Optimizing the Research Process

Dr Jean Claude Guédon
Université de Montréal
Open Data and Science
Jean-Claude Guédon

In the beginning:
The ethos of science
Open Data and Science
Jean-Claude Guédon

Building on the shoulders of giants
(and others, not so giant…)
Open Data and Science
Jean-Claude Guédon

The scientific process is a mixture of

Collaboration

Communication

Competition
Open Data and Science
Jean-Claude Guédon

Competition is not the unique motivation for doing science

Collaborating and sharing through publishing are also fundamental
Open Data and Science
Jean-Claude Guédon

Sharing research results is one of the hallmarks of the Scientific Revolution
Open Data and Science
Jean-Claude Guédon

It was not always easy!

Here is how Galileo told Kepler about Saturn's satellites:

smaismrmilmempoetaleumibunenugttauiras
In case you did not get it, it means:

Altissimum planetam tergeminum observavi
Still having difficulties?

Well, it simply means:

I have observed the most distant planet to have a triple form
Open Data and Science
Jean-Claude Guédon

Now, that is clearer...

Don't you agree?
Open Data and Science
Jean-Claude Guédon

You don't agree?

Well, you are right
This is the language of alchemists, not that of modern scientists.
Open Data and Science
Jean-Claude Guédon
Something better was needed
Open Data and Science
Jean-Claude Guédon

Thus, the modern scientific journal

was invented
Nowadays, a new revolution is breeding: Modelling!
An example?
Check out this weather model vs. Actual observations:
http://vimeo.com/117256684
Across the centuries, researchers have learned to share their papers

Now, they must learn to share their data
Open Data and Science
Jean-Claude Guédon

HOW?
Open Data and Science
Jean-Claude Guédon

Two problems related to this question:

- Scientific ethos
- Technical problems
Scientific ethos

The need to reward sharing by visibility and prestige

The need to preserve competitive edge where it matters
Open Data and Science
Jean-Claude Guédon

Technical problems

Retrieval

Interoperability

Cross-disciplinary capacity
First steps:

Data sets linked with published papers

e.g. OpenAIRE+ in Europe
Second steps

Curation and metadata

e.g. EUDAT in Europe
Open Data and Science
Jean-Claude Guédon

Third steps

Preservation
e.g. Locksss in the USA
Open Data and Science
Jean-Claude Guédon

Fourth Steps

The means to evolve:
Standards
procedures
Over all this:

A governance model
Possible models:

The Internet (e.g. IETF)
Free software
(e.g. Free software Foundation)
Open Data and Science
Jean-Claude Guédon

In the Works

Research Data Alliance
Possible dangers?

Privatization of data with dire consequences for the prospects of optimizing research for the common good
Reminders

Recording and Forum:  
www.dataone.org/previous-webinars

Registration opening soon for:

TBA
Dr Cameron Neylon
Public Library of Science
Tuesday March 10\textsuperscript{th}
www.dataone.org/webinars

#DWS2015
@DataONEorg