Protecting Your Data: Backups, Archives & Data Preservation

DataONE Community Engagement & Outreach Working Group

Lesson Topics

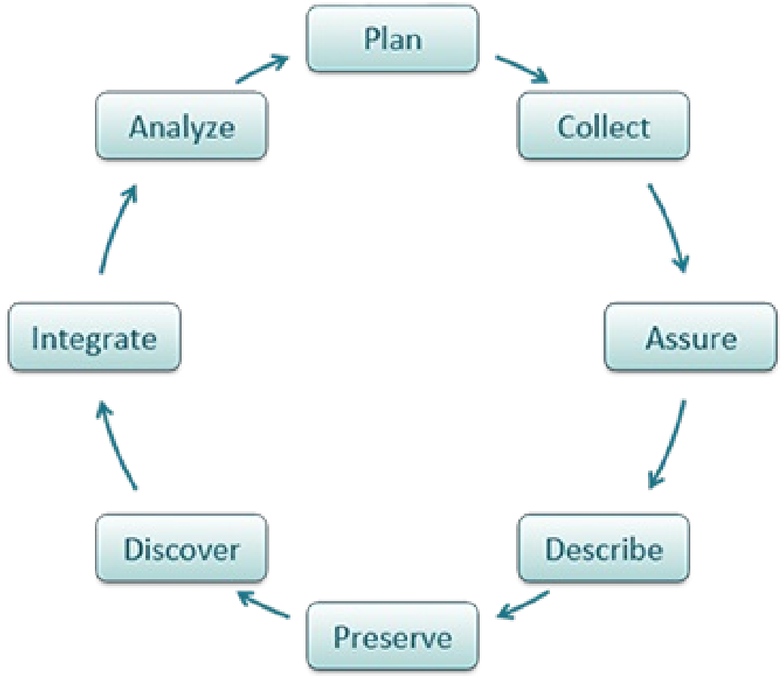
- Key Digital Preservation Concepts
- Backups: Things to Consider
- Data Preservation
- Recommended Practices

Learning Objectives

After completing this lesson, the participant will be able to:

- Define the differences between backups and archiving data
- Identify significant issues related to data backups
- Identify why backup plans are important and how they can fit into larger backup procedures
- Discuss what data preservation covers
- List several recommended practices

The DataONE Data Life Cycle



Data Protection, Backups, Archiving, Preservation

Differences at a Glance

- Data Protection
  - Includes topics such as: backups, archives, & preservation; also includes physical security, encryption, and others not addressed here
  - More information about these topics can be found in the “References” section

Data Protection, Backups, Archiving, Preservation (continued)

- Terms “backups” and “archives” are often used interchangeably, but do have different meanings
  - Backups: copies of the original file are made before the original is overwritten
  - Archives: preservation of the file
- Data Preservation
  - Includes archiving in addition to processes such as data rescue, data reformatting, data conversion, metadata

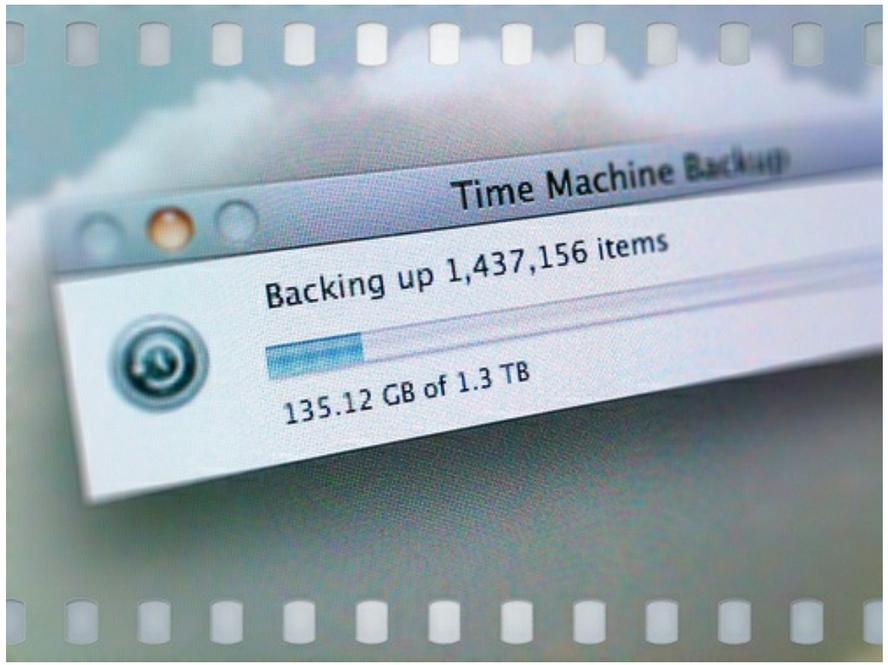
A Closer Look: Backups vs. Archiving

- Backups
  - Used to take periodic snapshots of data in case the current version is destroyed or lost
  - Backups are copies of files stored for short or near-long-term
  - Often performed on a somewhat frequent schedule

Why Perform Backups?

Limit loss of data, some of which may not be reproducible

- Save time, money, productivity
- Help prepare for disasters
  - Accidental deletions
  - Fires, natural disasters
  - Software bugs, hardware failures
- Reproduce results of past procedures (if they were based on older files)
- Respond to data requests
- Limit liability



Backups: Things to Consider

- What are the existing policies that might affect how and when you do data backups?
  - May be separate project, office, department, funding source, or organizational polices
  - Policies may differ between groups; which has precedence?
  - Are backups already part of a larger data management or contingency plan for your group?
- Who is responsible for performing backups?
  - Users? System administrators? Both?
- Do these various policies fit your needs?

Backups: Things to Consider (continued)

- How often should you do backups to capture significant change?
  - Cost versus benefit
  - Continually? Daily? Weekly? Monthly?
- What kind of backups should you perform?
  - Partial: backing up only those files that have changed since the last backup
  - Full: backing-up all files
  - How often and what kind will depend upon what kind of data you have and how unique it is
- What about non-digital files (such as papers)?
  - Consider digitizing files

Backups: Things to Consider (continued)

- Where will you backup your files?
  - May depend upon project requirements, etc.
  - Personal external disk, centralized computer storage (Dropbox), “cloud” storage (Amazon, Google)
    - CDs and DVDs, while cheap and convenient, are not good media for backups
  - What metadata is needed when using these systems?
  - Are the files backed up individually or as one large file?
  - Consider that not all backups may be immediately available, depending on how the files are packaged
  - Good practice to keep backups in different location than source data
  - If a disaster strikes, it can destroy both versions of data

Considerations

- How are backups carried out?
  - Manually may work for single files, but requires that the user remembers to perform regular backups and can be time-consuming
  - Automated backups can be run on a set schedule that doesn't require the user to remember
- What do I do if I need to get a file from backups?
  - Backup mode may determine how the file can be retrieved
  - You should know how to obtain files from backups , where they are located, and who to contact
  - You need to know this information beforehand, as often you need a file from a backup in an emergency!
- Understanding the backup process is part of creating good data management practices

Considerations

- How do you verify a backup has been successfully performed?
  - Most backup software will have a log file that contains details of the backup (which files, when the backup was created)
  - However, don’t rely solely on the log file
  - Even if a log file states the backup was successful, you still need to check the backup to make sure the files are there and accessible
  - Test by trying to pull a file off from backup and restore it to another location
  - Hardware and software failures can happen after backups and log files are made
  - Make sure your system is backing up the correct files

Considerations

- If you are working with someone, such as an IT group, who helps manage and perform backups, confirm and verify that the backup process has been successfully completed
- How do you verify a backup has been successfully performed?
  - Since manual checks of all of the files in your backup is probably not possible, you should utilize other methods such as checking file sizes, date stamps, and checksum values.
  - Checksum are mathematical calculations based upon a specific file. If the calculated checksums match between the backup copy and the original file, chances are the file is the same and was not modified when copied or stored.

Considerations

- Are there backups of the backups?
  - Necessary for high-value data
  - Usually different copies of backups are kept in different locations
- How long do you keep your backups?
  - Depends upon specific situation, and should be determined in concert with stakeholders and resource managers
  - Understand relevant guidelines, policies and rules for retention of data
- What are the long term storage and access solutions that are relevant for the project? What to do when funding ends or key staff depart?
  - Changes in the status of the project, funding, or key staff are important reasons to have a full understanding of related options and requirements for storage and access

Data in Real Life

A design firm was handling their own backups. The system was working and the backup software was reporting that the data was successfully backed up. The administrator checked the backups immediately after they were done and confirmed they were good.

Data in Real Life

After a computer virus erased most of their files, they went back to their backups. Unfortunately they found that the backups were all blank and all of the data was gone. Only after some investigation did they discover that the computer tapes (which contained the backups) were placed against a wall that had an elevator on the other side of it. When the elevator went past, the magnets inside erased all of the tapes.

- Had they checked their backups properly, they probably would have noticed this before there was an emergency

Final Considerations

- Can you read data from older backups?
  - Media changes. You may no longer be able to read older versions and formats such as floppy disks, Jazz and Zip drives, WordPerfect files, etc.
- Media can degrade quickly, unexpectedly, inconsistently
  - Even if you can open a file today, that doesn’t mean you can in a month from now
- How will you dispose of outdated data?
  - Make decision to copy, archive Remember: back up the data you can’t afford to lose!

Data Preservation

- By managing and preserving your data well, data rescue may not be necessary. Why?
  - Addition of relevant metadata, proper file naming (can help the file from getting lost in the system), utilization of proper file formats (lets you open the file without having to convert the file), backups (limits loss of files), and media types (limits degradation of files), you may limit or prevent the need for data rescue.
- A good data management plan is another tool to help limit the need for data rescue.

Processes Related to Data Preservation

- Includes backups and archiving in addition to processes such as data conversion, data reformatting, and data rescue
  - Older files may no longer be in a usable format and may require conversion or “rescue” before the data can be used.
  - Data reformatting, conversion, and backup becomes even more important as projects finish up and/or are no longer funded.
  - Data may have been kept at the end of the project, but if no one is managing the data, data may be left in formats that are no longer usable or in locations that are no longer accessible.
- Additionally, data preservation requires planning, structure, and ongoing management and assessment

Preservation Formats and Version Strategies

- Create useful, relevant metadata
- Data Conversions and Formats
  - Use non-proprietary, standard formats
  - Convert text files from .doc or .xls to .txt, image files to .tiff or .pdf
  - Be sure to check files after converting them, as data, metadata, and formatting loss can occur
- Versioning
  - Use consecutive numbers and letters to help keep track of changes to a file throughout various edits and revisions. This will help you quickly differentiate between files with similar names.
- File Naming
  - Use file names that are consistent, descriptive, and concise so that you can find and quickly identify the file the file at a later time.
  - Rename files that have a default file name when exported such as “image.jpg” or “archive.zip”

Recommended Practices

- Create a preservation policy that clearly identifies:
  - roles
  - responsibilities
  - where the data is backed up
  - how often the files are backed up
  - how to access the files
  - recommended file formats to be used
  - policies for migrating data to assure data are not lost due to media degradation or changing formats or programs
- Review your preservation policy and plan periodically to ensure it is still valid and applicable

Recommended Practices (continued)

- Minimize or remove reliance on users to perform own manual backups (if possible)
  - Implement standardized and automatic backups
  - If possible, put experts in charge of this task (computer staff) as they are more likely to keep up-to-date regarding software updates, hardware issues, best practices, etc.
- Don’t assume backups are being performed for you
  - You don’t want to find out after the fact that no backups have been performed
  - If you are using third-party software (like Yahoo or Google Mail), what happens if they lose your files?
- Use non-proprietary, standard formats
  - Convert text files from .doc or .xls to .txt, image files to .tiff, or .pdf

Recommended Practices (continued)

- Check your backups manually
  - Start with log files, as they may tell you the backup was unsuccessful
  - Do not rely solely on the log files – they may be incorrect or the data may have become corrupted after the file was transferred
  - Look at file dates and file sizes to see if they match; calculate a checksum on the original and archived file and make sure they match
  - Ensure you can read files off of older backups and archives.
- Have multiple versions of backups on multiple formats in multiple places
- Good data management will limit the amount of data rescue that needs to be performed on older data

Data in Real Life

In 2011, a software bug caused some Gmail users to lose access to their email. Fortunately, Google had backups!

Summary

- Data preservation is more than just backing up and archiving your files
  - organizational infrastructure, technological situation, resources
- When devising a preservation strategy, one needs to consider how often to perform backups, where to backup, accessibility to backups and how long to keep the files
- There are many reasons we need to perform backups, primarily to prevent data loss
- Check for backups on outdated media and test your backups often!

References

1. Stanford University Libraries, Data Management Plans, (Stanford University Libraries), https://library.stanford.edu/research/data-management-services, (accessed 9/21/2016)
2. Albanesius, Chloe, Google: Storage software update led to e-mail bug, http://www.pcmag.com/article2/0,2817,2381168,00.asp (accessed 09/21/2016)
3. Van den Eynden, Veerle, Corti, Louise, Woollard, Matthew, Bishop, Libby and Horton, Laurence, Managing and Sharing Data, http://www.data-archive.ac.uk/media/2894/managingsharing.pdf , and companion materials, https://www.ukdataservice.ac.uk/manage-data/handbook (accessed 09/21/2016)

For more information about physical security, encryption, and data disposal, visit: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

About

Participate in our GitHub repo: https://dataoneorg.github.io/dataone_lessons/

The full slide deck (in PowerPoint) may be downloaded from:
http://www.dataone.org/education-modules

Suggested citation: DataONE Education Module: Data Management. DataONE. Retrieved November 12, 2016. From
http://www.dataone.org/sites/all/documents/L01_DataManagement.pptx