BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 7: e111453
doi: 10.3897/biss.7.111453

OPEN ACCESS

Conference Abstract

Preservation Strategies for Biodiversity Data

Dmitry Mozzherin[‡], Deborah L Paul[‡]

[‡] University of Illinois, Champaign, United States of America

Corresponding author: Dmitry Mozzherin (mozzheri@illinois.edu), Deborah L Paul (dlpaul@illinois.edu)

Received: 21 Aug 2023 | Published: 22 Aug 2023

Citation: Mozzherin D, Paul DL (2023) Preservation Strategies for Biodiversity Data. Biodiversity Information Science and Standards 7: e111453. https://doi.org/10.3897/biss.7.111453

Abstract

We are witnessing a fast proliferation of biodiversity informatics projects. The data accumulated by these initiatives often grows rapidly, even exponentially. Most of these projects start small and do not foresee the data achitecture challenges of their future. Organizations may lack the necessary expertise and/or money to strategically address the care and feeding of this expanding data pile. In other cases, individuals with the expertise to address these needs may be present, but lack the power or status or possibly the bandwidth to take effective actions. Over time, the data may increase in size to such an extent that handling and preserving it becomes an almost insurmountable problem. The most common technical challenges include migrating data from one physical data storage to another, organizing backups, providing fast disaster recovery, and preparing data to be accessible for posterity. Some sociotechnical and strategic hurdles noted when trying to address data stewardship include funding, **data leadership** (Stack and Stadolnik 2018) and vision (or lack thereof), and organizational structure and culture. The biodiversity data collected today will be indispensable for future research, and it is our collective responsibility to preserve it for current and future generations.

Some of the most common information loss risk factors are the end of funding, retirement of a researcher, or the departure of a critical researcher or programmer. More risk factors, such as hardware malfunction, hurricanes, tornadoes, and severe magnetic storms, can destroy the data carefully collected by large groups of people.

The co-location of original data and backups create a "Library of Alexandria" where a single disaster at this location can lead to permanent data loss and poses an existential threat to the project.

Biodiversity data becomes more valuable over time and should survive for several centuries. However, SSD (solid-state drive) and HDD (hard disk drive) storage solutions have an expiration date of only a few years. We propose the following solutions (Fig. 1) to provide long-term data security:

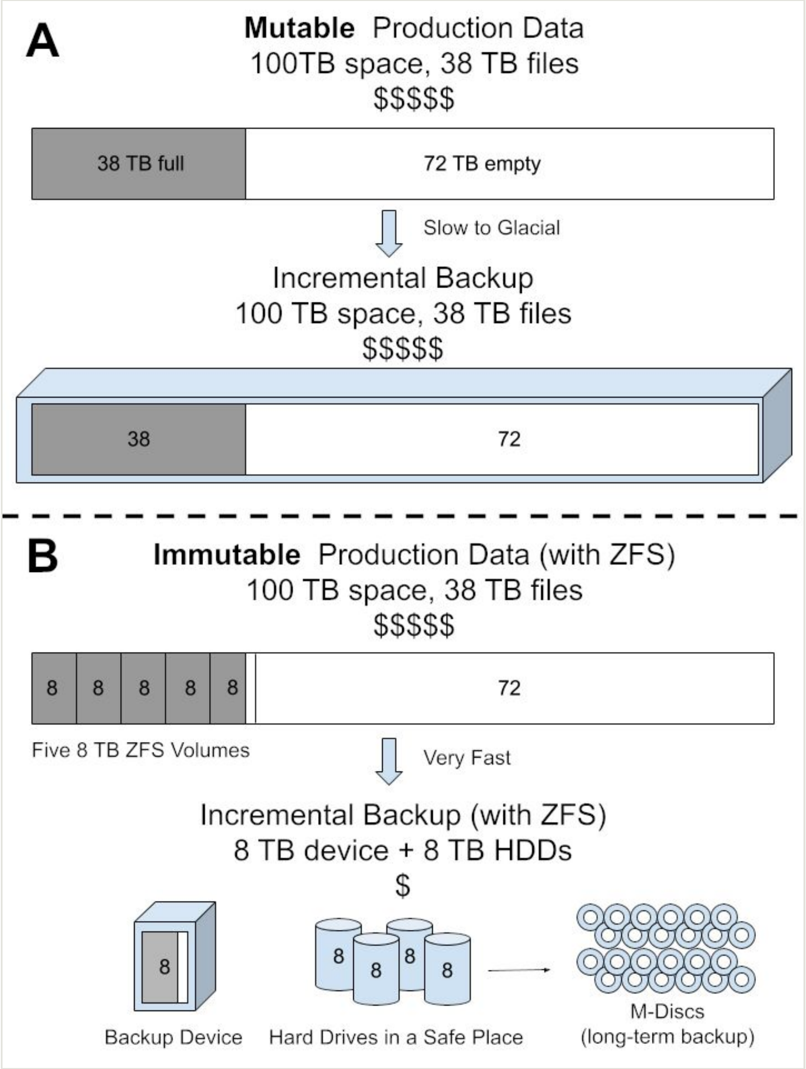

Figure 1.

A) A commonly observed pattern of 'monolith' data handling. B) Suggested data handling.

Technical tactics

1. Use an immutable file storage for everything that is not entered very recently.

Most of the biodiversity "big data" are files that are written once and never changed again. We suggest separating storage into a read-only part and small read/write sections. Data from the read/write section can be moved to the read-only part often, for example, daily.

1. Use a Copy-On-Write file system, such as ZFS (Zettabyte File System).

The ZFS file system is widely used in industry and is known for its robustness and error resistance. It allows efficient incremental backups and much faster data transfer than other systems. Regular incremental backups can work even with slow internet connections. ZFS provides real-time data integrity checks and uses powerful tools for data healing.

1. Split data and its backups into smaller chunks.

Dividing backups into cost-effective 2–8 terabyte chunks allows running backups using cheap hardware. Assuming that the data is read-only, such data organization always splits the backup into chunks, with hardware costs changing from tens of thousands of dollars (US) to less than two thousand dollars. We recognize that with time data storage costs drop, and larger chunks will be used.

1. Split the data even further to the size of the largest available long-term storage unit (currently an optical M-disc).

The write-once optical M-DISC is analogous to a Sumerian clay tablet. Data written on such discs does not deteriorate for hundreds of years. This option addresses the need for last resort backups because the storage does not depend on magnetic properties and is impervious to electromagnetic disasters. Optical discs can be easily and cheaply copied and distributed to libraries worldwide. In the future, discs' data can be transferred to a different long-term storage medium. We also trust these discs can be deciphered by those in the future, just like clay tablets.

Sociotechnical insights

The above example of a comprehensive strategy to preserve data epitomizes "LOCKSS" (lots of copies keep stuff safe) and makes it clear that these copies need to be in multiple media types. Our suggestions here focus on projects that experience data growth pains. Such projects often look to see how others address these data needs. Recently, The Species File Group (SFG) did this exercise to evaluate and address our data growth needs (Mozzherin et al. 2023). We recognize and emphasize here the need for

- personnel with the knowledge and skills to build, maintain, and evolve robust strategies and infrastructure to make data accessible and preserve it,
- funding to back the most suitable architectural strategies to do so, and
- people with expertise in long-term data security to have a seat at the leadership table in our organizations.

We encourage our colleagues to evaluate the status of data leadership at your organizations (Stack and Stadolnik 2018, Kalms 2012). Implementing these suggestions will help ensure the survival of the data and accompanying software for hundreds of years to come.

Keywords

information loss, data leadership, data stewardship, exponential data growth, data architecture, data backup, ZFS

Presenting author

Deborah L. Paul

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Kalms B (2012) Digitisation: A strategic approach for natural history collections. Atlas of Living Australia. URL: https://www.ala.org.au/wp-content/uploads/2011/10/Digitisation-guide-120326.pdf
- Mozzherin D, Pereira H, Yoder M, Paul D (2023) Dealing with an Exponential Data Growth. URL: https://github.com/gnames/papers/blob/master/devops-blog-2023/blog.md
- Stack C, Stadolnik EM (2018) Data leadership: Defining the expertise your organization needs. URL: https://www.spencerstuart.com/research-and-insight/data-leadership-defining-the-expertise-your-organization-needs