“Digits: for Good”; How Patterns Can Help

World Digital Preservation Day is held every year on the first Thursday of November, which falls this year on November 5. The theme of this year is “Digits: for Good,” which allows the data preservation community to highlight the good effects that data preservation has had on society. The data community (creators, curators, and consumers) from around the world are invited to share their own Digits: for Good stories.

Digital preservation is a key issue for Patterns. As a scientific journal, we are responsible to curating our content and for preserving the scientific record. This includes data, and as we’re an online-only journal, even our articles need digital preservation.

I should confess that 2020 has been a strange year that proved to us how much open data sharing is important. During the pandemic lockdown, remote access became (and still is, as the lockdown comes back every now and then) an essential part of everyday life. Teaching and working moved immediately online in the countries that have the infrastructure for it, meaning that digital preservation and communication are now more important to more people than they have been ever before.

Remote access to data is one of the most valuable aspects of digital preservation that we may not have been much aware of before March 2020. I’m not just talking about data as in scientific results but also data as resources to help understand the spread of the pandemic, how to treat it, and also for more everyday uses, like helping children learn from home and communicate with their socially distanced friends.

The pandemic showed us that data observed or created by any of us may or can save lives. This situation has moved discussion of open data sharing to a new level, as now that spread of COVID-19 is a problem of every single one of us, we all have decided to share data openly. This question bothers me a lot, because if this wasn’t a pandemic but only an epidemic situation (and would not affect us directly), would we all willingly share our data openly? Data scientists have known for many years how important it is to save lives by sharing our valuable data, immediately. Policy makers and government officials are now starting to realize that too.

What is usually forgotten in the discussions about open science is that data sharing requires data preservation. While we share data openly, we must guarantee the safety and integrity of it and during the time. It’s more than just the data too—we need to capture the metadata and other information about the context the data were created in, so we can look to the future for long-term preservation and understanding.

Although we are aware of open science, we still need to work on how to best preserve data for possible future use, e.g., in machine learning. For instance, one of the best methods to share data and guarantee their digital preservation is depositing them on a trusted data repository. Many researchers who choose to share their data on a repository share them in a format that is not possible to be (re)used by other researchers or by machines. In many cases to be able to even just visualize and inspect the data, one needs to have special software. Or the data are in a format that is not easily readable by machines.

PDF is used as a common format to preserve text data in repositories, but this requires the use of a text extraction tool to convert the PDF documents into machine-readable text format if the data are to be used for other research. Best practice for data archiving includes thinking about how to share your data openly but also reusably, and so, for digital preservation of your data, consider using an open format that is usable by the majority of interested people and also by machines.

Researchers are mainly focused on the content of their research and not necessarily preserving the outcomes. Without digital preservation, there is a clear and present danger that the data could be lost. Even details that are clear and obvious to the data creators now can be easily forgotten given time. The loss of data creates the situation that research cannot be validated or reproduced. History is made based on what was preserved from a hundred years ago. Preservation takes the value of data through time! Putting digital data aside does not guarantee preservation, and without preservation, it cannot become part of the history for future researchers.

Data creators produce hundreds of gigabytes of data, sometimes without realizing that they are not equipped with a proper digital preservation method. The output of research nowadays is mainly digital, so preserving that output is comparable with the function of libraries or archives in the past. In the digital age, data preservation must also evolve from traditional methods of data management to newly developed methods of digital preservation. One of these new methods of digital preservation in research is preprint servers, or which “arxiv” was the first. During the COVID-19 pandemic, thousands of research papers appeared in bioRxiv and medRxiv prior to peer review, providing a great and speedy resource to researchers but at the risk of a lack of validation, as none of the articles were peer reviewed. With fast-moving research, digital preservation is very difficult but more important than ever, and we must make sure to collect and preserve all the information created and lessons learned during this pandemic. The Research Data Alliance COVID-19 Working Group considered digital preservation while addressing the key issues with data during a pandemic and have produced guidance for others to use (https://zenodo.org/record/3932953#X6Hlhp5Kg2x).

Digital preservation is not only conserving the data safely but also maintaining them while the technology is changing. The new digital preservation methods should allow updating the data with new outcomes and migrating the data to new environments as technology moves forward. As the data issues and the definition of “big data” are changing with the rapid change of technology, preserving digital data is a whole new branch of data science that demands specialists to dedicate time, effort, and expertise to it.

Data preservation is not only important for scientists and data creators but also for data consumers such as industry and policy makers. Let us think about health data and how they are archived in different countries. Checking the responses to the digital preservation community survey for 2019–2020 (held by the Open
Preservation Foundation) shows that responses are from 98 organizations of 31 countries. Of these countries, 51% are in Europe, 35% are in North America, 5% in Africa, and 9% in South America, Asia, and the continent of Australia. If you check more closely, the response from Asia is mainly from India and Turkey. This shows that either the organizations that are not participating in the survey are not aware of this community or they do not have the infrastructures or ability to join.

We must consider diversity and inclusion in digital preservation of data to have a complete and better picture of the global health in human society. In many developing countries, little if any of the patient’s information are digitalized. Instead they are still archived using a traditional paper-based method. To have a good global preservation system, we need to have good data preservation methods and tools and to oblige suppliers to use data packing and validation systems. We also need easy-to-use systems to help developing countries digitalize their valuable data. Doing such huge projects needs collaborative teams and partnership and support of political authority.

**Patterns** supports digital preservation by being an open access journal. We not only publish the results of high-quality research, we also oblige authors to share their data and codes in trusted repositories and encourage them to standardize data and store it in formats that are community appropriate and future proofed (as much as is possible). By sharing best practice across different science domains, **Patterns** aims to bring diverse minds together to help them solve big issues of digital data. We published our first issue in April 2020, at the beginning of the COVID-19 lockdown in the US and Europe, to be the voice of researchers, data scientists, data managers, digital archivists and librarians, digital preservation officers, data stewards, data users, and decision and policy makers. Our aim is to help all our readers and authors understand the whole data life cycle in order to share the knowledge and experience they’ve gathered.

**Patterns** exists to communicate the value of research, of data science solutions, and of implementing those solutions in real life, to empower researchers and decision makers to make good, well-informed decisions to improve the quality of life for everyone. When data are preserved digitally, we all benefit. To paraphrase a common saying, yesterday was the best day to think about digital preservation, but today is the second best. **Patterns** is here to support and bring the data science community together so that we can make Digits: for Good a commonplace reality.

Happy World Digital Preservation Day everyone!

**WEB RESOURCES**

RDA COVID-19 Recommendations and Guidelines on Data Sharing, https://zenodo.org/record/3932953#.X6Hlhp5Kg2x.

Sahar Farajnia  
Scientific Editor, **Patterns**  
Sarah Callaghan  
Editor-in-Chief, **Patterns**  
https://doi.org/10.1016/j.patter.2020.100134