Harnessing the Power of Scientific Data

Michael Ernst

Information and Communication Technologies are the most recent transformational factors in science. They enable close and almost instantaneous collaboration between scientists all over the world and they provide access to unprecedented volumes of scientific information that can in turn be processed on powerful computational platforms. With robust infrastructure for data transmission and data processing in place, we can now start to think about the next step: data itself. A high level goal is a scientific community that does not waste resources on recreating data that have already been produced, in particular if public money has helped to collect those data. Scientists should be able to concentrate on the best ways to make *use* of data. Data become an infrastructure that scientists can use on their way to new frontiers.

Our stock of intangible knowledge, expanding at today's hyper-speeds, needs to be thought of as a new kind of asset in itself that serves all. As such, it requires professional analysis and engineering. Its contents are heterogeneous –different data formats, value and uses. There is tremendous value in having the data made seamlessly available, to use, reuse and recombine to support the creation of new knowledge. And the data must be available to whomever, whenever and wherever needed, yet still be protected if necessary by a range of constraints including licenses, time embargos, community or institutional affiliation.

To collect, curate, preserve and make available ever-increasing amounts of scientific data requires new types of infrastructures. The result will be a vital scientific asset: flexible, reliable, efficient, cross-disciplinary and cross-border.

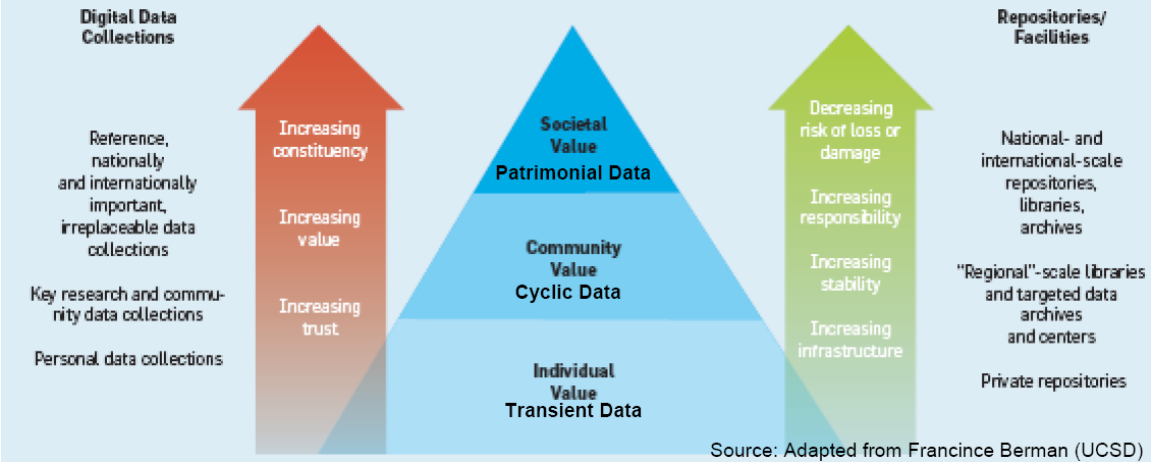
The anticipated infrastructure is supposed to support seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance.

- All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.
- Researchers from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.
- Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. Federated repositories work to international standards, to ensure they are trustworthy.

The focus of this white paper is on scientific data because, when the information is so abundant, the very nature of research starts to change. A feedback loop between researchers and research results changes the pace and direction of discovery. The "virtual lab" is already real, with the ability to undertake experiments on large instruments in other continents remotely in real time. Researchers with widely different backgrounds - from the humanities and social sciences to the physical, biological and engineering sciences – can collaborate on the same set of data from different perspectives.
Just how will we train people to work in this environment? What tools do we have or will we need to move, store, preserve and mine these data? How to share them? How to understand them, if you are in a different scientific discipline than that in which they were created? As a researcher, how will you know the data you access remotely are accurate, uncorrupted and unbiased? What

if those data include personal details – individual health records or financial information? These are just a few of the profound policy questions posed by this new age of data-intensive science.

A data pyramid (below) suggests the complex data ecology. At the bottom of the pyramid lie the most abundant, transient forms of data – billions of personal data files across the planet, on private disks and storage services, of obvious value only to the few who create or use them. At the top of the pyramid is patrimonial data – high-value, irreplaceable data of importance to an entire nation or society, redundantly stored in national or international trusted archives. In the middle is cyclic data – a mid-range of data created and used in a specific task, community or region. The data infrastructure must cope with all these data classes.



Suggested properties of a data infrastructure for science include

- Open deposit, allowing user-community centers to store data easily
- Bit-stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years
- Format and content migration, executing CPU-intensive transformations on large data sets at the command of the communities
- Persistent identification, allowing data centers to register a huge amount of markers to track the origins and characteristics of the information
- Metadata support to allow effective management, use and understanding
- Maintaining proper access rights as the basis of all trust
- A variety of access and curation services that will vary between scientific disciplines and over time
- Execution services that allow a large group of researchers to operate on the stored data
- High reliability, so researchers can count on its availability
- Regular quality assessment to ensure adherence to all agreements
- Distributed and collaborative authentication, authorization and accounting
- A high degree of interoperability at format and semantic level