Science in a data-intensive age

Geoffrey Boulton

1. The last two decades have seen unprecedented growth in the capacity to acquire, store, manipulate and instantaneously transmit data. It is a world historical event that is changing the lives of individuals, societies and economies. It has major implications for science, research and learning that are far more profound and pervasive than those of the earlier, analogous revolution in data storage and human communication, that of Gutenberg's invention of the printing press in the 1430s.

2. These developments offer profound challenges to science, and because of this to CODATA in adapting its historic role as the principal coordinator of data initiatives at the international level to focus on new, data-intensive ways of doing science.

3. Much of the challenge comes from so-called "big data", which are "big" because of the volume that systems must ingest, process and disseminate; because of their diversity and complexity; and because of the rate at which data streams in or out of the systems that handle them. Terabyte-sized data sets are now common in Earth and space sciences, physics and genomics etc. The exploitation of these opportunities depends on the development and use of many technical solutions.

4. The data explosion and our capacity to combine, integrate and analyse large, varied and complex datasets offers powerful new ways of unravelling complexity, improving forecasts of system behaviour and detecting patterns in phenomena that have hitherto been beyond our capacity to resolve. It is the Google way of doing science. Such data-intensive science will at least complement the classical approach of hypothesis-theory-test. Some even argue that it will replace it. In any case it requires that we understand the mathematical and statistical basis of data manipulation. It is essential to develop new tools and new techniques to exploit this understanding, and to adopt new habits of working that have an ethos of open access to data in order to facilitate re-use, re-combination and re-purposing. Openness also facilitates more effective dissemination of scientific concepts and the evidence for them, in society and in education. It has the potential to change the social dynamics of science, and contribute towards the evolution of science as a public enterprise, rather than one conducted behind closed laboratory doors.

5. These issues naturally pose major questions about the way science is done and also define the major issues to which CODATA should apply itself:
   a) How do we understand the deep mathematical basis of "data science" and how can we articulate this with clarity for scientists, technologists, businesses, policymakers and the public? What does it mean to be a scientist and researcher in a digital age?
   b) How can we maintain the open data principle in a data-rich world to ensure that the data underlying scientific concepts are open to scrutiny and replication or invalidation? (The danger is that data manipulation takes place within a black box that is not open to scrutiny, in which case, science ceases to be science).
   c) How can we incentivize and enable the data sharing, re-use and cooperation required to efficiently use multiple data sources and to address global challenges effectively?
   d) How should we exploit the capacities of machines to use learning-based algorithms to match patterns and to interpret complex information in ways that are accessible to human cognition and thereby aid and not by-pass human creativity?
   e) How is cyber-security to be maximized and how is personal privacy to be respected

in the use of data for research?

f) How should the interface between publicly-funded science and rapidly advancing commercial data science be stimulated and exploited?
g) How are universities, institutes and other places where science is done best encouraged to be develop proactive and creative management of their data and their support for data-intensive science?
h) How should principles and processes of data science be embedded in scientific education and training?
i) What opportunities does the digital age offer for more inclusive, democratic ways of producing scientific knowledge and in playing a transformative role in society?

6. These major tasks require boldness, vision and organization. They are so central to the issues of data use and integrity that the credibility and relevance of CODATA would be seriously in question were it not to address them. The priorities for CODATA should be the analysis, articulation and communication of answers to these high-level questions, and engagement with the national and international bodies that have the capacity to implement them. CODATA should continue to collaborate with bodies such as the Research Data Alliance (RDA) - which works to build solutions to promote data interoperability - and with the World Data System (WDS) - which ensures that data are managed and made available for the long term - with the objective of identifying and advocating those international norms and standards for which there is a need.

7. CODATA should be a forum to advance understanding of data-intensive science and to advocate solutions to questions that this developing science raises. It should be a means of stimulating a response to them within national science systems. It is well placed to do these things through the structures of the International Council for Science (ICSU), to which it reports, with its two orthogonal axes of membership:
   - The scientific unions in ICSU represent international science communities and articulate the principles and priorities of their disciplines. CODATA should work with them to promote understanding and change and to advocate good practice from those that have adapted to the data-intensive challenge.
   - The national representatives in ICSU are well placed to influence national scientific structures, priorities and education in data-intensive science. The work of scientists and their institutions is embedded in national systems of organization and funding to which the development of national data-intensive science needs to adapt. Collaboration with national members of ICSU will be a key in ensuring that national needs are respected whilst meshing with the international nature of science.

8. It will be the role of the President, working with the Director and key committees to ensure that the above priorities are deeply embedded in the activities of CODATA.