Data - unbound by time or discipline – challenges and new skills needed

David Giaretta

Giaretta Associates Ltd, Yetminster, Dorset, UK
david@giaretta.org

Abstract. We live in an exciting information age, where the deluge of data enables the 4th paradigm to be used by the greatest number of scientists who have ever lived, able to connect to hundreds of thousands of sources of information which are encoded digitally and used in an ever changing technological network.

To take advantage of these opportunities presents challenges. The most obvious involves simply coping with the volumes of data with which one has some familiarity, from familiar sources.

However in order to combine data from multitudes of unfamiliar sources, covering a variety of disciplines, created over timescales which are long compared to technological and even many conceptual and terminological cycles there are new challenges both for the researchers and the infrastructure needed to support them.

This presentation will focus on these challenges raised by the need to ensure we can deal with the unfamiliar and outline the resources, both human and technical, which will be needed to address them.

1 Opportunities

The term "4th paradigm" was coined by Jim Gray and colleagues to express the idea that in addition to the empirical, theoretical and computational paradigms we now have data exploration enabled by the vast amount of data that is being produced. This has been explored in the literature as a source for scientific progress. However there are far broader opportunities which those who fund the research are interested in.

The Riding the Wave report provided a vision for 2030 which addressed the question, as part of the EU Digital Agenda, "How Europe can gain from the rising tide of scientific data".

The starting point was the observation that "A fundamental characteristic of our age is the raising tide of data – global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge."

The vision was of "a scientific e-Infrastructure that supports 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."

2 Challenges

An underlying challenge was sustaining the availability and usability of the digitally encoded information across disciplines and over time. An associated, fundamental, question was "who pays and why". While data is newly created and of obvious use there will be resources available, but as the Blue Ribbon Task Force pointed out, the value of much data is potential – it may be useful in the future, but this is not certain.

Resources are needed to address the many V's[1] which are normally discussed in terms of big data – but which are also relevant to small data, since as noted[2] the real revolution, which is the mass democratisation of the means of access, storage and processing of data – small as well as big.

In this presentation I divide these Vs into two groups. The first consists of Volume, Velocity, Variety and Volatility which are ones more related to data management – i.e. issues which arise even if the data is being used by the researchers who created it and over just a few years. The other group consists of Veracity, Validity and Value, which this presentation will focus on for the following reasons.

Veracity, including Understandability and Authenticity, is vital for using data from unfamiliar sources and with which the researcher is unfamiliar – otherwise how can a researcher use the data and trust that it is what it is claimed to be? The challenge will be exacerbated by the data management "Vs" noted previously, in particular scaling with Variety.

Validity (including correctness, data quality and legality) is vital interest to researchers if they wish to undertake scientifically useful work.

Value (or potential value) must be identified in order to justify keeping the data in the long term – and even in the short term (related to Volatility) – because keeping data requires resources. The minimum, relatively easily identified, costs are those for storage which tends to scale with Volume and are very front-loaded. Other costs, which are less obvious and more uncertain are those associated with maintaining Veracity and Validity.

3 Solutions

The bulk of the presentation will look at practical solutions to the challenges presented by the second group of V's. These solutions involve underlying consistent concepts, technology and widely agreed procedures, all supported by skilled and well trained humans, across the whole lifecycle of data from conception through to and including curation.

They will help put in place the data infrastructure which can be used across disciplines and across time for the benefit of science, technology, the economy and society.

---

1 http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/

2 http://www.theguardian.com/news/datablog/2013/apr/25/forget-big-data-small-data-revolution