Data Centre View: Data Publication

Wim Hugo

In recent years, the concept of data publication as a valid, measurable output of scientific endeavour has solidified to the point where it is clear that (1) it will provide heretofore unavailable incentives to data-focused scientists for peer recognition, (2) it will lead to increased availability of data, including quality assured data, (3) society in general and science as a whole will benefit from the compound effect of such availability, and (4) that it is now a question of when it will happen, and not if it will happen.

Data Centres will play a critical role in this new infrastructure in support of science, and to some extent the WDS, is pioneering the creation of such infrastructure. Based on its criteria for membership, the wider context of what constitutes viable data centres has been defined. In broad terms, it addresses technical, scientific, governance and financial feasibility for the longer term, underpinned by standards and policies that limit risk and engender trust.

To be effective, data curated by such centres should be

- Discoverable;
- Capable of being understood in respect of scope, quality, and usability, even if the data sets are large;
- Preserved and made available for the long term;
- Standardised in view of use by both humans and systems.

Data Centres face several challenges in providing the fabric of permanence and reliability that is required to make data publication a success. These include (1) funding for the indefinite[1] preservation of data, (2) universal access to prospective data providers, irrespective of data quality, (3) interoperability to the point of seamless integration with the journal publication industry, (4) lack of capacity, know-how, and incentives amongst the producers and providers of data.

Positive trends need to be reinforced to support Data Centres in this endeavour:

1. Support for the provision of ring-fenced, grant-linked funding for the preservation of data;
2. Policies that support the publication of grant- and tax-funded data sets;
3. Inclusion of data management and informatics training as part of honours-level degree study;
4. Entrenchment of standardized data publication metrics and data management plans in proposal assessment processes;
5. Due recognition in science and publication rankings.

We foresee some potentially negative outcomes from a more formalised data publication environment, including (1) an initial flood of low-quality submissions, (2) competition between Data

[1] Which, for now, is still to be defined …

Centres for funding of long-term preservation, and (3) competition from the established publication industry for what estimates show could be a doubling in size of the market.

Research questions and best practice development should at least include (1) scientific basis for the decision to terminate or alter the on-line availability of a data set, (2) extent to which derived works are unique, (3) minimum data and visualization services required to support interoperability with published articles and reports (4) integration of citation, quality, and traditional meta-data records for a more comprehensive view of the characteristics of a data set.