03/10/2012

ICSU
WORLD DATA SYSTEM

WDS and Open Access

Discussion paper submitted by ICSU World Data System
Prepared by the International Programme Office

1. Historical background

The World Data System (WDS) builds on the 50+ year legacy of former World Data Centres (WDCs) and Federation of Astronomical and Geophysical data Analysis Services (FAGS), established by the International Council for Science (ICSU) to keep and organize data generated by the International Geophysical Year (IGY) in 1957–1958. While they readily served and preserved the IGY data legacy for more than half a century, the WDCs and FAGS were facing tremendous challenges. The International Polar Year (IPY 2007–2008)—another data-intensive international research effort launched by ICSU and the World Meteorological Organization—revealed that they were not equipped to anticipate and respond to the sheer multidisciplinary scope of IPY-sponsored research. In 2009, the ICSU General Assembly discontinued both bodies and established WDS as a new ICSU Interdisciplinary Body to incorporate re-applying WDCs and FAGS, as well as state-of-the-art data centres and services.

WDCs and FAGS were created as part of ICSU's longstanding commitment to promote access to scientific data and information, which is enshrined in its Principle of Universality of Science. One of the guiding principles of the WDCs was indeed to '*provide data to scientists in any country free of charge, on an exchange basis or at a cost not to exceed the cost of copying and sending the requested data'*. Therefore, the concept and spirit of open access to data were already put into practice by these organizations. In 1996, an ICSU resolution was adopted by its General Assembly that recommends '*…as a general policy the fundamental principle of full and open exchange of data and information for scientific and educational purposes'*. In actuality, several ICSU bodies embraced this recommendation, most notably IPY, which adopted a full, free, and open access Data Policy[1] for all of the data generated through its sponsored projects.

2. WDS Data Policy

ICSU WDS is an evolution from WDCs and FAGS and its mission is tightly linked to the full and open access to scientific data and information. Its objectives are to '*Enable **universal and equitable** access to quality-assured scientific data, data services, products, and information, and to ensure long-term data stewardship*' and '*Foster compliance to agreed-upon data standards and conventions, and to provide mechanisms to facilitate and improve access to data and data product'*.

To promote an Open Access Policy to data, the WDS Data Policy[2] derives from the GEOSS Data Sharing Principles[3] developed under the auspices of the Group on Earth Observations and under the guidance of ICSU CODATA. The WDS Scientific Committee (WDS-SC) decided that it was unnecessary to not only reinvent a new data policy but also go into further details, since the default position must be full and open access and any exception should be dealt with on a case-by-case basis as such. It is obvious that WDS will directly benefit from a wider adoption of the principle of 'Full and Open Access' to scientific data as this would allow an expansion of its membership and ultimately better serve the scientific community and multidisciplinary research. WDS is also contributing to its adoption through the recruitment of relevant data and service providers.

3. Open access challenges

3.1. For WDS

The largest challenge for WDS has been the implementation of a WDS Data Policy as decided by the WDS-SC whilst respecting ICSU's recent terminology. On the WDS website and within our documentation (Constitution, Data Policy,

[1] http://ipy.arcticportal.org/images/uploads/final_ipy_data_policy-1.pdf
[2] http://www.icsu-wds.org/organization/data-policy
[3] http://www.earthobservations.org/geoss_dsp.shtml

etc.), we use both the term '**universal and equitable access'**—as employed by ICSU—for scientific information (published journals) and '**full and open access**' for data. ICSU's phrasing, used for scientific information as opposed to data, was adopted to avoid any confusion with open access journal publishing. However, the distinction between scientific information and data is relatively fuzzy, especially in the context of emerging concepts such as data publication, making this distinction cumbersome. Moreover, the publication landscape has dramatically changed, with an influx of open access publishers, and a widespread advocacy of open access practices. As a result, ICSU's definition must be reconsidered. Although there has been movement towards full and open access, ICSU's meaning of open access is still not in line with that generally accepted. WDS would appreciate clarification on this.

3.2. For WDS Members

Although WDS Members share the principle of providing open access to their holdings, in general, there is huge variability in attitudes towards data sharing across research disciplines, and only 25% of research data collections are thought to be openly available. This discipline-dependency is a microcosm of national implementations of sharing principles, which even at this level are highly inconsistent, no matter internationally.

In part, this inconsistency is considered to be due to unrestricted data sharing being contradictory to the highest bodies in the scientific community, as well as the majority of academic journals and funding agencies, that value scientific novelty over long-term data stewardship. The success of open access depends on the ability to address this discrepancy.

a. Incentivising scientists—Data citation?

WDS Members identify the biggest challenge to open access as convincing scientists to share their data. Scientists often feel that there is a cost and no immediate benefit or reward for data sharing (others are perceived to receive most of the benefits). In particular, scientists focused on the production of datasets do not receive adequate professional recognition for their efforts. Consequently, there is increasing support from the scientific community for peer-reviewed data publication and data citation when using published data. Several WDS members state that they will only provide data if they are recognized as the data source, while others insist on 'rewarding' others through citation.

It is unclear, however, how effective an incentive data citation is and whether citation is a reward in itself. Traditionally, the culture of science has been to predominantly judge a researcher's merit—and hence their evaluation and funding—on the number and quality of their peer-reviewed publications. This culture encourages restriction of data access such that researchers maximize their number of publications, and provides little enticement to compile and document data beyond the needs of the original research. Hence, academia at large needs to endorse fair and formal credit and attribution for data producers, which must in turn be recognized by evaluation and funding bodies (in a manner similar to the h-index). By doing this, not only is a clear incentive established for data providers to share data but also efforts to acquire new high-quality data by providers are promoted and reinforced.

Attempts at data citation are not considered to have been particularly successful thus far, with citation of datasets often missing or incorrect. Internationally correct citation of both literature and datasets is vital for the reproducibility of science. Areas of responsibility for ICSU to rectify this include development of standards for citing data, such as where in a journal article data be should cited; and assignment, management, and versioning of unique (digital or analogue) data identifiers to create stable links for cited data. Data identifiers are recognized by WDS members as being particularly pertinent, and scientists must be encouraged to cite datasets by referring to those data identifiers when they publish their results. Existing DOIs can be used as data identifiers or a set of new identifiers can be applied to existing data alone.

b. Technical difficulties—The need for infrastructure

The above implies that an infrastructure exists to curate, disseminate, and publish research data. Data curation requires dedicated data centres; but, a lack of professional data management training and resources, as well as inadequate support for data archives, is evident within research programmes. Moreover, so that it can generally be reused by others, data must be properly described. Making data available thus involves significant effort and cost; even more so for real-time and provisional datasets, which require further processing to use in precise scientific analyses. The cost of long-term data preservation is hence one of the impediments to continued access, and the most appropriate business model appears to be linked to the awarding of a predictable percentage of grants.

It is obviously not possible to keep everything indefinitely, however, and some scientific grounding for the practice of 'what to keep and what to throw away' is necessary. Furthermore, it is often difficult for open archive managers to know what practices are allowed and what are not. An increasing need hence exists for informatics specialists, and basic data management training should be included in the core scientific curriculum such that researchers of all types take a 'data class' as part of obtaining an advanced degree.

Most current journals are patently not intended for data publication, and the issue of journals especially designed for publication of scientific data should be encouraged by ICSU.

c. The role of funding agencies

Many institutions and projects with relatively short-lived funding are not in a position to provide the required infrastructure for open access. It is clear that funding agencies have a strong part to play in encouraging open access. The strongest incentive for data sharing is thought to be when data deposition into an identified, financially supported open archive is a requirement for on-going research funding. Although, whether such a hard requirement leads to lower quality data being deposited is open to debate. Agencies that fund data collection must therefore assume responsibility for data management and develop the necessary infrastructure to support their preservation and use. To achieve this, a basic requirement of project proposals—which do not normally identify individual datasets to be delivered—should be the inclusion of data management plans. Moreover, if data is to be available for free (or at minimal cost), all funding agencies must agree to policies that cover author publication charges in their grant awards.

d. The extent of open access—Embargos and exclusions

Some exceptions and qualifications to the extent of providing unlimited open access are considered valid. Current restrictions to open access again often stem from pressures on researchers to publish, which has fostered a practice of embargoing data until formal publication. While an ICSU general policy of full and open exchange of data and information for scientific and educational purposes is lauded, WDS Members feel that this right has to be balanced by reasonable measures to allow researchers to exploit the academic value of their work without undue competition, typically in the period leading up to publication of a paper or thesis.

One WDS Member grants a one-year retention period to allow investigators time to properly analyse, document, and publish their data before submitting them in standardized format. Another Member permits an embargo period that exists for the life of a project (in most cases, less than 4 years) because papers are typically not written until towards the end of a project. However, with different research domains moving at different speeds and having different lengths of relevance, the correct period for embargoing data is a contentious issue. What works for one field may be completely inappropriate for others, and a uniform embargo period (e.g., 12 months) would be too long in well-funded and fast-moving disciplines and too short to be practical in fields such as mathematics, in which articles have citation lifetimes of years.

Natural restrictions may also apply when data centres are not the primary source of the data. In this case, the data exchange policy of the originator should be applied and the contract with that provider regulates the accessibility of each dataset. Unfortunately, this can lead to conflicts since data upon which research results are based cannot be made open because of the originator's wishes. Open access to data can therefore depend on the willingness and efforts of individual members. In fact, WDS Members with open access data policies state that this can sometimes cause concern for potential international collaborators.

e. Lack of trust? A 'common' answer

A significant barrier to data sharing seems to be a general lack of trust between data providers and users. Providers do not trust others with their datasets, fearing that users will apply data incorrectly, misuse it for ethical or legal breaches, or gain financially from it. In contrast, users do not trust the accuracy or believability of data or results that they find in journal articles. Such prejudices must be removed and data producers, curators, and users need to establish closer working relationships based on mutual trust. To aid in this process, more research is needed to create effective incentives that cultivate trust and motivate data sharing. An evolving legal and social science research agenda is also required to balance society's need for open data and protection of people, heritage, endangered species, and cultures from misuse.

Although in reality there appears to have not been great take-up of the concepts behind the Polar Information Commons, the most popular solution to problems of trust is the creation of an open 'commons' that allows researchers to place data in the public domain with limited retention of intellectual property rights through permissive licenses. Historically, there has been an absence of data-sharing norms and traditions within research communities. To develop such a commons, expected norms (not legal requirements) on the behaviour of data users and providers as regards ethical, collaborative data sharing must be defined and asserted. Publishing of standard global licenses (similar to Creative Commons licenses) is beneficial since when, for example, data is classified because of conservation concerns, everyone will be served by a globally accepted and understood license.

f. (Inter)National Policy

Implementation of an efficient scientific data commons may be difficult in practice because of the proliferation of open access copyright licenses used by governments. Such licences should be standardized to streamline permissions and facilitate efficient sharing and reuse. However, standardization requires support at the international level, and work is needed to harmonize governmental policies across national jurisdictions in accordance with common principles of openness and ethical use.

Although data policies should include clear identification of roles, responsibilities, and resources, the details of data policy will necessarily vary with discipline and research culture. Therefore, policy development must be a dynamic and interactive feedback process that involves all stakeholders and with active leadership by the sponsors of data collection. Again, the most adaptable policy approach is one based on community-accepted norms of ethics and behaviour rather than rigorous enforcement of licenses and contracts.

Unfortunately, tensions often exist between the drive towards free and open access, and current or future national legislation, especially those dealing with intellectual property rights and protection of information. The intent of these Acts is typically not of concern; nevertheless, they can be abused to restrict access and thereby harming knowledge creation and socio-economic development.

Tensions are particularly harsh between the public and private use of publicly-funded research. Ideally, all data, information, and research outputs generated by state-funded means should be included in national policies of open access. However, developing countries often regard research data as sensitive, based on (possible) future commercial value or on conservation implications (natural resources). In the case of developed countries, for areas of publically funded research in which they have a competitive advantage, it can be considered to be against national interest to publish everything since industry is supported by the margin created through intellectual property. Specifically, there is a 'public' obligation to create wealth for the domestic economy. One solution to this problem would be to mitigate those that derive a significant proportion of their income from the sale of publicly funded data.

4. Final comments

The diversity of approaches to open access reflects the complexities of its implementation. The public domain status of factual data is an especially complex subject in terms of legality. For this reason, it is highly desirable for ICSU to take a leading role in both open access of scientific data and its publication. It is also very important for the entire scientific community that ICSU show a strong commitment to this issue. To achieve this, it needs to encourage broader research focused on producing valuable scientific data rather than prestigious publications. ICSU WDS must be consolidated and promoted as the system of choice to achieve open access to ICSU-sponsored research data. The chances that a data management plan is successful are increase by including data centres and data service providers in the initial phases of research planning. In concert, CODATA and WDS can help maximize the chances of open access to the ICSU research data legacy by providing an appropriate mix of lawyers, data experts, and practitioners.

In the future, we hope that ICSU research programmes and their funders will have increased interest in, and active involvement with, ICSU data bodies (CODATA, WDS, INASP) and other partner organizations (ICSTI, GEO, etc.). Future Earth, the current ICSU flagship programme, is the first of its kind to be designed by both scientists and their funders, and unfortunately, it is still not addressing the important issues of data management and data legacy.