**Upstream – Downstream Subgroup**

***Issues in the Relationships of Not-for-Profit Research Data to Innovation***

**CODATA Data Policy Committee (DPC)**

**Membership:** **Chair**, LEE Jeonghoon. **Members**: Kevin ASHLEY, (Mariel BOROWITZ), CHUANG Tyng-Ruey, Usha MUNSHI, (Jerome REICHMAN), Paul UHLIR, YARIME Masaru.\*

**Scope:** A white paper will be drafted in order to define the scope of several issues/problems and what useful follow-up activities by the DPC or others could be.

**Topic:** There are many topics of potential interest on the relation of public data with innovation.

The following specific issues have been articulated and may be addressed in the white paper:

1. **Release of semi-commons data.** Upstream research data can be made available through a broad, public-domain commons or a more limited “semi-commons” (research project or consortium), with some of the data released more broadly either during the project or after it is completed. We may want to develop rules for the financially sustainable release of the data and data-related tools (such as software) in semi-commons projects.
2. **Broadening the fair use doctrine for research.** There are different approaches in the world’s legal systems to copyright. For example, the 3-step test in “fair dealing” tends to shut down uses, whereas the four-step, fair-use doctrine is much more liberal for users. Also, many legal systems have no such exceptions at all. There may now be a trend towards a global fair-use approach. For example, Brazil, South Africa, and some other countries are raising it at WIPO. What can be the implications of this for research databases? How does the exception differ for literature (text) versus for databases, which may not even be copyrightable? For example, it is generally recognized that text data mining (TDM) is legal under the fair-use exception in the U.S., whereas there are significant problems with TDM in the EU and elsewhere. May want a fair use-like exception mandated under other IP laws for non-copyrightable databases. Also, may want a fair-use exception for data obtained by the government from the private sector.
3. **Use of commercial data for public interest uses.** This has several dimensions. One is where the private sector data are mandated to be disclosed to some regulatory agency for regulations, which may or may not be available to the public. Another is partial disclosure. Another may be the diminution of the public domain through government contracts. Need to consider the raw versus more processed data, since the re-use rules should be different.

As part of this, we should examine and attempt to minimize the restrictions on open data due to commercial confidentiality. What does “commercial interest” mean? What limits should it impose?

1. **Pooling of data with mixed restrictions for integrating diverse data sources.** The greatest restrictions can infect the new whole. Need a common agreement and a waiver of the restrictive rules. It is a legal interoperability issue (see the RDA Interest Group on Legal Interoperability’s Principles and Implementation Guidelines). It is also a financial sustainability issue. Need suggestions for governmental policymakers or the research community (i.e., universities).
2. **The emerging legal equivalence of digital and physical data sources.** Over the past decade there has been a rise in the creation of life from a genetic blueprint, which has been patented. This has profound legal, socioeconomic, ethical, moral, and religious implications that are sure to be addressed from now on. What are the implications of the redefinition of what is a database and what is a genetic resource, from a purely legal perspective? What are the research data access and use issues? What’s being obscured here is that everything in the CBD, for example, is from a physical genetic resource. This may have major trade implications. We are also expressly excluding synthetic biology and “3-D printing” (everything is a database). IP rules are different for digital and physical objects, which greatly complicates matters.
3. **Research data as a non-monetary benefit under the Convention on Biological Diversity (CBD).** There are rules under the Access and Benefit Sharing (ABS) provisions of the CBD that have expressly considered the sharing of research elements as a “benefit” from more economically developed nations to less economically developed ones. The sharing of research data is a part of that benefit. What exactly does that mean? This could have serious implications for freeing up resources and countries should be encouraged to apply that.

One may need a “change of intent” clause and to define what that means. One should not prohibit commercial uses (whether by private-sector or public research entities) of public research data, but need to have some equitable sharing of profits, if profits are generated. “Intent” can be a big factor.

1. **Examine the existing and potential uses of public data in innovation.** Many public research data are subsequently used in private-sector innovation. What is known generally about such uses? For example, in the United States (US), well-known instances are the use of National Institutes of Health clinical trials data for the development of pharmaceuticals or the use of public geospatial data for developing many commercial applications. What rules should govern such activities?
2. **The commercialization of knowledge (research data) at public universities.** In 1981, the US passed the Bayh-Dole Act, which sought to promote the commercialization of university research, especially biomedical results, for broader societal applications. This was soon followed by the Thatcher revolution in the UK, which broadened that approach to all forms of university research, including completely non-commercial forms of inquiry. A whole institution to perform these tasks then sprang up at the universities in the form of Technology Transfer Offices (TTOs). Public-private partnerships (PPPs) were also encouraged, as was IP protection and enforcement of various forms of public research at the source. Since then there has been a debate about the conflicts this approach has had with the traditional roles of public education, knowledge creation, and dissemination; and between the public good (economic), public domain (legal), and public interest (political) versus the privatization and commodification of research and knowledge. Many other countries have adopted similar legislation. **Research data** are one of the central inputs and outputs of this debate, and this should be explored.

**Process:** A series of monthly calls by this subgroup of the DPC will be held to guide its work. The substantive and procedural aspects will be agreed in the January-May 2018 time frame. In the June-December period, one or more of the subgroup participants will each draft a description of one or more issues or problems that would benefit from further analysis and next steps for solutions to those problems. Each of these will be a maximum of two pages of text plus a selected bibliography that surveys (with key words) at least the main sources of information and is limited to 2-3 pages as well.

The text will include the following elements in a common template, each consisting of a very short description, that will:

(1) summarize the issue/problem and describe why it is important;

(2) identify who the main players are and where they are located;

(3) provide a brief timeline of past activity and future prospects; and

(4) recommend next steps for how such problems might be addressed and a solution implemented.

The drafts will be shared with the members of the subgroup and discussed in (approximately) monthly calls and in email exchanges. The white paper also was discussed in a face-to-face meeting of DPC members at the International Data Week conference in Gabarone, Botswana in early November 2018. A determination of how each issue/problem will be followed up, and how the compiled set of issues/problems will be published, will be made by early 2019.

**Output and Outcome:** The subgroup will produce a white paper by early 2019 (including internal review) that substantively explores the issues and recommends where there is a need for further work and how a solution may be implemented. All subgroup members who contribute text and remain engaged in the full process will be listed as authors.

\* NOTE: Group members in parentheses are ad hoc members of this project and not members of the full CODATA Data Policy Committee (DPC).