Freedom to Research: Keeping Scientific Data Open, Accessible, and Interoperable

Thinh Nguyen
Counsel
Science Commons, a project of Creative Commons

According to Thomas S. Kuhn, scientific revolutions occur when a sufficient body of data accumulates to overthrow the dominant theories we use to frame reality—a so-called paradigm shift. At a certain point, we can no longer ignore the fact that the old models don't appear to be working or producing the results we want.

As an outgrowth of our work with the scientific community, we at Science Commons have had our own paradigm shift. The result is the Science Commons Protocol for Implementing Open Access Data, a set of principles designed to ensure that scientific data remains open, accessible, and interoperable. Creative Commons' announcement of the beta CC0 waiver is another milestone in this shift; the waiver is a new legal tool, along with the Open Data Commons Public Domain Dedication and License (PDDL), that implements the Protocol.

The Protocol came about as a response to the growing recognition, increasingly among scientists themselves, that the trend towards applying licenses and click-wrap agreements to data can not only threaten innovation and productivity, but also scientific freedom. Any researcher who needs to draw from many databases to conduct research is painfully aware of the difficulty of dealing with a myriad of differing and overlapping data sharing policies, agreements, and laws, as well as parsing incomprehensible fine print that often carries conflicting obligations, limitations, and restrictions. These licenses and agreements can not only impede research, they can also enable data providers to exercise "remote control" over downstream users of data, dictating not only what research can be done, and by whom, but also what data can be published or disclosed, what data can be combined and how, and what data can be re-used and for what purposes. Imposing that kind of control threatens the very foundations of science, which is grounded in freedom of inquiry and freedom to publish.

A growing community of scientists and researchers, mindful of these dangers, are struggling to find ways to ensure that data remain free and open. However, finding the right model has been difficult. They are confronted with hard questions: Should we guarantee freedoms by borrowing ideas and tools, such as "copyleft" licenses, from the

open source and free software movements? What kinds of property rights apply to data and databases in different jurisdictions, and what happens to those rights when we share data globally on the Internet? What's the best way to ensure the interoperability of data at a technical and legal level?

These are issues Science Commons has been exploring in collaboration with data licensing experts from around the world. In 2006, we hosted the *Information Commons for Science* congress at the National Academies of Science in Washington, D.C., where eminent scientists and scholars from the United States and other countries gathered to discuss data sharing strategies. In September of 2007, we co-sponsored with CODATA the *Workshop on Common Use Licensing of Scientific Data Products* in Paris, which included representatives from the Global Biodiversity Information Facility, and leading legal scholars, scientists, and CCi collaborators actively involved in working on data sharing policy. It was through this collaborative process that the Science Commons Data Protocol emerged as the best—and possibly the only—solution to the challenges we collectively identified.

Before developing the Protocol, Science Commons offered guidance to scientific database providers through the Science Commons Database FAQ. This document explained how and when it was best to use Creative Commons licenses for databases. In general, it encouraged providers to apply Creative Commons licenses only to copyrightable content, while also encouraging them to clarify that no restrictions or obligations were asserted on facts, ideas, and other uncopyrightable content.

There were two problems with that approach. First, it proved very difficult, not only for scientists but also for lawyers and legal scholars, to provide useful guidance on when copyright stops and the public domain facts begin. This problem is compounded when multiple jurisdictions are involved, as is the case with collaborative online global databases. Second, facts and ideas may also be protected as such in some jurisdictions under a database copyright theory, or under *sui generis* database rights, or both.

For example, consider a biodiversity database that has collaborators contributing from all around the world, implicating the laws of many countries. Under U.S. law, databases are protected by copyright if there is some degree of creativity in its compilation, while facts and ideas themselves are generally not protected. A Canadian contributor, on the other hand, might be entitled to copyright protection under a slightly different standard: whether sufficient “skill and judgment” is involved. An Australian contributor is entitled to copyright protection for databases as a consequence of “industrious collection” or “sweat of the brow.” A European contributor may have both copyright protection as well as *sui generis* protection for data and databases—a special protection enacted by directive of

the European Union that grants protection for database owners from unauthorized extraction and re-use of data.

Not only do different legal standards apply in different countries, but within any legal standard, it can be very difficult to distinguish between what is and is not protected. For example, what is the level of creativity needed to protect a database in the U.S.? What is the level of skill needed in Canada? Or the levels of industriousness or economic investment required in Australia and the E.U., respectively? These are questions that can only be resolved over time, through individual court cases. Unfortunately, that's of little comfort to the data provider who must decide on the right policy today, and whose expectations and assumptions may be upset by future court cases.

Therefore, while our guidelines were intended to accurately reflect the law, they were extremely difficult to apply in practice, because the risk of a legal misclassification is irreducibly high. This left scientists and other non-lawyers with little practical guidance on what steps to take when they wish to signal their intention to keep their databases open and free of restrictions.

The patchwork nature of legal protection challenges the coherence of any scheme of database licensing, not just the Creative Commons licenses. Any license is premised on the existence of underlying rights. However, when those underlying rights are highly variable and unpredictable, a dilemma exists for both data providers and data users. Data users will not be able to predict accurately when compliance is necessary, and may under-comply or over-comply, both of which has its problems. Data providers, on the other hand, may be given a false sense of security when providing access to data, only to find it impossible later to enforce the terms of the license consistently on a global basis. Collaborative global data projects may face a different danger: different contributors may hold different, unpredictable rights, with potentially unwanted consequences for the continued openness of the project.

Similarly, Web site terms of use, click-through contracts, and other online contractual restrictions on data give rise to similar uncertainties. Some jurisdictions may enforce them and others may not. Another problem with contractual schemes of data protection is that they can sometimes create conflicting or overlapping obligations. These not only produce administrative burdens, but they have the potential to render entire datasets non-interoperable with others for the purpose of data aggregation and transformation. For example, copyleft or share-alike licenses often require the user to distribute derivatives only under an identical license. But when combining two datasets under different copyleft agreements, there would be no way to comply with both at once. Thus,

such a system would only work if everyone in the world used the identical license, a situation that seems unlikely in our current environment.

The core insight behind the Science Commons Data Protocol is that the solution to these problems is to return data to the public domain by relinquishing all rights, of whatever origin or scope, that would otherwise restrict the ability to do research (i.e., the ability to extract, re-use, and distribute data). The goals of the Protocol are to keep data open, accessible, and interoperable, and its virtues lie in its simplicity, predictability, and consistent treatment of users and data. This approach is consistent with established scientific norms, expectations, and practices with data, and so, in some ways, we are not proposing anything new. There are already many data sharing initiatives that have long been doing what the Science Commons Data Protocol recommends: waiving all rights and placing data in the public domain without restriction. What we seek is to map out and enlarge this commons of data by seeking out, certifying, and promoting existing data initiatives as well as new ones that embrace and implement these common principles, so that within this clearly marked domain, scientists everywhere can know that it is safe to conduct research.

Many data providers have valid concerns leading them to consider licenses or other contractual mechanisms. Such concerns can range from protecting the integrity of the data or project, ensuring appropriate credit and attribution, or identifying provenance. However, for reasons discussed above, using legal mechanisms is an unreliable strategy, with potential for creating additional problems. That is why a critical component of the Science Commons Protocol is promoting the development of community norms and best practices that help to achieve these goals through voluntary, technical, or informal strategies.

So in Kuhn's terms, this is a revolution, in the sense that it departs from a trend towards placing increasing restrictions on data. But in another sense, it is far from revolutionary, because it is how science has worked best in the past and, we believe, how it must work in the future if we want science to help us find cures, meet the challenge of global threats to our welfare, and promote economic development and shared prosperity.
