Comment

Public domain, copyright licenses and the freedom to integrate science[*]

John Wilbanks

From the life sciences to the physical sciences, chemistry to archaeology, the last 25 years have brought an unprecedented shift in the way research happens day to day. The traditional cycles of research, beginning with a study of the relevant journal articles and books, moving into experimental design and data gathering, hypothesis formulation and testing, and finally republishing new knowledge into the scholarly canon, remain in place. But the quantity of information now available at each step of the cycle has exploded, and the average scientist is now simply awash in data.

This explosion of data brings with it many attendant issues: software to manage data, annotation systems to make sense of data, nomenclature standards, visualization, and on and on and on. Each of these issues deserves its own research. The annotation of data alone is an incredibly dense and complex field, with standards that can take years to emerge for relatively well-evolved experimental tools. I cannot hope to do justice to all of these issues in a single paper of any length.

For the purposes of this paper, I will instead focus on a relatively specific use case: the integration and federation of an exponentially increasing pool of data on the global digital network. I will furthermore explore the question of the legal regimes available for use on this pool of data, with particular attention to the application of "Free/Libre/Open" copyright licenses on data and databases.

The application of such licenses has the potential to severely restrict the integration and federation of scientific data. The licensing approach is likely to result in high costs and a confusing fog of interlocking contracts, and at best slow, but at worst prevent, the emergence of an integrated web of data. The public domain for science should be the first choice if integration is our goal, and there are other strategies that show potential to achieve the social goals embodied in many common-use licensing systems without the negative consequences of a copyright-based approach.

Open licensing

Copyrights and their licensing have become a central part of the network world. Copyrights are a powerful form of socially constructed property rights that grant sweeping rights to creators – authors, photographers, software developers – to control the copying and distribution of their work. Open licensing has emerged as a social movement enabled by computer networks, in which the strong powers created by copyright are used to further the goals of individual freedom. Open licensing is also a methodology to resolve the inherent conflict between the capabilities of digital technologies, which allow for essentially zero-cost copying and distribution of digital content, and copyright laws that have expanded over time to provide monopolies lasting 70 years and more to content owners.

Free/Libre/Open licensing (FLO) comes to us from the world of software. A free software license is a contract granting to users the freedom to make modifications to software, and to then copy and distribute the software, in a way that copyright law by default makes illegal.[1] This freedom is famously embodied in the Free Software Foundation's Four Freedoms:

- The freedom to run the program, for any purpose (freedom 0).
- The freedom to study how the program works, and adapt it to your needs (freedom 1). Access to the source code is a precondition for this.
- The freedom to redistribute copies so you can help your neighbor (freedom 2).

[*] The following document is a derivative work of other works by Thinh Nguyen and John Wilbanks. This document is licensed to the public under Creative Commons Attribution 3.0 unported.

- The freedom to improve the program, and release your improvements to the public, so that the whole community benefits (freedom 3). Access to the source code is a precondition for this.[2]

A variation on the free/libre software license is the open software license. This is also a contract based on copyright that allows users to make modifications to software and then to copy and distribute the software, but typically represents a focus on development methodology rather than the social goals of freedom and autonomy embodied in the Free/Libre licensing movements.[3]

FLO licensing has worked very well in the software world. It has been central to the development of the Apache Web Server and the GNU/Linux operating system. Companies as traditionally conservative as IBM invest heavily in software running under these licenses, and millions of users worldwide use the Mozilla Firefox browser under a FLO license.[4]

However, the ideas behind the FLO licenses for software have propagated beyond the boundaries of source code and compilers. The GNU General Free Documentation License (GFDL)[5] was one of the first FLO licenses intended for regular text instead of of computer code, and the Creative Commons project launched a suite of copyright licenses designed specifically for use on cultural works in 2001.[6]

These licenses have also achieved real success. GFDL is the license on Wikipedia.org, and the Creative Commons (CC) licenses now cover more than 100,000,000 digital objects worldwide, with more than 67,000,000 digital photographs at Flickr alone.[7] The CC licenses have also been "localized" linguistically and legally in nearly 50 countries.[8]

Managing the law around data

As data explodes, it becomes more and more clear that scientists, users, and providers need a clear legal regime to manage data and databases.

There is a growing trend towards applying licenses and clickwrap agreements to data, forcing a researcher who needs to draw from many databases to deal 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 reused 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?

Porting the FLO licenses for data

It is very tempting to use FLO licensing regimes for data. These regimes provide a variety of choices, ranging from the "share-alike" quid pro quo to the non-commercial restrictions of the Creative Commons suite. FLO licenses provide a clear signal to the user of the provider's intent – if you use this, you are using an open system, and if you violate the terms of the open system, you can be pursued in court. And the FLO licenses also signify the provider's membership in a social movement, a social commons, marked by iconography (the GNU gnu or the "CC in a circle logo") and common language ("free as in speech" and "I Like To Share").[9] [10]

However, any approach built on licensing databases under a CC license or the GFDL is marked by a difficulty: what part of the database is copyrightable, and thus subject to the license terms, and what part of the database is not a creative work? If the database contains measurements about the height of hills in and around Boston, then no matter what the license terms on the database, a Creative Commons license will not prevent the extraction and republication of the global positioning data about those hills.

It is difficult for seasoned attorneys skilled in database practice to determine with accuracy where copyright begins in and ends in many databases – much more so for non-lawyers. As Abraham Lincoln famously noted in the United States, a house divided against itself cannot stand – it must become all of one or the other. A database divided into copyrightable and non-copyrightable elements suffers a similar fate: the user tends to assume that all is under copyright or none is under copyright. And the decision dictates which part of the "license" the user decides to comply with.

There are at least three significant problems with a licensing approach based on using intellectual property rights to enforce norms of attribution, share-alike, or other terms.

First, any solution based on rights will result in categorization errors: the application of obligations based on copyright in situations where it is not necessary (for example, a share-alike license on the copyrightable elements may be falsely assumed to operate on the factual contents of a database). In the reverse, a user might assume that the "Facts Are Free" status of the non-copyrightable elements extends to the entire database and inadvertently infringe.

We do not know what courts will decide in the future. But it is conceivable that in 20 years, a complex semantic query across tens of thousands of data records across the web might return a result which itself populates a new database. If intellectual property rights are involved, that query might well trigger requirements carrying a stiff penalty for failure, including such problems as a copyright infringement lawsuit. It's also probable that simply running a Google query across databases might trigger licensing requirements, even if the query returns *negative results* from underlying databases.

These interpretative problems are exacerbated by differences among countries over the standards for copyright protection for databases, by the existence of sui generis database rights,[11] and by the difficulty of interpreting contractual language.

Second, there is also the problem of false expectations. Many users choose to apply common-use licenses such as the GPL and CC in order to declare their intent: thus, a user might choose to apply a "copyleft" term to the copyrightable elements of a database, in hopes that those elements result in additional open access database elements coming online. But a user would be able to extract the entire contents (to the extent those contents are uncopyrightable factual content) and republish those contents without observing the copyleft or share-alike terms. The data provider, having thought she was protected, is likely to feel "tricked" by this outcome. That is not a desired result.

Last, there is a problem of cascading attribution if attribution is required as part of a license approach. In a world of database integration and federation, attribution can easily cascade into a burden for scientists if a category error is made. Would a scientist need to attribute 40,000 data depositors in the event of a query across 40,000 data sets? How does this relate to the evolved norms of citation within a discipline, and does the attribution requirement indeed conflict with accepted norms in some disciplines? Indeed, failing to give attribution to all 40,000 sources could be the basis for a copyright infringement suit at worst, and at best, imposes a significant transaction cost on the scientist using the data.

Users deserve a clear set of answers

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 undercomply or overcomply, 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, clickthrough contracts[12], 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 noninteroperable with others for the purpose of data aggregation and transformation. This conflict drove the HapMap project[13], a follow-on to the Human Genome Project, to abandon a registration and click-through strategy created to preserve freedoms to utilize data.[14]

For example, copyleft or sharealike 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.

Organizing principles for data regimes

Rather than attempting to fit a cultural or software key into a data opening or depend on the collapsing of the world into a single licensing format, we can instead cast a different net for data management. We developed a methodology for evaluating legal tools for an open data sharing with three key principles in mind: legal predictability and certainty, ease of use and understanding, and low costs to users.

These principles are motivated by our experience in distributing a database licensing Frequently Asked Questions (FAQ) file. We found that scientists were uncomfortable applying the FAQ, in which we advised on how to use a CC license on the copyrightable elements of a database and how to manage the uncopyrightable facts stored in that database, because they find it hard to apply the distinction between what is copyrightable and what is not copyrightable, among other factors.

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 reuse 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 nonlawyers 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.

We concluded that any usage system must both be legally accurate while simultaneously very simple for scientists, reducing or eliminating the need to make the distinction between copyrightable and non-copyrightable elements.

The terms also need to satisfy the norms and expectations of the disciplines providing the database. This is different from software, where archaeologist uses the same browser as the football enthusiast, or culture, where the same license governs a photograph of a church as a photograph of a laboratory. This normative diversity makes a single license approach difficult – archaeology data norms for citation will differ from those in physics, and yet again from those in biology, and yet again from those in the cultural or educational spaces. But those norms must be attached in a form that imposes the lowest possible costs on users (now and in the future).

The public domain as first recourse

We concluded after a lengthy investigation that, for the purposes of the sciences, the conflict between simplicity and legal certainty can be best resolved by a twofold measure: 1) a reconstruction of the public domain and 2) the use of scientific norms to express the wishes of the data provider.

We did not reach this conclusion lightly, or alone. 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 cosponsored 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. We were also informed in our work by our participation in the Global Information Commons for Science Initiative, or GICSI, a multi-stakeholder initiative emerging from the World Summit on the Information Society at Tunis in 2005. It was through this collaborative process that the reconstruction of the public domain emerged as the best—and possibly the only—solution to the challenges we collectively identified.

Reconstructing the public domain can be achieved through the use of a legal tool (waiving the relevant rights on data and asserting that the provider makes no claims on the data). Given the significant amount of fundamental data in the public domain in the United States and elsewhere, we elected to not write a single license, but instead to craft a protocol for evaluating database terms of use, in hopes of providing a unified framework for users to evaluate if any given database may be integrated with any other database.

The protocol[15] calls for data providers to waive all rights necessary for data extraction and re-use (including copyright, sui generis database rights, claims of unfair competition, implied contracts, and other legal rights). The protocol further requires the provider place no additional obligations such as copyleft or share-alike or attribution. Requesting behavior, such as citation, through norms and terms of use rather than as a legal requirement based on copyright or contracts, allows for different scientific disciplines to develop different norms for citation. This allows for legal certainty without constraining one community to the norms of another.

In many jurisdictions there are other rights, in addition to copyright, that may apply. For example, sui generis rights apply in the European Union, and uncopyrightable databases may be protected in some countries under unfair competition laws. Thus, the protocol calls for waivers of sui generis and other legal grounds for database protection.

There is always the possibility of using contract, rather than intellectual property or statutory rights, to apply terms to databases. This fails to provide legal certainty, ease of use, or low transaction costs, as it forces scientists to either hire a lawyer or interpret contracts themselves. The protocol therefore calls for providers to affirmatively declare that no contractual constraints apply to the database.

There will be significant amounts of data that is not or cannot be made available under this protocol. In such cases, it is desirable that the owner provides metadata (as data) under this protocol so that the existence of the non-open access data is discoverable.

Reactions to the protocol

We have encountered a wide range of reactions to our protocol, from strong enthusiasm[16] to strong distaste[17], across a swath of user communities from biology to chemistry to geospatial. I will provide a brief overview of the most common reactions in disagreement (those who agree with the protocol are represented well by its description). We can divide the disagreements into three classes: the first represents a desire to enforce behavior on users, the second a legal uncertainty about the internationalization of the public domain, the third a skepticism of the risks of attribution stacking.

The first class of reaction comes from users. The most common reaction can be encapsulated as "I want to ensure that downstream users have to re-contribute – freedom isn't created without a share-alike legal provision." This is not an unexpected response, and indeed, represents where we began our research in 2006. The success of the GNU GPL, the Creative Commons share-alike licenses, and other copyleft licenses has played a vital role in the creation of a vibrant commons of free/libre software and content. From Rufus Pollock comes an interesting argument that the data-licensing in the "GPL style" is similar to the use of the GPL in dynamic programming languages such as Python.[18] To our point that every database query produces itself a data product, and thus that even the results of Google queries would need to carry licensing in a world of GPL-style database licensing, he points to Python, where programs write new programs using the GPL.

It is natural that users wish to project a successful experience, and indeed an ideology of freedom, from those worlds into the data world. We have received this response most strongly from groups of geospatial data users, but it is common in general in communities whose primary goal is not interoperability with other databases. This response also correlates to nationalities that have data protections embedded in their own countries – again, unsurprising, as these are users whose native experience of data is as being socially constructed out of the public domain.

Another reaction in this class is the fear of data "theft" or use without attribution. Similar to the share-alike desire above but based more on the desire to maintain proper rights of the depositor than to ensure downstream freedom, this reaction also presumes that the legal agreement creates the mandate more effectively than any other tool. This is a common reaction more from the scientific communities than the distributed, user-driven community projects.

The second class of reaction comes from our colleagues in the international Creative Commons and in open science. The public domain is a concept that is very strong in the United States, especially in the scientific data space. The human genome, geographic data, NASA photographs, and more all are naturally in the public domain. And in the US, there is no tradition of moral rights in copyright (footnote on moral rights) – but outside the US, moral rights on copyright cannot always be waived. And the previously mentioned variety of data rights, from the European sui generis right to Crown copyright to "sweat of the brow" rights, present a tempting source of property right "hooks" with which to enforce behavior on users.

The second international reaction comes primarily from the developing world. The history of the 20th century brings too many instances of the expropriation of knowledge from the less developed nations to the more developed nations. The scholarship around the commons and the developing world teaches us to tread carefully in proclaiming the "global digital commons" to those whose traditional knowledge might be exposed in culturally offensive ways, or in ways that violate international treaties on biological diversity.[19] There is a real desire in these spaces for non-commercial restrictions on knowledge products and data and databases.

The third class of reaction tends to come from lawyers and programmers. This response states agreement on the risks that share-alike and non-commercial pose to data integration, but question our assertion that attribution stacking will be a problem in the future. In one set of arguments, lawyers state that the method of attribution can be specified in contract in a way that prevents stacking. In another, programmers note the ability to use metadata and other technical means to automate attribution, even at very large scales heading out into the federated web, using means akin to the trackback-ping mechanisms of the blog world.

The irony of the protection instinct

Our inspiration in the protocol was to create, at a minimum, at least one zone of legal certainty: the freedom to integrate databases. This is a tough goal, and the public domain was the only answer we found. And in each of the responses we see a common thread: the recapitulation of the control instinct common in the non-free/libre content and software worlds. The public domain relinquishes much of the control – even control in the service of freedom – to which we as users and producers of culture and software have become accustomed. This is a difficult thing for many to embrace, on both sides of the freedom debate.

The moral rights issue is a real one, as is Pollock's point about programming language similarity to data. These are issues to examine in depth and beyond the scope of this short paper.

Our instinct is that in many classes of scientific data, the mixture of moral rights and interlocking, but not interoperating, national data protection regimes, represents primarily a perceptual problem – it is difficult to imagine a "moral right" in the sense of an author or creator over a fact of nature such as the height of Mount Everest,[20] or the amino acid sequence of the p53 protein.[21] And in the absence of the moral right, even in a "sweat of the brow" or sui generis regime, the public domain can be reconstructed with a waiver of all relevant rights such as that enabled by the Public Domain Dedication and License.[22]

Pollock's point is more complex and deep. Our fear is that the encoding of norms from different data disciplines in different formulations of share-alike to create major barriers to data federation, query, and combination. Even a single word of difference in a definition can freeze share-alike interoperability, as we see in the need for GFDL-CC BY SA dual licensing at wikipedia[23] – the risks of killing the web of federated science data is too high to justify the benefits of "enforced" freedom on downstream users, which might or might not be enforceable depending on data type, jurisdiction, and the class of user.

The public domain in most of these reactions is portrayed as be a toothless choice – a place where rampaging companies and unscrupulous users can free-ride without fear of retribution. Users and providers alike feel more comfortable using social constructions based on copyrights and contracts than with the idea of a true public domain, a zone of absolute freedom. Alternately, the public domain is portrayed as a naïve dream – nice if you can get it, but unachievable. This rhetoric is reminiscent to that of the advocates for increased patent protection in the United States before the Bayh-Dole legislation in 1980.[24]

There is great irony in this position, of course. The idea of the public domain has been subjected to relentless erosion by corporate lobbying, legislative action, judicial decisions, international treaty adjustments, and more. Copyrights now last so long that Windows 95 will not enter the public domain until the latter half of this century at the earliest – and if current patterns of term extension continue apace, the 1995 operating system won't be public domain until 2100 or later.

In the arguments against the public domain in the cause of freedom we see recapitulated many of the arguments made against the public domain in the cause of protection. The unruliness of an unregulated system. The tragedy of the commons. The fear of theft, and the desire for control. These great ghosts familiar to us from the pro-control forces see echoes in the "open licensing" discussions, reaching closest to each other with the promotion of technical protection measures as a necessary complement to contract. There is absolute similarity in strategy between closed and open in the placing false data in data systems as a necessary complement to licensing, as "Easter Eggs" to prove violation of contract.[25] This is the irony of the control culture; it has even promulgated itself into the discussion of freedom.

One of the common reactions we hear in our arguments for the public domain in contrast to a copyleft-inspired license is that "this is the GPL v. BSD debate all over again"[26] – but it most definitely is not. That would be a copyleft-style license v. an attribution-requirement license, each for data. That would not be the public domain – the public domain is at its core an absence of control, a one-to-many grant of true freedom, not contractually constructed "freedom."

A different path built on norms

The public domain as a legal strategy for the data itself does not however leave us bereft of overall strategies to manage behavior. While the public domain does rule out the idea of using a contract or a copyright strategy to pursue the violators of "freedom" in the sense that anyone can come along, make a copy of a public domain database, and do whatever they wish – sell it, improve it, attach a viral open contract to it or a vituperously closed contract to it – and they know they cannot be sued, no matter what they do.

The original public domain database will just sit there, as it was, as copies propagate. That's what allows the endless transformative use that science is going to require if we are to achieve the kind of integration for databases that the web has brought to documents. We give up the right to patrol and enforce infringement, to sue and shame, in return for that integration.

But is that threat of enforcement the true core of an open community? Is openness truly based on fear?

This is a question that I also feel deserves another paper. But I will explore a few closing thoughts on other methods that a community can use to be open, to reward and punish without the use of copyright licensing and the threat of lawsuits.

The first is the use of normative systems. We can look to scientific communities that were very successful in creating robust public domains of data while using *norms* to control behavior rather than *contracts*. The human genome is perhaps the clearest example of this method – as DNA was converted from physical molecule to digital sequence, the scientists involved in the public Human Genome Project hammered out the agreement that those sequences would flow into the public domain within 24 hours.[27] Each sequencing center would retain some rights to publish the first papers on certain configurations, but the data went online without legal restrictions.

The system was not perfect, and some actors broke the norms. Given the hyper-competitive nature of genome studies funding in the 1990s and the pressure to publish, damn the consequences, that is not surprising. But on an overall basis, the genome flowed online without enormous complication and the sequencing centers got their first papers out, all in the face of a strong, incredibly well funded, well marketed private company that was not only competing with the public genome but *integrating all of the public data into the private genome*.

The public domain won anyway, despite its poverty of licensing. And thank goodness we have a public reference genome, because it would be decidedly hard to collaboratively annotate and explore a closed source genome. Another great irony is that, even without a share-alike contract, the private genome competition wound up simply depositing its own effort into the same public domain.[28] Sometimes the goals of the open movements can indeed be achieved without the force of contract.

This can be argued as a *sui generis* case – massive public funding, a clear public interest, the need to operate internationally and indeed synchronize a global database daily all leave no choice other than the public domain. And there were many other factors at play in genomics.[29] But it remains the impossible honeybee of the power of the public domain – despite the arguments that the public domain is too weak, here in front of is proof positive of its power, a legal method that eventually outcompeted a company worth billions of dollars at its peak.

We can build on this concept. If we abandon the idea of the "bad actor" and the concomitant requirement to constantly pursue and prosecute, we can articulate methods that reward the good actor – the person who obeys the norms.

Trademark is a form of property right that is ideally suited for experimentation here. One of the reasons many users join open communities is simply a passion for openness, rather than the desire to enforce. If a community applies a public domain strategy to its data, and articulates the desire for share-alike and attribution and so forth as normative statements, then it can also promulgate a set of trademarks for those who follow the norms. This rewards the "good guys" and prevents the capture of the community's identity or appropriation of the community's spirit. This is captured in the idea of the norms of the Public Domain Dedication and License's accompanying "Community Norms"[30] and the CC0 + Science Commons norms pilot project.[31]

Closing thoughts

The public domain is not a license. It cannot be made "more free" – only less free. It is not an "unlicensed commons" – it is the original commons. The key here is to understand that our choice is not a choice between GPL and BSD, or closed licensing versus open licensing. It is the public domain versus everything else. And if your answer must be "everything else" – whether because you're a company with trade secrets, or a virally open community of data harvesters that feels the public domain is a pit of quicksand – make your metadata open, so at least the public domain has a pointer to your information.

One point of congruence that we have found, again and again, even with those who disagree with our position within the open data communities – it is a sad commentary on the success of the control culture that even conversations around freedom rely on the vocabulary and ideologies of those who emphasize protection, and that freedom isn't free unless someone can get sued.

But nothing other than the public domain really works from the perspective of data integration. And data integration is coming at us at exponential speed.

It's the human genome, which must be mixed with annotations and protein structures and expression profiles and biochemistry and clinical trials. It's the petabytes of data generated by the satellites in our skies will have to mix with the petabytes of data that will come off sensor networks in the brunt of a hurricane, and again with geothermal data, and again with hydrological data, and again and again and again, if we are ever to understand climate change.

If we begin to create a culture in which we attach control at every layer, every time, we'll never get there. We'd be adding the problem of "too many contracts" to the problem of "too much data" and we'd be preventing the emergence of the power of the crowd to make the idea of "too much data" a quaint anachronism.

We have the power, those of us in the movements around open data, to drive the rhetoric. The choices we make about first principles will resonate. If we establish the license, the contract, and the lawyer as the core of free data, it's unlikely that we'll see a Wikipedia of databases, or a GNU/Linux equivalent for data federation. We will only see that day come if we focus on a core freedom, one more powerful than the freedom to bring suit – the freedom to *integrate*.

Notes and references

[1] http://en.wikipedia.org/wiki/Free_software_license

[2] http://www.gnu.org/philosophy/free-sw.html

[3] http://en.wikipedia.org/wiki/Open_source

[4] http://marketshare.hitslink.com/report.aspx?qprid=1

[5] http://www.gnu.org/copyleft/fdl.html

[6] http://creativecommons.org/

[7] http://wiki.creativecommons.org/License_statistics

[8] http://creativecommons.org/international/

[9] GNU Head by Victor Siame <vcopovi@wanadoo.fr>, http://www.gnu.org/graphics/official%20gnu.svg

[10] Meet the Licenses, http://creativecommons.org/about/licenses/meet-the-licenses

[11] Directive on the Protection of Databases, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML

[12] From *Specht v. Netscape Communications Corp.*, 150 F.Supp.2d 585 (S.D.N.Y. 2001), *aff'd*, 306 F.3d 17 (2d. Cir. 2002), via Wikipedia: “A click-wrap license presents the user with a message on his or her computer screen, requiring that the user manifest his or her assent to the terms of the license agreement by clicking on an icon. n12 The product cannot be obtained or used unless and until the icon is clicked. For example, when a user attempts to obtain Netscape’s Communicator or Navigator, a web page appears containing the full text of the Communicator / Navigator license agreement. Plainly visible on the screen is the query, "Do you accept all the terms of the preceding license agreement? If so, click on the Yes button. If you select No, Setup will close." Below this text are three button or icons: one labeled "Back" and used to return to an earlier step of the downlod preparation; one labeled "No," which if clicked, terminates the download; and one labeled "Yes," which if clicked, allows the download to proceed. Unless the user clicks "Yes," indicating his or her assent to the license agreement, the user cannot obtain the software.”

[13] HapMap Project: About the Project, http://www.hapmap.org/abouthapmap.html

[14] From the Wellcome Trust’s Sanger Center press release, at http://www.sanger.ac.uk/Info/Press/2004/041213.shtml:“One consequence of the license requirement was that the temporary click-wrap license prevented HapMap data from being integrated into major public databases, which require that data deposited carry no conditions on use...” – it is worth nothing that the reason for the click-wrap was to prevent patents, but it was the combination of disclosures and new technologies that rendered patents irrelevant, not the contract. Thus, from the same release: “Therefore, the original reasons for imposing the requirement to obtain a license to see the data no longer exist, and the licensing requirement has been dropped by the HapMap consortium.”

[15] Protocol for Implementing Open Access Data, Science Commons, http://sciencecommons.org/projects/publishing/open-access-data-protocol/

[16] http://www.plausibleaccuracy.com/2008/05/12/data-should-be-public-domain-and-more-esoteric-blog-based-rasslin/

[17] Open Street Map, Legal-talk list archives, in particular post on “Deconstructing the loss of data claim” by SteveC at http://lists.openstreetmap.org/pipermail/legal-talk/2008-February/000710.html

[18] Drawn from a private email by Pollock to Wilbanks, used with permission.

[19] Indigenous peoples and the commons, by Preston Hardison. http://www.commonsdev.us/content.php?id=962

[20] 8,848 metres (29,029 ft) according to Wikipedia

[21] According to the authoritative Ensembl database at http://www.ensembl.org/Homo_sapiens/protview?peptide=ENSP00000269305 , it’s
“MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPA
APTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTR
VRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHY
NYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTS
SSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPD
SD”

[22] Public Domain Dedication and License, by Open Data Commons / Jordan Hatcher, http://www.opendatacommons.org/odc-public-domain-dedication-and-licence/

[23] “CC in Review: Lawrence Lessig on Compatibility” on the Creative Commons blog, http://creativecommons.org/weblog/entry/5709

[24] “in this new vision, public ownership of research results was equivalent to "*dead-hand*" control” – from “Public Research and Private Development: Patents and Technology Transfer in Government-Sponsored Research,” Rebecca S. Eisenberg, *Virginia Law Review*, Vol. 82, No. 8, Symposium on Regulating Medical Innovation (Nov., 1996), pp. 1663-1727

[25] Copyright Easter Eggs at Open Street Map is the best explanation available. I do not mean to implicate OSM in the implementation, as the wiki makes clear – Easter Eggs are a reason *for* open data. http://wiki.openstreetmap.org/index.php/Copyright_Easter_Eggs

[26] chem-bla-ics blog, “John Wilbanks replies to my post on ChemSpider/OpenData Discussion” at http://chem-bla-ics.blogspot.com/2008/05/john-wilbanks-replies-to.html

[27] “Community spirit, with teeth” by Eliot Marshall. Science 16 February 2001: Vol. 291. no. 5507, p. 1192. http://www.sciencemag.org/cgi/content/full/291/5507/1192

[28] “Celera to End Subscriptions and Give Data to Public GenBank” by Jocelyn Kaiser. Science 6 May 2005: Vol. 308. no. 5723, p. 775 http://www.sciencemag.org/cgi/content/summary/308/5723/775a

[29] “The Public Domain in Genomics” by Rebecca Eisenberg, http://www.law.nyu.edu/ili/conferences/freeinfo2000/abstracts/eisengberg.html

[30] Open Data Commons Community Norms, http://www.opendatacommons.org/odc-community-norms/

[31] Promoting the Public Domain with Creative Commons' CC0 Initiative by Terry Hancock, Free Software Magazine, http://www.freesoftwaremagazine.com/columns/promoting_public_domain_creative_commons_cc0_initiative

Author

John Wilbanks runs the Science Commons project at Creative Commons. He came to Creative Commons from a Fellowship at the World Wide Web Consortium in Semantic Web for Life Sciences. Previously, John was the first Assistant Director at the Berkman Center for Internet and Society at Harvard Law School. John holds a Bachelor of Arts in Philosophy from Tulane University and studied modern letters at the Universite de Paris IV (La Sorbonne). He also serves on the Advisory Boards of the U.S. National Library of Medicine's PubMed Central, the Open Knowledge Foundation, and other organizations.
E-mail: wilbanks@creativecommons.org.