Scientific Data Commons Underutilisation: Causes, Consequences and Remedial Strategies

Paul A. David

Professor of Economics (Emeritus), Stanford University
& Senior Fellow, Stanford Institute for Economic Policy Research

Professorial Fellow , United Nations University-MERIT
Knowledge and Industrial Dynamics Group

(pad@stanford.edu)

W. Edward Steinmueller

Professor of Information and Communication Technology Policy
SPRU – Science and Technology Policy Research
University of Sussex

(w.e.steinmueller@sussex.ac.uk)

Paper Prepared for ISNIE 2013

Panel 4.E: IN HONOR OF ELINOR OSTROM: GOVERNING NETWORKED KNOWLEDGE COMMONS

Florence, Italy

Version: 16 June 2013

Note: This is a draft paper not yet suitable for attribution or citation.

1. Introduction

The collection, retention and stewardship of scientific data are activities that in publicly funded science more often than not involve the construction of 'knowledge commons' based upon agreements that afford both suitable access to research resources and facilitate the development of governance structures which assure that those participating in both the assembly and use of data are able to benefit from the resulting data stores. The agreements underlying such knowledge commons are typically informed by the norms of 'open science' in which data disclosure is at the heart of an incentive system based the recognition and rewards accompanying 'priority' – being the first to discover and to publish the results of theoretical or experimental investigation Dasgupta and David (1987). Publicly funded science almost universally adheres to open disclosure norms as a matter of principle[1] and it is possible for research funded from other sources to subscribe to the same principles. However, in practice the tightness of coupling between disclosure in the form of published claims to priority and the revelation and provision of the data underlying these claims varies across scientific disciplines. Arrangements for the governance of 'knowledge commons' are complicated by interpretations of what constitutes 'suitable' access and the ways in which those generating and assembling the data are able to derive benefits, including recognition and, possibly, the ability to further exploit the data to create additional discoveries. These complications cause practice to differ from principle and as in other situations of commons governance involve negotiation and consensual development of institutions (rules, norms and standards).

Outside of 'open science' the arrangements for access and the nature of governance of the collection, retention and use of data are varied although often entangled with issues of intellectual property or other methods of 'appropriation.' Appropriation may be seen as necessary for securing the returns on investments in data, but it also interferes with the principle of the 'knowledge commons' to varying degrees depending upon the governance arrangements that are made for access and reuse of this data. In extreme form, appropriation can create an anti-commons, a state in which common access and use arrangements unravel and ultimately diminish value of the knowledge commons fostering duplicative investments and impeding the progress of scientific understanding by reducing the inputs available for discovery and invention. Although appropriation can be toxic, the degree to which it is toxic depends upon the nature of arrangements for access and reuse of data.

In short, between the poles defined by complete public disclosure and complete private appropriation, there exist a spectrum of practical arrangements for the collection, retention and use of scientific data.[2] The scope of these arrangements and the various accommodations that may be made to improve social welfare given differences in these arrangements is the central theme of this

[1] An obvious exception is classified or secret military research which may involve subject matter and research methods that would be commonly understood to be 'scientific' but does not adhere to either the disclosure or public recognition precepts generally applying to science.

[2] Referring to data that is entirely privately appropriated as 'scientific' may be seen by some as a misnomer to the extent that science is taken to be synonymous with open disclosure. In this paper, we use the term scientific data to refer to data that is conceivably of value to scientific understanding.

paper. Much has been written on the potential hazards of extending the private or IPR domain for scientific knowledge, e.g. David (2000), Lessig (2003). This discourse amounts to interpreting the poles as involving a clash or conflict between institutional arrangements which creates a dilemma that might, in principle, be resolved by more comprehensive or even universal rules favouring one or the other poles in the above spectrum. As important, and unresolved, as this dilemma is, our present purpose is to examine the intermediate area between these poles rather than rehearse the arguments in favour of complete public disclosure or an extension of private appropriation. As in any conflict between institutional arrangements, a variety of accommodations are made for pragmatic purposes and this is no less true in the case of the conflict between public and IPR domains for knowledge.

To achieve the aim of examining intermediate arrangements between complete public disclosure and complete private appropriation we begin in the next section (Section 2) with a further elaboration of the concept of a 'knowledge commons' and how this may be defined in contexts where rights of access and use are negotiated and contingent rather than based upon entitlement and mandate. Section 3 takes up a series of specific issues which serve to define the scope of arrangements needing to be considered in this view of the 'knowledge commons.' Section 4 concludes with a summary of the key issues, most of which constitute opportunities for further research situated within the domains we identify in Section 3.

2. Defining the Knowledge Commons

This paper employs the term 'knowledge commons' in relation to institutional arrangements for the collection, retention and access to scientific data. This section considers the uses and abuses of the term 'commons' and explains the ways in which scientific knowledge in general and scientific data in particular can be seen to constitute a particular type of commons with similarities to and differences from other types. Understanding the use of this term and its relation to the other types of commons is a preliminary step to exploring its extent and boundaries as well as the institutional arrangements themselves.

A central idea, and one which is familiar from the work of Elinor Ostrom, e.g. Hess and Ostrom (2003) Ostrom and Hess (2005) is the tension between global or centralised forms of governance and contingent (or negotiated) decentralised forms of governance for resources that are constructed through collective action or which are collectively owned. The need for governance arises from the variety of opportunistic behaviours that are possible in collective action as well as the need to preserve the incentives to engage in collective action. The tension between the forms of governance arises from different understanding of the source of breakdown or dysfunction as well as different pre-conceptions regarding appropriate solutions. One of Ostrom's major contributions to the understanding of this tension is the observation that centralised intervention often does not lead to outcomes that are superior from a social welfare perspective and are often inferior to governance arrangements constructed through decentralised negotiation among the actors who participate in the use of a commons.

The claim that scientific data may be viewed as a commons is not entirely straightforward. On the one hand, processes of scientific investigation and discovery certainly benefit from the availability of

data to explore and test hypotheses. In many cases, data may be reused and reinterpreted to inform different facets of scientific theory or hypothesis, combined or re-combined to strengthen the empirical foundations for findings or conclusions, or create a more robust foundation for scientific exploration and discovery. In each of these cases, collection, retention and access arrangements that expand the common use of data has the potential to improve the quality of scientific investigation and may also make discoveries or insights possible that would not be available if common use opportunities were absent.

On the other hand, those generating the data may wish to extend their use of the data beyond findings and results that are immediately supported by the data they generate. Further exploration of data may lead to publishable or even more profound findings and results through serendipity or hard thought. Moreover, many modern scientific investigations are team efforts where groups cooperate in the generation of a large body of data which may be used different purposes and whose release is therefore governed by some form of internal governance to projects.

The resolution of these tensions is the social norm within the general domain of science (the republic of science as some would say to emphasise the significance of such norms within scientific research communities) is that publication ordinarily (but not universally) requires investigators to make their data accessible to others. Even if this norm is not specifically required by a publisher, a researcher's reputation and possible standing with funding bodies and employers may be influenced by a failure to cooperate with those seeking the data underlying published claims. The extent of disclosure required to support a published claim differs across scientific fields but rarely involves the same data resources as are available to the researcher or research team who publishes the claim.

Claims of priority may be contested if data is unavailable or withheld and the withholding data may suggest possibilities ranging from scientific misconduct involving the fabrication of supporting data to scientific errors, scientific claims that would be rejects if the data supporting them was subject to independent scrutiny. Thus, within the norms and institutions of science, failing to disclose or disclosing incompletely the data underlying scientific research imperils the perceived quality of scientific claims and findings. Thus, there is a tension between the investigators wish to publish results in a timely way in order to support claims of priority (and to support those claims by providing the evidence on which they are based) and the wish to retain control of the data in order to make further or stronger claims before other investigators are given access accounting for some measure of reticence in disclosure.

From a social welfare viewpoint, disclosure is valuable in assisting scientific investigation (e.g. by cost saving in research that duplicates existing findings or that might be conducted differently if prior results were known), but also in monitoring and policing the quality and veracity of scientific claims. These tensions have persisted throughout the history of science. They have been heightened in recent times by the widespread use of information technologies that in principle, and often in practice, is able to greatly reduce the costs of making data available Ostrom and Hess (2005) David and Steinmueller (1991). Moreover, these same technologies change the nature of the negotiation between those holding data and those seeking to access it. In earlier times, investigators might engage in a variety of behaviours which would impede the speed and scope of data disclosure and give greater priority to other investigators that were trusted to not infringe egregiously on the territory of those supplying the data. Or, more innocently, in previous times, the costs of

distributing (as well as using) data served as an impediment to transfer. Reticence is seriously challenged by publication requirements requiring the deposit of digital data underlying publication or the increase in expectations that data files should be freely transferred. It is not entirely surprising, therefore, that practices of disclosure are often seen to be at odds with principles.[3]

An alternative to a knowledge commons in scientific data based upon widespread public disclosure is a market-oriented system based upon appropriation and intellectual property rights. Several authors have taken the view that information appropriation using intellectual property rights offer a viable means for the social organisation of the exchange of knowledge, e.g. Grandstrand (2000) Arora, Fosfuri et al. (2001; Carrier (2009). While these authors differ in the scope of information which they argue should be governed by intellectual property institutions, on what institutions might best serve this co-ordinating function , and how these institutions might be governed, they share the view that market incentives are needed to fund the costs of generating as well as collecting, retaining and distributing the information underlying scientific or technological knowledge.

The role of intellectual property rights is taken as facilitating markets for this knowledge by establishing enforceable property rights and defining boundaries between 'extant' knowledge and 'novel' or innovative knowledge, thus providing incentives to add to the existing accumulation of knowledge. In the case of patents, disclosure reveals the principles of operation and novelty of new knowledge to trigger the granting of an exclusive right to use this this knowledge during the period of protection while other intellectual property rights such as copyright define the boundaries of original expression of authors or designers to exclude others from direct reproduction. . This distinction is important in relation to information because of the way in which each form of protection structures the resulting market. In many areas, the various sets of information that are useful in knowledge exchange may have important complementarities requiring the combination of exclusive property rights to achieve an outcome and therefore suggest the possibility of market exchange to achieve this end. However, these same complementarities may create the possibilities of achieving dominant positions through pooling patents and require companies to create substantial patent portfolios to resist being excluded from access to useful knowledge. Because copyright protects expression, the novelty of ideas is virtually impossible to protect using copyright. For example, efforts to use copyright to protect software expression provides very limited capacities to build portfolios of complementary knowledge *expressed* as information and substantially greater incentives to obfuscate and conceal the workings of software in order to secure intellectual property protection through copyright.

[3] Commentary on the problems and opportunities of data access are common in several scientific communities. For example, Vickers (2011) expresses frustration with the state of enforcement of data disclosure in the medical research community. A 2005 editorial in Nature expresses the hopes that such access might become more common, restating the ideal surrounding publication above Nature (2005:531) "Scientists may be justified in retaining privileged access to data that they have invested heavily in collecting, pending publication — but there are also huge amounts of data that do not need to be kept behind walls."

'Markets for knowledge' that are created through institutions of appropriation can be interpreted as a particular form of managing the 'knowledge commons' in which 'suitable access' is defined by non-discrimination in the willingness to sell rights to access and reuse scientific data and in which the benefits of those contributing knowledge are addressed by the possibilities of generating revenue from this access. In other words, appropriation need not, in essence, be antithetical to the advantages of collective knowledge accumulation However, there are several reasons to be cautious about a market solution to the access and benefit issues involved in creating and maintaining knowledge commons.

The first reason for caution is that the principle of non-discrimination in the willingness to sell rights of access and reuse is far easier said than enacted. Establishing institutions for appropriation creates market power and enables price discrimination according to ability to pay. Although it is possible, in principle, to regulate the resulting markets by principles such as those present in American antitrust law that forbid discriminatory pricing by firms with market power, the practice of such regulation is costly and its results uneven, in part due to inherent conflicts between the exclusivity of IPRs which grant exclusivity (or limited monopoly) and competition laws which were established in different jurisdictions for different political reasons. Broadly, in the US, a populist concern with the concentration of economic power and its political implications and in Europe a concern with the abuse that might be undertaken by a dominant competitor. Behind each of these motives, a sustaining belief that is partially independent from the practice of jurisprudence is the view that competitive markets are rarely extinguished by forces other than government intervention favouring market concentration, the existence of an 'essential' facility that can be appropriated and used as a means of excluding competitor or the effects of anticompetitive business practiceswhose implicit premise is that markets can only be monopolised through the use of inappropriate business practices.

A second reason for caution is that the costs to those accessing the data have to be financed with considerable uncertainty regarding the ultimate value of any particular body of data. Scientific investigation is inherently uncertain and therefore the value of any body of data used to support investigation must also be uncertain. Costs of scientific investigation may therefore be increased to the extent that investigators seek to receive more data than they contribute. Yet, because of uncertainty in the value, the capacity to consider and analyse more data than is generated is at the very heart of scientific investigation and the basis for establish a knowledge commons in the first place. Moreover, this concern is not allayed, as in the case of technology, by potential rules that might be made with regard to re-use or application of information since the value of acquiring data stems not only from its direct use but from its value in helping to avoid unfruitful lines of investigation. Based upon this reason alone, market arrangements for access and re-use can be seen as toxic to the methods of scientific investigation which have proven to be productive over the last several centuries. The extent of the toxicity of market arrangements will depend upon the extent to which they occupy the commons, i.e. the proportion of relevant data subject to market arrangements. The question of extent is important because it gives rise to the problem of multiple marginalisation in which each of a potentially large number of actors, each with their own 'piece' of the information commons may impose access restrictions (including the pricing of access) that 'stack' to create insurmountable barriers in the collective assembly of data. Such royalty or licence 'stacking' problems exclude uses of information and weaken the potential contributions of these commons to the advance of scientific understanding.

Although there are other reasons to be concerned about the extension of markets to the institutions of scientific data, these will suffice for present purposes to indicate a legitimate bias against arrangements based upon appropriation and IPR as a fundamental principle for maintaining the processes of scientific data collection, retention and distribution.

Nonetheless, there remain significant problems in constructing and maintaining sustainable knowledge commons. For example, it does not follow that mere exhortations regarding the value of data collection, retention and access will necessarily generate the resources for accomplishing these tasks any more than exhortations concerning the social value of complete public disclosure of data will overcome researcher reticence to make such disclosures. In practice, a variety of contingent and negotiated arrangements are likely to be necessary. This is a situation that is very consistent with the observations that Ostrom has made with regard to the general theory of governing commons Ostrom (1990; Ostrom and Hess (2005). Once one rejects centralised solutions such as the granting of property rights in scientific data or a universal mandate (with effective enforcement) for assuring complete public disclosure, one is still left with the problem of crafting institutions that will be sustainable for the purposes of generating and maintaining commons.

In the search for effective arrangements, it is important to note some differences between the nature of the knowledge commons and other forms of commons that have previously been addressed. Knowledge resources are not common pool resources in the sense that Ostrom studied the classic problems of natural resources.[4] The fundamental distinction is what Ostrom calls 'subtractability' – that an additional use of the resource reduces the amount available for use by others. Although subtractability is a feature of knowledge artefacts such as books or journals lodged in libraries, our principal concern is with digital data resources.[5] With regard to digital data, it is limitations in access that create the potential for scarcity rather than limitations in the stock of the available resource (because of the negligible costs of information reproduction compared to other resources). In the digital era, the potential for scarcity is created by institutions (rules, norms and standards) rather than physical constraints. These institutions may create 'subtractability.'[6] However, as previously noted it is more common for them to be used simply as a means of price discrimination, to create scarcity in accordance with the 'ability to pay' governing access.

---

[4] Hess and Ostrom (2003) recognises this distinction and responds by creating a trifold division (artefact, facility and ideas) for analysing institutions surrounding information governance. The application of this framework in Hess and Ostrom (2003) is with the problem of scholarly publishing in which the scholarly outputs previously embodied in artifacts (paper copies of journals whose use is rival, but where access is permissive) have become embodied in digital libraries (where use is generally not rival, but access is either potentially or actually restrictive). Although we do not employ this framework explicitly here, this framework is a useful method for considering changes in institutions arising from digitisation and points to contemporary problems of scholarly publishing, open access, and possible reforms in the systems for scientific communication.

[5] There are also other interesting cases of information commons that have subtactability properties that are intermediate. For example, microbial commons in the cultivation and exchange of micro-organisms rely upon more cumbersome processes of replication (culturing colonies of such organisms) which raise the costs of maintaining the commons represented by institutions for exchange of biological organisms and materials. The boundary between these commons and digital information commons is one of the most significant technological challenge of our age – the question of when it will be possible to reconstitute organisms from digital genetic information.

[6] For example, when the number of simultaneous users or 'desktops' with access to a particular data resource is limited by license.

In the following section we identify and examine five issues that influence the construction and maintenance of 'knowledge commons' devoted to scientific data. Our aim is to make a clear identification of the problems involved, some of the potential ways in which these problems may be mitigated and the need for further research to better understand the scope of the problems and the efficacy of various solutions.

3. Issues of Under-Utilisation and Institutional Gaps

The spectrum of behaviours from incomplete disclosure, what we have referred to in previous section as reticence in disclosure, to actions aimed at more explicit appropriation of scientific data, however motivated,[7] are likely to have the effect of constraining access to scientific data. These constraints may occur either from under-investment in collection, retention and distribution or more directly by imposing obstacles to access. This section examines five key issues contributing to these constraints and the resulting problem of the under-utilisation of scientific data as well as some actions that might be taken to address these issues.

1. Asymmetries in the costs and benefits to participants preparing dataset for use by other parties, which gives rise to confusions or opportunistic behaviours that impede shared use:

The first of the issues stems directly from what was previously described as user reticence, but may be more accurately described as an asymmetry between the costs and benefits to participants in a knowledge commons of collecting, and retaining scientific data. Scientific data is collected in a very wide range of institutional contexts: from publicly funded research, from commercial research activities and even in activities associated with everyday life where systematic and structured observations may be made that capture information of scientific value. The motives of specific actors and these actors' assessment of the value of collecting, storing and curating data differ. This diversity creates many possibilities for the costs associated with data to be borne unevenly and for different expectations to emerge about its value.

Ideally, the potential barriers that this diversity, and the asymmetries that they engender, may be overcome by the observation that the value of a data or knowledge commons depends more upon the investments in the uses made of it than the costs of its creation and maintenance. But, in practice, this observation does not always prevail in discussions concerning resource allocation questions that must be settled in advance of the commitment of further resources in order to exploit the data for particular, valued end-purposes.

As we noted in previous section, it is sometimes argued market solutions will provide appropriate mechanisms for resolving such problems. In eliciting the users' valuations of information-goods, however, markets are often plagued by transactions costs, non-transparencies regarding the nature

[7] It is important to note that forms of appropriation may be the only response in the conduct of public science when resources are not provided for collecting, retaining and providing access to data. Such activities are recurrently under-funded despite oft-stated intentions to improve their quality and availability and the observation that, in terms of the last units of resource invested in science, greater social welfare might be obtained by investing in distributing knowledge (including the prior steps of collecting and retaining it) than in generating more of it David and Foray (1995).

and quality of the goods to which access is being purchased, hold-out behaviour by suppliers of critical data elements (aimed at extracting the highest possible bid from prospective users), and still other "pathologies" arising from the distribution of ownership (exclusion) rights to different pieces of an extensive body of data. In practice, the elements of a body of scientific knowledge are subject to many possible complementarities, few of which may be known *a priori*. Scientific findings often involve assembling these complementary elements to uncover new knowledge or to substantiate theories and hypotheses that might otherwise be mere speculation.

The absence of market features and their specific pathologies does not assure that structurally funded or voluntary activities will suffice in assembling (collecting) and curating data generated by a community of researchers. The perceived value of contributing data to a commons as well as the net benefits of maintaining the commons have to be sufficient to motivate a sufficient number of actors to generate the positive externalities that will cause others to wish to contribute. A centralised mandate for participation by funders is, in principle, a means to overcome reticence or reluctance to participate. However, to be effective such mandates must have some means of enforcement and must be funded unless those funding scientific research (e.g. the state) are willing to accept a reduction in the amount of scientific research that can be conducted (and to accept this reduction in a global context in which other governments may not follow a similar course of action). If unfunded, the costs of data collection, retention and distribution have to be covered by researchers diverting their time to these activities rather than conducting research. The tendency toward unfunded mandates is already present in the related sub-system of scientific communication where funding based upon publication in peer-reviewed journals generates a demand for the voluntary contribution of editors and peer-reviewers to the publication process.

More optimistically, the activities associated with collecting, retaining and distributing scientific information might, in some contexts, be complementary with research processes. For example, a facility in the use of scientific databases may be generated by these activities which improve the researcher productivity in scientific investigation. Nonetheless, asymmetries area likely to remain between individual participants contributing different amounts of data and 'free riding' on the voluntary efforts of others will persist without some specific means of governance.

There are two stumbling blocks to improving upon this situation. The first is the problem of recognition. Since scientific data commons are likely to result from extensive collective efforts, the potential for individual recognition is diluted by the collective nature of the effort. In situations where it is difficult to award individual recognition, a useful way forward is the promulgation of a social norm. The rapidly declining costs of data storage make it possible to address the issues of social norms by retaining the provenance of contributions so that it is possible to individually and collectively observe the contributions of individuals. In itself, the retention of data provenance will not generate the social norm of adequate recognition for individuals who make contributions; there must also be explicit efforts to recognise such contributions. Such recognition might include measures of contribution that are recognised within a community such as the emerging forms of citation counting that appear in efforts to asses scientific publication records of individual researchers. The possibility of identifying one's contribution as a factor to be considered in research funding may be appropriate in particular communities.

The second stumbling block to improving asymmetries is structural in nature. So long as scientific research is evaluated exclusively in terms of a narrow range of outputs for purposes of funding and status, there will be strong incentives to minimise participation in the ancillary or after publication activities of collecting, retaining and distributing scientific information which incur the opportunity costs of diminishing performance on established indicators of performance. Faced with this structure of incentives, it is not surprising that, in some fields, researchers will more seriously assess the opportunities to appropriate data resources and attempt to monetise their efforts by generating resources that can be appropriated and sold. Although it may be difficult to do this for data that is generated in projects where funders require disclosure, activities which improve the quality of such data (e.g. annotation, generation of meta-data, structured curation) is, at present, unlikely to attract direct funding and therefore, absent of the complementarities in research investigation noted previously, likely to create incentives for market-based solutions.

This, like most of the issues considered in this paper, is a not a stable or static situation. The capacities to store and improve the organisation of repositories of digital data are increasing at a rapid rate. Data that is made freely available can simply be taken as a free input into processes that add value to it and thereby transform it into a proprietary product with greater value to some users than the original data collection. It is therefore possible, at least in some scientific communities, for the most useful resources using data to be generated through market processes rather than in terms of voluntary (or mandated) contribution. To the extent that this sort of process becomes the social norm, some of the pathologies noted above are likely to ensue.

2. Uncertainties arising from 'latent' property rights in scientific data contributed to knowledge commons:

The compilation of scientific data often involves assembling data that in some legal context is considered to be 'property' or an asset. A particularly thorny issue stems from compilations of data derived from biological materials whose provenance in the natural world is subject to the Nagoya Protocol of the Convention on Biological Diversity (CBD), which seeks to secure 'rights' related the utilisation of genetic materials. At present, the Nagoya Protocol can best be understood as a statement of intention rather than as an established institution governing the exchange of scientific data since the scope of its application and capabilities of enforcing its intent are yet to be resolved. Nonetheless, its existence raises questions concerning the assembly and retention of data based upon organisms which are taken to have been generated as an evolutionary patrimony by the biological environments within modern states.

The issues surrounding the Nagoya Protocol are unusual in their implicit scope and complexity. However, they are similar to questions that arise from the intersection between the public domain and claims that may be made under various systems of intellectual property protection. For example, the European Database Protection Directive (1996) creates the possibility of asserting an intellectual property right in the assembly of data, even though the data itself may be entirely in the public domain David (2000). In addition to enhancing the likelihood of the market based route for funding the collection, retention and distribution of scientific information mentioned in the previous case, this is an example of the possibilities for compounding uncertainties in the use of scientific data which may have been combined or re-combined in a variety of ways in its journey through

various network databases. As received by a research in performing tasks of collection or retention, it may be not be clear whether any body of scientific knowledge has passed through processes of assembly and arrangement that give rise to claims of intellectual property protection under the Directive and the legislation that has been enacted by the European member states in response.

Under these conditions, those undertaking the task of collecting and curating scientific data are exposed to various liabilities of uncertain magnitude that their activities may infringe upon intellectual property rights claims asserted in particular jurisdictions. To our knowledge, there has not yet been a legal case testing the limits of such liability or its applicability in the case of scientific data resources, but the existence of such systems of protection should give pause to those considering providing global access to assembled databases. It also suggests that a prudent individual or group seeking to develop such databases should incur additional costs in ascertaining the possibility of infringement (of current claimants or possible future claimants in the case of the Nagoya Protocol which has yet to be translated into specific bases for claims), costs which further reduce the incentives for such activities regardless of their social utility.

What has previously in some countries, and the US in particular, been referred to as provisions for 'fair use' in academic or other research contexts (such as public research institutions or research conducted by charities) is under increasing challenge from efforts to extend the means of appropriating knowledge. These extension efforts are underpinned by the claim that the lack of appropriation constitutes a market failure for which the only solution is clearer assignment of property rights. While such assignment may contribute to the incentives to collect, retain and distribute scientific information, it also creates unintended effects and potential pathologies which are likely to effect the utilisation of scientific data and the maintenance of a 'knowledge commons.' Perhaps the only way forward in this area is to enact (where not present), re-establish or strengthen provisions for 'fair use' and similar exceptions to the assertion of intellectual property rights in information.

3. Tensions arising from the interests of employers in controlling and tightly circumscribing access to the knowledge (capabilities) being gained by employees engaged in scientific and technological research activities, particularly the tacit knowledge and un-codified information that has not been made the basis of legally protected intellectual property rights.

Capabilities to use scientific insights, observational and experimental data, and information (codified signals extracted from structured data) are developed by researchers in the course of their employment, and those capabilities (as well as the skills and capabilities with which they entered employment) are a source of potential benefits for the individual researchers, their current employers and also for future employers should the researcher move away from the organization in which they have been engaged. Such capabilities are often generated in relation to the use of scientific data resources, some of which may be within the public domain while others are proprietary to the employer or have been acquired by the employer under specific terms and conditions generally lodging access in the purchaser rather than the user.

At first impression, this issue may seem out of step with the earlier discussion of this paper. However, the institutions surrounding the use of scientific data are certainly relevant to the nature and extent of knowledge commons. The use of data in scientific investigation depends upon

individual capacities to understand and comprehend it. Individuals' capacities to build upon and further articulate their skills are part of the larger system of knowledge creation in which repositories of data are a key element. These capacities may be constrained or enhanced depending upon the arrangements made for access during the period of their employment by a particular organisation and it follows that a change in employment may lead to futher changes, either for better or worse, in the capacity to use and further develop these capabilities.

In many contexts, the employment relationship encompasses understandings (including legal contracts) that aim to restrict the future use of employees' capabilities – specifically those that can be shown to have been acquired in the context of employment – when they leave one employer for another.

Often employment contracts contain so-called "non-compete" clauses, intended to prevent --for a defined period of time (e.g., 3 years) -- the exiting employee from taking up a position with another employer engaged in the same industry, or line of business, or within a defined geographical region that is centred on the current place of employment. Yet, it is quite possible that the match between an employed researcher's specialized capabilities (and interests) and the capabilities of the organization in which he or she is presently employed is not very effective. In such cases, it would be socially more productive for the individual to move to a different employment position where the individual's contribution to research (and the benefits derived therefrom) in the new position would exceed the costs incurred by the current employer in mitigating the disruptive impact such a move might have on-going research projects in which the person had been working.

Clearly, in such situations, legally enforceable contractual restraints that prevent dissolution of one "match" in order to move to a new position to form a new match which is more productive is not only a barrier to the sharing and wider dissemination of scientific and technical data and information, it is also a barrier that is difficult to justify on social welfare grounds. (Moreover, in this case, it would be possible for the new employer to benefit even after compensating the former employing firm for the disruption of existing work that had been incurred as a result of the researcher move and actual payment of the compensation would render removal of the restraint not only social efficient but "equitable" among the concerned employers.)

It might also be the case that by trying to prevent business competitors from benefitting from the capabilities and special information that their employees gain in the course of their R&D activities, the new employer may suffer from the symmetric loss of access to the services of researchers that would be more productively employed in their firm, where they not restricted from moving to it by the non-compete clauses applying to potential employees.

Non-compete clauses in employment contracts are only one source of impediments to wider access to scientific knowledge and data through the mobility of researchers. Pension rights that are not vested, or only become vested in the employee after many years another source of mobility impediments that inhibit the beneficial reallocation of research talent – not only in the private sector, but among the "public" sector comprised of non-commercial (academic institutions and government) research institutions.

These impediments to researcher mobility seem likely to be a source of social loss – they will precipitate duplicative training and research efforts and incur additional costs in the accumulation of

experience by employees not constrained by pre-existing employment and correspondingly, not enriched by that earlier experience. It is certainly possible that this flaw in the commons can be largely mitigated by contracts allowing subsequent employers to compensate former employers for the human capital investment made (and the experience gained). Of course, this is not a simple transaction to negotiate even in the example suggested above where the value gained exceeds the value lost. Nor might it be entirely separated from anti-competitive practices such as efforts to raise rivals costs by claiming larger losses than actually experienced. Nonetheless, public policy as implemented in employment law is capable of providing enabling structures for fruitful negotiation of such disagreements such as institutions for arbitration of disputes over compensating payments. A more difficult issue is constraining an individual's capacity to transfer specific proprietary knowledge. However, in this case, the more specific instruments of non-disclosure and the possibility of individual liability for violating trade secrets or commercial confidence would already appear to offer some redress to the former employer. This line of argument may, however, be over-optimistic concerning the potential for harm in specific industries where key knowledge is not easily protected by such agreements and is clearly an area where more research is needed before attempting a centralised intervention by the State.

Even without this evidence base, however, it may be mutually beneficial for private and public employing entities to develop a "research mobility commons" that would remove the restraints on serially shared access to the accumulating knowledge and experience-based capabilities of knowledge-workers. How might such commons be implemented in sustainable voluntary agreements, rather than by political action to make "non-compete" clauses in employment contracts unenforceable at law – as is the case in some jurisdictions, notably the State of California?

4. Possibilities for increased 'data hoarding' in the emerging 'big data' era:

The exploitation of scientific data for either commercial or scientific takes time while its generation and storage, although in many cases rapidly decreasing, remains costly. It is natural that those funding and managing the creation of scientific data will wish to achieve direct benefits from its creation before it is made more available. As the variety and amount of data generated becomes larger, the possible means of exploiting data expand and improve. However, the time required to make a reasonably full exploitation of data is not falling at the same rate as data is being accumulated. This is, of course, the canonical form of "data hoarding" that arises in the world of small data projects, when the data collection or experimental data generation process is concentrated at the early phases of the research, and followed by an extended period of analysis, interpretation and preparation of findings for publication.

Even where there is pressure to arrive quickly at and publish some "findings," the resources available to the research team more often than not will leave some hypotheses untested, and other questions un-investigated; especially when the latter lines of inquiry have emerged clear only after the project got underway and had not been anticipated in the original funded proposal, the pursuit of research is likely to be deferred in favour of the promised "deliverables." Although the policy of funding agencies calls for the rapid release of data gathered by projects, rarely is the funding provided for the personnel time needed to prepare datasets that are thoroughly documented and structured for easy use by researchers outside the projects. As we have noted above, this is itself a

central issue leading to the under-utilisation of knowledge commons. Hence, the research team's leaders have a rationale for deferring the data's release that will, incidentally, afford an opportunity to exploit them in future research.

Seen from this perspective the problem of "data hoarding" in scientific research, deplorable as it may be, is not a novelty peculiar to the era of 'big data." But the sheer increase in its scale does make a difference, insofar as resources are being invested in the technologies and infrastructure to facilitate the creation of massive data resources to tackle problems in specific scientific domains, and to make it accessible to community of researchers working in that and related fields. Moreover, it is an issue that is further complicated by the potential in some fields for asserting intellectual property rights as the consequence of investigations that are ancillary to scientific disclosure. This is a particular problem in the field of biotechnology where scientific investigation is directed at generalised knowledge whose 'scientific character' makes it ineligible for patent protection but from such knowledge it is possible to derive further specific knowledge about particular applications of knowledge that might relevant to disease diagnosis or treatment or methods for bio-engineering organisms of commercial value, applications that are subject to patent protection.

The public commitment to large scale data generation of is being undertaken in the expectation that there will be a commensurate return in the form of an accelerated pace of scientific discoveries, and the disappointment of those expectations coupled with the appearance that research groups that nominally favour open access to scientific data and information are delaying the releases of high quality high quality data that they have been responsible for collecting, will jeopardize the readiness of public funding agencies to continue supporting the "big data" enterprise. Moreover, delays arising from the underfunding of the costs of making the reliable and well documented datasets quickly accessible could create pressures for more limited, more fragmented, and increasingly duplicative data generation projects that would reduce the lags between data collection and systematic exploitation. Recognizing the limitations on future funding of overhead support in many research domains, it is important to ask: What alterations in existing incentive structures and governance methods might be devised to address the incipient problem of "big data hoarding"?

This is another area where it is more likely that centralised solutions are likely to be either unenforceable or dis-incentivising. More localised solutions involving negotiated solutions are more likely to establish the conditions for mutual advantage. There are a number of possible templates or models undertaking such negotiations. A traditional solution in some scientific communities is to share recognition (e.g. through co-authorship) with those providing the specific data from which additional research results are derived. This may be an effective method where the *ex post* outcome is a relatively minor advance in scientific understanding or practical consequences.

For *ex post* outcomes bringing greater recognition or status, however, there remain significant issues. In these situations, the composition and capabilities of research teams is an important issue. If two research teams have similar capabilities the mutual sharing of data may be a means of risk sharing, a case of the prisoner's dilemma problem in which sharing a joint outcome may be better than gambling on being first.[8] As the size of coalitions of teams increases, however, it is likely to be

[8] In the original statement of the prisoner's dilemma, the possibility of being first to confess to a crime conveys an advantage. However, a team of scientists cannot be certain that it will be first and may be able to gain advantage over third or further teams by bilateral sharing, hence the co-operative outcome.

both more difficult to make arrangements for sharing recognition and more likely that some sub-group will form an exclusive club because they believe their probability of their joint success is sufficiently greater that they would prefer not to share recognition with groups that are less likely to succeed.

These examples apply to cases of data sharing in which the probability of success is markedly increased by data sharing between a limited number of groups which is a corollary to the nature of many projects that involve 'big data,' at least in present conditions where the costs of gathering data are relatively large. It is likely that these conditions will systematically change in the next decade as the technologies for data acquisition in environmental, meteorological, and social sciences become cheaper and the resulting flood of data from multiple sources creates additional opportunities for making gains by more co-operative arrangements for data sharing.[9] Under these conditions, the advantage that may be gains by any single or small group of research teams by withholding their data sets from general use will be reduced. These developments will once again test the evolution of social norms discussed in relation to the first of our cases since data are unlikely to organise themselves or be made inter-operable without incurring costs for which the present social systems offer inadequate rewards.

5. Governing 'hybrid' commons in which freely accessible data is combined with data that is proprietary and licensed commercially on a restrictive basis:

In many fields, both public and private sector research is engaged in generating scientific data. While it is not universally true that private sector sponsors of research seek to retain ownership and control of the data or to generate revenues from its use, doing so is a common practice. Substantial gains are possible by data structures and data management techniques that identify the existence of useful privately held data in data collections that are openly accessible. Private sector data providers have incentives to cooperate in efforts that would make their commercial data offerings identifiable because this would serve as a means for generating the revenues they seek. Identifying and seeking to exploit situations in which new 'hybrid' commons could be created -- where the holders of proprietary information and data do not have strong enough incentives to make it freely accessible in exchange for access to other, open access content of the "commons" – could be a socially beneficial step.

Interestingly in this connection, commercial uses of software --whether distributed by the vendor in diskette, or downloadable digital files, or used to generate on-line services provided for a fee – can and do make use of software that contain both open source code packages that are licensed under GPL terms, and proprietary code packages for which the source code is not disclosed. The GPL license permits this, so long as the vendor also freely distributes the source code for the GPL elements in the program (with attribution to the authors and copyright holders), along with the API's that have been created to enable the proprietary, undisclosed code packages to work with the open

[9] In the social sciences there are many unresolved issues involving the rights to privacy that constitute a potential barrier to the accumulation of large data sets on individual behaviour. However, many of these issues are being circumvented by various means of making individual data anonymous, even if the behavioural models created using this data promise more precise predictions any individual's behaviour. Privacy concerns obviously do not apply to other biological organisms or atmospheric conditions although the possibilities of creating better ecological or meteorological predictive models may result in some disturbing opportunities of misuse.

source packages. In this case interpretation of, and compliance with the GPL terms are in principle sufficient to solve the hybridization problem – but it is also the case than in the case of large and widely used open source programs, there are associated foundations than create a governance structure and contract monitoring and enforcement capability supporting that part of the hybrid software.

Although not socially optimal, 'hybrid products' and 'hybrid' commons organization represent important socio-economic structures whose design should be examined with a view toward obtaining the greatest social benefit that are feasible when "first best" of enlarging the public domain cannot be attained. One might seek also to determine the extent to which offering the "hybrid" option would pre-empt some opportunities for eventual formation of a "pure' commons holders – if the voluntary expansion of free sharing increased the attractiveness of the common resource pool to the point that IP holders recognized it would be beneficial to join, and license their property freely for common use. This consideration suggests that the 'hybrid" form could be viewed as a "back-stop" solution to be deployed after voluntary formation of a socially beneficial pure commons was found to be infeasible.

This suggests the need for better understanding of specific conditions in various research communities of the costs and rewards (both direct recognition and indirect benefits of being more productive) from being engaged in data collection, retention and distribution activities. Such an understanding is a pre-condition to developing feasibility and performance measures in efforts to create voluntary solutions. It would provide a foundation for asking when it is necessary from the outset to form a hybrid solution and what observations would be needed to conclude that a hybrid solution would perform better after initial efforts at voluntary cooperation founder.

The possible existence of hybrid structures (e.g. in areas such as bioinformatics) raise questions of what guidance might be useful in crafting rules governing cooperation for mutual advantage and how such rules might assessed from the perspective of public interest for those parts of the hybrid effort that are funded from public sources. As with the other issues identified in this section, the challenges is to devise ways forward in which the parties directly engaged in scientific information collection, retention and distribution have an interest in reducing the under-utilisation of knowledge commons.

4. Conclusions

This paper began by examining two ideals in the governance of the knowledge commons represented by 'open science.' The more commonly understood application of the term 'commons' in this area is the open disclosure of scientific data so that scientific data becomes public good. However, like other public goods, the problem of providing the resources to create it imply the existence of a credible commitment by funders to this aim. In practice, the commitment is often not credible, as discussion of the absence of resources and recognition for scientific data, collection and retentions are common and persistent across many fields of scientific investigation. Approaches that involve funders mandating the provision of open data resources without providing resources run the risk of constraining the performance of national science systems and, in practice, rhetoric about such mandates is not matched by enforcement provisions that would force participants into compliance. Perhaps this is for the best. There are plausible reasons for those generating data and contributing to its collection, retention, and distribution to be reticent about complete and open disclosure, some of which are clearly antithetical to social interest while others appear to be legitimate efforts to enhance the returns for the effort in generating and curating scientific data.

We also considered a less common definition of commons based upon appropriation and arrangements for access governed by market principles. In this case, the 'commons' nature of such arrangements could, in principle, be made operational by effective enforcement of non-discrimination rules. While such arrangements would impose an additional overhead cost on the conduct of scientific investigation and weaken the data resources available to many researchers, it would also address the problem of underfunding and perhaps also the lack of recognition to data collection, retention, and distribution activities. We conclude, however, that there are inherent flaws in this ideal type that stem from the nature of scientific information in use – that the combination of scientific resources involves uncertainty and market arrangements for access would lead to even further under-utilisation of this knowledge commons than current access arrangements. Moreover, market solutions make possible a number of pathologies that are accentuated and amplified versions of the opportunistic behaviour already present in the system. These pathologies have the potential not only to enhance under-utilisation but also to weaken the very foundations of the scientific social system. Although we conclude that market arrangements are toxic, and potentially fatal, under certain conditions they are necessary, in some cases, for the development of sustainable arrangements for data collection, retention and distribution.

With the recognition that neither ideal is entirely feasible, we note the existence of a spectrum in which it is desirable and legitimate for policy to be biased in favour of open access arrangements but to also recognise that such arrangements may prove infeasible, unenforceable or unsustainable. In the resulting 'gray area' between these two ideals we have identified five specific issues. In each of these issues, we have identified the need for localised and contingent or negotiated solutions, an approach that is largely consistent with Elinor Ostrom's observations on the governance of commons in which she identifies the dysfunctions created by centralised control. Like Ostrom, we find that decentralised agreements aiming to create mutual advantage which evolve through experience and reflection to become more effective and inclusive arrangements offer a potentially more useful

means for governing access and use of common resources. For the specific areas we have identified, the existing evidence base is inadequate to provide prescriptive solutions either for the participants or for science policymakers. Nonetheless, in each of these areas we have identified potentials for developing more localised solutions and likely changes in conditions arising from the increasing importance of digitisation of scientific information.

The first of these issues we identified is the need to redevelop the social norms governing the 'knowledge commons.' For several reasons, the current system is incapable of creating consistently mutually beneficial arrangements for the activities of scientific data collection, retention and distribution. While researcher reticence to prematurely disclosing the full extent of the data underlying their investigations is an important issue, the system of recognition and rewards for data-related activities are, in many fields, an inadequate basis for the development of mutually beneficial arrangements. The resulting shortcomings in the creation and under-utilisation of data resources may be addressed by participants agreeing to stronger norms of contribution and participation but may also be improved by policies that recognise and reward these activities as an essential feature of the science system.

The second issue arises from the existence of latent or possible claims to intellectual property in scientific data arising from the general trend toward 'market solutions' to 'market failure' problems in which the centralised solution of more complete property rights is seen as a solution. In this area, researchers have much less scope for localised solutions because the liability risks they face are external to their communities. This issue is the direct product of centralised policy and it is only through reform of centralised policy that this source of shortcomings in production and under-utilisation are likely to be mitigated.

The third set of issues stems from employment arrangements that impede the mobility of researchers and hence the distribution of capabilities for using scientific data. We focus on non-compete clauses and similar arrangements that bar former employees from working in the same industry for some period of time but also note the constraining effects on mobility of pension systems that are tied to specific employers. In this area, there is very little evidence from which to derive any possibility of centralised intervention and some reason to believe that centralised intervention might be counter-productive. Instead, there may be some value in a common recognition of the mutual desirability of mobility among a group of employers and agreements to streamline and standardise arrangements for compensation to former employers. Such arrangements would offer clear opportunities for social gains, better incentives for individuals to contribute to collective efforts to create scientific information commons, and might well provide employers with mutually beneficial outcomes compared to the present system in which such mobility bars results in the loss of productive inputs. Employers are affected not only in losing employees but also in the costs of being unable to acquire individuals with relevant accumulated skills and expertise. We stress, however, that the full extent of this issue and its consequences requires more research.

The fourth issue we identified is the possibility of data hoarding. This issue involves the lack of willingness to participate in data commons because of the uncertainties surrounding the possible use of the increasing stocks of data associated with modern information technology systems including the digitisation of experimental results and various record keeping systems as well as the

growing array of sensors populating human and natural environments. We note that this issue is likely to become more severe over time as the technologies for data acquisition continue to improve and become more extensively deployed and the costs of data retention are rapidly falling. In this area, mutually beneficial arrangements for sharing risks and rewards appear to offer superior opportunities to any centralised intervention. Nonetheless, it is clear that such arrangements are likely to be contingent on the composition of the actors curating 'big data' resources. A consequence of the technological developments, however, is likely to be that actors that are currently in possession of such 'big data' resources are likely to be joined by many others, making it more desirable to craft the necessary institutions for achieving data commons and reducing data hoarding behaviour.

The fifth and last set of issues considered is the interaction between open scientific data resources and various proprietary data. Based on our observations that the 'second best' solution of proprietary data is often a necessary outcome, we suggest several means for mitigating the problems of such commons including the mutually beneficial advantage of disclosing metadata and other forms of partial disclosure which reduce the social (and private) losses which may result from duplicative research and serve to identify research opportunities which would remain clouded without such partial disclosure. We also noted that acknowledging the potential value of hybrid communities was a means of dealing with the breakdown or non-formation of voluntary arrangements and suggested the need for better methods of analysing feasibility and performance of scientific data collection, retention and distribution activities as a way to assess when it would be desirable to encourage the formation of hybrid forms of the scientific data commons.

Together, the issues that we have identified define a substantial set of research challenges for the social science research community interested in science policy and the operation of scientific communities. Individually, these issues suggest the need for a greater consideration of the mechanisms of governance within scientific communities that should be considered by scientific organisations as well as research teams. Hopefully, for policymakers, these cases provide some justification for concluding that the problems of governing the collection, retention and distribution of scientific data are not best addressed by a single ideal model – in practice, neither complete open disclosure nor market solutions as they can actually be implemented are likely to foster the most complete utilisation or the most robust system for generating scientific data resources.

References

Arora, A., A. Fosfuri, et al. (2001). *Markets for Technology: The Economics of Innovation and Corporate Strategy*: MIT Press.

Carrier, M. A. (2009). *Innovation for the 21st Century: Harnessing the Power of Intellectual Property and Antitrust Law*: Oxford University Press.

Dasgupta, P. and P. A. David (1987). Information Disclosure and the Economics of Science and Technology. *Arrow and the Ascent of Modern Economic Theory*. G. R. Feiwel. New York, New York University Press**:** 519-542.

David, P. A. (2000). *A Tragedy of the 'Public Knowledge Commons'?: Global Science, Intellectual Property and the Digital Technology Boomerang*. Stanford Institute for Economic Policy Research Discussion Paper No. 00-02. Available from Oxford Intellectual Property Research Centre-Electronic Journal/Working Paper Series at: http://www.oiprc.ox.ac.uk/EJWP0400.pdf.

David, P. A. and D. Foray (1995). "Accessing and Expanding the Science and Technology Knowledge Base." *STI Review* **16**(Autumn): 13-68.

David, P. A. and W. E. Steinmueller (1991). "The Impact of Information Technology upon Economic Science." *Prometheus* **9**(1): 35-61.

Grandstrand, O. (2000). *The Economics and Management of Intellectual Property*. London: Edward Elgar.

Hess, C. and E. Ostrom (2003). " Ideas, Artifacts, and Facilities: Information as a Common-Pool Resource." *Law and Contemporary Problems* **66**: 111.

Lessig, L. (2003). *The Future of Ideas*. New York: Random House.

Nature (2005). "Let Data Speak to Data." *Nature* **438**(7068 (1 December 2005)): 531.

Ostrom, E. (1990). *Governing the Commons: The evolution of institutions for collective action*: Cambridge University Press.

Ostrom, E. and C. Hess (2005). A Framework for Analyzing the Knowledge Commons. *Understanding Knowledge as a Commons: from Theory to Practice*. E. Ostrom and C. Hess. Cambridge MA, MIT Press**:** 41-82.

Vickers, A. J. (2011). "Making raw data more widely available." *BMJ British Medical Journal* **342:d2323**(DOI 10.1136/bmj.d2323).