Information & Management 47 (2010) 237–245

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

Governing the data commons: Policy, practice, and the advancement of science

Joshua B. Fisher *, Louise Fortmann

Department of Environmental Science, Policy & Management, University of California at Berkeley, 137 Mulford Hall, Berkeley, CA 94720-3114, USA

ARTICLE INFO

Article history:
Received 2 October 2007
Received in revised form 9 October 2008
Accepted 1 April 2009
Available online 21 April 2010

Keywords:
Common pool resources
Common property
Commons
Data sharing
Design principles
FLUXNET
Intellectual property
Remote sensing

ABSTRACT

Property rights of shared scientific data are often analyzed by applying formal intellectual property law. However, the scales at which these laws apply are not necessarily relevant to data sharing practice among individuals (e.g., in virtual, social communities). Rather, the data sharing communities and their members often form their own policies, practices and norms governing data sharing. Using common property theory, our research objectives were to determine: (1) parallels between data sharing communities and natural resource-sharing communities; (2) factors that lead to successful data sharing; (3) circumstances under which rules, laws and policies govern data sharing. We used cases from two emerging data sharing communities in the environmental sciences—a micrometeorology community (FLUXNET) and a satellite remote sensing community. Cases of data sharing without conflict or irresolvable disputes had more principles in common with a successfully managed natural resource commons than did those characterized by conflict. Successful data sharing requires that biophysical scientists be more attentive to the social nature of data sharing.

1. Introduction

The expense and complexity of data required by researchers working across multiple disciplines at increasing spatial scales have often exceeded the means of a single scientist or institution. Thus, data sharing in distributed collections and virtual communities has become increasingly common. While the literature supports the benefits of data sharing, scientists have given relatively little attention to the property aspects of shared data outside of website statements or institutional guidelines. The literature on data sharing has focused mostly on its technical infrastructure rather than end-users. Our focus was on the property issues that arise when data are shared. We asked how data sharing communities govern themselves.

This is an issue with practical implications. One concept on data sharing in the field of biotechnology demonstrates the idea of an "anticommons" in which too many patents (over-privatization) have blocked advancements. Another potential problem may occur when cooperation or collective action ceases before a product achieves its full potential. For scientific data, misuse (use without permission and/or poor analysis) by users and mistrust by data producers can damage collaboration and interfere with scientific advancement. Understanding how rights to the shared data commons are effectively governed may help avoid adverse effects.

The governance of rights to shared data occurs at three levels. At the *macro*-level, rights to data are governed by international treaties and national laws. At the *meso*-level, rights may be affected by policies of formal and informal associations of data producers and users or institutions with which they are affiliated. At the *micro*-level, rights may be the result of negotiations among individual producers and users.

The research framework for this paper was based on both common property theory of common pool resources (CPR) and intellectual property rights (IPR) theory of intellectual property. Common property theory provides a framework for analyzing how groups develop, implement and sustain collective management regimes for *natural resources* they use in common. Those resources include fisheries, groundwater, and forests, as well as large-scale resources such as the seas, space, and the atmosphere (i.e., the "global commons"). Certain rules or design principles tend to be strongly present in successfully shared natural resource commons and less so in CPR that are in conflict [1]. We define "successful" as data sharing that takes place without conflict or unresolved disagreements. In short, common property theory aims to describe potentially universal principles for sustainable resource-sharing communities.

IPR describes different types of exclusive powers (government-granted) of control and use over *intellectual resources* (non-physical and non-exhaustible). Intellectual resources include inventions (patents), creative works (copyrights), identifying insignia (trademarks), and some forms of confidential information (trade secrets). IPR are created through statutory law and treaties. In short, IPR are

* Corresponding author. Fax: +1 818 354 9476.
E-mail address: joshbfisher@gmail.com (J.B. Fisher).

0378-7206/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.im.2010.04.001

policies, rules, laws and property relations. However, IP scholars are generally concerned with formal entitlements and do not consider how the rules affect behavior.

Data is a form of intellectual property, and the properties of shared data have been compared to CPR [2]. Open source and free software communities have been considered in the context of common property theory [3]. Questions of the public domain have been addressed in the IP literature but without reference to common property theory. Scholarly communication has been compared to an information commons. Our study was oriented towards the governance of data sharing.

We assessed practices at three levels of scientific data sharing, applying CPR theory's design principles for sustainable common natural resources [1] to practices governing commonly shared data. We were particularly interested in the emergence of rules of data sharing written or unwritten, formal or informal, by explicit or implicit understanding and the effect such rules may have on interpersonal conflict. Because individual outcomes affect the pool of data available to the community, we examined individual relationships between data users and data producers. What rights do data producers have over their data? What rights do data users have over the data analysis, manipulation, distribution, modification, and its publication? Do the rights provide enough protection from misuse or data ownership and provide reward (such as monetary, prestige, and publication) resulting in the continuation of data production? What are the sanctions against those who break the rules? Questions about property rights in shared scientific data are located at the intersection of IP theory and common property theory (Fig. 1).

2. Theoretical approaches: common property and intellectual property rights

2.1. Common property

A shared resource system, or CPR, has been defined [1] as a system that is large, making it costly (but not impossible) to exclude potential beneficiaries from obtaining benefits from its use. A CPR within the context of a legal regime of shared rights is referred to as common property. For the purposes of analyzing the complexity and variability of knowledge or information resources, a commons is defined as a resource shared by a group of people attempting to solve social problems. Once a dataset has been released into cyberspace, it is very costly (if not impossible) to control its use. Shared data in a digital form is thus a global resource.

The data producers and users differ from most commons communities in two respects. The group is scattered around the globe rather than being contiguous geographically and thus regularly involved in face to face interaction. Thus, the governance of the shared data commons is complicated by the difficulties of communication among individuals and groups with differing power, languages, cultures, and knowledge of their rights and responsibilities in data sharing. Second, a producer provides a particular dataset. While the data sharing community may have a set of appropriation rules, a user's access ultimately depends on negotiation with its producer.

Traditional analysis has shown that a sustainable commons often requires small, relatively homogeneous groups [4]. The question therefore is not whether a data sharing global community of disparate people, cultures, and institutions represents a small and homogeneous group, but whether the definition of "small" and "homogeneous" has changed in cyberspace. It may be faster to send an email to a hundred people on the internet than to say "hello" to ten people in a firm. People in the global commons all share the same transport on the information superhighway.

While shared data meets the definition of CPR, it differs from natural resource CPR because it is not subtractable: the use does not diminish the quality or value. The possibility that the supply of a particular dataset or datasets will be exhausted is not at issue in the management of the shared data commons. Indeed, as long as data users acknowledge the source and as long as their analysis does not reflect poorly on the quality of its data, users enhance the quality of the resource.

The real stakes are threefold. *First*, professionally valuable credit resulting from the analysis of shared data is limited. Publishing data without crediting the producer constitutes "free riding": using a resource without contributing to its production and management or taking more than permitted. *Second*, poor analysis reported in papers with fatal flaws, poor assumptions, etc. can reflect on both the data user and producer because the reader may not be able to identify what was the reason for the bad paper. Such an activity constitutes resource degradation in the natural resource commons. Data producers attempt to ensure that they are credited when a data set they produced has been used and thus to control publication rights. *Third*, although the data are not exhaustible, the analysis might be. Specifically, users should not publish the same data twice to answer the same question twice, thus prior use conditions future use.

CPR design principles are a useful tool for analyzing governance of property. They are conditions associated with the success of voluntary collective management of natural resources by groups of users in accomplishing their outcome. They have been used to analyze common property systems as well as large-scale resources in a global commons. The importance of design principles in successful management has been shown to be highly dependent on the situation.

Each of the these design principles can be applied to shared data:

1. *Clearly defined boundaries* specify who has a right to use a dataset and describe the dataset itself. Datasets should be easily identifiable. Thus poaching can be detected. However, like mobile natural resources, they move or are moved through a territory (cyberspace), which is not easily bounded. The users of

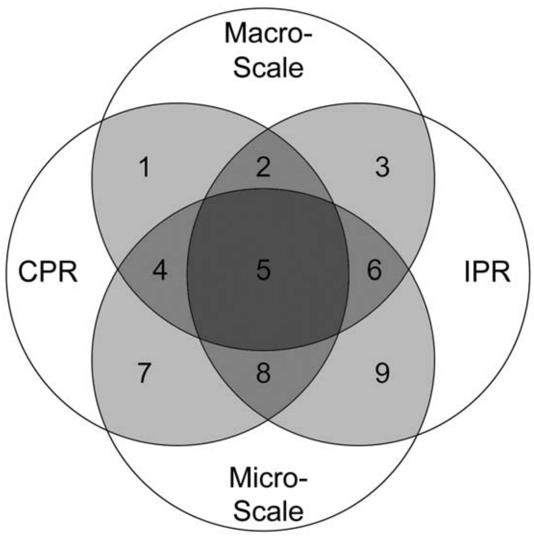

Fig. 1. The theoretical framework of data sharing that links CPR and IPR across the levels. Key:**(1).** CPR at the macro-level (global commons);**(2).** CPR and IPR at the macro-level (global commons with laws and policies);**(3).** IPR at the macro-level (international laws and policies);**(4).** CPR at the meso-level (institutionally-shared);**(5).** CPR and IPR at the meso-level (institutionally-shared IP with policies);**(6).** IPR at the meso-level (institutional policies, rules, and contracts);**(7).** CPR at the micro-level (individually-shared);**(8).** CPR and IPR at the micro-level (individually-shared with rules, understandings and contracts);**(9).** IPR at the micro-level (individual contracts).

shared data may be bounded socially through associations but seldom geographically. A global community may require identifying information before allowing access to data as a means of enforcing boundaries.

2. *Appropriation rules related to local conditions* occur through national and international laws and treaties. Local communities may augment local rules by using *de facto* rules.
3. *Collective-choice arrangements* allow individuals to contribute to modifying the rules. These ensure that each party (data producer and user) benefits from the arrangement as well as being protected. Because data producers have the power to refuse or delay data distribution, they have a major input on how data can be used.
4. *Monitoring* is needed by the users or people who are accountable to them. Because they are generally physically dispersed, monitors are unlikely to catch poachers in the act of downloading, copying and printing data. Thus unauthorized use can generally only be detected when a manuscript is circulated, reviewed or published. When journals are indexed online, post hoc monitoring becomes relatively easy. Data producers can contact journal editors or publishers with complaints before manuscripts go to final print.
5. *Graduated sanctions*, depending on the seriousness and context of the offense, should be applied to appropriators, by officials accountable to them, or both. These readily lend themselves to the application of shared data.
6. *Conflict-resolution mechanisms* should be rapid and low-cost. The global nature of a data sharing community complicates this issue. Universities and institutions in countries with different political regimes may or may not act in response to the actions of local data users and producers. The mechanism of one country or culture may differ from that of another. Intellectual property rights may be ill-defined or even unrecognized. Larger institutional or governance rules may not have a large impact, whereas local norms may have different weights.
7. *Minimal recognition of rights to organize*. The rights of institutions and informal groups to devise their own rules is usually recognized but is complicated globally. The recognition of rights, rules, and conflict-resolution mechanisms varies internationally. However, a virtual community may be able to develop and follow its own norms without coming under governmental scrutiny or sanction.

The eighth principle, *nested enterprises*, is not relevant here.

2.2. Intellectual property

Scientific data, analyses, and results have traditionally been in the public domain [5]. In the US, for example, the US Supreme Court has held that facts that are "part of the public domain [are] available to every person" and cannot be copyrighted (U.S. Supreme Court, "Feist Publications, Inc. v. Rural Telephone Service Co.," vol. 499: U.S., 1991, p. 340; U.S. Supreme Court, "Harper & Row Publishers, Inc. v. Nation Enterprises," in Feist: U.S., 1985, p. 348). Government policies have encouraged scientific data to be freely available in the IP-free public domain. Recently, however, federal appellate courts have broadened copyright protection of data compilations. The EU, on the other hand, has enacted an IP regime to copyright and protect the contents of scientific databases; the US has not [6].

The idea of transforming ideas into property has been argued to be justifiable based on three propositions from Locke's labor theory of property. In the US, IP rights are embedded in the Constitution. The theory is that labor is required to produce ideas and it should be encouraged. Baird argued that people "have the right to enjoy the fruits of their labor, even when the labors are intellectual". Reichman and Uhlir [5] proposed that research communities form contractually-bound commons to preserve their data in the public domain.

IP law primarily protects production of intellectual content rather than access for data users. IPR are supported by four general types of law that focus primarily on knowledge property and the distinction between what is and is not IP: copyright law, trade-secret law, patent law, and trademark law. Copyright law protects "original forms of expression" such as computer software and scientific data for publication. Digital copyright protection has been argued to be a legal "monopoly and [to have] limited the scope of fair use rights" due to click-through licenses [7], which are required to gain access to many scientific data sharing portals. But, data users are forced to accept them to gain access to the data. In a non-digital sense, users could gain access to the resource without having to agree to the copyright license as long as they did not publish or disseminate the resource. Trade-secret law protects proprietary information that data users employ in scientific analyses and data producers in data production; it provides data producers with an advance period to analyze and publish their data, but does not provide legal mechanisms to deter direct competition. This limited protection is forfeit once data are disseminated to the public in print.

The primary IPR question for scientific data is: who has the rights to the data produced? The anticommons concept suggests that multiple right holders can block scientific advancement through intentional or inadvertent gridlock, which occurs when the bundle of rights is divided among different people with no clear communication or cooperation. Overlapping claims can create a bottleneck, especially when a user needs access to multiple claims to create a single useful product or analysis.

At the macro-level, the US has had perhaps the most activity in legislatures and courts in terms of addressing data access in copyright, federally funded science, and fair use. In 1998 the US Congress passed the Digital Millennium Copyright Act (DMCA) updating the Copyright Act (Title 17 of the US Code) to include digital data. The Digital Consumer Right to Know Act required data producers to inform consumers about technological features that would restrict their use of content (e.g., encryption). The Public Domain Enhancement Act ensured that abandoned copyrighted works passed into the public domain, though this does not necessarily mean that the public does gain access. The Public Access to Science Act, which aimed to exclude works from copyright protection that are substantially federally funded scientific research, is still the subject of debate on the grounds that removing copyright protection could hamper academic publishing and negatively affect research funding. The Consumer Access to Information Act of 2004 attempted but failed to allow "fair use" of data comparable to that under copyright law and duplicated database protection under other federal laws. The Digital Media Consumers Rights Act of 2005 focused on defining fair use (copies for purposes such as criticism, comment, news reporting, teaching, scholarship, or research). The legal definition of fair use is the core of the FLUXNET data sharing policy and many remote data sharing policies.

Meso-level governance of data sharing started in the mid 1980s when the National Academy of Sciences was among the first US national organizations to discuss the benefits of data sharing and ethics with publication policies involving shared data. IRM, in general, is aimed at sharing information to increase efficiency and effectiveness for institutions; data sharing increases the success of IS integration [8–10].

At the micro-level, gridlock may occur when conflict arises from personal conflicts. A data producer has the power to exclude or suppress findings for personal reasons or due to the constraints of multiple claims. To overcome these barriers, norms and unwritten relationships can, but do not always, facilitate work and progress.

3. Methods and context

Our case study used examples from two data sharing communities: (1) FLUXNET and (2) the remote sensing community. These are the primary scientific global research communities working on large-scale, global phenomena, such as climate change, land use, and natural resources monitoring. Although many of the questions and problems addressed by the two communities are similar, they are different in size, organization, and structure. The expense and complexity of the data required by the individual community members necessitate data sharing. Data and analyses are the primary products of these communities.

FLUXNET is a global network of over a thousand scientists associated with micrometeorological towers that measure the exchange of carbon dioxide, water vapor and energy between terrestrial ecosystems and the atmosphere [11]. At present, over 400 tower sites in 45 countries across 5 continents in a wide range of ecosystems and climatic zones operate on a long-term, continuous basis. The organization provides researchers with infrastructure for compiling, archiving and distributing the data to the science community and supports synthesis, discussion and communication of ideas and data through support for scientists, workshops and meetings. NASA's Terrestrial Ecology Program and NASA's Earth Observing System Data and Information System sponsor FLUXNET. The individual tower sites are funded by institutional support associated with the AmeriFlux, AsiaFlux, CarboEurope, ChinaFLUX, FLUXNET-Canada, KoFlux, LBA, and OzFlux sub-networks. Their overarching goal is to provide information for validating remote sensing products for net primary productivity, evaporation and energy absorption.

The remote sensing community is not formally centralized. Its organization is manifested through funding agencies like NASA, the Brazilian Space Agency, peer-reviewed journals such as *International Journal of Remote Sensing*, and various regional organizations (see http://en.wikipedia.org/wiki/Space-faring_nations, October 2008 for a complete list). This community has four main participants: (1) scientists, who provide new sensors, techniques, phenomenology, and applications; (2) industry and business, who develop new platforms and sensors; (3) browsers, who employ remote sensing products; and (4) governments, who want to develop the technology to improve and protect their societies while promoting commerce and defense.

We undertook a micro-level case study of the practices of FLUXNET and NASA researchers at the Berkeley campus of the University of California (UC), conducting semi-structured interviews with 17 researchers who produced and/or used shared data. These respondents included all FLUXNET and remote sensing faculty and staff researchers and all graduate students working with them. Ages ranged from 25 to 60, and the gender ratio was approximately equal. We coded these cases for the presence of the CPR design principles.

4. Results

4.1. Macro-level governance: international treaties and national law

There was no official external recognition of the FLUXNET fair use policy, but its organizers hoped to be involved in a larger international treaty that would connect the parts of FLUXNET and institutionalize fair use practices internationally [12]. The remote sensing community in the US is commonly linked with NASA, which, as a Federal agency, must adhere to national IPR legislation and international IPR treaties to which the US is a signatory.

Data user, Ryan Riyousha (a pseudonym) was involved in two cases of international remote sensing data sharing involving macro-level governance. In a successful case, he received data from Australia, performed an analysis and sent the analysis back to the data producer. There were clearly defined boundaries in the licensing and collective arrangements between him and the data producer. Conflict-resolution mechanisms followed international copyright law. Their licensing agreement was thus valid under international copyright treaties.

In a case involving conflict, he witnessed multiple purchases of rights between Chinese and Taiwanese organizations. Data were generated and published in Shanghai, China. Then the paper-publication and -distribution rights were sold to a Beijing, China company, who sold the *digital*-publication and -distribution rights to an institute in Taiwan, who then distributed a digital version of the data in a book. Later, the original Shanghai researchers decided to publish and distribute their data. The Taiwan institute discovered this (through an international consortium on collaboration) and threatened legal action. The Shanghai researchers went to court, but differences among Taiwanese, Chinese and international copyright law made it difficult for parties to agree on jurisdiction, much less on a settlement. The Shanghai researchers released their digital version of the data, but with some changes so that it could be said that the two publications were different. Although the Taiwan institute continued to object, the international consortium saw the Taiwan institute as obstructing research and sided with Shanghai in the dispute. The Taiwan institute promptly resigned from the consortium. Riyousha, as part of the international consortium that ruled against Taiwan, has not been able to access the Taiwan publication.

4.2. Meso-level governance: institutional policies and scholarly journals

At the meso-level the most detailed and directly relevant policies were those of the professional networks, while those at UC (our study site) had little immediate effect. The Office of the President of UC had published memos[1] adhering to the US copyright acts. Sponsored works and copyrights produced by or through UC were owned by UC unless an agreement said otherwise. Although UC recognized copyright for original works in a tangible form of expression, the terms *original* and *tangible form of expression* are open to interpretation, especially with regards to shared data.

NASA's Distributed Active Archive Center (DAAC) managed the archival and distribution of NASA data through the Oak Ridge National Laboratory (ORNL). Their data citation policy requested that authors include a bibliographic citation to ORNL DAAC rather than to the individual scientists who provided the data.[2] The DAAC also included a policy for data producers on quality control for data dissemination (e.g., metadata, defined parameters, and consistency). The constraints on data producers were more detailed than for data users. NASA expected peer review rules to govern the actions of the researchers and accepted negotiated collective-choice arrangements.

According to an official at NASA Headquarters, a White House Circular governed the policy [13] for data collected by or through the federal government. Circular A-130 directed the government to implement an open and free data sharing policy after an initial *check out* period, so that government officials could review the quality and accuracy of the data. Agencies were also required to provide access to agency records under provisions of the FoIA and the Privacy Act. Data dissemination should be on *equitable and timely terms*, but could be limited if the cost to the government and public outweighed the usefulness of the information.

[1] http://www.ucop.edu/ucophome/coordrev/policy/12-01-99.html, October 2008.
[2] http://daac.ornl.gov, October 2008.

Table 1
The FLUXNET/AmeriFlux and OzFlux fair use policies.

	FLUXNET/AmeriFlux	OzFlux
Requirements & restrictions on users	(1) Inform PI of use plans (2) Acknowledge &/or cite PI & agency support (3) Discuss co-authorship (4) Provide reprints (5) PI right to first-publication	(1) Accept liability (2) Temporary copies for browsing (3) Single copy only for download or print (4) Cannot "change" material
Ownership	PI initially, then public	CSIRO
Cost	Free	Free
Design principles	(1) Appropriation rules related to local conditions (2) Clearly defined boundaries	Minimum recognition of rights to organize
IPR	Trade-secret law (head start)	(1) Copyright (2) Unspecified "law"

The official saw the Circular as a guide that should apply to data collected by a principal investigator through a grant. There was no formal monitoring mechanism in place, but, in theory, if a rule-breaker were to be caught, NASA could sue him or her and the courts could assign penalties based on the severity of the offense. "The success of the NASA data sharing," according to user/producer Barbara Beide, "depends on the government payment of production costs, so that conflicts over money are minimized." Essentially, the data are subsidized, because the user does not have to pay for the development of sensors and satellites, but merely the cost to maintain data acquired from the sensors.

The fair use policies of FLUXNET/AmeriFlux[3] and OzFlux (Australian and New Zealand subset of the global FLUXNET)[4] are compared in Table 1. Data producers wrote the fair use policy. The FLUXNET/AmeriFlux fair use policy requested communication and agreement between users and producers, and this should lead to micro-level collective-choice arrangements. The OzFlux statement did not allow for any change to the data, though the definition of "change" is unclear. For instance, calculating and reporting an average of two values does not change the original data, but does change the reporting of the data. OzFlux data were owned by the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia's national science agency, giving OzFlux official recognition. FLUXNET/AmeriFlux was not part of any governmental agency, and ownership of the data followed the stipulations of the funding agencies. Generally, they released the data to the public domain after an initial ownership period by the producer. "Payment" for the data came in the form of acknowledgment, citation or co-authorship. The OzFlux statement included a section on copyright, which gave the owners certain rights, but the fair use policy placed a limitation on them.

In terms of access, data download sites could be password-protected and only accessible to a small group of users. Though fair use and first sale doctrines have been described as precarious, the FLUXNET/AmeriFlux policy is consistent with trade-secret law because the producers are allowed immediate use in analyzing and publishing results from the data.

Scholarly journals influence access to the shared data commons by sometimes requiring researchers to provide the data used in their analysis. This exchange may conflict with prior arrangements between data users and producers under which data users cannot redistribute or share the data. Thus data users must wait for producers to publish the data first or request approval to share the data with the journal. Data users Paul Polza, Kim Kayttaa, and Ben Bruk reported having been unexpectedly denied data access because producers had difficulties or delays in publication. The data producers had trouble benefiting from the priority use rights due to the lengthy review processes, slow collaboration, or case-specific delays. Because data sharing could not start until the producers first published the data, users had to wait to start analyzing the data.

4.3. Micro-level governance: individual practices

4.3.1. Cases of successful data sharing

Clearly defined boundaries, appropriation rules related to local conditions, conflict-resolution mechanisms, and minimal recognition of rights to organize were strongly present in successful cases, with collective-choice arrangements and graduated sanctions present in just over half. Those that were mostly successful generally lacked clearly defined boundaries and graduated sanctions but were characterized by the strong presence of appropriation rules, monitoring, and conflict-resolution mechanisms.

Producer/data users, David Deux and Barbara Beide, both received data (as users), which resulted in successful outcomes. Deux received data from a producer who had much data, but did not possess an extensive set of analytical tools. Deux was analytically experienced, but did not have a good dataset. By collaborating, they produced several publications. Beide acquired remote sensing data from NASA's DAAC, which shares data with users on a formal basis with clearly written rules; there was no conflict. The successful cases tended to be characterized by the presence of more design principles than less successful cases.

4.3.2. Data sharing cases with problems

Clear appropriation rules were often lacking in cases with problems. Respondents had not been formally warned of the rules of ownership, data sharing, and co-authorship. Due to the general absence of written guidelines, scientists in our study learned the rules through oral tradition (conversations and lab meetings) or from mistakes.

This is illustrated by the case of data user, Bob Bhoga, who did not fully understand his own rights and responsibilities nor the rights of the data producer. First, Bhoga put the data producer on a paper as co-author without informing him. Upon learning this, the producer was upset. Second, Bhoga and another data user were under time-constraints to submit an abstract about an analysis with shared data. The data producer could not be informed in time and, having learned not to include a data producer as co-author without prior consent; Bhoga did not include him as co-author. Bhoga contacted the data producer the following day, but the producer was upset that the abstract was submitted without prior consent. In both cases, Bhoga felt that although boundaries were clear at the institutional level, the rules at the individual level were not. There were no clear agreements, no monitoring mechanism in place, and no formal conflict-resolution mechanism.

Conversely, although monitoring is often considered to be one of the most important principles leading to successful commons

[3] http://public.ornl.gov/ameriflux/data-fair-use.shtml, March 2008.
[4] http://www.dar.csiro.au/lai/ozflux/disclaimer/index.html, March 2008.

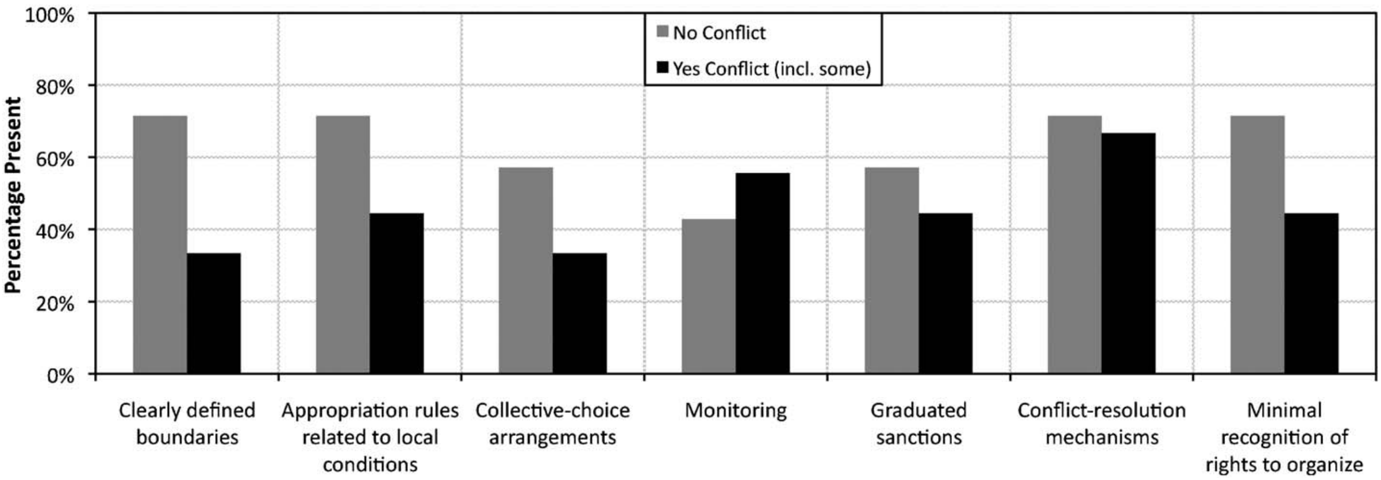

Fig. 2. The percentage of design principles present in the conflict and conflict-free cases.

management, monitoring was present in a higher percentage of the unsuccessful cases (50%) than in the successful ones (43%) (Fig. 2). This breakdown shows both the prevailing culture of data sharing and the way monitoring is effected. The culture of the FLUXNET and NASA communities was based on an "honor system". It is possible that the lack of monitoring indicated the presence of trust. Monitoring shared data currently relies on the peer review process, which is, of necessity, after-the-fact, with the resolution of any issues at the editorial stage. Thus misuse of data can be detected only when an article has been submitted for publication. This can lead to producers experiencing publication surprise. Data producers Ian Ilabas and Patricia Producer each made their data freely available on the web with the stipulation that any users contact them. Both have seen publications (sometimes as reviewers) that used their data although they had not been informed about its use. Sometimes they were acknowledged, sometimes not.

At least two of the design principles rely on effective monitoring: graduated sanctions and conflict-resolution mechanisms. No one can be sanctioned unless he or she is caught. Nor can conflicts be resolved if they are not identified.

David Deux shared his data with a researcher who reported a poor analysis. He clearly owned his data and was able to monitor its misuse, but felt unable to do anything about it. However the process of peer review had the effect of sanctioning the researcher, because he was not able to publish the work.

The data producers in our study did not act directly against the offending papers after they were published, though they did implement low level graduated sanctions against the authors. Following the offense, they scrutinized the offending authors more carefully in the course of review of publications, grant proposals, and collaboration. Perhaps of equal or even greater importance, they reported that data *mis*-users were stigmatized at conferences, avoided in social settings, or even suffer a loss of reputation. Some respondents mentioned the possibility that if a data producer felt that data users were not following fair use policies and that the monitoring and sanctioning systems were inadequate, s/he might deliberately "flood the community" with excessive or incorrect data (possibly the form of an initial data set and subsequent series of "corrections" that would require a potential user to contact the producer personally to ensure s/he was getting an uncorrupted data set). No respondents said they had done this so it may have been an unfounded fear or empty threat. If the peer-review system is working well, an unacknowledged data producer may be selected as the most qualified reviewer. The cases of Ilabas and Producer demonstrated that this does not always happen. On the other hand, because the research fields are often quite narrow, according to data user, Keith Kutimo, monitoring of published infractions can be effective when researchers provide literature reviews. The mixed results for the presence of monitoring in this study suggest that post hoc meso-level monitoring is not sufficient for effective governance of the shared data commons.

5. Discussion

5.1. Applying the design principles

There were more CPR design principles in cases without conflict (4.4) than with conflict (3.2). The application of these principles to all of the data sharing cases is shown in Table 2 and summarized in Fig. 2:

- Clearly defined boundaries were present in eight of the 16 cases, more in cases without conflict. Of the seven cases without conflict, they were present in five (71%). Of the nine cases with some or much conflict, they were present in three (33%).
- Appropriation rules related to local conditions were present in nine cases, more in cases without conflict: in five of seven cases without conflict (71%), and four of nine with conflict (44%).
- Collective-choice arrangements were present in seven cases (least frequent), more in cases without conflict: in four of seven cases without conflict (57%), and three of nine with conflict (33%).
- Monitoring was present in eight cases, the only design principle in a higher proportion of cases with conflict: in three of seven cases without conflict (43%), and five of nine with conflict (56%).
- Graduated sanctions were present in eight cases: in four of seven cases without conflict (57%), and four of nine with conflict (44%).
- Conflict-resolution mechanisms were present in 11 cases (the most frequent), more in cases with conflict, but a higher proportion of cases without conflict: five of seven cases without conflict (71%), and six of nine with conflict (67%).
- Minimal recognition of rights to organize was present in nine cases, more in cases without conflict: five of seven cases without conflict (71%), and four of nine with conflict (44%).

Because it was not possible to code cases for the actual presence or absence of design principles, we relied on the perceptions of data sharers. Respondents characterized their experiences as successful, mostly successful, or in conflict. Success was defined as lack of conflict or irresolvable disagreement. "Mostly successful" (some conflict) cases were defined as resolvable conflicts or disagreements. The "mostly successful" category was created to

Table 2
Ostrom's CPR design principles as applied to 13 data sharing case studies. The rows are sorted by amount of conflict.

	Conflict?	Total # of design principles present	Clearly defined boundaries	Appropriation rules related to local conditions	Collective-choice arrangements	Monitoring	Graduated sanctions	Conflict-resolution mechanisms	Minimal recognition of rights to organize
NASA DAAC (Beide)	No	6	✓	✓	✓		✓	✓	✓
Combining skills and data (Deux)	No	6	✓	✓	✓	✓	✓	✓	
FLUXNET user (Polza)	No	5	✓	✓	✓	✓		✓	
International remote sensing collaboration (Kutimo)	No	5		✓		✓	✓	✓	✓
Australian collaboration (Riyousha)	No	4	✓		✓			✓	✓
Remote sensing user (Fayda)	No	3		✓			✓		✓
Remote sensing user (Bruk)	No	2	✓						✓
DOD data request (Guna)	Some	5		✓		✓	✓	✓	✓
FLUXNET publications, but no prior consent (Ilabas)	Some	4		✓	✓	✓		✓	
FLUXNET publications, but no prior consent (Producer)	Some	3		✓		✓		✓	
To co-author or not to co-author (Bhoga)	Some	2	✓						✓
Reluctance to share ideas (Gebruiker)	Some	1			✓				
Authorship exclusion (Deux)	Yes	6	✓	✓	✓		✓	✓	✓
Poor analysis of her data (Deux)	Yes	4	✓			✓	✓	✓	
China-Taiwan conflict (Riyousha)	Yes	3					✓	✓	✓
Shared measuring device (Beide)	Yes	1				✓			

encompass the somewhat subjective continuum between conflict and success, including respondents' differing definitions of minor or major conflict. Respondents also characterized the presence or absence of design principles.

The fewest design principles (one: collective-choice arrangement) were present in data user Gabe Gebruiker's reluctance to share ideas and Barbara Beide's case of a shared measuring device. Gebruiker's unwillingness to share data resulted in a reciprocal withholding of data by potential collaborators—a negative feedback. Beide, however, participated in a case with the most design principles present (six of seven, missing only monitoring)—her interaction with NASA's data archive center. Both FLUXNET producers told of similar cases where their data were used for publications without prior consent. Neither the FLUXNET nor the remote sensing cases showed any clear difference in the presence of conflict.

Fig. 2 shows that data sharers perceived the presence of design principles in more than half (up to 71%) of the cases with no conflict except for monitoring, which they saw as present in less than half (43%) of the cases. Only monitoring and conflict resolution were perceived as present in more than half the cases characterized by any degree of conflict.

5.2. Respondents' concepts of property

The FLUXNET/AmeriFlux community had been struggling with the issue of data sharing, rights and obligations of data users and data producers, who, in our study, were aware of how fast, cheap, and inconspicuous the internet has made dissemination of information. They were trying to find an appropriate balance between fluid data sharing and retention of rights to acknowledgement and authorship [14]. Data producers did not want their datasets to be so difficult to obtain that no one would analyze them, but neither did they want data users to forget the source because they were so easy to acquire. In general, because there were so many potential owners in scientific research, none of the respondents knew who owned what. Conflicts with funding sources, sharing results, and co-authorship were often due to differing perceptions of rights among the parties.

Most respondents did not know IP law. Many were confused about when they could copyright data. In the mid-1970s, publication was required for copyright protection. The 1976 Copyright Act eliminated this requirement, but, because they almost never copyrighted anything or purchased data, few of them paid attention to the change. Some researchers were interested in copyrighting their work to obtain a "copyleft" (the application and protection of copyright law with the intent of free redistribution). Thus a copyright was sought to ensure that no one else could obtain the copyright and shut off access to others.

We asked respondents for their opinions on rights to data, personal rewards, observed ownership, and their view of patents/copyrights (Table 3). All indicated that the most important rewards were publication or acknowledgement. Respondents identified measuring devices and processes, algorithms and analysis methods but not data as subject to copyrighting.

In general, data producers did not claim many rights over the data they had produced, nor did data users claim many rights over the data they were using. The number of rights claimed ranged from a minimum of none or first-use to more rights that depended on agreements or the cost of the data. The more users paid for the data, the more rights they felt they should have. Data producers tended to claim priority use rights for publication using their own data. Most data producers indicated they were responsible for or required to share data. Data users expressed a right to use publicly funded data. They sometimes saw their rights as being abused by data producers who were unwilling or slow to share. Data producers felt that they should always be asked about their opinion of the analyses. Ian Ilabas explained, "Data producers enjoy the relationship between data users and data producers because the data users give interpretation and feedback to the data producers to increase understanding on the data produced." If a data producer allows data users "first-look" or "first-use" of the data before public access, the data user should reciprocate by offering co-authorship on publications. Data producers should also limit the time delay before they transfer the data to the public domain.

82% of the respondents felt that either the researcher or the funder owns the data. 35% identified the funder as the owner; 29%, the researcher; and 18%, both. Personal viewpoints on the practice of property were more complex. Ilabas and Producer felt there was a "fine line" between who owns the data and whether or not a user needs to ask permission or offer co-authorship or even acknowl-

Table 3
Summary of interviewee statements on rights, ownership, and copyright. Data producers are indicated by ^{P}; data producers/users are indicated by $^{P/U}$; and, data users are indicated by ^{U}.

Person	Perceived use rights	Perceived ownership rights
Ilabas^{P}	First-use	Funder
Producer^{P}	First-use	Data producer and public
Ambos$^{P/U}$	Right not to share	Inapplicable
Beide$^{P/U}$	None	Funder
Deux$^{P/U}$	Unsure	Researcher unless otherwise stipulated
Bhoga^{U}	Use publicly funded data; none to data; limited for analyses, must inform DP	Funder, university
Bruk^{U}	Use publicly funded data; publication, not redistribution	Funder
Fayda^{U}	Use publicly funded data; analysis and publication, but must give opportunity for co-authorship	Funder and researcher, proportion depends on funding proportion
Gebruiker^{U}	Use publicly funded data; depends on agreement	Researcher, funder, collaborators in that order
Guna^{U}	Use publicly funded data; depends on supervisor	Researcher—based on intellectual input, funder is merely thanked
Kayttaa^{U}	Use publicly funded data; ownership to analysis and reformatting	Researcher and private investigator
Kutimo^{U}	Use publicly funded data; publish without offering co-authorship if data are not major part of analysis, distribute with acknowledgement	50% funder, 50% researcher and collaborators
Polza^{U}	Use publicly funded data; producers' first-use	DP for data, DU for analysis
Riyousha^{U}	Use publicly funded data; depends on agreement	Stipulated by funder
Sayoung^{U}	Use publicly funded data; vary depending on cost	Purchaser
Utente^{U}	Use publicly funded data; only usufruct, no redistribution	Unsure
Yong^{U}	Use publicly funded data; if any, then absolute	Funder

edgement if the shared data only constitute a "small" percentage of the analysis. Data producer/user, Aaron Ambos, believed both that "the concept of data or analysis ownership is inapplicable in most situations, except in specific agreements with funders" and that he owned his dataset and therefore could "horde" it. Ambos felt "no obligation to share data," although he did try to share as much as possible when requested and expected others to do the same. Beide noted that few producers actually made their data readily available, even if publicly supported, because of heavy competition for publication. Beide pointed out that shared datasets appeared to be finished products [to the user], but in fact were quite raw with "flaws and errors." She felt that data sharing was guided by "personal responsibility," but the data and analyses were completely separate. The funder owned the data and analyses, but funding agencies did not monitor this carefully. For instance, the funder might request monthly data, but the researcher records and kept the hourly data.

David Deux, was clearer about his rights as a user than as a producer. As a user he felt he had the right "to use data under the stipulation that [he] must keep the data producer informed of [his] work." Deux noted that "upfront conversations on rules and policies are wise" but rare—laying down some basic understanding of ground rules without going through any legal or contractual venues could result in conflict-free interaction. In his experience, discussions were more often carried out under initial friendly settings than under settings of miscommunication or conflict. As a producer, Deux preferred to share data given a beneficial reciprocity, but had two concerns. First, he was unsure if a data user would give him authorship even at his request. Second, he was concerned about people using his data to produce poor quality publications without his oversight. Deux wanted to know the plans of the data users before he distributed his data, but noted that this demand conflicted with the fundamental scientific tenet of replication whereby anyone should be able to replicate the results of a previous study. Frequently because of the cost and contingent nature of remotely-sensed data, replicating a study is only possible if the same data are available. Deux felt that he owned his data and analyses unless otherwise informed, particularly within grant stipulations. He noted that local funders tended to want more ownership than state or federal funding agencies.

Some data users expressed interest in partial ownership. Data user Sarah Sayoung's sense of ownership of data depended on payment. She felt that if she paid for data, then she owned it, though rights, protection and rewards were variable, depending on the ease of acquisition and cost of the data. In the case of shared data, it is possible to assign specific rights to the producers and users, rights that might be subject to change over time (for example, The US established a copyright apportionment doctrine in "Sheldon v. Metro-Goldwyn Pictures," 106 F.2d 45 2d Cir. 1939; and, "Sheldon v. Metro-Goldwyn Pictures," 309 U.S. 390, 1940). Data users Gabe Gebruiker, Fran Fayda, and Keith Kutimo—had ideas of relative ownership in the "bundle of rights" down to exact proportions or percents. One breakdown by Gebruiker specified 50% ownership in the funders depending on proportion of total funding, and the remaining 50% ownership in the researcher and collaborators. Another example included the ownership order of first the primary researcher, then the funder, and last the collaborators. Other specifications for ownership included "intellectual input," "ideas," "time," and "energy." Gebruiker noted that the benefit of including others in ownership was that they would share some of the responsibility for publishing fees. In practice this might range from an actual percentage allotment for fees or a chain of approval for use of the analysis, to being wholly dismissed by funders and collaborators.

"I feel that I own my data or analysis, but I realize that my point of view may conflict with other policies," said data user, Pam Polza. Gebruiker felt reluctant to share ideas of analysis or methodology (rather than data) for fear of intellectual theft. Data user, Uma Utente, stated that her rights are specified in formal agreements at either the individual or institutional levels, but that these never include the redistribution of someone else's data. She felt that if a dataset was paid for, the rights were more likely to be clearly specified. She pointed out that "data sharing rules are not codified in [her] university rules," nor has she "ever received a hard copy or email about these issues." Data user, Lee Yong, felt that international laws recognize the rights of data producers and data users in a "general human rights sense"—seizure of data or analyses is prohibited without due process and "everyone has the right to own property alone as well as in association with others" nor "shall be arbitrarily deprived of his property" (for example, "Universal Declaration of Human Rights," vol. resolution 217 A (III), 1948).

5.3. Additional factors

Miscommunication/personal relationships—As the owner of a shared measuring device, Beide expected the user to share data obtained from the device, but the user refused. Beide had not initially specified clear boundaries, rules, nor made any collective-choice agreements with the user on data sharing, leading to miscommunication. Because conflict-resolution mechanisms had not been included, the conflict was addressed through personal communication. The case was further complicated by tensions between the researchers. Despite negotiations, the user continued to refuse to share the data and was not subject to any sanction.

Sayoung viewed data producers as taking one of two stances: sharing data at-will, depending on connections or personal relationships, or creating data for users who had absolute rights. Thus data sharing success depended largely on the personalities of the researchers involved, in particular the specific rapport that one researcher or group had with another. Hierarchies, level of respect, and trust, may influence interactions. Timing appeared to be important in relationship outcomes. Upfront conversations, asking for permission early and establishing common understandings could address the first three design principles and lead to successful data sharing, even in the absence of the other principles. The choice of whether or not to initiate these factors appeared to be influenced by the social network in which the actors were embedded.

Scale—Governance of the shared data commons was characterized by complex and shifting cross-scale relationships in which institutional and geographic scale were not necessarily closely coupled. For example, in the case of the Shanghai researchers involving two countries, governance was controlled by a macro-level institution, Chinese copyright law. Monitoring micro-level behavior often occurs in the normal course of operations for meso-level institutions such as peer-reviewed journals.

6. Conclusion

We have shown how common property theory can be extended to understand the governance of shared scientific data. Although it is subject to national and international law, governance of a data sharing commons also emerges from the micro-level daily practice of data sharing. In some situations, this may have a greater effect on how data sharing is done than formal law.

Governance of the emerging data sharing commons is clearly a work in progress. At best, present practice constitutes improvised, temporary solutions rather than a consistent and coherent set of governing rules and practices, in part because data producers/users are scattered around the world. A set of clear and mutually understood boundaries and rules (with local variations as needed) has not yet been developed; without clarity and mutual understanding, the likelihood that the effective monitoring, appropriate graduated sanctions, and accessible conflict resolution is needed for effective commons governance.

We have shown that rights and sanctions are scale-dependent, varying across macro, meso, and micro policies and practices. The relative importance and function of each level are highly contingent—an international data sharing collaboration will invoke more macro-level rights than will a question of who is a co-author of a paper (micro-level).

Finally, it is particularly important to note that the finite resource is not just the data, but publications and/or acknowledgement. Publications advance careers, and generally aim to impact the scholarly community and society. When data users bypass data producers' input, they may feel they have been treated unfairly. When data producers do not share their data or restrict data redistribution rights, data users are prevented from producing publications and earning this academic currency.

Our primary research contribution was to: (1) demonstrate that common property theory can be applied to data sharing; (2) show how the principles present in successful and unsuccessful data sharing cases differ; (3) demonstrate the connection of IP rules, laws, and polices to data sharing practices, and the scales at which this connection varies in importance.

Acknowledgments

We thank D. Baldocchi, G. Biging, C. Hess, C. Levitan, Y. Malhi, D. Winickoff, and anonymous reviewers and screeners for helpful comments; and to international colleagues Q. Chen, J.-H. Lee, K. Mennell, Y. Parag, R.M. Roman-Cuesta, R. Vinya and P. Zelazowski for internationally-focused internal review. Due to this journal's formatting requirements, many references were deleted; the fully-referenced work may be obtained from the authors. This work was supported by NASA Headquarters under the Earth System Science Fellowship Grant NGT5-30473, and by the University of California at Berkeley.

References

[1] E. Ostrom, Governing the Commons: The Evolution of Institutions for Collective Action, Cambridge University Press, Cambridge, 1990.
[2] C. Hess, E. Ostrom, Ideas, artifacts, and facilities: information as a common-pool resource, Law and Contemporary Problems 66, 2003.
[3] R. van Wendel de Joode, Explaining the organization of open source communities with the CPR framework, International Association for the Study of Common Property, Oaxaca, Mexico, 2004.
[4] E. Ostrom, C. Hess, A framework for analyzing the knowledge commons, in: C. Hess, E. Ostrom (Eds.), Understanding Knowledge as a Commons: From Theory to Practice, The MIT Press, Cambridge, Massachusetts, 2007.
[5] J.H. Reichman, P.F. Uhlir, A contractually reconstructed research commons for scientific data in a highly protectionist intellectual property environment, Law and Contemporary Problems 315, 2003, pp. 317–325.
[6] P. Samuelson, Enriching discourse on public domains, Duke Law Journal 55, 2006, pp. 783–834.
[7] D. Bollier, The missing vocabulary of the digital age: the commons, The Common Property Resource Digest 65, 2003, pp. 1–4.
[8] A.C. Stylianou, C.J. Jeffries, S.S. Robbins, Corporate mergers and the problems of IS integration, Information & Management 31, 1996, pp. 203–213.
[9] K.E. Kolekofski, A.R. Heminger, Beliefs and attitudes affecting intentions to share information in an organizational setting, Information & Management 40, 2003, pp. 521–532.
[10] H.A. Artail, Application of KM measures to the impact of a specialized groupware system on corporate productivity and operations, Information & Management 43, 2006, pp. 551–564.
[11] D. Baldocchi, 'Breathing' of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems, Australian Journal of Botany 56, 2008, pp. 1–26.
[12] C. Okoli, W. Oh, Investigating recognition-based performance in an open content community: a social capital perspective, Information & Management 44, 2007, pp. 240–252.
[13] Office of Management and Budget, Management of Federal Information Resources, 2000.
[14] J.N. Lee, The impact of knowledge sharing, organizational capability and partnership quality on IS outsourcing success, Information & Management 38, 2001, pp. 323–335.

Dr. Joshua Fisher is an ecosystem scientist at NASA/JPL who works across plant, ecosystem and global scales with a combination of field ecology, biometeorology, modeling, GIS and remote sensing. He received his BS in Environmental Sciences and PhD in Environmental Science, Policy and Management from UC Berkeley.

Professor Louise Fortmann is a rural sociologist. She has spent eleven years doing research on agriculture and natural resource management in east and southern Africa. She currently works on the sociology of science. Prof. Fortmann came to UC Berkeley as an assistant professor in the Department of Forestry and Resource Management in 1984.