Astrocommons and the Evolving Futures of Scientific Research

Michael J. Madison*

Abstract

The constructed cultural commons framework is applied to two cases of commons governance. Both are situated in the context of data-intensive science in astronomy and astrophysics. One, the Nearby Supernova Factory, is an interdisciplinary collaboration among several groups of professional scientists. The second, Galaxy Zoo, is a leading example of a citizen science project, in which volunteer non-scientists have been recruited to participate in large-scale data analysis via the Internet. Each project has grappled with the challenges posed by enormous volumes of astronomical data by using adopting a kind of commons, but the particular solutions employed vary, in part to accommodate the different demands of managing professional and volunteer participants in scientific research.

Introduction

Humans have long looked to the stars to understand how they should look at each other and at their world. Humans likewise have long looked to each other to understand how they should look at the stars. That reciprocal relationship gave us the disciplines of astrology, astronomy, and now astrophysics, and to ever greater understandings of both literal and metaphorical influence and force. Along the way, and beginning with early astronomers, cultures of scientific inquiry and research emerged, with their own influence and force both on scientists themselves and on the institutions of science and related public policy.

Twentieth-century science has been defined conventionally as a network of institutions for collaboration and knowledge-sharing among individual researchers. Policy analysis of science, particularly in recent decades, has focused on tensions between underlying norms of open science, on the one hand, and political and economic pressures to embed scientific research in market-based institutions, such as modern patent law. How has that tension improved scientific research; how has that tension impeded it? This paper offers a study of the norms of scientific research as expressed in a handful of contemporary astrophysics research institutions, using as its lens not the conventional framework of proprietary rights exchanged in markets, but instead a framework grounded in the idea of commons governance. The intuition is that commons, not the market, offers a superior analytic framework for understanding the changing futures of science.

While the primary goal of the study is to discern the mechanics and functioning of the research resources described below, certain tentative conclusions are offered. First, at a high conceptual level, there exists a dynamic relationship among scientific practice, forms of knowledge and knowledge structures, and social organization. That postulate was offered by Thomas Kuhn nearly 40 years ago; it appears to be accurate today. Second, at a lower and more concrete level, relevant forms of social organization – that is, both the shape of astronomy and astrophysics disciplines and the character of their commons governance – are dependent on elastic conceptual and material (technologically-grounded) understandings of the data that scientists generate and use. In light of that elasticity, those disciplines are not static.

The particular knowledge institutions under review are the Nearby Supernova Factory (SNfactory) (http://snfactory.lbl.gov/), a formal inter-institutional astrophysics data collection, curation, and distribution enterprise, and the Galaxy Zoo (http://www.galaxyzoo.org/), an online citizen science project for classifying galaxies, using observations from a terrestrial telescope and the Hubble Space Telescope.

"Commons" as used in this chapter is an umbrella term that refers to a broad array of possible institutional arrangements for sharing information and knowledge (that is, products and sources of human culture) and for sharing legal rights that might pertain to that information and knowledge. Commons refers to openness, but to structured openness, with formal and infromal institutional mechanisms in place to manage that structure. Commons is governance. In this sense commons should be distinguished from the unrestricted formal openness which defines the concept of the public domain in intellectual property law and which is sometimes attached to the term "commons" in casual or political usage. I refer to commons as "constructed" because of the important sense in which commons are human institutions, often produced purposefully but sometimes emerging from or evolving out of historical happenstance. Below, I refer to the "constructed commons framework" as the analytic framework for researching constructed commons that is presented in *Constructing Commons in the Cultural Environment* (Madison, Frischmann & Strandburg (2010)).

Part I below provides an overview of the constructed cultural commons concept in greater detail, with an emphasis on points of distinction between specific commons instances as well as on similarities. Part II describes the Nearby Supernova Factory and the Galaxy Zoo in greater detail, both by supplying a brief narrative of their histories and functioning in light of histories and practices of astronomers and astrophysicists, and by breaking their components down in light of the detailed clusters of research inquiries suggested by the constructed commons framework. Part III offers some tentative analysis.

I. The Constructed Commons Framework

The constructed commons framework builds on a series of related intuitions. The first of these is that structured openness in the management of both natural and cultural resources is likely to lead to socially-beneficial and/or socially-productive outcomes, focused especially on the production and generation of infrastructural resources, as well as system outputs. Salient among this class of cases are contexts where social interest in positive spillovers from bilateral, market transactions is high; commons may sustain the production of spillovers when the market otherwise may not. The second intution is that such constructed commons are observed widely in practice and are in broad use; their relatively marginal status in policy discussions often stems from individual commons institutions not being collected and treated as a body of related phenomena. The final intuition is that a standard framework for identifying and assessing commons across a variety of domains can support the development of more sophisticated tools for realizing the potential for commons solutions in new institutional settings, and for distinguishing commons solutions from other solutions in settings where some other approach, such as an approach grounded in proprietary rights, should be preferred.

Analysis of commons in the cultural context builds on the Institutional Analysis & Development (IAD) framework pioneered by Nobel Laureate Elinor Ostrom and her colleagues

(Ostrom (1990)), but it adds some important modifications. The IAD framework was developed to structure analysis of solutions to collective action problems in natural resource (i.e., physical environmental) contexts such as forests, fisheries, and water management systems. IAD analysis is premised on rational behavior by self-interested individuals, and it looks to explain sustainable collective action that produces measurable, productive outcomes. The IAD insight is that commons solutions can be as robust as market-oriented solutions to classic "tragedy of the commons" scenarios. Shared governance can lead to more fish, more trees, and more usable clean water.

The cultural commons framework differs in certain key respects. It does not assume rational, self-interested individuals as the only key actors. It accepts the role of historical contingency in the evolution of collective or commons institutions. It does not measure the success of a commons regime solely or even primarily by measuring the regime's outputs. And at the front end of the analysis, it requires understanding the contingency of the underlying resources themselves. Natural resource commons largely take their resources for granted: fish, trees, water, and the like. Cultural commons identify resource design as one of the variables to be analyzed. Patents, copyrights, and underlying inventions, creations, and data are shaped by a variety of institutional forces, rather than by nature. Critically, the cultural commons framework does not assume that commons resources are rival and depletable. The framework generally assumes precisely the contrary: that intangible information and knowledge resources are non-rival, non-excludable public goods. The "tragic commons" problem to be solved is not, accordingly, a classic overuse problem. Instead, it is an underproduction problem: in the absence of a governance mechanism to moderate consumption, producers of resources will fail to invest in creating new goods, because of uncertainty regarding their ability to earn returns that justify the investment.

Against that background, the cultural commons framework proposes to undertake comparative institutional analysis by evaluating a series of buckets of questions, or clusters, in the case of each instance of commons. Several of these are borrowed or adapted from Ostrom's IAD framework. Some are developed specifically for the cultural commons context. The commons examples and illustrations described in the next part are not measured in each case against each of these buckets. Rather, the full list is described here, and an abbreviated version is applied to the illustrations that follow.

The initial question is whether the relevant commons is characterized as an initial matter by patent rights or other proprietary rights, as in the case of a patent pool, or by a legal regime of formal openness, as in the case of public domain data or information collected in a government archive. A particular commons might involve sharing data and information, or sharing rights in information, or sharing both.

Answering that question sets a default baseline against which a commons regime is constructed. Within that regime, one next asks definitional questions: What are the relevant resources, what are the relationships among these resources and the relevant legal regime (for example, what a scientist considers to be an invention, what patent law considers to be an invention, and the patent itself are three related but distinct things), and what are the boundaries and constitution (membership) of the community or communities that manage access and use of those resources? How is membership acquired (this may be informal, formal, or a blend of the two), and how is membership governed?

Next in order are questions concerning explicit and implicit goals and objectives of the commons, if there are any. Is there a particular resource development or management problem that the commons is intended to address, and what strategies are used by the commons to address that problem?

How "open" are the resources and the community of participants that create, use, and manage them? Some commons and commons resources have precise and fixed definitions of both resources and community membership. In some commons, either resources or membership or both are more fluid, with boundaries defined by flexible standards rather than rules.

A large and critical cluster of questions concerns the dynamics of commons governance, or what Ostrom refers to as the "rules in use" of commons: the interactions of commons participants and resources. Included in this cluster are (i) stories of the origin and history of the commons; (ii) formal and informal (norm-based) rules and practices regarding distribution of commons resources among commons participants, including rules for appropriation and replenishment of commons resources; (iii) the institutional setting of the commons, including the character of the commons' being "nested" in larger scale institutions and dependent on other, adjacent institutions; (iv) relevant legal regimes, including but not limited to intellectual property law; (v) the structure of interactions between commons resources and participants and institutions adjancent to and outside the commons; and (v) dispute resolution and other disciplinary mechanisms by which commons rules, norms, and participants are policed.

At this point it becomes possible to identify and assess outcomes. In Ostrom's IAD framework, outcomes are assessed in terms of the resources themselves: Has a fishery been managed in a way that sustains fish stocks over time? Do commons participants (fishermen) earn returns in the commons context that match or exceed returns from participation in an alternative governance context? In cultural commons contexts, equivalent outcome measures are difficult if not impossible to assess. Outcomes take different forms; in the cultural context framework, it will often be the case that patterns of participant interaction constitute relevant outcomes as well as relevant inputs. In a patent pool, pooled resources may constitute components of larger, complex products that could not be produced but for the pooling arrangement that reduces transactions costs among participants. But participant interaction in the context of a pool may give rise to (or preserve, or modify) an industrial field, or a technical discipline.

Having identified relevant outcomes, it becomes possible to look back at the problems that defined the commons regime in the first place. Has the commons in fact solved those problems, and if not, then what gaps remain? And on the other side of the assessment ledger, has the commons created costs or risks that should give policymakers pause? Costs of administration might be needlessly high; costs of participation might be high; and the commons might offer a risk of negative spillovers that offsets the initial instinct that commons produce positive spillovers. A collection of industrial firms that pool related patents in order to produce complex products may produce those products but may also engage in anticompetitive, collusive behavior. A commons regime may facilitate innovation; it may also facilitate stagnation.

II. Astrocommons: Nearby Supernova Factory and Galaxy Zoo

Two distinct but related astronomy and astrophysics research projects were chosen for study, on the following basis. First, the decrease in cost and explosion in the capacity and speed of information technologies has led to a grand reformation in the scale and pace of scientific research across a broad variety of disciplines, as more data than ever before has become available for study. Astronomy is among the first of these fields to begin to develop a rich array of institutional responses. Scientific research in general has been characterized as a commons, but as the technologies and techniques of that research change, and in many cases as they change dramatically, fields of research have begun to pause and consider anew how to construct appropriate governance regimes for new data and research outputs. In the emerging era of data-intensive science, once a commons, always a commons? Or, are other shifts observed, whether large or small, in governance of knowledge and information? The Nearby Supernova Factory is an example of a collectively-managed resource for sharing research data, created and managed by a group of high-level research institutions for the use and benefit of professional researchers. The Galaxy Zoo, by contrast, is a successful example of a citizen science project, developed and implemented by professional astronomers and physicists for the use of a broader public interested in astronomical research as well as for the use and benefit of professional astronomers.

The results below are derived primarily from examining the publicly accessible online materials that describe and implement each project. In the case of the Galaxy Zoo, I have also relied on a small number of research papers and conference posters reporting on studies of (and by) Galaxy Zoo participants. In each case, in the main I have described the project according to the buckets or clusters of questions framed by the constructed commons framework.

A. Nearby Supernova Factory

The Nearby Supernova Factory (SNfactory) is an international astrophysics experiment "designed to collect data on more Type Ia supernovae than have ever been studied in a single project before, and in so doing, to answer some fundamental questions about the nature of the universe." (http://snfactory.lbl.gov/snf/snf-about.html) Specifically, it is "an experiment to develop Type Ia supernovae as tools to measure the expansion history of the Universe and explore the nature of Dark Energy. It is the largest data volume supernova search currently in operation." (http://snfactory.lbl.gov/) Planned in 2001 and launched in 2002, SNfactory has six participating institutions (three in France, two in the US, and one in Germany), and about 30 participating individual members, about half of whom are in the U.S. and the other half in France (a very small number of members are located in other countries). Membership is interdisciplinary; it includes physicists, scientists and software engineers, among others. The project uses its primary telescope in Hawaii (Haleakala) and a second telescope in California (Palomar) to collect up to 80 GB of data each night, using specifications provided by a geographically distributed group of two to six people. That data becomes part of Sunfall (SuperNova Factory AssembLy Line), "a collaborative visual analytics software system to provide distributed access, management, visualization, and analysis of supernova data." (http://snfactory.lbl.gov/snf/snf-sunfall.html) The data is transferred via a high-speed network from Hawaii to the Lawrence Berkeley National Laboratory (LBL), where the search for supernovae is undertaken in a PC cluster. Follow-up Spectroscopic screening and analysis takes place at LBL, at facilities in France, and at Yale. Its distributed, interdisciplinary data curation and management strategies are regarded as integral to the project, and key contributors to its success. SNfactory and Sunfall have reduced false supernovae identification by 40%, improved scanning and vetting times by 70%; and reduced labor for search and scanning from six to eight people working four hours per day to one person working one hour per day. It led to ten publications in 2009 in both computer science and physics journals. (Tony Hey presentation; confirm source.)

The default baseline

The SNfactory presents no special challenges in understanding the default baseline regarding the status of the resources that are part of the scientific collaboration. The scientific data (images) collected at the front end of the SNfactory process are nonproprietary, public domain information. Each image is in principle a copyrightable work of authorship as an image. For at least two reasons copyright is unlikely to apply in this context. First, the source images are produced by the Near-Earth Asteroid Tracking (NEAT) program operated by the Jet Propulsion Laboratory, which is a facility of NASA, a U.S. Government agency. (http://neo.jpl.nasa.gov/programs/neat.html) (The source images are produced by a CCD camera, with a connected computer system, that is attached to the telescope in Hawaii.) The "authors" of the work are at least in part employees of the U.S. Government – Air Force contractors who operate the telescope and CCD combination. Second, in the context of the SNfactory the source images are considered to be data rather than shareable or exploitable works of authorship. In a manner of speaking, they are functional things rather than expressive objects. In various combinations, individual observations or data objects might be combined into collections of data that could be treated as copyrightable works of authorship, but given the interests of researchers and disciplines in data organized (at least initially) in patterns dictated by disciplinary and/or technology needs, it is unlikely that such combinations or collections would demonstrate sufficient "originality" to be treated as such. Moreover, in practice, SNfactory researchers do not treat their data as proprietary in any relevant legal sense. There is no evidence that any commons resources have been dedicated to clearing, combining, or exchanging rights in the source images as part of the SNfactory governance structure.

The character of the resources and of the community

As noted, the initial commons resources are the images obtained via the NEAT program. Those images are compressed and transmitted via a dedicated Internet connection to SNfactory processors, where they are analyzed according to the project's protocols. The project's description of the process captures the scope and technique with greater precision:

> The imaging data [note the phrasing here] are compressed and transferred to the National Energy Research Science Center (NERSC) at LBNL and archived on a 2 Pbyte tape vault. In the case of the Haleakala data, the high-speed internet connection between the Air Force Maui Supercomputer Center and NERSC is used. In the case of Palomar, it was necessary to set-up a custom dedicated 48 Mbs wireless internet connection to relay the data from Palomar to the San Diego Supercomputer Center (SDSC), and then send the data on to NERSC via the Energy Sciences Network (ESnet). The images are processed and subtracted to search for SNe using the 390+ node Parallel Distributed Systems Facility (PDSF) at NERSC. (http://snfactory.lbl.gov/snf/pdf/spie_2002.pdf) (Overview of the Nearby Supernova Factory, Proc. SPIE 4836, 61 (2002); doi:10.1117/12.458107)

Membership in the SNfactory is defined in the first instance by the several institutions that are sponsors of the project: (1) Centre de Recherche Astronomique de Lyon (CRAL) (a Joint Research Unit of the University of Lyon 1 (UCBL), Ecole Normale Supérieure de Lyon (ENS-L), and Centre National de la Recherche Scientifique (CNRS)); (2) Institut de Physique Nucleaire de Lyon (IPNL) (a joint project of the Universite Claude Bernard de Lyon (UCBL) and the Institut National de Physique Nucleaire et de Physique des Particules (IN2P3) of the CNRS); (3) Lawrence Berkeley National Laboratory (LBNL) (owned by the U.S. Department of Energy and managed by the University of California); (4) Laboratoire de Physique Nucleaire et de Hautes Energies (LPNHE) (a Joint Research Unit of the IN2P3 and the Universities Pierre et Marie Curie (UPMC) and Paris Diderot. Il); (5) Physikalisches Institut Universitat Bonn (Bonn) (part of the University of Bonn); and (6) the Department of Physics at Yale University. (The University of Bonn is listed as an official sponsor of the SNfactory, and Bonn faculty are individual member researchers, but Bonn is not identified as part of the formal governing collaboration.) It is obvious that the project is embedded in a complex matrix of government institutions and public and private research institutions.

Individual researchers who are members of the SNfactory project are employed by these institutions or are students of faculty and researchers employed there (or both) and are academic scientists from a variety of disciplines: astrophysicists, computer scientists, and engineers from several engineering sub-fields (including electrical, mechanical, optical). The nature and breadth of the disciplines involved corresponds roughly to functional engagement with the SNfactory; the physicists work with the data and the computer scientists and engineers are responsible primarily for design and maintenance of the hardware and software facilities used to analyze the image data and are referred to by the project as "Builders" rather than as "Members."

That distinction is part of a formal governance structure that was put in place to manage everyone involved with the project. (http://snfactory.in2p3.fr/people/SNFactory_Organization_v4.1.pdf)

According to that document, the SNfactory is managed by an SNfactory Collaboration Board (SCB), which consists of one representative of each sponsoring group (each of the sponsoring institutions, aside from Bonn); plus the Principal Investigator of Supernova Cosmology Project at LBNL, a spokesperson for the French consortium, and a project manager. Operation of the project is delegated by the Collaboration Board to an Operations Committee (OC) (a small group of Member researchers, and the Project Manager). Individual graduate students and postdocs can be added to the project at the discretion of the leadership of each participating group. The Collaboration Board must approve admission of new faculty researchers or permanent staff.

The collaboration document includes some broad guidelines on matters of particular interest to the group but specifies little in the way of general disciplinary guidance. For example, the document contains the following statement regarding rights and duties of project participants:

> Participation in the project includes instrumentation development and maintenance, observations, data reduction, development of analysis tools, analysis and publication. All collaboration members have access to raw data and available calibrated data. Members will not spread the data, or information about unpublished results, beyond the membership of the collaboration.
>
> The OC will coordinate physics analysis. The SCB encourages the participation of all collaboration members in the definition and execution of relevant analysis concepts.

> Outlines of concepts for potential relevant analyses are requested to help guide this effort and to avoid unnecessary duplication of effort.

There is no express provision for identifying or resolving disputes regarding application of these standards.

Much more of the collaboration document is dedicated to identifying standards related to publication of works based on SNfactory data. "Any paper written by a collaboration member that uses data, software, or internal group knowledge that comes out of the collaboration's work is assumed to be a collaboration paper unless otherwise agreed to in advance by the SNfactory Collaboration Board." Including non-SNfactory authors on SNfactory-derived papers is permitted if their authorship is relevant. Authorship of conference proceedings may be expressed a single author, consisting of the Nearby Supernova Factory, with the project's names listed in a footnote. The governance document grants the Collaboration Board a certain (and somewhat unclear) power to designate project members to deliver talks on behalf of the SNfactory when invitations are received, even when invitations are received by individual project members.

The Collaboration Board is identified as the authority responsible for resolving disputes regarding authorship, though no dispute resolution standard is separately specified.

The governance document permits use of unpublished SNfactory data by non-project members with the permission of the Collaboration Board. In that respect, project data is treated as "proprietary" in a sense that is related to authorship and publication norms derived from scientific research practices, rather than norms derived from intellectual property or other formal law.

The style and tone of the collaboration document suggest that it was drafted by collaboration members and project staff, rather than by or with the aid of counsel. In other words, the document and the governance structure that it describes appears to be intended to coordinate the work of project participants rather that to grant any of them legally enforceable rights or obligations.

I have not yet talked with any project participants to learn more about how this governance structure has operated in practice.

Goals and objectives: the commons problem(s)

The commons problem here is simple to describe and to understand. Because of advances in information technology, astrophysicists now have access to previously unheard-of quantities of observational imaging data. Effectively accessing and analyzing that data requires the efforts of large numbers of researchers across a range of related but distinct fields. In this case, modern IT has made possible an experiment – observing a large number of supernovae in order to improve our understanding of the rate of expansion of the universe – that could not have been conducted previously. Conducting that experiment requires disciplinary, technological, and financial resources that, practically speaking, must be shared – just as the underlying data must be shared.

It appears that given the type and large quantity of data being analyzed, and the range of disciplines represented among the project's participants, the project can include a number of subsidiary goals, such as improving the specifications for supernovae selected for study, and will produce a wide variety of research results. For example, the project has identified supernovae other than the Type Ia supernovae whose standard brightness is the basis of their usefulness in measuring the expansion of the universe. The project has been able to determine the intrinsic standard brightness of Type Ia supernovae with much improved accuracy. A list of project publications is available at http://snfactory.lbl.gov/snf/snf-pubs.php.

Resource and community openness

The discussion above of the terms of the SNfactory collaboration suggest that the community of participants is quite closed, on the whole, in several ways: One must be a researcher in a relevant discipline to be eligible to participate; one must be a researcher (or a student or postdoc) at a sponsoring institution in order to participate; and one must have the approval of the governing body of the project itself in order to participate. There is no evidence that the identity or number of sponsoring institutions has increased since the project's inception; in fact, there is some suggestion that institutional sponsorship has declined (by one). The list of individual project members suggests that individual participation has changed over time. Some new members have joined the project. Other members have left.

The imaging data itself is presumptively open to anyone who partners with the facilities that generate it; the facilities host or partner with projects other that the SNfactory. More interesting here is the fact that the SNfactory draws an explicit boundary with respect to the data that it generates analyzing the source imaging data. The relevant issue identified by the project is accessibility and use of unpublished project data. That unpublished data is available to non-project researchers, but not at the discretion of individual SNfactory participants. Access to and use of that data has to be approved by the Collaboration Board, that is, by a small number of governing researchers.

"Rules in use" (narratives; appropriation, management, and replenishment; institutional nesting; relevant legal regimes; discipline)

- *Narrative(s)*

  To an observer outside the disciplines relevant to the SNfactory, perhaps the most significant thing about the narratives that accompany the project is their very normalcy. Within the astrophysics literature, the SNfactory is presented as a technical solution to a very important new scientific problem: understanding the properties of "dark matter" or "dark energy," which is now believed to be the key to measuring the rate of expansion of the universe. (*Overview*, *supra.*) The collaborative elements of the project and its governance structure are largely hidden from public view. They are not hidden outright, but they appear to be treated as scientific or research infrastructure (which, of course, they are).

  There is an important meta-narrative at work among astronomers and astrophysicists and other research scientists, who are working to solve policy problems associated with working with very large data sets, and their collaborators in information science, information technology, sociology, law, and institutional funders. SNfactory is offered as a central example of a large-scale collaborative project with unusually demanding technical, social, and perhaps legal and policy needs. (Tony Hey presentation; Alyssa Goodman presentation; *Science* issue on data-driven science Spring 2011; *Nature* on data-driven science; same in *Communications of the ACM*; *The Economist*.)

- *Appropriation and replenishment processes and rules*

  SNfactory's source imaging data is created via technological processes, which consist of the hardware and software that drive the telescopes that collect data every evening, then transmit it to the Sunfall processing system. Those processes include both hardware and software specifications and protocols, which means that data creation is not insulated from human activity; the data depends on criteria devised by humans. The data used by the SNfactory consist of frequently-refreshed images of each patch of sky. The "refresh rate" is as often as every six days. It is unclear whether that rate is specified by SNfactory or adopted by SNfactory based on the operation of the Haleakala and Palomar telescopes. Given the huge volume of imaging data generated nightly, so long as the machinery works as intended, there is an endless pool of new imaging data to be analyzed. The sky is a non-depletable resource.

  One might also look at intellectual contributions to data analysis as the relevant resource, such that the collaborative governance provisions regarding authorship and attribution of scholarly papers are the most important mechanisms for ensuring that the pool of analysis is "refreshed" via publications.

- *Nesting*

  The SNfactory is embedded within a large number of linked institutions, many if not all of which might fairly be characterized in commons terms themselves, as research institutes, universities, and government agencies charged in large part with supporting or sponsoring scientific research. It may not be accurate to characterize these relationships as "nesting," if that metaphor implies that a particular commons is typically subordinate to a particular higher level commons in the manner of Russian nesting dolls. "Nesting" might be interpreted to mean that the SNfactory commons is threaded throughout a nest, or web, of inter-related commons. Using that interpretation, the nesting of the SNfactory is quite complex. The imaging data come from telescopes operated by the U.S. Government. The analysis of the data is conducted at facilities owned by the U.S. Government and operated by an agency of a state government (the LBNL), owned and operated by a private research university (Yale), and owned and operated by institutes housed in government-owned and run universities (Bonn, and a consortium of French universities).

- *Legal regimes*

  There do not appear to be any formal legal regimes that bear on the construction or maintenance of the SNfactory project. Indirectly, the project depends on and benefits from an intricate array of institutional arrangements and funding mechanisms that permit the project's sponsoring organizations to operate and to participate in the SNfactory. The SNfactory also depends on and benefits from an intricate array of institutional arrangements and funding mechanisms that permit the construction, operation, and maintenance of the facilities at Haleakala and Palomar that generate the imaging data that feeds the project.

- *Discipline*

  There do not appear to be any formal mechanisms for disciplining violations of collaboration protocols and norms, aside from the limited mechanisms described above. Informal mechanisms related to general norms of research and scholarship are likely powerful in this context. What is specified formally likely operates as a supplement to rather than as a substitute for baseline expectations of research scientists.

Outcomes and assessment

This section awaits collection of further data regarding the operation of the SNfactory in practice. Three observations are worth making at this point. One, the project began with a burst of ambition, and judging from the formal data available so far – publications, new scientific discoveries disclosed in the popular media as well as in the scientific literature, and the updated state of the project website – much of that ambition has been realized. Two, the commons here is in part a typical, almost stereotypical research project involving scientists. It is also in part an intricately devised formal and technical collaborative structure that is built on the normative practices of those scientists, and a collaboration that does not simply track the use of shared research objects but instead coordinates the flow of massive quantities of scientific information, from raw imaging data to published scholarly literature. Three, the SNfactory has earned what appears to be a deserved place in the casual literature of scholars studying the mechanisms of new data-driven science. I heard about the SNfactory for the first time in early 2011 at a conference at the National Academies of Science on legal and policy challenges facing data-driven science; it was one of two large-scale astronomy and astrophysics projects that were identified then as representing two of the most successful large-scale data-related scientific collaborations now underway.

The second was (and is) the Galaxy Zoo.

B. Galaxy Zoo

Galaxy Zoo is a collection of online astronomy projects, launched by a team of astronomers in England in 2007 primarily as a method of out-sourcing (or, in part, crowd-sourcing) a daunting data analysis challenge to "the people" of the Internet, that has taken on important educational and follow-on research dimensions. The initial research challenge, and the focus of the discussion below, was morphologically classifying roughly 900,000 known galaxies. Public volunteers were asked to answer a set of questions about galaxy images displayed on a public website. Based on a brief online tutorial, users were asked: Is this an elliptical galaxy or a spiral galaxy? If it is a spiral galaxy, which way does it appear to be rotating? The project was a tremendous success almost overnight: "Within 24 hours of launch, the site was receiving 70,000 classifications per hour. More than 50 million classifications were received by the project during its first year, from almost 150,000 people." (http://www.galaxyzoo.org/story) Galaxy Zoo is now the world's largest database of galaxy shapes. Galaxy Zoo participants have made a number of important discoveries, such as the object known as Hanny's Voorwerp and so-called "Green Pea" galaxies, and data generated by the project has been the based for dozens of published and submitted scholarly papers. (http://www.galaxyzoo.org/published_papers) The initial Galaxy Zoo project relied on image data from the Sloan Digital Sky Survey (SDSS) at the Apache Point Observatory in New Mexico. Galaxy Zoo 2, launched in 2009, asked volunteers to answer a more detailed set of questions in the classification of 250,000 galaxies within the original SDSS dataset. The current iteration of the project is Galaxy Zoo: Hubble, which has enlisted Galaxy Zoo volunteers to aid in classifying galaxies in imaging data obtained via the Hubble Space Telescope, i.e., galaxies that are farther away and older. Galaxy Zoo is now part of Zooniverse, a portal for citizen science projects in other domains of astronomy. Nearly 200,000 people are registered users of the Zooniverse.

The default baseline

The resources here consist partly of the underlying SDSS image data, which are in the public domain but which are managed by the collaborative of public and private research institutions that manage the SDSS (http://www.sdss.org/collaboration/), and primarily of the classification data supplied by Zooites, the volunteers who access Zoo images and respond to classification questions. The responses generally take the form of responses to questions with a binary structure; a single response to a question that asks "Is this a spiral galaxy or an elliptical gallery?" is unlikely to be original or creative enough to warrant copyright protection. There appear to be no express rules or licenses in the public Galaxy Zoo documentation that refer to ownership or transfers of ownership of interests in Galaxy Zoo data. The collections of data assembled from the first Galaxy Zoo project (known as Galaxy Zoo 1) have been posted online for download by anyone (http://data.galaxyzoo.org/).

The character of the resources and of the community

The initial resources are SDSS galaxy images; after volunteers participate in the classification exercise, a second set of resources are generated: classification data. Those data are treated as effectively public both by the Galaxy Zoo team and by the broader community of Zooites, many of whom apparently have invested a large amount of time not only in participating in classification exercises but also in building the Zooite community, researching astronomy questions prompted by SDSS data available via the Galaxy Zoo, and in some cases collaborating on the publication of scholarly research papers based on their work with Galaxy Zoo.

The nature of the commons community here therefore has several dimensions. The primary commons community consists of the astronomers and cosmologists who devised the Galaxy Zoo in the first place as a method of data analysis and who now manage the Galaxy Zoo and the related Zooniverse. These were primarily faculty and graduate students at the University of Oxford, together some colleagues in the U.S., primarily at Yale. The Galaxy Zoo team has expanded somewhat over time, both geographically (to include specialists in other European countries and the U.S.) and technically (to include specialists in IT disciplines who have helped to build out and maintain the various online elements of the Galaxy Zoo. There is no public record of any formal or informal management structure among the team. Virtually all of the scholarly papers that have used Galaxy Zoo data have been authored by team members, though some have had non-team members as co-authors, and some scholarly papers have emerged from volunteer focus on "Irregular" data identified during the classification exercises. (http://www.wavwebs.com/GZ/Irregular/Hunt.cgi)

The rapid growth of the population of Zooites gave rise in late 2007 to the construction of an online forum where Zooites could communicate with one another and talk about the science involved in the Galaxy Zoo (thus relieving the team of some of the growing burden of answering a growing flood of emailed questions) and then to the construction of a blog, where the team (sometimes called "Zookeepers" by Zooites) could communicate with the broader group. The membership of the forum, where registration is required for participation, is large and diverse and constitutes a second, related commons community. These are citizen scientists of a sort: volunteer lay astronomers who have both built and relied upon the Galaxy Zoo forum to participate meaningfully in astronomy research and, importantly, in education. (The Galaxy Zoo now serves as a useful teaching resource.) The technical aspects of the forum are managed by the team, but the social and content-related aspects of the forum are managed by lead volunteers. I have not yet determined the history of the forum. No formal governance rules are posted online; there is no public account of forum leadership, of norms of civility, or of discipline for bad behavior. Informal governance structures are evident from the structure of forum topics and discussion threads. Forum participants evidence compliance with strong subject-matter-based norms. The forum appears at http://galaxyzooforum.org/. There is some semantic ambiguity in the phrase "citizen scientist" that remains to be explored.

There is a third commons community at work, which consists of the broad population of Internet users who have participated classifying Galaxy Zoo images. The Galaxy Zoo site does not require registration to participate in classification. This community is, accordingly, difficult to characterize as a "community" in any sense, because so little is expected of members and because there is little meaningful way to identify who is and who is not a member. Nonetheless, there seems to be considerable overlap between the population of people who simply click through the classification exercises and those who participate in the forum, and the Galaxy Zoo project itself has undertaken some studies of the general population of participants, suggesting there may be some rough boundaries that define a community of Galaxy Zoo participants. The most obvious of those boundaries is the Galaxy Zoo classification exercise itself: community membership requires responding to the classification questions. It is apparent that the questions invite isolated or occasional responses, but in practice many volunteers spend hours at the task. The most common explanation given by volunteers for their interest in participating is the desire to contribute to real scientific work. That norm provides a certain modest discipline for the broader Galaxy Zoo community. (Raddick, et al., *Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers*, Astronomy Education Review (2010).)

Goals and objectives: the commons problem(s)

It might be said that Galaxy Zoo created a commons problem rather than solved one. By the turn of the 21st century, astronomers had identified more galaxies – roughly 1 million in the Main Galaxy Sample of the Sloan Digital Sky Survey -- than could be classified ever by professional astronomers using visual inspection techniques. (Not all galaxies in the SDSS were included in Galaxy Zoo 1; the data sample included galaxies of a specified brightness or greater.) Astronomers had been looking for solutions in various computation-based methods: artificial neural networks, computational algorithms, and model-based morphologies coded into software. Oxford astronomers came up with the idea of out-sourcing classification to volunteers on the Internet in order to obtain data to improve the modeling approach.

The success of the project, in terms of the number of participants and the amount of data generated by them, gave rise to the commons communities described above, in addition to the existing research collective that designed the project. Galaxy Zoo is now perhaps more broadly known for its educational and outreach features, citizen science in the sense that lay scientists are using Galaxy Zoo to teach schoolteachers. But as one leading paper notes, the core of Galaxy Zoo and the genius of its design lie in a data processing and analysis method that spawned a broad community of volunteers to organize both a set of supplemental community resources (the forum, for example) and to adapt the Zoo infrastructure and methods to other, related data processing projects: the Zooniverse.

> It is tempting to think of Galaxy Zoo purely as an Education and Outreach endeavor with all its successes in garnering publicity and focus on a community of non-expert volunteers. And with that temptation, one might imagine applying the Galaxy Zoo method to an indiscriminate array of projects with the idea that the public would be engaged in the process so it does not matter if the scientific outputs were “real" or whether the data processing could have been better accomplished through standard computational methods. What must be made clear is that Galaxy Zoo turned citizen science into a data processing method - a data reduction tool for data intensive science which when applied correctly provides the best possible data product from a set of “raw" data. The genius in this method lies in the fact that the public actually prefer to participate in a meaningful set of tasks where they know their work is useful. Galaxy Zoo established this coupling between high-priority science output and the public engagement in science. Once it became clear that the appetite of the volunteer classifiers could crunch significantly more data the question became one of how this new citizen science method could be made available across different disciplines and data products. And how to begin the process of developing the

machine algorithms trained by the human classifiers. Fortson, et al., *Galaxy Zoo: Morphological Classification and Citizen Science, in* ADVANCES IN MACHINE LEARNING AND DATA MINING FOR ASTRONOMY (2009).

Resource and community openness

The three, related commons communities described above operate with different degrees of openness, both as to membership and as to creation and use of resources. The community of scientific professionals is not static on its own terms, and members of that team have shown their willingness to admit both non-team scientists to the task of analyzing the data and "volunteers" and/or volunteer-produced discoveries to the task of developing scholarly papers. (http://blogs.zooniverse.org/galaxyzoo/2009/07/02/the-story-of-the-peas-writing-a-scientific-paper/, describing this paper: http://arxiv.org/abs/0907.4155.) Published papers using Galaxy Zoo data are listed at http://www.galaxyzoo.org/published_papers. Forum participation is open to anyone who wishes to register and create an account in order to post. The classification tasks may be performed by anyone with an Internet connection or access to one.

"Rules in use" (narratives; appropriation, management, and replenishment; institutional nesting; relevant legal regimes; discipline)

- *Narrative(s)*

  Much of the relevant narrative has been disclosed already. The underlying scientific problem, the classification of galaxies, is a long-standing challenge for astronomers and was among the questions addressed by Edwin Hubble, for whom the space telescope is named. The out-sourcing solution emerged partly by plan and partly by chance. Astronomers at Oxford were looking for a solution to the morphological classification problem and were designing a larger-scale version of the Stardust@Home project, which had used Internet volunteers to identify tracks made by interstellar dust in samples that were flown in NASA's Stardust mission. The designers of that project happened upon an independent project underway in the Oxford Physics Department, which planned to set a computer in an Oxford cafeteria and invite users to classify galaxies by their "handedness." The teams merged their efforts. The Galaxy Zoo project was launched publicly on July 11, 2007, accompanied by publicity in the popular media in the UK and soon, around the world. (http://news.bbc.co.uk/2/hi/science/nature/6289474.stm)

- *Appropriation and replenishment processes and rules*

  As with the SNfactory, the sky and the associated imaging data is essentially non-depletable. Replenishment of the classification data has proved to be robust in practice, despite the absence of any formal or informal rules that govern how the classification data is created. For example, one might imagine that Galaxy Zoo would instruct volunteer classifiers to enter an answer only one to a classification question regarding a given image. But it does not. It may be that there is little need for a formal rule; given the number of galaxy images, it is highly unlikely that a given volunteer would see the same image more than once, and a volunteer gains nothing of value by repeatedly entering a response to a single classification task. So far as I can determine so far, there are no formal or informal rules governing collaboration on scholarly papers using the Galaxy Zoo data, aside from underlying norms of scientific collaboration generally.

- *Nesting*

  The nesting of Galaxy Zoo is complicated by the fact that the project began as a hierarchal data processing method and evolved over time into a more complex collection of inter-related commons communities focused on the series of Galaxy Zoo projects themselves. It has also evolved into a collection of like-minded citizens science-oriented Zoo projects that are collected as the Zooniverse. More research needs to be done to identify the formal and informal relationships among the several Zooniverse communities. The academic team members of the Galaxy Zoo each are housed in their respective universities, but the Galaxy Zoo project itself appears to be organizationally linked to and accountable to none of them.

- *Legal regimes*

  I have not found any legal regimes that bear on any features of Galaxy Zoo. The forum is nominally subject to the "safe harbors" for hosted content described in Section 512 of the Digital Millennium Copyright Act.

- *Discipline*

  I have yet to identify any formal or informal disciplinary mechanisms with respect to any of the commons communities described above.

Outcomes and assessment

In terms of the number of volunteer participants, their passion and focus, the amount of morphological classification data generated, and the number of scholarly papers published using that data, there seems to be little reason to question the success of the Galaxy Zoo project, both on its initial terms and on the terms that evolved over time, with the emergence of the forum and with volunteer-led discoveries. Two new and rare classes of object have been identified by volunteers.

Without further data, however, about the practices of the Galaxy Zoo team and volunteers, it is different to offer more assessment at this point.

III. Analysis

The purpose of this paper is to outline two distinct but conceptually related "big data" projects in terms of the constructed commons framework. It is not to undertake a comparative analysis in any detail. Nonetheless, some comparisons are ready at hand. SNfactory constitutes what might be characterized as an "ordinary" norm-driven scientific research commons. There is nothing new in the observation that scientists are sharing data and collaborating on scholarly papers, though it always interesting enough to see that expectation confirmed in practice and to see how a particular group of scientists have elaborated norm-based practices in explicit detail. Galaxy Zoo is the paradigmatic citizen science project, though the core of Galaxy Zoo is professional research science, with more implicit and less explicit reliance, one suspects, on underlying norms of research science. It may be the case that the citizen science dimensions of the Galaxy Zoo project off a means of sharing those norms with lay participants, and educating them in the ways of research scientists. If further research bears that out, then that outcome would be an interesting and perhaps novel outcome of this commons. Galaxy Zoo would look more like SNfactory than it appears to at first glance.

A second point of similarity may lie in the resources at stake in each setting and the commons tools used to manage them. If commons are generally understood as offering a governance solution to a resource management problem, then these two cases test that proposition. Neither SNfactory nor Galaxy Zoo offers a clearly defined set of resources to be managed. In the reviews above, I have stretched a bit to identify "resources" at stake, but in fact what appears to be at stake is less a discrete collection of "things" and more a flow of information, or a method of analysis.

That observation prompts the title of the paper: Astrocommons and the Evolving Futures of Scientific Research. Both of the cases offered in this paper describe blends of disciplines, technologies, and participant communities that have developed in response to opportunities created by prior blends of disciplines, technologies, and participant communities. Prior practices and tools created new data; that new data prompts the creation of new practices and tools. In one case (SNfactory), a new equilibrium was reached that required a refinement of existing professional protocols for authorship and publication, that is, the continuation of mostly traditional scientific practice. In a second case (Galaxy Zoo), the creation of new practices and tools spawned a largely new range of mechanisms for "doing" science at the intersection of professional and volunteer communities. In neither case have the formal disciplinary boundaries of what it means to be an "astronomer" or a "physicist" been tested meaningfully, but the depth of engagement among professional disciplines (in the SNfactory case) and between professional disciplines and volunteer scientists (in the Galaxy Zoo case) suggests that disciplinary boundaries are increasingly, if informally, porous.

Conclusion

[To be written.]

Partial bibliography of secondary sources

John Seely Brown & Paul Duguid, *Knowledge and Organization: A Social-Practice Perspective*, 12 ORG. SCI. 198, 198-213 (2001)

Brigham Daniels, *Emerging Commons and Tragic Institutions*, 37 ENVTL. L. 515 (2007)

Rebecca S. Eisenberg, *Patents and the Progress of Science: Exclusive Rights and Experimental Use*, 56 U. CHI. L. REV. 1017 (1989)

Rebecca S. Eisenberg, *Proprietary Rights and the Norms of Science in Biotechnology Research*, 97 YALE L.J. 177, 181-84 (1987)

Stephen Hilgartner & Sherry Brandt-Rauf, *Data Access, Ownership, and Control: Toward Empirical Studies of Access Practices*, 15 KNOWLEDGE 355 (1994)

THOMAS MANDEVILLE, UNDERSTANDING NOVELTY: INFORMATION, TECHNOLOGICAL CHANGE, AND THE PATENT SYSTEM (1996)

Robert P. Merges, *Property Rights Theory and the Commons: The Case of Scientific Research*, 13 SOC. PHIL. & POL'Y at 145 (1996)

Robert K. Merton, *The Normative Structure of Science, reprinted in* THE SOCIOLOGY OF SCIENCE: THEORETICAL EMPIRICAL INVESTIGATIONS 267 (Norman W. Storer ed., 1973)

Arti K. Rai, *Regulating Scientific Research: Intellectual Property Rights and the Norms of Science*, 94 Nw. U. L. Rev. 77 (1999)

J. H. Reichman & Paul F. Uhlir, *A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment*, 66 LAW & CONTEMP. PROBS. 315 (2003)

Katherine J. Strandburg, *Sharing Research Tools and Materials: Homo Scientificus and User Innovator Community Norms, in* WORKING WITHIN THE BOUNDARIES OF INTELLECTUAL PROPERTY (Rochelle C. Dreyfuss, Harry First, & Diane L. Zimmerman, eds., Oxford University Press 2010)