CHAPTER 6

Organizing
Social, Medial, and Epistemic Orders in Data-Centric Collaboration

6.1 This Chapter's Plan

This chapter examines data-centric collaborations in the natural sciences as "organizational experiments" (Sharrock 2011, 29). Distinct from collaborations that build and operate large facilities like high-energy physics particle detectors and space probes, many data-centric collaborations are relatively short-lived teams whose work centers on datasets produced at public facilities like observatories, research ships, field stations, spacecraft, and national laboratories. Researchers join together to submit proposals to these facilities and, if successfully reviewed, are granted time and opportunities to use their equipment. For a certain period of exclusive proprietary use, the data that such groups obtain and make are their epistemic and social "stuff" that members use to establish connections and collaborations. Other teams work with open data or combine open and proprietary datasets. Diverse in origin and structure, data-centric collaborations typically lack features of formal organization, such as organizational charts and legally binding contracts. They often emerge from ties of academic genealogy and friendship, but may also be instigated by meetings, such as at hack weeks.

I concentrate on how data-centric collaborations organize their work. As in previous chapters, I track social accountability and examine scientists' practical reasoning and its resources. These include medial formats, epistemic orders, and a team's joint orientation to a dataset's projected coherence. I will focus on what I refer to as the Multiwavelength Galaxy Survey (MUWAGS; a pseudonym), an international collaboration of around thirty astronomers, but consider less formal and spontaneous collaborations as well. This introduction to the MUWAGS team also serves as a background for the next two chapters, on normative expectations of data access in "open science" (Chapter 7) and attempts to encode a team's collective knowledge of its data (Chapter 8). But to locate the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

work of MUWAGS, we first need an overview of its epistemic and organizational context.

6.2 A Society of Collaborations? Complex Datasets, Large Teams, and "Open Science"

The increasing dominance of teamwork and large collaborations has been noticed in many sciences, raising issues of social organization, coordination, authorship, and identity.[1] Several studies examine long-lasting collaborations that have a privileged or even exclusive access to the data-generating facilities they build.[2] Some projects in astronomy are like this. More often, though, researchers do not build instruments themselves but form teams to use public facilities for limited, shorter-lived projects.[3]

In astronomy, the move to large teams is associated with surveys – observational censuses of object populations.[4] Many surveys observe galaxies to examine the universe's history and matter content. Such studies have been at the forefront of promoting the open access to data, assembling large datasets, as well as using computational techniques of data mining and machine learning. Deep field studies are a prominent kind of such projects.[5] Making long exposures of selected small fields on the sky at many wavelengths, they use the finite speed of light as a sort of time machine to infer the history of galaxies from observing populations of distant exemplars (cf. Chapter 3).

Deep field studies emerged from a conjunction of three technological developments: the availability of large charge-coupled devices to photograph substantial parts of the sky, the repair of the Hubble Space Telescope (HST) in 1994 (which enabled high-resolution optical and near-infrared imaging), and the availability of ground-based 8-meter telescopes (enabling deep, long exposures and sensitive spectroscopy). Robert E. Williams, then the director of the Space Telescope Science

[1] Hackett (2005), Shrum et al. (2007), Wuchty et al. (2007), Leahey (2016), and Wu et al. (2019).
[2] See Knorr Cetina (1999), Vertesi (2015, 2020), and Roy (2024).
[3] There are also intermediate cases, such as the Dark Energy Survey (Abbott et al. 2016) that built a dedicated camera for use at the 4-meter Blanco telescope at Cerro Tololo Inter-American Observatory in Chile, a tax-financed, national facility.
[4] Lahav (2001). Historically, the first large collaborations in astronomy were surveys, including the Carte du Ciel, a photographic sky atlas that used observatories in the northern and southern hemisphere (Jones 2000).
[5] In this chapter I use the term "deep field" inclusively as describing all fields on the sky observed with long exposures at multiple wavelengths for extragalactic studies. What follows is a condensed and partisan history. See Koo (1999) and Williams (2018) for earlier work.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

Institute in Baltimore (Maryland, USA), decided to dedicate the entire annual "Director's Discretionary [Observing] Time" for 1995 on the HST to make series of long photographic exposures of a "blank field" (a region without bright stars) in the northern sky in four optical wavebands.[6] This so-called Hubble Deep Field (HDF) revealed an unexpectedly large number of faint and distant galaxies with unusual shapes and became a resource for many subsequent studies of galaxy evolution.[7] A team of sixteen postdoctoral scholars and staff scientists at the institute processed the data and released calibrated images and object catalogs (tables of the measured properties of detected objects) to the public less than six weeks after the last exposure had been taken – much shorter than NASA's usual one-year proprietary period for the exclusive use of unprocessed data (Williams 2018).

The late 1990s and early 2000s saw a sprawling development of deep field projects. Many institutions that had built a large telescope or detector selected and observed their own target field at a unique location on the sky. Examples are the Subaru Deep Field, observed with the National Astronomical Observatory of Japan's Subaru telescope, the FORS Deep Field, observed by the makers of the FORS (Focal Reducer and low-dispersion Spectrograph) instrument at the European Southern Observatory's (ESO's) Very Large Telescope, and the Chandra Deep Field South (CDFS), defined by a very long and sensitive x-ray exposure with NASA's Chandra spacecraft.[8] Astronomers soon realized that beginning new deep field projects all over the sky was not an optimal use of available resources. It seemed advisable to focus on selected fields and combine expertise and capacities to make deep (sensitive) exposures at many wavelengths to study detected object populations.[9] Thus, the deep x-ray exposure of the CDFS has attracted further observations at many other wavelengths ever since, making this one of the most observed parts of the sky. The two Hubble Deep Fields (HDFs; in the northern and southern sky) and the so-called COSMOS field have been much observed

[6] In 1995, the Director's Discretionary Time at the HST amounted to 10 percent of its annual observing time.

[7] Prior to the HST launch, Bahcall et al. (1990) had predicted that no new distant galaxy population would be found with this area, arguing against investing large amounts of telescope time for such a search. Mindful of this, Williams (2018, ch. 8) emphasizes that he is a risk-taker, arguably not unlike daring entrepreneurs (cf. Appadurai 2015).

[8] Riccardo Giacconi, a key figure in space x-ray astronomy, and his team chose the CDFS for their (successful) attempt to resolve the x-ray background radiation discovered in the 1960s (Giacconi 2008).

[9] This required forming teams with members of distinct "instrumental communities" (Mody 2011).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

as well, being sometimes described as "magnets" that attract complementary datasets.[10]

This development is contingent both on what data are like (digital, additive, shared format; cf. Chapter 1) as well as on how astronomers conceive of the distant universe as an object of study. For many practical purposes it is stable over human lifetimes (one can add exposures taken at different times), as well as, on large scales, isotropic (the universe looks the same in all directions) and homogenous (it has a universal average density; cf. Chapter 3). Without these assumptions (and their observational support) astronomers could not make comprehensive claims about the universe's history from observing a few selected fields only. Medial and epistemic orders thus enable and shape the work in this domain.

Over the course of my fieldwork, conversations among researchers that I witnessed were punctured with acronyms that referred to various projects, including GOODS, VVDS, EDisCS, DEEP2, COSMOS, LEGA-C, CANDELS, CLASH, and, more recently, DES, DESI, JWST CEERS, JADES, COSMOS-Web, and RUBIES.[11] Such acronyms are not only shortcuts to avoid rehearsing long project names, to make projects recognizable, and their publications searchable on the internet, they are also markers of ownership, responsibility, and members' collective identity (cf. Chapters 7 and 8). Ideally, they help building a team's good reputation. The projects that these acronyms describe are organizations in the sense of having specific goals (that are described in observing proposals and on team websites) and a well-defined membership (as listed on team websites and author lists), assigning specific rights and duties (data access and work tasks), being organized hierarchically (led by a principal investigator whom a collaboration board may support), and being, within limits, autonomous in their decision-making.[12]

[10] Interview with Mark Dickinson, National Optical Astronomical Observatories, October 25, 2016.

[11] Here, CANDELS is the Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (Grogin et al. 2011), CLASH is the Cluster Lensing and Supernova Survey with Hubble (Postman et al. 2012), COSMOS is the Cosmic Evolution Survey (COSMOS, Scoville et al. 2007), COSMOS-Web is a JWST survey of the COSMOS field (Casey et al. 2023), DEEP2 is the Deep Extragalactic Evolutionary Probe 2 (Newman et al. 2013), DES is the Dark Energy Survey (Abbott et al. 2016), DESI is the Dark Energy Spectroscopic Instrument (Abdul-Karim et al. 2025), EDisCS is the ESO Distant Cluster Survey (White et al. 2005), GOODS is the Great Observatories Origins Deep Survey (Giavalisco et al. 2004), JADES is the JWST Advanced Deep Extragalactic Survey (Eisenstein et al. 2025), JWST CEERS is the James Webb Space Telescope Cosmic Evolution Early Release Science project (Ronayne et al. 2023), LEGA-C is the Large Early Galaxy Census (van der Wel et al. 2016), RUBIES is the Red Unknowns: Bright Infrared Extragalactic Survey (de Graaff et al. 2025), and VVDS is the VIMOS VLT Deep Survey (LeFèvre et al. 2005).

[12] According to Kühl (2011) these criteria define a formal organization.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

Researchers are typically members of one collaboration or another.[13] As in Charles Perrow's (1991, 481) "society of organizations," it seems that, in this field, large collaborations have "absorbed society." Examined closer, many of these collaborations resemble each other in their organizational structure and work programs. This is the case for teams that assemble multiwavelength datasets, that is, observations in several distinct wavelength regimes like gamma rays, x-rays, ultraviolet, optical (visible light), infrared, and radio waves. Each of these wavelength regimes requires the use of specific kinds of detectors. Members of distinct "instrumental communities" (Mody 2011) make and calibrate these data. As the electromagnetic spectrum marks their division of labor, these collaborations' social organization mirrors their research's epistemic order and, indeed, the natural order that researchers in the field agree upon.

Collaborations in the field also divide labor according to specific tasks and scientific objectives that they broadly share. They usually include a group that estimates photometric redshifts (a measure of cosmic distance), one that conducts spectroscopy, and one that constructs a multiwavelength catalog (a table of measurements of objects sorted by celestial coordinates). Many datasets are used for a set of well-established research topics, such as measuring galaxy merger and star formation rates over cosmic time or mapping dark matter by observing distortions of background galaxy images due to weak gravitational lensing. Shared, as well, is the mandate to release datasets to the public. This may be done quickly, as with the HDF, or after periods of exclusive proprietary use. These periods introduce constraints on the temporality of teamwork. A sort of organizational isomorphism marks collaborations in this field.[14]

But collaborations must also signal being different from others in specific ways. To make the case for using large amounts of observing time at public facilities, applicants argue with metrics that demonstrate their unique advance over existing work, in terms of sensitivity, range and number of wavebands, observed sky area, or number of objects included in their sample.[15] As suggested earlier in this chapter, previous studies of deep

[13] Some astronomers are members of two or even more collaborations, even competing ones.

[14] I borrow this notion from DiMaggio and Powell's (1983) institutional isomorphism.

[15] Members' narratives cast relations between collaborations variously as complementary or competitive. When I asked members of a California-based team about its relation to the work of the MAMBO (pseudo-acronym) team, one astronomer argued that MAMBO had scooped it on key science, while another emphasized that their complementary techniques (MAMBO using photometric redshifts whereas the California team was using spectroscopic redshifts) yielded complementary insights.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

fields could argue for the benefit of observing different locations on the sky to counter biases due to "cosmic variance," that is, deviations from cosmic isotropy and homogeneity. Mitigating effects of cosmic variance became an incentive for merging two deep field projects to concentrate yet more resources in one large team.[16]

The move to large collaborations is tied to the increasingly open access to large datasets. Most astronomers agree that there were two precedents for this development: the HDF and the Sloan Digital Sky Survey (SDSS). Robert E. Williams conceived the HDF as a critique of the "Balkanization of science" in the use of public facilities (like the HST), where time allocation committees allegedly try to split the available observing time into many small projects to "make everyone happy," instead of focusing resources on few significant, but possibly risky, observing projects.[17] Realizing that a broad range of new telescopes and digital detectors promised a new era of abundant data, Williams criticized many astronomers' secretiveness about their data and their hesitation to share it. The point, he argued, was to change the "culture" of astronomy.[18]

Williams insisted that large, complex datasets enable interesting projects for everyone and should therefore be openly available to all astronomers. The original HDF data were quickly complemented by other datasets, such as spectroscopic observations with the 10-meter Keck telescopes on Mauna Kea (Hawaii). Many makers of these complementary datasets followed the precedent that the HDF team had set and published their data soon after the observations. This novel openness to large datasets was also adopted in observations of the popular CDFS, arguably solidifying a normative expectation of data access in this domain (cf. Chapter 7).

Conceived in the early 1990s by James E. Gunn, the SDSS is an imaging and spectroscopic survey using the 2.5-meter telescope at Apache Point (New Mexico). It is not a deep field project but a panoramic survey that covers the northern and equatorial sky.[19] Begun by Gunn and his colleagues in the Department of Astrophysical Sciences of Princeton University, its complexity brought this project to the edge of failure until Fermilab (Chicago), a high-energy physics laboratory with previous

[16] This was an argument in the design of CANDELS (Grogin et al. 2011).
[17] Interviews with Ethan J. Schreier, October 7, 2008, and Robert E. Williams, October 7, 2007.
[18] Leonelli (2016, 144) notices that some open science projects in biology are associated with individuals that peers regard as charismatic. By many astronomers' accounts, Williams would match this characterization.
[19] The local galaxy population that the SDSS documents is often compared, as a "present-day view," to galaxy samples in deep fields observed in earlier stages of cosmic evolution.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

experience in running large experiments, took over its management. The contact of university academics with the hierarchical management of a national laboratory turned out to be riddled with conflict, but it ultimately rescued the survey.[20] This is commonly described as a major learning experience for astronomers in terms of project management, dividing labor, tracking accountability, and budgeting costs. SDSS has been operating since 2000 in a series of five-year phases, financed and codeveloped by a variety of national and international partners who pay for their early access to its data, which have been subsequently made public in a series of releases. Scientists within SDSS and beyond have used these datasets for addressing a wide variety of scientific questions.

As with the HDF, astronomers within SDSS and beyond experienced the availability of its images, catalogs, and spectra as a sudden abundance of data that had a "civilizing" effect on how astronomers cooperate in using large datasets.[21] Gunn called for the SDSS to practice a sort of "open science" within the collaboration, where "any member could do any science he/she wanted. Nobody owned any data or any project." Members had to announce project ideas within the collaboration and allow interested members to join them. PhD projects, however, were protected, assuring early career scientists the exclusive use of SDSS data. Gunn argued that these rules "resulted in the natural growth of widespread collaborations of common interests across the world, and a very happy project." Much like Robert Williams with the HDF, Gunn argued that SDSS "changed the sociology" of astronomy.[22]

6.3 From Organizations to Organizing

This account suggests that, in studies of deep fields, collaborative work and organization is oriented to, and informed by, a field of other projects and organizations. I have used the notion of "field" casually.[23] But much of what I noticed resonates with field theories in the social sciences that examine "individual action by recourse to position vis-à-vis others" (Martin 2003, 1). Economic and organizational sociologists have adopted this viewpoint when examining competing businesses (Davis 2005) and

[20] Finkbeiner (2010), Szalay (2018), and Gunn (2020).

[21] By alluding to the "civilizing" effect of data I have Hirschman's (1977) study of early capitalism in mind.

[22] These quotations are from Gunn's award lecture at the 2019 Kyoto Prize celebration. See: youtube.com/watch?v=z657rXUIEus at minute 34 (accessed July 7, 2025).

[23] Perhaps confusingly, this would be a field of deep field studies.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

it is tempting to compare scientific collaborations with firms: both operate in contexts where "isolated individuals hardly matter in the production of goods and services" (Granovetter 2005, 429) and both compete with similar units, mutually constituting their competitors' environment. I noticed the similarities of deep field collaborations in terms of their division of labor, calling this an example of organizational isomorphism. Some collaborations design their project to be comparable with others, thereby asserting a position in a field and making a claim for membership in it – arguably a case of organizational diffusion.[24]

Notions of field, isomorphism, and diffusion help to characterize important aspects of this domain of data-rich science. But that they describe the work of business firms equally well reminds us that invoking them ignores the specifics of either domain. For a naturalistic account such as mine this is unsatisfactory. I noticed earlier that discipline-specific medial and epistemic orders clearly matter to how these scientists work and cooperate. We can approach this by following Karl Weick and examine organizing as processes that interlock the actions of collaboration members through interpretive practices (Weick 1979; 1995). Egon Bittner suggested a related ethnomethodological approach, proposing to regard "the concept of organization as a common-sense construct" by studying how it is used in "real scenes of action by persons whose competence to use them is socially sanctioned" (Bittner 1965, 247).[25]

In the following, I speak of data-centric collaborations, since it is through the joint production and use of data that scientists develop "forms of togetherness and for-each-otherness" (Simmel 1992 [1908], 18–19; translated from German). Many data-centric collaborations in astronomy have idiosyncratic origins, beginning with small teams of supervisor-and-student, friends, or colleagues, and growing through adding members or merging with other teams. These are "organizational experiments" (Sharrock 2011, 29) that typically lack organization charts, diagrams representing workflows, and legally binding contracts. But they are not without documents and other resources for coordination. Take observing proposals, for example. Although prepared for submission to time allocation committees, proposals are also resources for organizing. They document medial, epistemic, discursive, organizational, and regulatory conditions

[24] See Strang and Meyer (1993). Note how comparisons are essential for defining this field and for shaping and legitimating members' decisions (cf. Heintz 2016).

[25] See also Anderson et al. (1989, 63–64) and Anderson and Sharrock (2018).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

and accountabilities: medial, since work with instrumentation and specific data formats shapes a project's "do-ability" (Fujimura 1987); epistemic, since the point of a project is to learn something new that justifies the proposed observing time; discursive, since peer reviewers unavoidably assess them in the context of competing proposals; organizational, since the doability of large projects hinges on team expertise and workforce; and regulatory, since proposals may specify promises about the release of data and data products (see Section 6.5).

Interesting as observing proposals are as organizational documents, the point of data-centric collaboration itself is to produce documents: digital image files, tables of measurements, other forms of data for addressing specific evidential contexts (Pinch 1985), and scientific publications. Going beyond Bittner, one may wonder whether a dataset, as an object of collaboration members' shared interest, may be an organizational resource even as it is being made. For examining this possibility, it is worthwhile to remember Harold Garfinkel's notion of accountability as the self-organization of social settings by its members who "make its properties as an organized environment of practical activities detectable, countable, recordable, reportable, tell-a-story-aboutable, analyzable – in short, *accountable*" (Garfinkel 1967, 33; emphasis in original).

In a study of the collaborative work of software engineers, Graham Button and Wes Sharrock (1998, 75) argue that this notion can be adopted in contexts where organizational accountability "involves describing how the engineers organize their work so that it is recognizable to relevant parties in the project's management and implementation as work-within-the-organization." The engineers that Button and Sharrock observed were members of teams that developed elements of software meant to be subsequently combined and operated. For this to succeed they "attempted to organize ways through which they could make their work visible to each other in the very course of its production" (Button and Sharrock 1998, 79). The orderliness of these engineers' collaborative work was commonly achieved not sequentially, but in "spheres of operation and accountability" (Anderson et al. 1990, 247). We can regard organizational accountability as a part of what Anselm Strauss (1988, 164) has termed "articulation work" in projects: "the specifics of putting together tasks, task sequences, task clusters – even aligning larger units such as lines of work and subprojects – in the service of work flow." It does not presume that organizational work is formally integrated (Schmidt 2011). These notions will help us to make sense of how the MUWAGS collaboration organized its work.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

6.4 MUWAGS

MUWAGS was a collaboration of around thirty astronomers from ten countries, including tenured senior scientists, postdoctoral scholars, and PhD students. Its primary aim was to make and use a multiwavelength dataset to study the distribution of dark matter in a supercluster of galaxies and examine how its environmental conditions control the evolution of its member galaxies.

Active from 2004 to 2009, MUWAGS grew out of an optical multicolor imaging dataset that the MAMBO team had prepared (cf. Chapter 3). Made with the 2.2-meter telescope at La Silla Observatory (Chile) it provided classifications and photometric redshifts (measures of cosmic distance) for more than ten thousand objects in the field of the galaxy supercluster A2713. This then-unique dataset was a strong reason to award the MUWAGS team 80 orbits of observing time on the HST to make a mosaic of optical high-resolution images with its Advanced Camera for Surveys (ACS) that would allow detailed morphological studies of cluster galaxies. Once these observations were approved, the team successfully applied for deep mid-infrared imaging using the Multi-Band Imaging Photometer (MIPS) on the NASA Spitzer Space Telescope. These would give insights into the cluster galaxies' star formation rates and masses. Thus, the collaboration comprised three major sub-teams in charge of making three constituent datasets that were to be made consistent in a public data release. Additional data that the team acquired at x-ray, optical, and radio wavelengths were less central to its analyses.

By supplementing MAMBO data with an HST mosaic, MUWAGS was a follow-up project of SAMGES (Spectra and Morphologies for Galaxy Evolution Studies), which had already combined MAMBO data with an HST ACS mosaic of another deep field. Junior scientists who had done their PhDs in SAMGES assumed more senior positions in MUWAGS. Some now supervised PhD thesis projects themselves. Like SAMGES, MUWAGS was led by a principal investigator, but while SAMGES was hierarchical (with a principal investigator and two full professors at its head), decision-making in MUWAGS was more consensual. In part this was due to differences in funding. The principal investigator of SAMGES had funds to recruit PhD students and postdoctoral scholars to work on science full-time, whereas MUWAGS members described their collaboration as voluntary, lacking funding for salaries.[26] Most work of MUWAGS was

[26] This voluntary nature became critical when personal circumstances intervened in the life of a team member who was meant to do the basic calibrations of HST dataset, thus delaying teamwork for several months.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

done in three regional clusters in Exeter (UK), Heidelberg (Germany), and Florida (USA), with individual members at Cambridge (UK), Portsmouth (UK), Bengaluru (India), Nanjing (China), Hamilton (Canada), Vienna (Austria), St. Louis (USA), and Göttingen (Germany).[27]

Senior SAMGES members had participated in the SDSS and GOODS projects. This inspired the MUWAGS management strategies, including the special protection granted to PhD projects. The broad outline of MUWAGS teamwork followed what the HST observing proposal had described, but it was refined at the first collaboration meeting following the completion of HST observations. A set of photographs of whiteboards, containing team members' agreements as reached in copresence, became a reference for the team's organizing and its publication policy. This was supplemented by agreements reached at teleconferences, as recorded in the principal investigator's meeting minutes. MUWAGS followed the SAMGES style of having annual or semi-annual in-person collaboration meetings as well as monthly or biweekly all-team teleconferences, apart from communications by email, telephone, video conferences, in-person meetings in sub-teams, and mutual visits of individual team members.

MUWAGS members consistently described their collaboration as a "happy family," "one friendly big family," a "group of friends," or a "gang."[28] When collaboration members became parents, they circulated photographs of their newborn children by email attachment to the team and these children were welcomed as new MUWAGS family members. Over the two years of my ethnography, I did not witness any significant conflict within the team but noticed a consistent orientation to mutual aid of work on the dataset. Abundant joking and common uses of irony marked conversations at team meetings (cf. Chapters 7 and 8).[29] When reviewing teamwork at a meeting four years into the project, Mallory, the principal investigator, remarked that "we're <u>still</u> a functioning and friendly collaboration and I think that's <u>also</u> an important thing. That's a <u>fairly</u> … Götz … that's a fairly <u>rare</u> thing!" This utterance prompted collective laughter and confirmatory nodding. It seemed to me that Mallory had a point with this

[27] These place names are pseudonyms.

[28] Occasionally these metaphors were invoked to mark a difference with other collaborations, that were deemed "armies," for example, or, ironically, "a group of generous people," when alluding to another collaboration's reputation of not sharing its data in a timely fashion.

[29] One could describe team members' shared exclusive access to data during the proprietary period, and perhaps their work on a shared dataset more generally, as a "gift of time" (Derrida 1992, 41) in which friendship is manifested and renewed. This includes a shared commitment to keeping secrets. On secrecy and friendship, see Simmel (1906, 464), who remarked that "what is withheld from the many appears to have a special value."

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

assessment. During the course of my ethnography I kept hearing gossip about conflict in other collaborations, such as fights over rights to use data for specific projects, competing interests, and complaints about the inadequate sharing of data between sub-teams or with other teams, often with the moral implication of not living up to the obligation of sharing what they had received as a gift from a public facility (cf. Chapter 7).

The shared experience of comfort in collaborating and the enjoyment of team meetings raised concerns among some MUWAGS members that their scientific analyses proceeded too slowly and that other researchers may use the team's data – openly available in unprocessed form one year after the observations – to scoop it on key science. This worry became acute when a research team based in Munich (Germany), known for its expertise in weak gravitational lensing, downloaded the MUWAGS HST mosaic, presumably for just such an analysis. Learning about this made Christina, the MUWAGS gravitational lensing expert, speed up her analysis and swiftly submit it to a journal (cf. Chapter 7). Concerns over being scooped flared up again when the team was about to release its core dataset, including the catalog. MUWAGS members recognized that it could be used to address several "low-hanging fruits"– projects that other astronomers could do in a "quick and dirty" way before MUWAGS members, stuck with their meticulous, time-consuming work practices, had time to do so themselves. Despite these concerns, team members decided not to delay the data release but use its publication to push themselves to speed up with their analyses before others did them.

Four and a half years after MUWAGS submitted its observing proposal and half a year after its data release, the principal investigator officially dissolved the team at its last collaboration meeting. By then ten journal articles and several conference proceedings had been submitted and more articles were still in progress. A core group around the principal investigator continued to work on the MUWAGS dataset, but after the team's official dissolution, the authorship of resulting papers became more selective. Two years after the dissolution of MUWAGS, this group began a follow-up project, now centered around a new spectroscopic survey of A2713, which soon chose a new name and acronym of its own.

6.5 The Proposal as Charter

In its members' narratives, MUWAGS began with a plan to develop an HST observing proposal. Collaboratively authored and submitted to peer-staffed time allocation committees, observing proposals can be curious

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

objects. They list the names of principal applicants and co-applicants, make the case for using a telescope to address important questions or problems, specify the planned observations (including celestial coordinates, exposure times, etc.), and demonstrate the planned observations' "do-ability" (Fujimura 1987) and the applicants' capacity to analyze the data and publish results. Proposals sketch the planned data analysis as well as an applicant team's division of labor. Including prominent scientists as co-applicants can be read as demonstrating one's political endorsements, particularly when a team applies for considerable amounts of telescope time at public facilities that are in great demand. By defining medial, epistemic, discursive, organizational, and regulatory conditions and accountabilities, proposals can become data-centric collaborations' foundational documents – charters, so to speaks.[30] The successful SAMGES and MUWAGS HST proposals inform the following account.

Calls for proposals that observatories issue annually are common catalysts for new projects and collaborations. When the deadline to submit proposals for the next HST or JWST observing cycle is announced, many astronomers wonder: "What do we do for the next round of HST/JWST proposals?" Strong science cases for galaxy evolution surveys are often those that enable statistical explorations of regions in parameter space that promise important new insights. This may be possible by adding observations to existing datasets, thus extending the range of question that can be addressed. Both SAMGES and MUWAGS succeeded in this way by dwelling on the MAMBO team's set of photometric redshifts and spectral energy distributions of more than ten thousand galaxies in each field. (As mentioned, MAMBO members formed a core group of both SAMGES and MUWAGS.) In the early 2000s, such large samples of distant galaxies were unprecedented. The MUWAGS proposal also lists SAMGES as a complementary dataset. The proposed optical high-resolution images taken with HST's ACS would enable the team to link the galaxies' visible shapes to the physical parameters that MAMBO had measured already.

As with grant proposals submitted elsewhere, successful HST observing proposals sketch and promote a team's anticipated division of labor and its capability to do the proposed work. It lists the applicants' skillsets and their relevance for data analyses. For SAMGES and MUWAGS, these

[30] Survey description papers (like York et al. (2000) for the SDSS) can have a similar function, but they are usually published far into a collaboration's work.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

included skills in reducing and calibrating HST images, automated object detection, modeling the light profiles of objects, and, of course, expertise in addressing various evidential contexts from tracing galaxy mergers to mapping cosmic dark matter. Leading team members had a track record of approved HST proposals and timely published results.

But to call a proposal a charter, as I suggest, it must also define rights and obligations. The SAMGES and MUWAGS teams used their proposals in this sense. Among the many documents that their members produced and worked with, the proposal stood out as a resource for organizing teamwork and reflecting on the scientific progress made. It marked the collaboration's boundary, defining its inside and, by implications, its outside. Those listed as co-investigators had a right to use the data and be coauthors of all resulting publications. This included co-applicants who ended up not contributing what the proposal had outlined. The SAMGES proposal exhibits a promissory character also in respect to outsiders by highlighting the team's decision to waive its period of exclusive proprietary access to the HST ACS raw data, making it immediately available to the public. Unable to work as swiftly as the SAMGES team, in which several team members were employed to work on the dataset full-time, the MUWAGS team decided not to do so.

6.6 Four Episodes in Organizing Data-Centric Collaboration

I have suggested considering data-centric collaborations as organizational experiments for the production and use of scientific datasets. For MUWAGS members, the proposal, photographs of whiteboards, and meeting minutes were resources for organizing their research. But these documents do not exhaust what they could, and did, use toward this end. In the following, I examine further resources, beginning with two episodes in the work of MUWAGS. At one moment the team used realist assumptions and mundane reasoning to settle a tension in its division of labor. At another, the team used its internal diversity of expertise to make their master catalog coherent. Then I move to smaller scales of social organization and consider friendship as a resource for mutual aid in accessing data. Eventually, I examine public hack weeks as venues for individual researchers to team up and learn how to use complex public datasets. In sum, I will have sketched elements of a continuum of size and formality in organizing research with large datasets.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

6.6.1 *The Projected Coherence of a Dataset as an Organizational Resource*

Along with organizational plans and documents, medial and epistemic orders can be resources for organizing collaborative work. What data are like (e.g., additive digital records) defines what a team can aspire to do (e.g., making highly sensitive records of deep fields), whereas scientists' shared understanding of the electromagnetic spectrum shapes many teams' division of labor, involving members of distinct "instrumental communities" (Mody 2011). But dividing labor for making a singular data product can lead to tensions in planning that require "articulation work" (Strauss 1988) for their resolution. The joint objective to make a coherent dataset can be one of its resources.

Much of the collaborative work in MUWAGS was dedicated to making a master catalog: a table of measured and estimated properties of galaxies detected in the A2713 field based on the team's ground-based optical data, the HST imaging mosaic, and the Spitzer MIPS infrared observations. This catalog was to be a resource for most team members' research projects.[31] As MUWAGS team members worked toward a consistent catalog, occasionally mutual dependencies between the three main constituent data subsets became a challenge. Some parts of the dataset did not depend on each other, such as measurements of galaxy images in the HST images and the Spitzer infrared photometry. But all absolute measures (galaxy luminosities, stellar masses, and star formation rates) depended on knowledge of galaxy distances, which the team derived from MAMBO photometric redshifts in conjunction with a cosmological model.

In March 2007, the detection of a mistake in the code that had been used to estimate the old (2003) photometric redshifts instigated the making of a revised draft MAMBO/HST catalog, called the J2007c catalog. It was circulated internally in late July 2007. Peter explained, in an email to the team,

> I have removed a bug in the photo-z software, which should make only a difference for rather faint objects. I also enlarged the redshift window considered – again, matters only for low-S/N objects. Should not matter for R<22 objects really much.

[31] Only Christina, the team's expert on weak gravitational lensing, needed a separate catalog for her work. In Chapter 8, I give a chronological account of how the MUWAGS team prepared its catalog for public release, focusing on a series of critical decisions.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

Here "photo-z" stands for photometric redshift, the "redshift window" pertains to the (Bayesian) probabilistic approach used, "S/N" for signal-to-noise ratio, and "R<22 objects" designates a magnitude range that includes relatively bright objects in the field which were targets for most of the team's research on the galaxy cluster.

Peter's apparently minor correction of a bug in computer code meant trouble for another sub-team. Two days after Peter's message, Eddie, the infrared sub-team's head, responded to the team in an email with the subject line "doh!," expressing his exasperation that

> My ((*galaxy*)) masses, SFRs ((*star formation rates*)), etc ALL USE THE OLD REDSHIFTS. I.e. NOT the J2007c redshifts. Worse than that, I do not have the code which does the masses, so the timescale to make new masses is LONG. This is a big deal – it's a major SNAFU to change photoz version 1/2 way through a project (…)[32]

Eddie continued his message wondering if revised masses and values could be "piggybacked" somehow from other parameters listed in the catalog.

Three days later, and after conferring with Peter, Mallory, the team's principal investigator, circulated her assessment of the situation in an email. She noticed that the revised photometric redshifts did not change the results that the team had already published. In this respect, the revised redshifts were at most "mildly irritating." However, Mallory agreed that the implications for the computed values of galaxy masses and star formation rates appeared to be serious. There was no straightforward way to "piggyback" them. Furthermore, the new catalog's completeness would have to be reassessed, implying considerable additional work for Peter.

Following consultations with the team members involved, Mallory decided to return to the original 2003 redshifts, deeming Peter's effort to correct the redshifts "worthwhile but ultimately (…) not enough of an improvement to justify the effort involved to bring everything else to the same system." The J2007c catalog was abandoned and replaced by the J2007d catalog, which listed the redshifts of the J2007b catalog, but was refined in other ways.

This episode suggests that preparing the MUWAGS catalog gave progressively less and less room for reprocessing its constituent datasets. Improvements to one of these could affect others and endanger the catalog's coherence. Thus, at this late stage the catalog could not be completed

[32] SNAFU: "Situation Normal: All Fucked Up," cf. en.wikipedia.org/wiki/SNAFU (accessed July 7, 2025).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

through a linear sequence of work. Rather, Mallory's coordination was attentive to what may be called a "sphere of accountability" (Anderson et al. 1990, 248) comprising the work of distinct MUWAGS sub-teams. Brokering an agreement with Peter and Eddie, Mallory was doing articulation work (Strauss 1988). It shows how, in this case, organizational accountability was an accountability to the catalog's coherence, its lack of apparent contradictions. By tying her assessment of the current catalog version to its agreement with the team's published work, Mallory exercised mundane reasoning and adopted a realist stance (cf. Chapters 3 and 5). Demanding the catalog be coherent and consistent with the team's earlier work is, here, a resource for the reflexive reproduction of a collaboration's social order.

6.6.2 A Collaboration's Diversity of Expertise as a Resource for Making a Complex Dataset Coherent

The previous episode was one of several retrospective efforts to maintain the MUWAGS catalog's coherence that followed the discovery of troubles. But the team also made prospective efforts to make its catalog coherent by engaging the team's diverse expertise.

The final MUWAGS master catalog assembled measurements in seventeen optical and one infrared wavebands and combined data made using one ground-based and two space observatories.[33] As a table of 88,000 rows and 200 columns, it was a remarkably complex object, made with a skillset that no single researcher possessed. It is practically impossible for any individual researcher to estimate photometric redshifts, process HST exposures, fit galaxy light profiles, combine mid-infrared observations with optical photometry, and estimate galaxy star formation rates.

When the team had merged the MAMBO, HST, and MIPS data into the first comprehensive MUWAGS catalog, Mallory asked its members in an email to "break" the catalog: "Please try to break it. Please recreate your earlier plots and make sure everything still works as it should." One afternoon I joined Antonio, a PhD student, in his effort to break the draft MUWAGS catalog. He began with uploading the catalog and selecting an object sample by specifying the celestial coordinates (Right Ascension and Declination) of the cluster field, explaining to me that "it is a good thing first to check RA and Dec." Antonio then plotted the positions of objects

[33] Additional constituent datasets, including optical spectra, ultraviolet, and x-ray images, as well as radio maps, were not used for the master catalog.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

in the field, assessing whether the distribution of objects looked reasonable. He explained to me that a reasonable distribution was one that showed the familiar pattern of the cluster galaxies with a relatively smooth, seemingly random distribution of objects in the cluster's background. Finding this plot acceptable, Antonio next recreated plots pertaining to his own project that he had made using the older draft catalog. The new plots looked almost identical. He concluded this work after about two hours, telling me: "I would say that in general the catalog is right," and communicated this assessment in an email to Mallory.

Two months later, at a collaboration meeting, a "catalog-breaking session" was held. At this session, team members did not reconsider plots and scientific results, but the proper assignment of sample selection and quality flags. These are numerical or textual descriptors that help users to select samples from the catalog (such as "All galaxies in A2713" or "All galaxies detected in the infrared, but not in the optical") and to recognize potential issues with its entries ("Object contains one or more saturated pixels," "Object was too close to edge of frame").[34] The catalog-breaking session turned makers into users of their collective work. Their task was to impersonate fictional future users and approach the catalog "from the outside," as users would engage it. In doing so, team members' limited mutual familiarity with the measurements of sub-teams to which they did not belong became a resource for examining the catalog's coherence. Finding contradictory entries or inconsistent quality flags and sample selection flags are examples of "catalog breakings." A few additional mistakes were found and the specifications of sample selection flags corrected. When the possibilities of breaking it seemed exhausted, the catalog was declared completed. It was "frozen in," as Mallory put it.

Mallory's call to attempt breaking the catalog was a call for testing its usability. Attempts to "break" and "freeze" the catalog figuratively assert its achieved closure and "hardness," its materiality as an object, a singular product of teamwork. With each team member recreating their plots, they assessed the catalog's use for examining diverse evidential contexts (Pinch 1985). As such, the catalog became irreducible to any individual member's work. Team members' expertise on these topics was a fair sample of expertise in the community. That the final master catalog was usable for these diverse projects without apparent logical contradictions, delivering results that peer reviewers trusted, is what arguably made up its perceived

[34] An exemplary list of flags used in the SDSS EDR catalog is table 9 of Stoughton et al. (2002).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

hardness.[35] Team members' expectation of working toward a catalog that would have these properties was driven by mundane reasoning (Pollner 1987) and their "realist strategy" (Barnes et al. 1996, 81). Their shared acceptance that a coherent catalog would result became an organizational resource for the team. As "catalog breaking" included all the team's projects on equal footing, it mirrored the team's flat hierarchy (cf. Rooksby 2009).

6.6.3 Mutual Aid and Relational Work in Accessing Data

Collaborations like SAMGES and MUWAGS are not facility builders, but groups of facility users with specific goals, rights, and duties, and defined membership, that are, within limits, autonomous in their decision-making. These features define a formal organization. But organizing data access matters beyond such an organizational form. In astronomy's diverse ecology of data production, many facilities accept observing proposals from researchers anywhere, but others restrict access to scientists working at institutions in member states. Networks of friends and ties of academic genealogy provide diverse ways of "mutual aid" to navigate such exclusions and organize access to data, assemble the expertise to analyze it, and establish coauthorship.

Let us consider Marcelo, during my fieldwork a postdoctoral scholar at the Heidelberg Institute. Hired to work mainly for COSMOS, a large collaboration, following his PhD at the University of Cambridge (UK), he also continued observational research on active galactic nuclei at multiple wavelengths (optical, infrared, millimeter radio waves) that he had begun in his PhD work. For him, this meant collaborating with former fellow graduate students, postdocs, and scientists in his supervisor's academic network. Marcelo explains:

> *Transcript 6.1*
> I guess I work in two groups. Here I work with COSMOS … and then there are the people that I worked with in Cambridge. Well … some of them are still in Cambridge … most of them are in the UK and some are in the US. (…) The way we've worked until now with the people from the UK … we write proposals … normally relatively small … in the sense of maybe one or two nights … maybe a whole week … but not months … like … They're not as big as things like COSMOS. And then …

[35] This understanding of hardness resembles Wittgenstein's (1978, 84, 2009 [1953], §97) notion of the "hardness of the logical must." For more on the catalog's fixation and assessments of its "hardness," and how a catalog can be said to fixate a team's collective knowledge, see Chapter 8 and Hoeppe (2021).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

> if someone gets the time to do these projects … but they have gaps … for example at the beginning of the night or end of the night … they come to observe their own objects … then we give them other objects to fill in. And we try and help each other with data like that.

For Marcelo mutual aid is not limited to filling available telescope time with observations for friends and colleagues. It extends also to enabling them to use telescopes which they cannot access otherwise:

> *Transcript 6.2*
> So … for example … being in Germany I have access to certain millimeter telescopes that people in the UK don't have access to … for example IRAM ((*a radio telescope in Spain*)). (…) When a colleague of mine who just finished his PhD (and) was studying a certain type of object I said: "Look … we can look at them with a millimeter ((*telescope*))." So I write the proposal because it helps if the first person is from one of the countries that pays the telescope. If you're … It kind of puts you into the normal time allocation committee. (…) But because it's his science … we will go and observe together … maybe … and then if I can reduce the data very quickly for him then I will do it for him. Or if neither of us knows how to reduce it … then one of the two would have to learn. But it's kind of his science … so I will let him lead the work … write the paper. So we try and help each other without stealing each other's work. (…) And it tends to lead to papers with maybe six or eight authors. And sometimes they help you and sometimes … So it's a kind of mutual help that works very well … but in a very informal way.

Marcelo's small, informal collaborations differ from teams like SAMGES and MUWAGS in lacking defined membership, formal organization, name, and acronym. Marcelo's teams may use observing proposals as charters that state rights and duties, but their participants are not in contractual relations, there is no website that list team members, and their structures do not mirror those of other teams.

A principal reason for Marcelo and his colleagues to collaborate resembles the now familiar incentive to form larger teams: the need to assemble a diverse set of technical expertise to process complex datasets. As Marcelo explains:

> *Transcript 6.3*
> If you go and use many bands … this is the thing with photometric redshifts or with multi-wavelength catalogs … one person cannot reduce twelve bands of data … right? You need teams to observe it … teams to reduce it. They are different … optical data … are different from mid-infrared. So you need different expertise. And that will inevitably … the more bands … the more hours you used … the more people you will end up with in your paper.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

That each article's author list differs alerts us to the inherent flexibility of this kind of mutual aid.[36] Its participants act in settings in which the rights and obligations characteristic of larger collaborations do not apply, enabling shorter-lived and often fluid forms of collaboration. These do not have formal contractual resources for holding others to account because they are based on different social relations – of friendship and academic genealogy – whose social accountabilities work differently. Human work lives are the characteristic timescale of these ties.[37]

6.6.4 Hack Weeks: Exploring Complex Datasets Collaboratively

Beginning with the Early Data Release (EDR) of the SDSS in 2002 (Stoughton et al. 2002), several facility-building collaborations have released large datasets to the public, including the European Space Agency's Gaia spacecraft's catalogs of the positions, proper motions, and spectra of more than a billion stars and galaxies.[38] In principle, such datasets allow individual researchers and small teams to do "cutting-edge" science. But because these datasets are so complex they can be hard to understand for those who did not make them. Using them properly can seem forbidding.

Hack weeks offer a remedy. In the words of regular organizers, hack weeks "combine structured, tutorial-style instruction with open-ended project work, providing opportunities for peer learning, networking, and building collaborations" (Huppenkothen et al. 2018, 8873). Groups of participants are recruited through open calls for meetings that never last more than five days. Hack weeks often follow the release of new large, public datasets. Participants are invited to bring and use their own code and data, provided they are willing to share access to them. Hack weeks are thus a social form contingent on openness, inclusiveness, and the abundance of data.

In May 2023, I joined a group of 35 astronomers at the "GaiaXPloration Workshop," a hack week at the Institute for Astronomy in Cambridge (UK) designed to train astronomers in using a recently released large and complex dataset that the Gaia Data Processing and Analysis

[36] Sophisticated practices of relationship-building and mutual aid have been documented among resource-poor actors in a variety of fields (cf. Collins et al. 2010).

[37] Several large data-centric collaborations in astronomy have developed from the friendship of actors in influential organizational roles. Thus, "old boys'" networks (McDonald 2011) certainly matter in this field.

[38] See Figure 4.4 for a representation of early Gaia measurements.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

Consortium (DPAC), a team of around 450 scientists and engineers, had produced. In June 2022, the DPAC had published the third Gaia data release, which included 220 million low-resolution spectra of stars and galaxies measured with the spacecraft's two spectrophotometers (Gaia Collaboration 2023). Due to Gaia's idiosyncratic observing and data-processing modes, these so-called XP spectra were not published in the conventional format as wavelength-dependent radiation flux densities (such as the continuous curve in Figure 3.2), but as a set of fifty-five parameters, coefficients of a mathematical (Hermite) function. This made their use nonintuitive for many potential users. Along with these data, the Gaia DPAC published GaiaXPy, a Python software package for calibration, transforming the coefficients into the conventional format and computing so-called synthetic colors in various wavebands – a means to compare the Gaia measurements with those of other projects. In their workshop announcement, the organizers wrote that "[o]ur goal is to bring together the collaboration that made this data product with the community that benefits from it, to create new scientific opportunities, by encouraging brainstorming and collective work between DPAC members and Gaia data users."[39]

Three days of morning lectures and afternoon work sessions were followed by two days of hacking. The workshop began with a round of introductions. Each participant had to contribute a "pitch slide," rehearsing "Who I am," "What I want to learn," and "What I want to do" with the Gaia XP spectra. Lectures included Gaia DPAC members explaining how they had processed the XP spectra and how to use GaiaXPy. Researchers gave brief presentations on their current work and how it may benefit from the XP spectra. At lunch and in the afternoon work sessions, participants met to discuss, collaborate, and explore selected problems and opportunities of data and code. This continued over the days of the workshop, each day ending with a daily round of wrap-up reports including all participants.

The workshop concluded with a more formal wrap-up session at which each participant presented a slide on "What I learned," "What I did," and "What I plan to do next." It demonstrated participants' enthusiasm for the data (e.g., "XP is much richer than I thought!," "OMG there is still so much to do with and learn from XP!") and their praise for the Gaia DPAC and its work. Responses to "What I did" revealed a variety of relations, new and preexisting, among participants. At the hack week, most had

[39] See ast.cam.ac.uk/events/gaia-xploration-discovery-and-measurement-low-resolution-spectroscopy (accessed July 7, 2025).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

worked with at least two other participants. They wrote: "Classification of WD [white dwarf star] samples (w/ Brigitte Nigellu)," "Devised stress tests for XP alpha measurements (w/ Zhuangzi Li),"[40] "Measured velocity dispersions in the ancient bulge (w/ Lilo Watkins),"[41] and "Quizzed the DPAC team members." On their "What I plan to do next" slide, several participants announced their intention to continue what they had begun together at the hack week.

Compared to the lifetimes of long-lasting formal collaborations and ties of friendship, this hack week was fleetingly short, but it gathered scientists with preexisting relations and ended with new connections. Whereas the MUWAGS team used its internal diversity of expertise in trying to "break" its catalog and assuring its coherence, at the hack week it was the diversity of participants' skills and experiences that helped them to discover new uses of the XP spectra and made their projects with this large and complex public dataset doable.

6.7 Conclusion

This chapter has defined and described a field of research with large datasets. Rather than focusing on types of organization, I have probed into organizing and its resources.[42] Many data-centric collaborations have formal resources to organize their collaborative work, but I took a broader view of their available resources, following the maxim that scientists, like all people, use whatever is available to them to organize their affairs (Liberman 2013). These resources include medial orders (such as the formatting of digital data) and epistemic orders (such as shared understandings of cosmology and the electromagnetic spectrum), which scientists engage through realist practices (cf. also Chapters 3 and 5). More than that, the MUWAGS team used its catalog-in-the-making reflexively as an organizational resource. Many of these resources are unavoidably discipline-specific, but ethnographers examining other disciplines can identify them by probing into the resources that realist practices engage.[43]

[40] In this context, "alpha" (α) is a measure of the abundance of chemical elements formed in stars by alpha process nucleosynthesis, including oxygen, neon, and magnesium.

[41] Work pertaining to the structure and dynamics of the inner Milky Way Galaxy.

[42] See Shrum et al. (2007) for a typology of scientific collaborations and Paine and Lee (2021) for a typology of data-intensive collaborations.

[43] Ankeny and Leonelli's (2016) notion of "repertoires" is useful to conceive of disciplinary-specific resources, but Ankeny and Leonelli do not address their role in uses of accounting practices and the ordering of work.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

Organizing data-centric collaboration means to establish, maintain, and care for social relationships. As organizational experiments, such collaborations often begin with ties of academic genealogy (supervisor–student, fellow graduate students) and friendship (relations among peers who have established mutual trust).[44] But data-centric collaborations also operate, and organize their work, in view of their obligations. Producing large datasets with public, tax-financed facilities obliges teams to share their data and offer data products that others can use. Observing proposals for large programs commonly specify such commitments, like waiving one's period of exclusive proprietary access. Thus, there is a promissory element reminiscent of what anthropologists have observed in gift exchange (Mauss 2016 [1923–24]; see Chapter 7). The perspective of both relationships and obligations is "internalist," that is, scientists locate initiatives, actions, and accountabilities within their professional world, typically within their discipline or subfield.

Most data-centric collaborations in astronomy begin with developing an observing proposal. Then a team exists before its dataset. But in the novel "ecology" of large public datasets, this order may be reversed. Smaller and shorter-lived teams may form to use such data. In 2007 – at the beginning of my work on this project, when many astronomers who still worked in small teams noticed the rising dominance of large collaborations –, Alexander Szalay, an astronomy professor at Johns Hopkins University and one of the founders of the "Virtual Observatory," gave a talk in which he argued that large databases would be the only way to work in small teams in the 2020s.[45] When I attended the GaiaXPloration workshop in 2023, this assessment rang true to me. Working with large databases like the 220 million Gaia XP spectra calls for novel forms of data-centric sociality that events like hack weeks enable.

Hack weeks may be the latest in a history of precedents that have shaped astronomers' work with large datasets since 1995, when Robert Williams and his team presented the HDF, followed by the SDSS in 2002. In setting precedents of data sharing with the HDF and the SDSS,

[44] Note that joint graduate training and early career interactions are rich in opportunities for demonstrating competence, honesty, and reliability – markers of trustworthiness that inspire well-placed trust (O'Neill 2014).

[45] Heidelberg Astronomical Colloquium, May 22, 2007. The Virtual Observatory was originally conceived to make all existing astronomical dataset available through one portal (Szalay and Gray 2001) but has evolved into a set of standards that a variety of data centers uses. See ivoa.net (accessed April 26, 2025).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008

Robert Williams and James Gunn were both late in their careers and operated from influential, central positions in academic astronomy. Williams aimed to change astronomy's "culture" and Gunn its "sociology." But how should other astronomers act properly in either? Facing expectations of the open access to one's data makes normativity a practical problem for scientists that I examine in Chapter 7.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 16 Nov 2026 at 17:24:51, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009686754.008