# Closing Reflections

## LOOKING AHEAD

# 19 The Next Generation of Challenges in the Curation of Scholarly Data

CLIFFORD LYNCH

## INTRODUCTION

Requirements for data curation are now well established across a wide range of scholarly disciplines, but particularly in the sciences and some social sciences, through a series of funder requirements for data management plans and policies mandating public access to large classes of research data. Institutional policies and journal editorial policies surrounding the management and availability of research data support and complement the funder-driven initiatives. The need for effective and affordable research data management services will only grow over the next decade.

The previous chapters of this book have covered the evolution of the policy environment and discussed some of the technical issues surrounding data curation. One of the most important and unusual contributions of the book is a series of case studies of pioneer and early adopter experiences in responding to the data curation challenges; in the next few years, we will see many more institutions following in the footsteps described in these leadership case studies, and designing services informed by the experiences documented here. The overarching priorities for the next few years will be to help faculty to develop credible data management plans, to appropriately document the datasets that they share and preserve, and to help them find platforms (either locally developed, through consortia or disciplinary centers, or even via commercial services) to share data and to prepurchase assured bit preservation for periods on the order of five to ten years. Until these three groups of services are in place and operating effectively and at

scale, many other challenges will have to wait or will be dealt with only on an ad-hoc basis when absolutely unavoidable.

And there is also a *new* group of challenges waiting in the wings, appearing as a result of the initial success in meeting these first three most urgent and most basic needs.

My purpose in this concluding chapter is to sketch a number of what I believe will be the key next-generation challenges. These are challenges that will ultimately need systematic engagement, and generally better sooner than later. Most of these are not very well understood at this point, and experience with them is very limited; however, it is certainly not too soon to put them on the strategic planning agenda for data curation work, to begin thinking about how to approach them, and in some cases, to start building experimental or prototype services. As with so many aspects of data curation, pure research is of limited value; it is necessary to actually build and deploy attempts at genuine operational services, working with real research data and real researchers, in order to make meaningful progress.

While some of my examples and specifics have been drawn almost exclusively from the United States, I believe that we will see very similar issues emerge in other national settings. Perhaps the area of greatest variation will be in the conflicts surrounding data that involves human subjects, where divergent national policies involving health care delivery—and thus health care records, privacy, and similar issues—may result in quite different outcomes from nation to nation (and thus, perhaps, even more formidable obstacles for sharing and reuse of such data *across* national borders).

## MOTIVATIONS AND DRIVERS FOR THE DEPLOYMENT OF DATA CURATION SERVICES

Before looking at these new challenges, it is useful to summarize the forces that are driving the various players—funders, scholars, the institutions that host these scholars, and journal editors—in their current actions related to research data. In general, these players are not calling for data curation, preservation, and sharing because it is abstractly the right thing to do as part of the creation, dissemination, and stewardship of knowledge; their motivations are much more specific and pragmatic. I do not believe that this array of driving forces will change significantly, at least over the next

ten years, so understanding them is essential to situating both current developments and the next generation of challenges.

Funders, and particularly public funders, are under great pressure to show how their funding contributes to broad economic growth, how it addresses the needs of society, and to demonstrate that the requirements that they impose on the work they fund makes discovery ever more rapid, extensive, and cost-effective. From this perspective, they are not interested in data preservation or even data sharing other than as a necessary precondition to data reuse; they are interested in conformance to their data management and sharing policies because it is the only way they can create the preconditions for data reuse. They are hungry for examples of how data reuse has improved the processes of scholarship and discovery, or contributed to economic growth, job creation, control of health care costs, or public policy.

While research libraries and other memory organizations do, I believe, have a deep and genuine mission in data stewardship as part of their commitment to managing the intellectual and cultural record and its underlying evidentiary base for the long term, at the broader level of research universities institutionally, the greatest pragmatic and operational interest is in ensuring conformance to funder requirements and managing institutional risk and liability. They certainly will provide some funding support for the long-term stewardship work of their memory organizations. They welcome improvements in the processes of research and scholarship, but usually they rely on the faculty to drive such improvements.

The vast majority of faculty will, at least in the near term, see little real benefit from making their data available for sharing. Despite work on data citation practices and on changing evaluation criteria for researchers, it will take a long time for faculty contributions of data for potential community reuse to make a compelling and widespread difference in tenure and promotion cases; the inertia and conservatism in this system is enormous. So developing and subsequently implementing data management plans will most often be viewed as just one more burden imposed by the funding agencies; faculty will want to satisfy these new requirements in the most time-efficient and easiest fashion. Some faculty (we don't know how many, or in what disciplines) will be very creative in exploiting the growing amounts of data available for reuse and will find their own scholarly work advanced. There will, of course, be some high-profile cases where faculty

who obtain important new results through data reuse gain important recognition (keep in mind that the funders are eager to identify, encourage, and recognize these scholars). Even researchers who provide data that is subsequently reused to significant effect may find their contributions honored—but there's a sizeable luck factor here, as it is not so much that they make data available for possible reuse as it is that they were lucky enough to have someone actually reuse it and then make an important discovery.

In a significant number of scientific disciplines, there is a growing crisis of reproducibility. With increasing frequency, papers report results that cannot be reproduced by other researchers. This is not new, and there are many reasons for it, not all of them sinister: inadequately documented methodologies; honest errors, sloppy work, or simply an incomplete understanding of new phenomena that are being reported and their causes (often compounded by a rush to publish); unavailability of data, tools, and/or materials to other researchers seeking to reproduce the work; and outright fraud and fabrication of data. As funding continues to decline and the number of researchers competing for funding (and tenure and promotion) continues to grow, this establishes a hypercompetitive environment that puts greater pressure on reproducibility. This is of great concern to all players—scholars, journals (doing more agressive and adversarial refereeing in response to a growing number of deceptive submissions and retractions), institutions, and funders. This crisis of reproducibility is starting to surface more frequently in political settings and broad public fora, and carries with it a very real risk of eroding public support for science and for scientific research. One easily can see a future where funders and institutions, assisted by journals and many individual scholars, introduce increasingly heavy-handed policies to root out irreproducible research; the retention of data and the sharing of data (perhaps as part of the refereeing process, but certainly effective as of publication) will be important elements here.

A final point on reproducibility: in most cases, it is a fairly short-term problem. Other researchers will try to reproduce results soon after their publication, and much of the practical thinking about reproducibility focuses on a relatively short time window, say five years or so. It is both very costly and very difficult (due to changes in experimental technology and methodology) to think in terms of reproducing a 50-year-old result, particularly without some fundamental rethinking about exactly what one is trying to reproduce.

## THE NEW DATA CURATION CHALLENGES

### Software

In a substantial number of cases, the interpretation and analysis of data is deeply intertwined with the availability of specialized software. Both the level of interdependence between software and data and the level of complexity of the software vary greatly. There is at least some reason to believe that software "decays" more rapidly than data, and it will require more frequent and more costly interventions to ensure that it continues to be useable over time (though there are promising developments in areas such as virtualization and emulation that offer some hope here, but these are certainly not a panacea).

Software is also vital when trying to reproduce published results. The good news here is the fairly short time horizon means that the software needs to be saved and made available for sharing, but it probably can be successfully maintained across the necessary time period.

It is clear that funders are going to have to develop a more holistic view of data management, and specifically address software as well as data in management plans; some of the major science funding agencies, at least in the United States, are already starting to think about this.

We will need to be able to offer researchers services that can preserve complex collections of interconnected data and software (and documentation), or simply preserve more general purpose software independent of specific datasets. In both cases, it will be essential to be clear about what it actually means to "preserve" the software in question and to understand the cost implications of various choices, particularly over a range of timescales. For example, there is a great difference between a preservation program that ensures that a given set of software is always ready to run on the most popular platform or platforms of the day, and a preservation program that simply makes it possible to launch an effort to resurrect a given set of software in future with a fairly high likelihood of success (given enough time and money). At the more demanding levels of software curation and preservation, the availability and development of the necessary skills and expertise in the workforce will be a serious problem.

### Conformance: Auditing the Promises in Data Management Plans

Today, short of a decision by a funder (or an institution) to audit conformance to a data management plan, which is most likely going to be specific to an individual contract or grant, or perhaps to the set of contracts and grants given to a specific institution, there is no way to track conformance to the promises made in a data management plan. Either such mechanisms will need to be developed or there will be reliance on occasional spot audits by funders, probably accompanied by increasingly draconian punishments in order to encourage compliance. This could shift the compliance monitoring burden at least in part to institutions (as is the case with many other funder requirements), but the problems of mechanisms and scale are conserved. Institutions will need to think very carefully about where to situate responsibility for audit and enforcement organizationally: if mishandled, it could easily poison the development of what everyone hopes will be collaborative and constructive relationships between institutional data curators and faculty researchers.

There are a few specific points in surrounding conformance that merit comment. In terms of data sharing for replication of results or reuse, there are two possible approaches. One is to say that data must be shared *upon request*, and to rely mainly on complaints from frustrated requestors to identify compliance problems. The other is to insist that data is placed in a transparent and public repository; it is then possible to just check that the data has been deposited and that the repository is being operated according to good practices. Clearly the second situation is much more tractable from the point of view of checking compliance, but if there are constraints on the data (for example, privacy constraints or a requirement that those who want to reuse the data contractually agree they will not attempt to deanonymize it) then considerations of control, accountability, and liability become complex. There are many implications here for how, why, and under what criteria we certify repositories.

A second issue is how often, and for how long, compliance needs to be verified, and what to do if there is a problem. Suppose that a data management plan promises to keep a dataset for 50 years. Is it enough to confirm that it has been deposited into a "reputable" repository that promises to keep it for 50 years, or do we have to periodically check that it is still there? Who certifies reputable repositories, and what happens if they fall upon

hard times or fail recertification? If there is a problem, who is responsible for dealing with it (particularly given that data can outlive the investigator who created it)? Implicit here is that the balance of responsibility between faculty investigators and host institutions in meeting commitments to funders, and how this balance may be shifted by time and circumstance, is going to be mapped out as part of the ongoing focus on compliance.

### Implications of Term-Limited Data Preservation Strategies: Managing Reassessment

One of the striking—and in my view overall very positive—changes in thinking about research data stewardship (and many other areas of digital preservation) over the past few years has been the move away from talking about taking a single decision and set of actions aimed at preserving data "forever" (or at least for a very long and indeterminate period of time). Instead, a stewardship organization makes a commitment to take care of a collection of data for a specific period of time—something on the order of 10 or 20 years, perhaps—after which it makes no further promises except that it will see that the collection receives a review and that it will ensure that if some other organization wants to accept responsibility for an additional period of time, it will cooperate actively in an orderly and well-thought-out transfer of the collection that will make every effort to preserve data integrity. This kind of thinking is prominent in the 2010 report of the *Blue Ribbon Task Force on Sustainable Digital Preservation and Access*, for example. This shift is driven in part by the recognition that we still have very limited experience in assessing the relative merits of various preservation choices about research data under constrained resources that say we cannot save everything (indeed, while discarding at this level is familiar to archivists in some other settings, it is relatively unfamiliar to research libraries operating as a system). Another motivation is a recognition that the uncertainties involved in very long-term commitments are not just technical; they are financial (in the sense of rates of return on funds) and organizational.

The upshot of this shift is that for data that does not fit into a disciplinary repository, it is increasingly common to find proposals to guarantee to preserve datasets (at the bit level) and make them publically accessible for a period of five or ten years (with the costs prefunded as part of the grant budget). At least by implication, and sometimes explicitly, there is

a reassessment process that will be conducted at the end of this period; if the data is viewed as of sufficient continued value, funding will be found for someone to sustain the data for an additional term (after which the process presumably repeats). Experience with the levels and purposes of data reuse during the earlier periods will help to inform the choices about whether to renew the data. This is perfectly reasonable—but there are no mechanisms in place to support this kind of periodic review and reassessment, or to gather funding other than individual institutional budgets to support ongoing stewardship.

Developing these mechanisms is going to become increasingly urgent over the next decade, and there are some very complex organizational challenges implicit in any successful approach. One is simply scale. Another is the way to balance the views of different disciplines, since the relevance and importance of a data collection to various disciplines may well shift over time. A third deals with tension between decisions that are local to a given institution and allocate institutional funds, and the need to think about research data as a shared asset and shared record that is held by the entire research and education community, nationally and internationally. A fourth challenge is to define the mechanisms and level of participation by funding agencies in the longer-term stewardship of research data. Addressing the challenges here will require both action at the institutional level, by frontline data curators and their institutional leadership, and also policy development and implementation of collaborative mechanisms and frameworks at the national and international levels.

### Understanding What Is Worth Preserving

Current trends suggest to me that over the next five or ten years we will collectively retain much more research data than we have the past. Some of this will be driven by the demands for reproducibility, but as already discussed, reproducibility typically supports only fairly short-term retention. Hopes that data will be reused are another driver, but beyond hope, we know very little in general about likelihood of reuse, or the time horizons within which that reuse is likely to occur, if it does occur. There are some classes of data where reuse is quite likely: data that is directly comparable to other data, which can be aggregated into some kind of time series or larger aggregate (for example, a set of medical records that can be combined

for greater statistical resolution of rare effects). A lot of this kind of data already goes into disciplinary databases or data repositories. Indeed, one of the powerful catalysts that funders can use to encourage data sharing and reuse is the identification of such classes of data and then the creation of databases or data repositories to facilitate aggregation and normalization.

Beyond reproducibility demands and hopes for near-term reuse, it will fall to our established stewardship organizations to allocate resources for the longer-term preservation and management of selected research data resources. The opportunities will doubtless vastly exceed available resources. The first round of these decisions will come quickly, more quickly than I think that many organizations realize, as short-term commitments funded through data management plans and associated grants expire. Institutions (individually and collectively) come to these decisions with a weak analytic framework to assist in decision making. Among the factors to be considered, and somehow balanced against each other, are: the very difficult to assess hope of reuse in future, perhaps in disciplines very distant from those that originally generated the data; the quality of the data and its documentation; the irreplaceability of many classes of observational (as opposed to experimental) data; the economic or ethical costs of regenerating experimental data (clinical trials, the use of animals, the cost of recreating experimental apparatus); and, of course, estimates of the cost of preserving specific collections of data. We will need good models, best practices, thoughtful analysis of experiences with case studies, and staff development opportunities to help with these critical decisions.

### Data Involving Human Subjects

There is an enormous emerging collision between the desires to share and reuse data, with all the benefits these practices can offer, and the very complex institutions and policies that have been established to protect the safety, privacy, and dignity of human beings who provide data to the research process. The landscape here is enormously complicated and problematic—there are very complex regulations from the Department of Health and Human Services in the United States (plus a massive set of revisions currently under review and discussion), inconsistent and sometimes idiosyncratic implementations of these regulations through local campus institutional review boards (IRBs), and a dearth of mechanisms for

facilitating multicampus research collaborations (much less international collaborations). The jurisdiction of the IRBs goes far beyond research related to medical and psychological experiments into social science surveys, and, on some campuses, oral history and other interview-based data collection.

One cornerstone concept in protecting human subjects is informed consent; this includes ensuring that potential subjects understand what data is being collected about them, how long it will be retained, who gets to use it, and an understanding of the specific uses to which it will be put (including the risks of those uses). Even if the potential subjects were willing to sign very general release forms that would facilitate sharing and reuse of data, the use of such consent forms would likely be rejected by the local IRB; at best, some specific and constrained kinds of data reuse, such as a meta-analysis, might be included in an acceptable consent agreement.

Another very problematic area here is the anonymization of data involving human subjects. For some kinds of reuse, an anonymized version of a data collection, which breaks the links between data and the individuals that provided it, is sufficient (though, of course, many other reuse scenarios will require the full data). But researchers in many fields and many contexts, from genomics to information science (query logs), have discovered that it is incredibly difficult to irrevocably anonymize data, particularly if data from multiple sources are merged together. So now we see researchers who want to reuse data being asked to certify that they will not attempt to deanonymize it; even more problematically, there may be some attempt to "qualify" the potential reusers and reuses as "legitimate" in some fashion, which quickly runs contrary to the goals of promoting broad and creative reuses, and engaging industry and the broad general public, not just the research community, in the reuse of data (and particularly data produced with public funding).

A broad and constructive conversation on the conflicts between the protection of human subjects and the advancement of scholarly work has been very difficult to advance; many scholars across the spectrum of disciplines conduct their research at the pleasure of the largely unaccountable IRB system, and thus, they are reluctant to challenge this system. There have been some recent promising beginnings, such as the National Research Council project titled "Revisions to the Common Rule for the Protection of Human Subjects in Research in the Behavioral and Social Sciences" (http://www8.

nationalacademies.org/cp/projectview.aspx?key=49500), which has recently issued a workshop report titled *Proposed Revisions to the Common Rule: Perspectives of Social and Behavioral Scientists* (http://www.nap.edu/catalog.php?record_id=18383). Additionally, there have been some very creative developments in the biomedical area (see the work of the Sage Bioinformatics Forum, http://sagebase.org, or the work of John Wilbanks on the Portable Legal Consent Framework, http://weconsent.us).

This creates many challenges for frontline data curation, beginning with the development of data management plans. Data management plans need to be synchronized with negotiations between investigators and IRBs about experimental protocols and the handling of data collected through these protocols, and often these negotiations will continue far into the actual conduct of the research funded under a given grant. As conformance is tracked more seriously, it may be necessary to develop ways to evolve and amend data management plans in light of these ongoing negotiations with IRBs. While there is often some form of support, education, and training available to investigators in meeting IRB requirements, there will be a need to provide these investigators with information and advice on how to *balance* IRB demands with the demands of funders to facilitate sharing and reuse of data. A good deal of the burden here is likely to fall on the institutional data curation staff, who will need to develop considerable expertise in these complex areas. There will also be a demand for flexible data publishing and curation platforms that can meet the IT security requirements imposed by IRBs and by other regulations such as the *Health Insurance Portability and Accountability Act* (HIPAA).

### Really Long-Term (Semantic) Preservation

Almost all of the practical research data management preservation work I am aware of has been about preserving bits across time, and ensuring that these bits are documented with sufficient metadata and other explanatory material that then can be understood and reused, today or tomorrow, by people other than those who created it. (In truth, while it may be possible to reuse data without communicating with the creator of the data, it can be perilous, particularly when there is not an active community already working with the data; access to the creator is often a great boon.) At least conceptually, bit-level preservation is fairly straightforward.

There are ideas about higher levels of preservation driven mostly by changing information technology practices and standards (e.g., moving from a proprietary format to an open standard as software evolves; updating ASCII or EBCDIC data to UNICODE; migrating from an older image format like JPEG to a newer one like JPEG 2000; converting from SGML to XML). These conversions will be much less frequent than migrating from one storage system or medium to a newer one and copying the bits over; they can be substantially more complex, however, and involve sometimes very subtle curatorial choices. Yet over extended periods of time, they are important in keeping materials meaningfully useable and interpretable. We have some experience with these types of conversions, though limited.

As we think about preserving research data across really long periods of time, however, it is clear that matters get very complex indeed. The underlying experimental methods or observational tools change as technology changes; understanding of the contexts surrounding data shift as new disciplinary paradigms emerge, and agreement on what data is actually significant and what characterizes objects or processes also changes. We have almost no experience in this area, or only very unsatisfactory partial analogies (trying to understand alchemical texts from the perspective of modern chemistry, for example). Understanding the limits of our ability to preserve, and our ability to reuse across long periods of time and the massive evolution of knowledge will be very important in making decisions about where to invest and what promises we can responsibly make to the present and the future. At the very least, it is important for our frontline data curators to inject a note of humility and caution about confidence in very long-term preservation.

## CONCLUSION

We are at the early stages of a genuine systemic and systematic response to the data stewardship challenges framed by the emergence of e-research, and to seizing the opportunities promised by more effective, broadscale data sharing and reuse. Key players in the system—notably the funders and policymakers—have made a clear commitment to addressing the issues and to forcing other players to do so as well.

Today intensive frontline institutional research data curation efforts are underway to respond rapidly to the most basic needs: documenting data man-

agement plans, setting up data documentation, bit preservation, and data publishing services. Some leadership institutions now have relatively advanced, robust, and comprehensive services in place; many others are following, with initial services either deployed or in the advanced planning stages.

I have not devoted much attention here to these three basic services, which are extensively covered elsewhere in this volume; however, it is important to emphasize that while they are reasonably clear conceptually, as the scale and depth of experience increases, some very critical operational issues are going to emerge. For example, data repositories are going to emerge as very attractive, high-value targets on an increasingly hostile Internet. Further, the security problems that we tend to emphasize here (because of incidents in other contexts) are data breaches: some attacker obtains access to data that was not intended to be generally available. But probably of greater concern in the data curation context is outright destruction or, even worse, deliberate corruption (perhaps quiet, unannounced, and subtle) of research data, potentially calling results and reputations into question through problems with replication, or leading to chains of erroneous conclusions or pointless investigations as data is reused and the corruption propigates.

Another area of great concern is ensuring that research data is appropriately documented to permit and facilitate reuse (which also implies discovery and assessment, but goes beyond these activities). It is easy to be glib about this, and to appeal to library and/or archival descriptive practices, which are by and large entirely insufficient to support the full cycle of reuse. We can certainly point to some real successes in documenting for reuse, ranging from social science survey data to remote sensing and geospatial data, clinical trials, or gene sequences, but this is often data that is collected with reuse in mind, and often comes out of fairly large-scale data acquisition projects. In other settings there is very little experience with data reuse; today's attempts at documentation are mostly best guesses and assumptions, unproven in actual reuse situations. And the documentation—and particularly automatic documentation (of parameters and readings from various kinds of experimental apparatus, data provenance, or computational workflows)—is still a very active research area. As we gain more experience with reuse in different domains and contexts, we will learn what documentation practices work and what is

needed to support the goal of reuse. It is essential that we feed this back into data curation best practices on a continuing basis, and that the curators and investigators who work together to document new data continually absorb these lessons.

The stewardship challenges do not stop with the three fundamental services, and there are specific and complex barriers that conflict with the goals of greatly expanded sharing and reuse. These are related to, but are not precisely the same as, stewardship challenges, and these are easy for curators to overlook unless they keep the mandate to facilitate reuse and not just preservation firmly in focus. Finally, it is clear that an enormous imbalance exists between the resources currently available to fund these efforts and the potentially almost infinite demands of a fully realized data stewardship program; a key strategy in managing this imbalance is the effective use of the *specific* policy goals, such as data reuse, as shaping and prioritizing mechanisms in shaping an overall stewardship effort.

It is my hope that this article has provided a better understanding of these emerging issues and the way they are likely to unfold over the next decade or two, and identified many of the key next-generation research challenges that are going to require attention in the not-very-distant future.