Title

Practices, Trends, and Recommendations in Technical Appendix Usage for Selected Data-Intensive Disciplines

Permalink

https://escholarship.org/uc/item/9jw4964t

Authors

Kunze, John A
Cruse, Patricia
Hu, Rachael
et al.

Publication Date

2011-01-18

PRACTICES, TRENDS, AND RECOMMENDATIONS IN TECHNICAL APPENDIX USAGE FOR SELECTED DATA-INTENSIVE DISCIPLINES

version 2011.01.18

John Kunze, Trisha Cruse, Rachael Hu, Stephen Abrams, Kirk Hastings, Catherine Mitchell, Lisa Schiff

California Digital Library[1]

[1] With generous support from the Gordon and Betty Moore Foundation

CONTENTS

1 Executive Summary

There is a need to establish a new publishing paradigm to cope with the deluge of data artifacts produced by data-intensive science, many of which are vital to data re-use and verification of published scientific conclusions. Due to the limitations of traditional publishing, most of these artifacts are not usually disseminated, cited, or preserved. These latent artifacts consist largely of datasets and data processing information that together form the foundations of the reasoned analyses that appear in the published literature. But this traditional record of science increasingly represents only the tip of the scientific iceberg.

One promising approach to this problem of data invisibility is to wrap these artifacts in the metaphor of a "data paper"[2], a somewhat unfamiliar bundle of scholarly output with a familiar facade. As envisioned, a data paper minimally consists of a cover sheet and a set of links to archived artifacts. The cover sheet contains familiar elements such as title, authors, date, abstract, and persistent identifier (e.g., a DOI or ARK) — just enough to permit basic exposure to and discovery of data by internet search engines; also just enough to build a basic data citation, to instill confidence in the identifier's stability, and to be picked up by indexing services such as Google Scholar.

This simple format represents only the first stage of the evolution of the data paper. There is room for the format to increase in complexity with the incorporation of other valuable elements, both general-purpose and discipline-specific, to enrich discovery, re-use, and archiving. An exciting potential outcome of this development of the data paper as publication is the parallel emergence of a new kind of "data journal". Like regular journals, data journals would spring up around disciplines and sub-disciplines as needed, and we could expect that some of them would also be peer-reviewed. The data journal is envisioned as an "overlay" journal; an editor would assemble an issue by selecting data papers from any number of sources and archives, and combining them with front matter, a table of contents, editorial policies, submission guidelines, etc.

This new data publishing paradigm promises to strengthen the scientific community practices of data sharing, re-use, and preservation. Scientists want to do science, get credit for it, communicate about it with their peers, and improve the measurable outputs by which their funders and employers evaluate their performance. The elements of the data paper create a recognizable and standardized form for previously unpublished data artifacts, making them easier to approach, evaluate, and automatically index for basic discovery purposes. Those same elements can easily be repurposed to create familiar-looking citations suitable for reference in CVs and all manner of publication. Finally, unique persistent identifiers for data papers and data artifacts greatly facilitate automatic discovery of a data paper's impact and re-use.

[2] This use of the term is different in detail but similar in spirit to various emerging uses of "data paper" (cf. [ESA][Rees]).

2 INTRODUCTION

We are living through a quiet revolution in the practice of science. As articulated in "The Fourth Paradigm" [Hey], over the last few millennia science has changed from a largely empirical to a theoretical endeavor, in which generalizations and modeling have been added to observation. In the last few decades, experimental models have become too complex to analyze without simulation, which has led to the development of computational science. The concepts of data-intensive research and "eScience" have emerged in the past ten years, characterized by huge amounts of data that are now both captured by instruments and generated by simulation, and that subsequently need to be processed, stored, and analyzed by computers before they can be used by scientists.

In the midst of this transformation in scientific practice, the scientific community has begun to realize that traditional journal publication is no longer adequate by itself to represent the scientific record.

> *"[T]he published literature is just the tip of the data iceberg. … [P]eople collect a lot of data and then reduce this down to some number of column inches in Science or Nature … So what I mean by data iceberg is that there is a lot of data that is collected but not curated or published in any systematic way." [Gray]*

Researchers also increasingly recognize the deep importance of fully integrating insights and practices from computer science and information science into their work. Across all the disciplines, eScience is the place where IT (information technology) meets science [Gray]. This meeting place involves two parallel branches that are emerging within each scientific discipline: a computational branch for collecting simulation data (computational ecology, computational biology, etc.) and an informatics branch for collecting observational data (eco-informatics, bioinformatics, etc.). Achieving standardized, codified, algorithm-ready data and metadata at strategic points within the junction of these branches, both within and across disciplines, is a key challenge. Once overcome, it could provide the basis for automated processing that would simplify the remaining challenges in data capture, curation, citation, analysis, and visualization.

Another key challenge faced by the new scientific paradigm is its continued dependence on an old publishing paradigm. The primary tool that motivates and records data-intensive research is still the published scientific journal article. Scientists want to do science, get credit for it, communicate about it with their peers, and improve the measurable outputs by which their funders and employers evaluate their performance. Published journal articles have traditionally supported these goals, but there is growing concern in the scientific community that the many artifacts generated by data-intensive research — data, techniques, software, etc. — are at risk of being neither saved nor shared. Moreover, motivation to expend the extra effort required to save and share these artifacts is limited by the absence of suitable rewards, such as publication or credit mechanisms.

At stake is the future of science itself. When the artifacts of this new paradigm don't enter the scientific record, published conclusions cannot be verified, challenged, corrected, or built upon. Phil

Bourne, Editor in Chief PLoS Computational Biology, has seriously questioned the possibility of scientific reproducibility from the paper alone [Bourne]. Research investments are also at risk, as data products created at great cost cannot be re-used cost-effectively without proper dissemination, documentation (e.g., metadata), and stable storage.

To meet the challenges requires involvement of many actors. At a high level these actors include research institutions (academic, government, private), publishers, data archives, and libraries. Dealing more directly with the data are data creators, including collectors and modelers, as well as data synthesizers, users, and managers.

The present study is both a cross-disciplinary investigation of and a set of recommendations on how to protect these technical artifacts, and in particular, on how to make them easier to publish, share, and archive. The work began with a short white paper [Agarwal] observing that vital details of how published scientific conclusions have been derived from research data are often not published, easily shared, or preserved. That early paper grew out of experiences discovering, accessing, understanding and re-using data from disparate sources and disciplines to create Fluxnet CO_2 synthesis datasets. Reproducibility seemed impossible without preserving details of the complex analysis and processing and motivation for this work seemed unlikely without crediting all the people and sources involved in generating the data. One emerging practice was to capture this information in what the paper called a "data application appendix", or DAX. In our study we have explored not only the DAX but also the more general context of journal article publishing.

In the world of scientific journal publishing, it is helpful to see these details as forming part of an article's "supplemental material (SM)", or material that supports a publication's main conclusions but is not itself included in the publication. In the context of publisher Elsevier's "Article of the Future" initiative, "Cell Press" identifies three broad SM classes (although the boundaries between them are not always obvious) [Cell]:

- evidence that provides deeper support for the points made in the main paper,
- large data sets and multimedia that can only be presented online, and
- detailed information about the methods.

There are many voices calling for greater publishing, sharing, and archiving of data — activities that fall into the second SM class above. According to some, in fact "data is a second-class citizen" [Brase] within the scientific record. The voices include government advisors [Atkins] and publishers [Nature].

3 RECOMMENDATIONS

Our study identified a number of unmet needs for sharing and preserving the artifacts of data-intensive research.

3.1 CREATE ONE SOLUTION FOR BOTH DATA SETS AND DATA PROCESSING INFORMATION

From our starting point with the problem of capturing the processing and attributions that would go into a data application appendix ("DAX"), we see this appendix category as being intimately related to data and, due to the dualism between data and algorithms, assert that it makes sense to include both within the same solution space. Given the subjectivity required to identify the boundary between data and analysis in each scientific inquiry, an approach for one that does not comprehend the other would seem to solve one problem while creating another.

3.2 Establish a new publishing paradigm that includes data sets and data processing information

Our study also showed how deeply researchers across the board use and rely on traditional mediums of scholarly communication. Making data processing artifacts reusable on a broad scale would entail a revolution in scientific practices, one that would be easier to adopt if the outcomes looked familiar. In the research world, publishing is a familiar concept but archiving is unfamiliar. New practices analogous to existing publishing norms could provide a gentle ramp into this new world, bringing with it analogous forms of preparation, submission, review, dissemination, and archiving.

3.3 Establish a "data paper" publishing paradigm

There is a need for a new scientific publishing paradigm that encompasses both data sets and data processing information. For a given "data paper" this pairing defines a data product that is the result of applying recorded processing information to stored data. Note that this use of the term "data paper" is different in detail but similar in spirit to various emerging uses [ESA; Rees].

The data product is a neutral, argument-free platform from which to draw scientific conclusions, presented in the form of a document with title, authors, date, and abstract, followed by a package of narrative and references that captures the entire data product. Regardless of origin as supplementary to an envisioned article, it is nonetheless independent and ready for reference and re-use by others. The first step is to firmly establish the "data paper" metaphor.

3.4 Libraries and open archives should lead in establishing the new publishing paradigm

In our study, we frequently encountered barriers to sharing not only data but also publications in general. We know that some publishers are interested in a solution to SM, but it is important that traditional publishers not be the only interested parties. Academic and national libraries are increasingly engaged in data archiving and are emerging as non-traditional publishers with a strong commitment to supporting open and accessible information. Paul Courant has observed irony in the suggestion that libraries have no business being in publishing when in fact mounting a digital library itself is an "act of publishing". [Courant] With centuries of experience in developing open access services and ensuring long-term access to information, libraries represent a valuable and interested party.

4 What is a "data paper"?

Like a traditional paper, the data paper would have a title, author list, date, abstract, and references section, but the data paper would be distinguished in a number of important ways.

1. A data paper is a package that includes data, its metadata, and any processing information (techniques, formulas, software, etc.) that together define the state of a data product that was used and is ready for re-use in traditional journal articles. The data paper does not argue toward conclusions, leaving that to the traditional literature, which it complements, in order to achieve "the most principled separation of concerns" [Rees].

2. With today's technologies, a data paper is likely to be a virtual born-digital package, in other words, an electronic document that includes larger components by reference via persistent hyperlinks.

3. A data paper can either be peer-reviewed or non-peer-reviewed, but must declare this status clearly, and if peer-reviewed, must describe the peer review process undertaken. By defining peer-review to be optional, publishing (i.e. sharing) this data product can proceed before peer-review or without it entirely. Indeed early sharing might become a pre-requisite to peer-review given that often the value of data isn't known until it is widely used and annotated. There are many parallels to traditional preprints that can inform strategies for data papers.

4. A data paper and its products can be corrected without compromising data integrity. Correction is a beneficial consequence of use, but requires that data products be versioned and that corrections and accompanying versioning be clearly exposed to those discovering such data papers and their products. "[S]tandards for ensuring integrity of research data" [NAS] are crucial both for the conduct of science and to counter "science denial" [Specter].

5. As a digital publication, the data paper is essentially unencumbered by physical page limits and formatting limitations of print. Of course there would still be limits on physical storage size and access to data that require special hardware or software to render or use. Regardless, digital publication enables full inclusion in one paper of many hundreds of data producers, data sources (e.g., in the references section), and funding sources. Especially apropos for a data paper is the term, "data author", that we picked up in the course of this study. [interview 13]

6. As a virtual package, the data paper does not need to contain all data artifacts in one location. Large artifacts would routinely be included by references (e.g., actionable links) to data and to processing information that would be distributed across a number of archives. A suitable label, such as "Included artifact", would accompany such links to distinguish them from links to material that was merely related but not meant to be included in the virtual package.

7. Selected data papers can be reassembled by reference to form derivative objects and new types of publication, from a journal issue (continuous or closed) to an edited

volume. It suffices for the journal editor to create a presentation structure that can encompass the complementary material — from covers, to front matter, tables of contents, references, editorial policies, submission guidelines, etc. These measures would render such collections intelligible, useful, appealing, and familiar to journal readers, and is more or less the concept of the overlay journal [Gray].

8. The data paper and its core artifacts would be available from stable online storage via persistent actionable identifiers. Simple, robust, and transparent institutional and disciplinary digital repository and identifier maintenance systems make this task easier.

Given all these differences, it is all the more important that the data paper present a surface similarity to a traditional paper in order to ease acceptance and validity in the eyes of scholars and research institutions. In fact, despite their significant differences, the data paper genre as proposed here bears a strong resemblance to the traditional paper genre, and the standardized structure proposed here should facilitate its adoption.

4.1 What is a data paper cover sheet?

A data paper as envisioned here would clearly indicate that it is different, while at the same time being encapsulated within a familiar HTML or PDF document with a "cover sheet" containing such elements as

Element 1:	*Title*
Element 2	*"Data Paper" (required label)*
Element 3	*Persistent Identifier (DOI, ARK, URN, stable URL, etc.)*
Element 4	*Date*
Element 5	*Abstract*
Element 6	*Authors*

If the data paper were co-created with a published article, beneath the "Data Paper" label would appear an additional label,

Element 7	*"Supplemental to <link-to-article>"*

As we learned in this study, the author list, together with affiliation and sponsor support labels, could easily run on for what would amount to several printed pages. Nonetheless it provides the credit so vital to motivating the entire data paper proposition. Attribution would thus be available to data gatherers, data sources, consultants, programmers, technicians, integrators, funders, and institutions that supported the effort required to produce the data product.

While we observed the practice of updating dataset splash pages over time to include links to other literature that re-use the dataset, the data paper is not an ideal place to add such backlinks, as this would alter the data paper, probably necessitating changes in its version number. A less invasive alternative includes adding backlinks to a splash page that is not itself part of the data paper but acts as more of an envelope. Another alternative is to let backlinks be computed and displayed periodically by internet-wide indexing engines (e.g., via "page rank").

A data paper whose cover sheet declares it to be "Supplemental to" an article would consist of two sections: an "Integral Information" section and a "Non-Integral Information" section, either of which could be empty. The former would contain those data products required to reproduce the conclusions in the article and the latter would contain all other SM. Each section would have its own reference sections. By finding a place for related but non-essential SM, we can accommodate all three classes of SM that we encountered in our study; this optimizes peer review away from the non-essential while reducing the need to create a third publishing paradigm.

Data discovery is still difficult. The above elements included in an HTML or PDF document are sufficient to expose data papers to internet-wide search engine indexers. For data papers containing no other indexable text, these core "bibliographic" elements provide at least a starting point for discovery. Of course some data papers will have extensive indexable processing narratives. Given that these elements are also required by academic indexing services such as Google Scholar [Gscholar], such additional components would result in increased exposure for researchers and institutions.

4.2 Return incremental value for incremental effort

A new "data paper" publishing paradigm would be a fitting accompaniment to the new eScience (4th) paradigm of science. The data paper concept could go far in addressing many of the impediments, incentives, and requirements that we found in our investigation. Moreover, it can be seen as a set of steps, as shown in Fig. 1. At each step, an increment of value is obtained by an increment of effort; there is solid gain to be had while resting on any step, and for a given data paper the climb can be paused, resumed, or abandoned at any time.

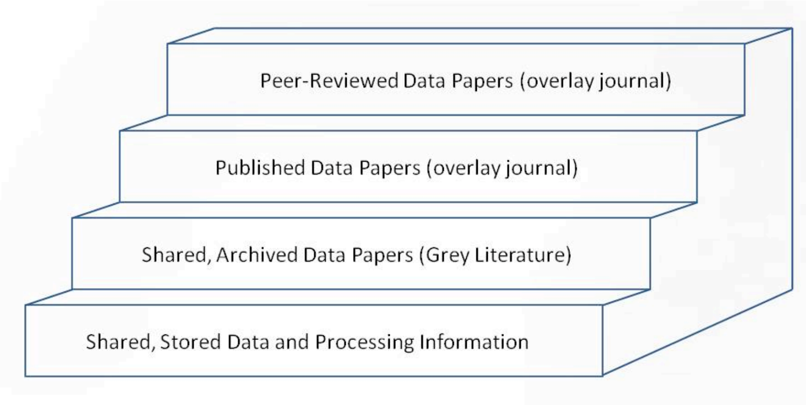
Peer-Reviewed Data Papers (overlay journal)
Published Data Papers (overlay journal)
Shared, Archived Data Papers (Grey Literature)
Shared, Stored Data and Processing Information

Fig. 1. The first step in building a data paper is to share all the artifacts of the relevant research from stable storage. Preparing at least a cover sheet and an assemblage of references is the next step, which provides some real exposure to internet search engines. Submitting for publication in a peer-reviewed or non-peer-reviewed overlay journal is next.

5 ANALYSIS

This study used predominantly qualitative methods based on a review of literature and on interviews with data scientists. Literature reviewed included formally and informally published material in data publication, citation, and preservation, including relevant national scientific guidelines and emerging whitepapers and international standards work in the realm of data-intensive science. This work provided a broad snapshot of the data artifacts landscape and the main actors involved.

We also interviewed scientists from a range of disciplines. While the scientists were predominantly from the earth sciences, permanent or strong identification with any one discipline was notably lacking. For example, skills from meteorology, ecology, environmental science, climatology, hydrology, and even computer science and genetics could have been appropriate specializations depending on what study was underway and where they were in their education plans and careers when they were asked.

To gain perspective and a better understanding the boundaries of our study, we also interviewed a social scientist and an archaeologist. We found [interviews 10 and 14] that archiving was relatively better-organized in both disciplines, with established networks of data depositories (e.g., DataVerses and clearinghouses for completed archaeological site studies). We speculate that this is due to the relatively smaller datasets and slower data collection processes in those disciplines — although they are data-driven, they are not data-intensive. We also interviewed a neuroscientist, which gave a glimpse into a culture whose data and processing (image analysis is quite processes that special problems

We found a variety of stakeholders, including research funders, research institutions, journals, professional societies, and individual scientists. We also found a variety of people who handle scientific data in their professional lives, including data collectors, modelers, synthesizers, facilitators, as well as data repository administrators, information scientists, and data standards specialists.

5.1 WHO OWNS THE PROBLEM?

It is unclear who "owns" the SM problem. Traditional publishers lay some claim by virtue of proximity to the main (i.e. published) article, although there are few uniform publisher policies with regard to SM and one publisher has recently taken a firm stance declining any interest in SM [Maunsell], and this stand has been applauded by some scientists [Piwowar]. However, if a dataset is produced solely for the purpose of publishing one or a few specific articles, at least the data portion of SM might appear to be a publishing problem. To complicate this view, datasets stored in most data centers, such as the DAACs, Dryad, or PANGAEA, are generally seen as secondary or supplemental to published articles, but at the same time, because they archive data and metadata, such centers must also attend to the problems of making SM available and usable

While largely unpublished, SM is not typically called out in discussions of "grey literature" [GreyNet], even as institutional and discipline-based repositories begin to stake out their roles in the SM problem. Some SM is informally published by making it available on a website, but we find few organized, cross-disciplinary attempts to collect, preserve, and disseminate SM. One international project that informs this area is the NISO/NFAIS Supplemental Journal Article Materials Project, which represents the clearest indication we found of wide interest among publishers to understand the SM problem. It aims to develop a recommended practice for publisher inclusion, handling, display, and preservation of supplemental journal article materials.[3]

Yet, there is also a trend to regard data as publishable in its own right. Sometimes this is done without fanfare, as when a traditional peer-reviewed journal publishes articles that focus primarily on the data instead of conclusions drawn from it. [Bond-Lamberty] Other journals, such as *Ecological Archives*, *Earth System Science Data*, and the *International Journal of Robotics*, are explicit about accepting an emerging genre of "data paper". For example,

"Data Papers are compilations and syntheses of data sets and associated metadata deemed to be of significant interest to the ESA membership and the scholarly community…. Ecological Archives provides a reward mechanism (in the form of peer-reviewed, citable objects) for the substantial effort required to compile and adequately document large data sets of ecological interest." [ESA]

Although not yet considered as valuable as traditional papers in academic promotion reviews, some senior faculty members we interviewed are actively trying to change this attitude in their institutions. It has been suggested [Rees] that even if a data set has been linked to as SM to a peer-reviewed article, it should nonetheless be published as a free-standing presentation of "materials and methods" independent of conclusions to enable re-use. If data is to be used by researchers who neither produced it nor conceived articles from it during its production, this data looks less like a supplement to a traditional published article and more like a publication in its own right.

What is still missing from the discussion of SM are the first and third classes enumerated at the beginning of this section. Here we are especially interested in the latter: detailed information about the methods of data creation, capture, processing, and analysis. The present study originated with a focus on this third class, which includes

- additional narrative that describes processes and techniques, such as MODIS Algorithm Theoretical Basis Documents (ATBD)
- program code that implements algorithms used,
- source lists that precisely identify subsets of data used,
- attribution lists that credit data producers and integrators,
- field definitions that bound sets of values found in tables,
- metadata that describes conditions applying to an entire table,
- processed tables that support published graphs, and

[3] We are fortunate enough to be participating in and helping to shape and this project. [http://www.niso.org/workrooms/supplemental]

- recorded data itself/themselves4

5.2 THE SUPPLEMENTAL MATERIAL (SM) LANDSCAPE

Across disciplines and publishers, we found no consistent definition of terms such as “data”, “dataset”, “data package”, “data publication”, and “supplemental material”. In this section we introduce the foundational terminology and a framing of the problem that we think points out some promising directions.

In journal publishing, supplemental (or supplementary) material (SM) is evidence that supports a main article (MA) but that cannot be published as part of it due to various constraints. Sometimes called “data-driven research”, an eScience journal article resembles the tip of Jim Gray’s “data iceberg”, resting on a large body of research artifacts. Using a less menacing metaphor, we can see this as the pyramid in Fig. 3.

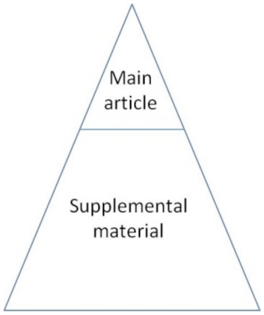

Fig. 3. A main article in eScience is the distillation of lots of data and lots of data processing, represented by large amounts of supplemental material. It can also be seen as the tip of Jim Gray’s metaphorical iceberg.

Reasons vary for excluding supporting evidence from the main article and relegating it to SM. Some SM is too big; for example, extended tables, interview transcripts, and explanations of processing techniques. “Too big” can bump into different limits, including file size, reviewer time, reader time, and printed page limits. An important constraint concerns medium or format. In print-only journals, for example, audio and video evidence is automatically SM, even if it is integral to the central argument of a publication. Online journals are not as constrained by format limitations of the printed page or by page counts, but are still constrained, for example, by reviewer time.

Another reason for the relegation of supporting evidence to SM has to do with its conceptual distance from the main argument. This distance exists when material is deemed to be supporting but secondary or not essential to the central message of the MA. What concerns us especially in preserving the scientific record, however, is evidence that is in fact essential to the article’s central message but that due to other constraints cannot be included in the MA.

[4] We follow NYT and IEEE practice in this exposition and use “data” as a singular or plural noun as seems most appropriate in a particular context.

5.3 DATA, INFORMATION, KNOWLEDGE

As mentioned earlier, although “data” obviously plays a huge role in data-intensive research, in our study we encountered no consistent definition of the term. As we look for strategies to publish, share, and archive SM across eScience disciplines, it is at least instructive to go through the exercise of looking for a useful single definition of “data”.

We can frame the inquiry by leveraging a traditional information science continuum [Ackoff] in which data is considered unusable by itself until interpreted to create information, and information combined with insight, context, and value systems is used to create knowledge. There are clear parallels for eScience if we think of supplemental material (SM) as including both data and the analysis by which it is digested, processed, normalized, visualized, reduced, and otherwise prepared for direct assimilation into a published article. Incorporating the first and third SM classes in a “processing” layer, the bottom “data” layer of Fig. 4 refers to data that may be raw, derived, or recombined from different sources (cf. [Gray]).

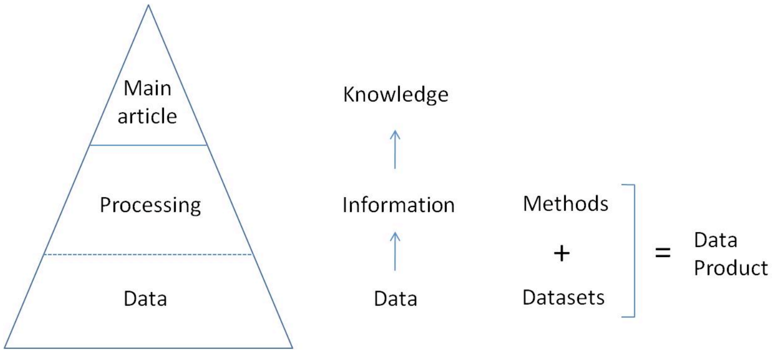

Fig. 4. Supplemental material supporting a main article in eScience includes data and processing information in a way that parallels a traditional data-information-knowledge continuum. A dualism between methods and data permits one to grow and the other to shrink while supporting the same platform from which to begin reasoning to conclusions in a main article.

Determined by evolving notions of “publication”, the boundary between main article (MA) and supplemental material (SM) is in flux. Even less certain is the boundary between processing and data. Simple tests — format for example — do not suffice. A table of 64 numbers could be either raw data or the end result of analyzing and reducing millions of raw data points.

Uncertainty in the boundary between processing and data is a natural consequence of a classic computer science metaphor captured in the book title, *Algorithms + Data Structures = Programs* [Wirth]. If we define “data product”[5] to be the combination of processing and data, an eScience version could be written

[5] Note that the ORNL DAAC uses this term more narrowly to describe an archived package of data and metadata.

```
Methods + Datasets = Data Product
```

Fig. 5 illustrates components of this equation. “Methods” comprises all the processing (techniques, formulas, software, credits, sources, etc.) that were or could again be applied to one or more archived datasets. A “Dataset” is a stored package of data (tables, images, etc.) and metadata that fully describes the stored data (e.g., variable names, conditions of its collection, instrument calibrations).

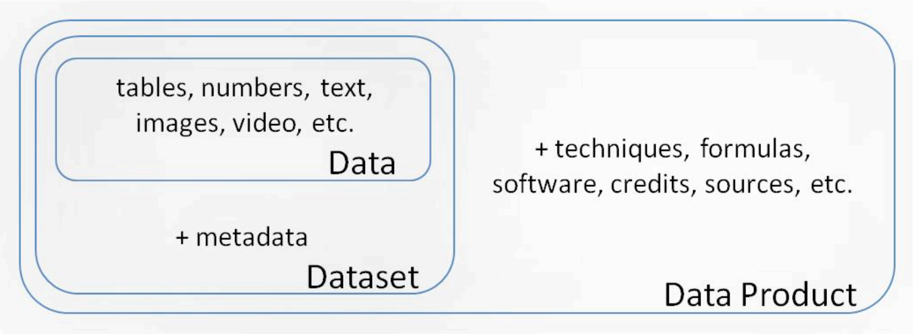
tables, numbers, text, images, video, etc.
Data
+ metadata
Dataset
+ techniques, formulas, software, credits, sources, etc.
Data Product

Fig. 5. Components of datasets, methods, and data products.

There is an inherent dualism between methods and data that plays out in eScience. As suggested by the equation, equivalent data products can be composed of different sums; the cruder the data, the more sophisticated the processing required to achieve the same level of resource, and vice-versa.

In an idealized situation, no processing information is required to support published conclusions based on a single dataset that, as archived, is already normalized, cleaned, and prepared specifically for the MA. But the need for processing often arises late in preparation of an MA, after data is already stored. Moreover, stored datasets are increasingly re-used, which means processing them in preparation for other MAs. Processing is more complicated when a data product is synthesized from multiple sources because each source generally calls for source-specific processing. and especially as scientific datasets are increasingly re-purposed for other MAs, the processing gap between data and MA needs to be documented to complete the scientific record. Fig. 6 illustrates some possible combinations of data artifact re-use.

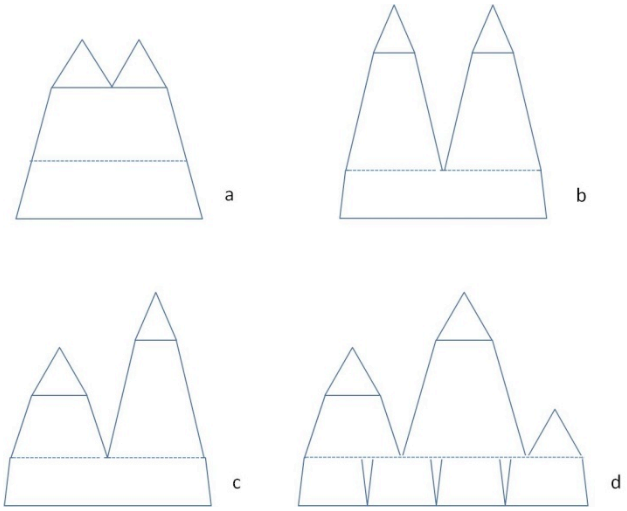

Fig. 6. Example data artifact re-use combinations: (a) two articles built on one top of the same data product (dataset and processing), (b) two articles built on one dataset but two different kinds of processing, (c) two different kinds of processing, the second more extensive than the first, and (d) two articles built on two data products that synthesize four datasets and one idealized article built directly (no processing) on the fourth dataset.

5.4 DEFINING DATA

As suggested by the above continuum, "unusable by itself" may be a fundamental characteristic of what it means to be data. In any case it certainly suggests a strong documentation requirement if re-use is a goal. Opacity of course, depends on context. For example, the narrative of this study is not data because it is not entirely opaque; on the other hand, the words representing this narrative could be seen as data points in an input stream for a concordance generator, or as a single collective data point in a 9-million-document replication queue for a preservation archive. As opaque representations in the right context, any of the following could be seen as data:

- Counts (specimens, people, particles)
- Measurements (temperature, humidity, luminance)
- Images (from satellites, x-rays, electron micrographs)
- Time series (any of the above sampled over time)
- Video, audio, seismograms, and other recordings (time series with high-enough sampling rate to appear continuous in time)

One definition that has gained some currency but is too restrictive for our purpose here is as follows: "Scientists regard data as accurate representations of the physical world, and as evidence to support claims." [Cronin] First, the physical world appears to exclude the important class of data comprising human opinions, beliefs, and attitudes. Second, in modern data-rich science many of the representations being collected will need to organized, curated, and accessed, perhaps for months or years, before they are known to be accurate or useful. From the point of view of data management, all representations not yet known to be inaccurate or useless need to be treated exactly as if they were real "data".

The excellent 2009 National Academies report, “Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age” defines research data as

“[I]nformation used in scientific, engineering, and medical research as inputs to generate research conclusions…. It includes textual information, …, equations, …, diagrams…” [NAS]

This is still too broad for our immediate purpose because publication and sharing of “data” sometimes needs to take place before its value in generating research conclusions is fully understood. Second, we want to preserve some distinction in this study between “data” and “information” (processing).

A more general definition of data that suggests opacity and tolerance for potential inaccuracy comes from an operations research specialist:

“Data are symbols that represent properties of objects, events and their environments.” [Ackoff]

Like the National Academies report, however, we exclude physical objects (but not their descriptions) from the definition of data. [NAS]. Ironically perhaps, intangible digital symbols are easier to deal with in terms of use, storage, and dissemination because they conform to widely adopted and economical digital information practices.

At the lowest level, these symbols (numbers, shapes, captured waveforms of noise, light, vibration, etc.) are “raw data”. It is difficult to know if data are truly raw given the increasing sophistication of sensor instrument hardware, such as consumer-grade cameras, that intercept physical signals and intervene with corrections and normalizations by default before ever recording the rawest form of the data. We can, however, refer to data that is more raw than other data.

Data may originate from instrument observation or model simulation. The latter can be used to generate hypotheses and the former to test them. Simulation encompasses modeling, as it can supply inputs to test and debug observational systems. Weather statistics result from observations and weather forecasts result from simulations. Simulation data may be from any source that is not strictly observational, such as what is predicted by a computer or by a mathematical formula. It is not uncommon to see simulation data complementing observation data; for example, one year of precipitation data in a Canadian study was missing due to malfunctioning equipment, and was replaced with data from a nearby airport gauge to simulate (approximate) it [interview 2]. At a high level is “synthesis data”, a combination of data elements taken from multiple sources and integrated after any processing steps required to achieve smooth data integration. It is possible for data to be synthesized from just one source, effectively transforming it (e.g., unit conversion) in combination with other data or metadata from that source.

Interpretation of the symbols comprising data is fundamentally the role of processing. Some scientists see data as “news” or as “confirmation”, [Birnholtz] depending on whether they are looking to form a hypothesis (discover events) or to test a hypothesis (verify events).

5.5 Complications and mitigations

Over the course of this study, we learned that there are many ancillary social, environmental, and process-based issues that may work against the kind of data publishing concept proposed here. They complicate the scientific community's ability and motivation to share and re-use data in any long-term way. Some of these issues are listed below, followed by a list of potential mitigating strategies.

- Not all scientific disciplines have central holding places for perpetual storage and archiving of data and/or supplemental material.
- Not all scientific disciplines have central holding places for the discovery, access, and display of datasets and supporting material.
- Even for a discipline that has a holding place, this was not necessarily known across the discipline; for example, some interviewed US earth science data archivists had no specific knowledge of their European counterparts.
- Despite a number of publisher mandates, some interviewed scientists are not motivated or do not have enough time to submit data into a central repository due to the extra work necessary to gather and write up supporting information and metadata from emails, lab notebooks, etc.
- Some interviewed scientists admitted that although they had never gotten around to archiving data for a publisher that stipulated data archiving as a requirement of publication, and they received no further requests nor were subject to any penalty from the publisher for failing to archive the data.
- In some disciplines there is no cultural norm to submit data for retrieval and storage in a central repository or website.
- Based on interviews, the primary way that most scientists discover datasets that they might possibly re-use will likely still be through published journal articles. Publishers often do not support the inclusion of datasets or supplemental materials along with these journal articles. Nor do they provide links to datasets or supplemental materials stored off of the publisher's site.
- Accessing such datasets often requires personal contact (email or phone) with the article's author, and using the dataset and interpreting errors often requires going back again to the author [interview 2]. These inefficient practices are likely to continue for data products.
- Another way that non-discipline expert scientists or users discover datasets for re-use is through commercial search engines. Currently, many datasets (and hence data products) are not available online or do not have sufficient metadata for easy discovery.
- A very difficult area is developing data descriptions (e.g., narrative and metadata) that are adequate for external re-use. No one we interviewed had been able to re-use data from an external source "as is" without either contacting the original data producer or running extra analyses on the data to try to "reverse engineer" missing schema information
- An important disincentive among scientists to share their data is loss of exclusive control in professions where competition for reputation is intense. Preparing data for re-use is hard enough without having to worry about lost opportunities to build "scientific reputation and its accompanying benefits, such as publications, grants, and students" [Birnholtz]. It can be

argued that reputational benefits should accrue directly from recognized acts of sharing, or that data produced in one discipline can lead to non-competing discoveries in other disciplines; nonetheless, short of a socio-cultural shift it is hard to imagine the scientific community requiring scientists to share their data freely until they have had sufficient time to extract what immediate benefits may be had from further analysis.

This complex of issues surrounding preparation and maintenance of data for re-use is unlikely to yield to one approach by itself. Some multi-faceted and community-based approaches involving many different players follow.

- University and research libraries should offer data archiving services and promote them to the communities and disciplines that they serve. Greater visibility and choice in archiving will foster more sharing, re-use and preservation generally.
- Educators should include sound data management practice as a part of the scientific curriculum early in the scholarly careers of future scientists at the high school and undergraduate level.
- Research funding agencies should continue to refine their mandates for a data management plan as part of the grant application process.
- Developers should create and extend tools to help capture supporting information, such as analysis and metadata information, at the point of data creation, analysis, and throughout the dataset's lifecycle. An additional function of these tools might be to submit data and accompanying support information to existing central data repositories; an example could be an Excel add-in that captures standardized metadata and allows for semi-automated submission into a data repository. A variety of such tools will need to be created to account for diverse types of data capture, such as streaming data, observational data, model data, etc.
- Research institutions should extend promotion and retention policies to take data papers and data journals into account. Recognizing individuals for the skill and effort that goes into producing and documenting data is an important motivator, and "research sponsors should acknowledge that financial support for data professionals is an appropriate component of research support in an increasing number of fields." [NAS, Recommendation 4]
- Publishers should create new visualization components that prominently display or link between data papers, data products (datasets, metadata, processing information), and traditional articles.
- Indexing services should expose the elements of data papers that enable citation tracking and bring credit to data producers. This means parsing and indexing possibly hundreds of data "authors" per data paper. The current practice (except by Google Scholar) of throwing away an article's unique persistent identifier (e.g., DOI) makes citation tracking difficult in traditional publishing; this practice should stop, especially as data papers enter the publishing mix.
- Tool builders should make it easy to create and administer unique persistent identifiers for datasets, data products, and data papers. This will allow for easier linking to data products from published articles. Strong identifiers are especially important when the data and published article are in different locations.

- Internet search engines should be able to index data products more easily because of information found in the data paper. Still, the data paper as presented here is only the beginning. It will be natural to want to search for data based on temporal and spatial coverage metadata, which is not yet called out in the cover sheet in any cross-disciplinary way. Scientists and archivists should work with search engine designers to capture metadata that will enable easier discovery of datasets for general non-expert users who might re-use data in unexpected ways.
- Memory organizations should support the rapid and responsive evolution of metadata for discovery and re-use. Controlled vocabulary maintenance as traditionally practiced in the library community (e.g., MARC/AACR2) and beyond (e.g., Dublin Core) is not sufficient. "Agile" evolution could be assisted with suitable application of modern crowd-sourcing techniques and platforms (e.g., Wikimedia).
- Data repository infrastructures and policies should support configurable embargo periods. If nothing else, deposit with or without public access achieves three critical archival goals – stable storage, curatorial oversight, and redundancy. If assurance of embargo is what it takes to obtain the data, the trouble to implement embargoes is worthwhile.

5.6 OUTCOME

Once established, the new data paper publishing paradigm will fundamentally change the way in which data-intensive science is conducted:

- data authors will be motivated routinely to deposit in stable public storage both data products (datasets and processing information) and the data papers that reward them with authorship credit
- data products will be more routinely re-used, annotated, corrected, and precisely linked to from traditional publications
- data products in general, including those resulting from synthesis efforts, will enter the scientific record instead of being lost
- data journals will spring up around disciplines, even if disciplinary data papers are scattered across geographically distributed repositories
- peer review will be optimized by authors' ability to indicate which information is essential for reproducibility and which information is not
- relevant but non-essential information will be available for the interested reader but will not interfere with peer review

The following scenario and Fig. 7 illustrate this outcome.

A climate scientist at the University of California is researching changing rainfall patterns in the Sierra Nevada watershed. During a preliminary literature review she discovers a journal article describing evidence of water usage in human and animal populations in a portion of the ecosystem under investigation.

Through the use of a persistent, actionable DataCite DOI [DataCite], issued and managed through the UC3 EZID system [EZID], the article references a peer-reviewed data paper held in the CDL eScholarship open access repository [eScholarship], with the underlying dataset in the lightweight, scalable UC3 Merritt curation repository [Abrams] for long-term preservation.   She downloads the paper and the dataset for further review.

Using the information in the data paper she is able to select and normalize an appropriate subset of the data for correlation with new observational data collected as part of her research program.   She plans to present her findings to her wider research community in a conference paper. As part of that work she prepares a new data product by first obtaining a preservation-ready identifier from EZID. That new data product is the synthesis of her own data, the previous data, and all the analysis and processing she performed to integrate the two sources.

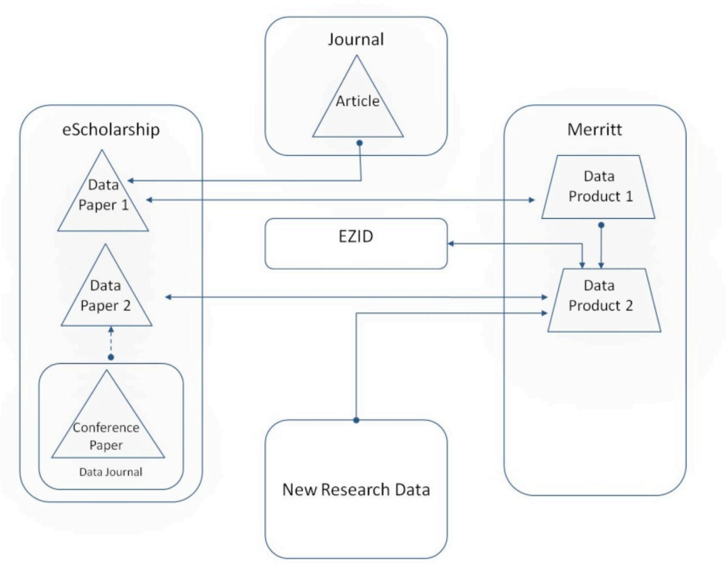

Fig. 7. A scenario illustrating the outcome of the new data paper publishing paradigm.

She deposits the new data product in Merritt for access and safekeeping, prepares a data paper to describe it, and deposits it in eScholarship. The new data paper is now ready for reference and annotation by others when she writes the conference paper, which she also deposits in eScholarship. At some later date, the conference paper is accepted and published in an eScholarship climatological data journal.

BIBLIOGRAPHY

Abrams, S., Kunze, J., & Loy, D., “An emergent micro-services approach to digital curation infrastructure,” International Journal of Digital Curation 5:1 (2010): 172–186 [http://ijdc.net/index.php/ijdc/article/view/154/217] . See also [http://www.cdlib.org/uc3/merritt] .

Ackoff, R.L. (1989). From Data to Wisdom. Journal of Applied Systems Analysis. 16:1989. p 3–9.

Agarwal, D., Cruse, P., & Kunze, J. (2009). White Paper: Technical Appendix Publishing to Support Dataset Usability. Submitted to the Gordon and Betty Moore Foundation.

Atkins, D. et al (2003). Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Retrieved from [http://www.nsf.gov/od/oci/reports/toc.jsp].

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: supporting sharing in science and engineering. Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work, 380. Retrieved from [http://portal.acm.org/citation.cfm?id=958215].

Bond-Lamberty, B. & Thomson A. (2010). A global database of soil respiration data. Biogeosciences Discuss 7, 1321–1344. Retrieved from [http://www.biogeosciences-discuss.net/7/1321/2010/bgd–7–1321–2010.html].

Bourne, P. [Personal email] (2010.10.17). Message-Id [E2591076–747C–4538–92D6-DBEC176B9D11@sdsc.edu].

Brase, J. (2010). Bridging the gap between data centres and publishers [PowerPoint slides]. Retrieved from [http://www.icsti.org/documents/winterworkshop2010/JB.ppt].

Courant, P. (2010). Economic Transformation of Libraries and Scholarship in the Digital Age [PowerPoint slides]. Retrieved from [http://research.microsoft.com/en-us/um/redmond/events/fs2010/presentations/Courant_Opportunities_for_Libraries_RFS_71310.pdf].

Cronin, B. (Ed.) (2008). Annual Review of Information Science and Technology. Medford, NJ: Information Today.

Data papers, supplements, and digital appendices for ESA journals [Web page]. (n.d.). Retrieved from [http://www.esapubs.org/archive/].

Nature — Data’s shameful neglect [Editorial]. (2009). Nature 461(145). Retrieved from [http://www.nature.com/nature/journal/v461/n7261/full/461145a.html].

DataCite. [http://datacite.org/].

ESA — Ecological Society of America [Policy] (n.d.). Retrieved from [http://www.esapubs.org/archive/].

eScholarship — CDL Publishing Program. [http://www.cdlib.org/services/publishing/escholarship.html].

EZID identifier service, University of California Curation Center. [http://www.cdlib.org/uc3/ezid].

Gray, J. (2007). Jim Gray on eScience: a transformed scientific method. Retrieved from [http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf].

GreyNet. [http://greynet.org].

Gscholar. [http://scholar.google.com].

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research. Retrieved from [http://research.microsoft.com/en-us/collaboration/fourthparadigm/].

Jingfang, N., & Hedstrom, M. (2008). Documentation evaluation model for science data. Proceedings of the American Society for Information Science and Technology 45(1): 11–11. Retrieved from [http://deepblue.lib.umich.edu/bitstream/2027.42/63090/1/1450450223_ftp.pdf]. Marcus, E. (2009, October 2).

Cell [Editorial] (2009.10.02). Taming Supplemental Material. Cell, 139. 1:11.

Maunsell, J. [Editorial] (2010). Announcement Regarding Supplemental Material, Journal of Neuroscience 30(32):10599–10600. Retrieved from [http://www.jneurosci.org/cgi/content/full/30/32/10599].

NAS, National Academy of Sciences. (2009). Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. Washington, D. C.: National Academies Press. Retrieved from [http://www.nap.edu/catalog.php?record_id=12615].

NSF, National Science Foundation. (2005). Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. Retrieved from [http://www.nsf.gov/pubs/2005/nsb0540/].

Parsons, M. A., Duerr, R., & Minster, J. (2010). Data Citation and Peer Review. EOS, Transactions American Geophysical Union, 91(34): 297.

Rees, J. (2010, March). Recommendations for independent scholarly publications of data sets. Retrieved from [http://sciencecommons.org/wp-content/uploads/datapaperpaper.pdf].

Piwowar, H. (2010, August 13). Supplementary Materials is a Stopgap for Data Archiving [Blog entry]. Retrieved from [http://researchremix.wordpress.com/2010/08/13/supplementary-materials-is-a-stopgap-for-data-archiving/].

Specter, M. [Video] (2010). The danger of science denial. Retrieved from [http://www.ted.com/talks/michael_specter_the_danger_of_science_denial.html].

Van de Sompel, H., Lagoze, Carl, Nelson, M. L., Warner, Simeon, Sanderson, Robert, & Johnston, P. (2009). Adding eScience Assets to the Data Web. Retrieved from arXiv:0906.2135v1.

Wirth, N. (1976). Algorithms + Data Structures = Programs Prentice-Hall.

Zimmerman, A. (2003). Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists. Retrieved from [http://deepblue.lib.umich.edu/handle/2027.42/61844].

APPENDIX A: DATASET USER INTERVIEW GUIDE

An interview guide was developed based on a questionnaire originally created by Ann Zimmerman of the University of Michigan for her work studying data re-use among ecologists. [Zimmerman] The questions in our interview guide covered three main categories: data location, data analysis processes, and dataset and supplemental material publication efforts. From July through November, 2010, we conducted 14 one-hour interviews with a sampling of people who handle scientific data in their professional lives, including data collectors, modelers, synthesizers, facilitators, standards specialists, as well as data repository administrators and information scientists. The interview guide we used follows.

I. Background

Can you tell me a bit about your background, your area of study, and your role when you work on projects? Please describe the other members of your project team or lab. How do you all work together?

1. How many years have you worked in this field?
2. Are you a data gatherer? Dataset user? A dataset re-user of previously gathered datasets?
3. What kind of datasets do you work with? Formats? Subject matter?

II. Locating the Data

1. How do you normally locate the datasets that you use in your work? Did you know what you were looking for when you started the search process? What process do you follow?
2. Do you normally utilize a specific dataset platform or repository? Do you use commercial search engines such as Google? Word of mouth? Personal communication (eg, phone calls)? Citation hunting through applicable published papers?
3. What is your level of satisfaction with your current data location process? Do you wish it were more automated? Do you like talking to other researchers about their data?
4. Have you ever not been able to locate data that you were looking for?

III. Working with the Data

1. Please describe how you work with these datasets. Do you synthesize data from a number of different datasets?
2. Do you capture the processing work that you do with these re-used datasets? How? Where do you put this information? What do you call it? Do you share this with others in your lab?
3. How is working with re-used data different from working with your own data or data captured for the project that you worked on?

4. If you were receiving an ideal dataset technical appendix (or supplemental material that contains information on how the dataset was originally gathered and processed) from a colleague what would it look like, what would it include that would make it easier for you to use?
5. Have you ever not been able to re-use a dataset? Why? What happened?
6. Has anyone ever contacted you about a dataset that you have gathered and synthesized? What did they ask for?

IV. Dataset and Supplemental Material (Technical Appendices) Publication

1. Have you ever submitted a dataset to a repository or for publishing with a paper? Have you ever submitted a "data paper"?
2. Have you ever had to create metadata for supplemental material or for technical appendices that describe the way in which a dataset was used for a published article? If you have worked with a metadata standard that you've had to apply, which one was it? Did you find it useful?
3. If we formulated a basic set of best practices for these supplemental materials or technical appendices, whom would we have to get on board to make "DAX" (technical appendix) publishing more mainstream or known or adopted?
4. What do you feel is of the greatest importance in creating these supplemental materials or technical appendices?

V. Follow Up

1. What are some general characteristics/features (good or bad) of the data repository systems that you use?
2. Are there other team members who are part of the dataset creation and re-use process that you feel would be useful for us to talk with?
3. Is there anything that you feel we should be aware of or pay special attention to in this investigation?
4. If we think of additional questions, would it be ok if we contacted you for additional information?