Value and Context in Data Use: Domain Analysis Revisited

Nicholas M. Weber, Karen S. Baker, Andrea K. Thomer, Tiffany C. Chao and Carole L. Palmer
Center for Informatics Research in Science and Scholarship
University of Illinois Urbana-Champaign
501 E. Daniels St
nmweber, ksbaker2, thomer2, tchao, clpalmer @illinois.edu

ABSTRACT
"Context" is an elusive concept in Information Science – often invoked, and yet rarely explained. In this paper we take a domain analytic approach to examine five sub-disciplines within Earth Systems Science to show how the context of data production and use often impacts the value of data. We argue simply that the value of research data increases with their use. Our analysis is informed by two economic perspectives: first, that data production needs to be situated within a broader information economy; and second, that the concept of anti-fragility helps explain how data increase in value through exposure to diverse contexts of use. We discuss the importance of these perspectives for the development of information systems capable of facilitating interdisciplinary scientific work, as well as the design of sustainable cyberinfrastructures.

Keywords
Research Data, Domain Analysis, Context, Cyberinfrastructure Development.

INTRODUCTION

Investment in cyberinfrastructure is proceeding apace with the development of trustworthy repositories and other digital environments that support the collection and discovery of digital research data. A primary objective of this investment is to make data a common, shared resource that can be used in new ways to answer increasingly complex scientific questions (NSF, 2007).

Previous studies of scientific data practices have demonstrated that even when data are made broadly accessible, the continued use or re-use of research data is highly dependent on knowing the context in which they were originally produced (Zimmerman, 2007). It is widely recognized that structured metadata can document some of this context by formally recording technical, structural, and methodological dimensions of data production. However, *formal* metadata often does not adequately capture the more *detailed* aspects of a research process, including indicators that have proven important for re-users to understand the potential for data to be used over time – such as the reputation of the producer, repository or unique details about the site of collection (Faniel & Jacobsen, 2010).

Studies of data practices have also shown that, as with most intellectual pursuits, the processes involved in producing and using data can be particular to local conditions and do not necessarily generalize across broad disciplines or fields of research (Cragin, Palmer, Carlson & Witt, 2010). Data practices are shaped by disciplinary norms, educational backgrounds, available infrastructure, and local cultures (RIN, 2008).

The qualitative studies of data practices reported on here have been conducted as part of the Data Conservancy (http://dataconservancy.org/), an initiative aimed at developing tools and services for data preservation, sharing and discovery across disciplines. Our role in the project has been to investigate how researchers in the earth and life sciences produce and work with data, as well as the cultures of sharing and norms of re-using data that influence what and how Data Conservancy systems and services should be developed.

One of the outcomes of this work was a theoretical framework for understanding the "analytic potential" of data, or the value of a dataset to be used over time, especially beyond its original intended purpose (Palmer, Weber, & Cragin, 2011). The notion of "fit-for-purpose," a fundamental concept in data curation (Lord, MacDonald, Lyon & Giaretta, 2004), is key to the analytic potential of data. Data tend to be appropriate for application to particular problems using particular methods or processes of analysis, and they may need to be represented, transformed, or enhanced in a specific way to be made fit for a new purpose. Assessing the re-use value of a dataset requires a domain analytic understanding of potential user communities, including the salient research problems, primary methods of analysis, and types of evidence (data) that can contribute to answering research questions for that community.

Here we apply a domain analytic approach to examine the production and use of data in five sub-disciplines that collectively represent the field of Earth Systems Science (ESS), explicating the ways data are used and reworked as they are applied to new contexts. Addressing the assertion that "the value of data increases with their use" (Uhlir, 2010, p.1), we ask: How do data change in value as they are used over time? Can datasets actually *gain* in value? If

so, how and under what conditions? And lastly, how do these changes effect the design of cyberinfrastructures mean to support scientific work?

Our approach differs from previous domain analytic studies in information science in its emphasis on economic perspectives, appropriate to our interest in understanding the phenomenon of data that change in value. First, we draw on Vertesi & Dourish's (2011) conceptions of data production within a broader information economy context. The "context of production" for data refers to the practices and the setting unique to the time, space, and people performing the work of data generation or collection. In our domain analysis, we shift "our perspective back to the point of [a dataset's] production," as suggested by Vertesi and Dourish (2011) to document practices that occur across sub-disciplines in ESS. With an emphasis on production within an information economy, interesting challenges arise in establishing a single "point" of production or a finished "data product."

We also consider the notion of "anti-fragility" in relation to change in value of data. Coined by economist Nassim Taleb (2012), the term refers to the property of a good or system capable of not only withstanding dramatic variations in use but actually increasing in value as it is exposed to stress. In our analyses, we demonstrate that some highly valuable research data appear to have an anti-fragile characteristic; they generate a resilience-through-use, which allows them to actually gain in value as they are exposed to broad contexts of use.

DOMAIN ANALYSIS REVISITED

Our use of a domain analytic framework is aligned with Hjørland and Albrechtsen (1995) who argued that Information Science (IS) should turn from the study of individual users to broader "knowledge domains." They discussed three approaches within the domain analytic paradigm:

> 1. The Social Approach – IS should be promoted as a social science in which theories of knowledge and sociology are applied.
>
> 2. The Functional Approach – studies of formal and informal communication might be used to study mechanisms underlying information behavior.
>
> 3. Embracing Philosophical Realism – a foundation for IS must be based on objective practices that reflect a world independent of our personal beliefs. This is a qualified sense of realism, as information systems are not built from discovered 'information laws' but in reaction to patterned practices observed over time (1995, p. 400).

Applications of domain analysis have been critiqued for being too focused on the "Social Approach" – especially in IS where researchers unfamiliar with epistemology often fail to distinguish between qualified and naïve realism (Frohmann, 2004). However, we've found that studying scientists and their data practices necessarily requires a sensitivity to important differences and cultural dynamics that influence the production and use of data. These "social" observations have been crucial to our interpretation of communication patterns (e.g. how information is transferred, or how knowledge claims are made and accepted) within each individual sub-discipline of the ESS domain. In order to simultaneously consider these unique sub-disciplinary cultures and those of Earth Systems Science as a whole, we've combined a Social approach with a Functional approach to domain analysis.

Previous Domain Analytic Studies

Talja and Maula's (2003) investigation of e-journal use across four disciplines took a largely functional approach to domain analysis. Examining the information seeking behavior in online databases for nursing, literature/cultural studies, history, and ecological environmental science, they aimed to generalize 'practice' at a disciplinary level. With this unit of analysis they could then look comparatively across discipline specific search strategies. Interestingly, their findings supported the work of Bates (2002) that found both the amount and supply of scholarly material in a domain had a profound effect on the search strategies a scholar employed.

Our own domain analytic framework aligns more closely with the blended approach taken by Fry (2006), in terms of her attention to both the social and functional aspects of information practices. As Fry observed, scholars are "producers of information just as much as they are users" (2006, p. 309). Thus the design of Information and Communication Technologies (ICTs) supporting these information practices should be informed by both the *production* and *use* of scholarly materials. Fry also diverged from many previous domain analysis studies by adopting what Chubin (1976) called an "intellectual field" as the unit of analysis for information practices, noting in particular that, "…specialist fields of enquiry are feasible cultural entities whose numerous representations more effectively capture the process of research than the more conventional use of disciplines as a unit of analysis" (Fry, 2006, p. 305).

Likewise, we examine Earth Systems Science by applying a sub-disciplinary unit of analysis. As noted in our previous work (Cragin, Palmer, Carlson & Witt, 2010), this level best captures generalizable data practices in small, field-oriented sciences. Focusing on sub-disciplines, instead of traditional disciplinary structures, has also been important for understanding the context of data production for a highly collaborative field like ESS where distinctions between data use and re-use (or re-purposing) are often unclear.

METHODS

Participants in our qualitative study of data practices are researchers active in the earth and life sciences with diverse

research agendas- studying phenomena such as climate change, magmatic dynamics, and nutrient cycling. Our interaction with a participant begins with a pre-interview worksheet used to orient participants to future, detailed discussion about their data practices and has also proven helpful for identifying domain-specific curation requirements. This is followed by a semi-structured research interview on data practices with both principal investigators, and other personnel involved with data management. Cases may also include a follow-up interview when appropriate for clarification, to fill gaps, or address questions that arose in the initial research interview. Where possible lab visits were conducted to observe the sites of research, collect sample data sets, and to conduct data inventories (Cragin, Chao & Palmer, 2011).

Data collection and analysis are ongoing, with twenty-one research interviews recorded, transcribed, and then coded by our research team using Atlas.ti. Our initial code list was based on terms and vocabulary from the data practices segment of the "Data Curation Framework" (see Cragin, Palmer, & Chao, 2011). These codes were then further refined to reflect emergent themes from the interviews. Extensive discussion among our research team also helped build a common interpretive understanding of the data and assured that we maintained inter-coder reliability. Team members' previous experience, including fieldwork in earth and life sciences, contributed important domain specific context to our analytic process.

In the following sections we include selective interview excerpts, modified slightly to improve readability by removing "um's", "uh's", or short affirmative responses (e.g. "OK"). The first section presents profiles of five ESS sub-disciplines. These profiles are meant to give a brief overview of the research methods and unique data practices of each sub-discipline, and have been derived from interviews with multiple participants, on-site observations, and analysis of relevant research artifacts (publications, datasets, specimen samples, etc.).

The next section examines data practices in greater detail, using both a social and functional domain analysis. Generalizing across the five sub-disciplines, we discusses how an economic perspective on data practices is essential to understanding how data are valued and judged within ESS.

THE DOMAIN: EARTH SYSTEMS SCIENCE

As a domain, ESS integrates knowledge from geology, meteorology, oceanography, and biology- combining traditional disciplinary methods with diverse collections of observational data to study the varied dynamic cycles of the earth (Lawton, 2001). ESS emerged from a belief that "…the global earth environment can be understood only as an interactive system embracing the atmosphere, oceans, and sea ice, glaciers, and ice-sheets, as well as marine and terrestrial ecosystems" (Asrar, Kaye & Morel, 2001, p. 1309). Increasingly, knowledge from this domain is being applied to the study of human activities and their effect on earth systems, which are of particular interest and concern for understanding an anthropogenic role in climate change and biodiversity loss (e.g. Ellis & Ramankutty, 2008).

Work in ESS involves the dynamic integration of data gained from a microscopic view of the earth to create and calibrate a macroscopic, holistic model of the planet and its processes (Schellnhuber, 1999). The following profiles describe the context of production and use of data that are particular to each sub-discipline. We look specifically at methodological approaches to data collection that each sub-discipline takes, how each independent pursuit of knowledge contributes a set of research products to ESS, and finally, how the norms of use and re-use drive the sub-disciplines' data practices. Table 1 summarizes the diverse data, methods, practices, and requirements.

Sub-discipline Profiles

Soil Ecology

Soil ecology is concerned with dynamic interactions between biological organisms and the physical environment, relying on investigations of biotic and abiotic aspects of soil. Physical soil samples are the primary objects of study, supplemented by maps and data loggers that provide environmental context.

The dynamics of soil and organismal processes are captured at targeted and multi-scale sites to understand how these dimensions interact in various ecosystems, of particular value to ESS as a domain. In conjunction with site-specific data collection, lab-based analyses are used to determine specific physical, chemical, and biological measures of soil properties that reflect these relationships. Results are recorded in field and laboratory notebooks and may correspond to digital files based on instrument output or the implementation of a research protocol.

The aggregation of raw data for analysis typically relies on manual input of measurements from several sources, including field collection, lab protocols, and instruments, into a unified structure such as a database table. Verifying the data is an iterative process that relies on consultation of the original recorded source (i.e. quantitative outputs from instrument readings, laboratory notebooks, etc.). Informal requests for specific measures and methodological protocols are common in this sub-discipline, however the re-use of physical soil samples is often limited due to their deterioration during lab processing.

Volcanology

Volcanology is the study of volcanic and magmatic systems: the dynamics, processes and underlying physics driving the flow of molten rock beneath (and sometimes through) the earth's crust. Much of volcanology is focused on determining crystallization rates in igneous rocks (rocks formed through the solidification of magma or lava), and using these "micro" level views of rock samples to inform "macro" level models of volcanic system structure. These

models inform other sub-disciplines within ESS such as structural geology, geobiology, and even climatology.

Though researchers draw from a diverse range of data to create mathematical models and maps of magmatic systems, data are primarily generated from physical rock samples or thin sections (thinly sliced sections of rock on glass slides). While secondary data like chemical analyses, high-resolution images of thin sections, crystal size distributions, and measures of phenocryst abundance are likely to be re-used, the physical samples and thin sections have the highest potential value for reanalysis.

The process of integrating these various types of data is highly dependent on understanding the data's provenance; researchers consult field notes, field photographs, and in some cases the data producer when performing secondary analyses. The collection of samples involves substantial analysis of already published work, resulting in high amounts of data re-use. Thus, because old and new data are constantly compared, the boundaries between data collection and analysis are somewhat fluid.

Stratigraphy

Stratigraphers study the ordering, composition and relationships of rock strata in order to understand geological history. Much of stratigraphy is fundamentally concerned with describing and documenting the order of sedimentary layers, assigning dates to those layers, and then extrapolating the depositional events or environments that would create those layers. This is done by iterating between the mapping of outcrops (visible exposures of rock faces in the field), and analyzing samples from these outcrops to better ascertain dates via radioactive isotope analysis.

The understanding of the Earth's geological history provided by stratigraphy is an important foundation on which other sub-disciplines in ESS build; stratigraphers construct a time-line by which other earth scientists date their data. In turn, stratigraphy relies on the data generated by these same sub-disciplines to refine the geological time scale. This is done by comparing qualitative data describing sedimentary layers to quantitative chemical analyses and other complementary signals of age or time (e.g. evidence of astronomical cycles). The idiosyncrasies of the materials being studied (e.g. the composition, location and context of the sediments) dictate what signals of age will be most useful or appropriate in studying them.

Because of the large amount of data needed in this work, and the high cost or difficulty involved in collecting new samples, stratigraphers rely heavily on existing data to construct complete maps and geological time scales. These data include rock samples, chemical data, isotopic data, and even numerical data extracted from printed graphs. The process of using one type of data to calibrate another signal is highly iterative, and involves a substantial amount of computational work on the part of stratigraphers. As in volcanology, this process requires frequent consultation with field notes, original data collectors, and other means of understanding the data's original context of production.

Sensor and Network Engineering

Sensor and network engineering uses coordinated arrays of instruments that allow scientists to remotely carry out autonomous field measurements of environmental properties. Sensor studies is a research area drawing upon engineering, computer science, telecommunications, and various domain sciences in order to optimize the performance of these technologies in recording and communicating data. A network is typically made up of a set of spatially distributed instruments equipped to monitor the environment at programmed intervals, to record measurements onto a data logger, and eventually send recorded data to a center for further analysis. Investigations use data collected about the sensors as well as about the environment in which instruments are deployed.

Development and deployment of individual sensors and sensor networks involves a high degree of interface with domain scientists regarding data collection and the arrangement of sensors. As a result network development proceeds in an iterative cycle of configure-prototype-analyze. Important tools include data loggers for data capture and databases for data storage, processing, and query. With large-scale collaborative programs funded to "instrument the field" and steward time-series datasets as community resources, autonomous instrumentation becomes an important component of many ESS sub-disciplines. Within-project and multi-project comparative studies are often carried out to improve understanding of a network. For dissemination, data may be available online via an ftp file repository or a website providing visualization. There is often informal sharing of the whole database or selected tables with an individual to whom the rationale for data arrangements is explained.

Ocean and Costal Modeling

Computational models in the earth sciences are increasingly used to forecast and now-cast events in natural systems like the ocean, atmosphere and climate. As the reliability and accuracy of modeling techniques have improved they've been adapted to study increasingly specific research questions, in ever more targeted settings. As a sub-discipline in ESS, these groups attempt to create formal mathematical models for dynamics in estuary, limnological and oceanic systems. This process is computationally intensive, depending, almost exclusively, on re-using data that has been gathered by field researchers with whom modelers have little to no direct communication.

While specific fields, funding agencies, and repositories are known to serve high-quality data, the process of finding useful, accurate data to develop ocean and coastal models is typically ad-hoc and dependent on informal channels of communication. Work practices are exceptionally varied in

	Soil Ecology	*Volcanology*	*Stratigraphy*	*RS Engineering*	*C/O Modeling*
Study approach	Biotic and abiotic properties of soil	Chemical and textural properties of rock samples combined with geospatial data	Range of signals compared to refine the geological time scale	Prototyping and designing field sensors to optimize field data collection	Computational or mathematical modeling of aquatic dynamics.
Kinds of data used	Physical soil samples, maps (paper & digital), biological species inventory, lab-based outputs	Whole rock samples; thin slices of rock samples on glass slides; chemical data; maps	Numerical data and graphs pulled from papers; physical samples; chemical, radioactive isotope, and astronomical cycle data	Autonomous field measurement of sensor and environmental data recorded on data loggers or transferred directly to a database	Water sample, meteorological, and remote sensing data downloaded; diverse models' output at many spatial & temporal scales
Patterns of data use	Systematic review of data for quality where values are checked against multiple sources	Iterative reference to & comparison of data sources, including chemical data, field notes, papers & maps	Highly iterative comparison of datasets and modeling of signals of time	Regular review of data for investigating various sensor configurations and contexts of data collection	Irregular patterns of use, based on need for model calibration or benchmarking for reliability
Norms of data re-use	Informal sharing of processed data and methods, though perceptions on re-use vary	High expectation of data re-use, particularly with physical samples and thin sections	Moderate expectation of re-use aiming to find new ways of determining geological time scales for re-use	Diverse, informal re-uses: optimizing sampling design; providing data to project researchers; or for public posting	Informal sharing of data inputs and software code; Informal and formal mechanisms for re-use and sharing of model

Table 1. A matrix of ESS sub-discipline data practices relating to the production and use of data

this sub-discipline; modelers iterate over observational data to create exploratory visualizations, or assimilate data to fit model parameters using a number of different granularities and grid densities as well as a variety of interpolation methods. There is often a near constant monitoring of experimental procedures because of a sensitivity of models to data and of data to empirical methods used in gathering field-based observational data. Although this sub-discipline is largely dependent on open sharing, and high-quality metadata that accompanies the data they re-use in building a model, their own practices of describing and sharing personal archives of processed data are highly irregular and infrequent.

THE ANALYSIS: SOCIAL & FUNCTIONAL

Working from Hjørland and Albrechtsen's original framework, our analysis first explores the Social and then the Functional approach to domain analysis.

The Social Approach: Context of Production

Across the sub-disciplines of ESS, our participants repeatedly described the context of gathering or collecting data as being the most important indicator of value. As one researcher put it, with data, "context is everything." How, when, where, and under what conditions data are produced have enormous implications for their regenerative value.

In stratigraphy, for instance, well logging data collected through industrial drilling activities are known to be of higher or more detailed resolution and are thus of potentially greater value than samples gathered by informal methods (as detailed below). However, because these datasets are proprietary and often unpublished, they are typically more difficult to obtain and may not be in an easily usable form. One stratigrapher reported only being able to find this type of "high resolution" data in graph form within a journal publication. Therefore, she had to resort to using software to semi-automatically extract the numerical data from scanned graphs, essentially recreating tabular data from an image. The elaborate and time-consuming process of converting these data to a usable form compounds their worth. In this instance, value was also tightly coupled with the reputation of the data collectors or institutions responsible for its production.

In some cases, value was determined by the uniqueness of the place where the data were originally gathered. For

example, if a site is ecologically unique, or requires special permitting to study, then the rarity of this context, quite intuitively, will positively affect the value of the data. As one soil ecologist explained, collecting soil samples from a foreign country requires not only a government-issued permit but also a specific space within their own laboratory to process, analyze, and store samples that might otherwise contaminate related experiments. Volcanologists reported placing similarly high value on samples collected at rarely studied sites or politically volatile regions. Value is indicated by the uniqueness of the scientific site, but it is also inherently tied to the broader political and geographical location in which the data were originally gathered.

A unique context of production can also prompt innovation, for which the researchers who originally collected the data have an "at-hand" advantage in processing and cleaning them for secondary analysis. For instance, participants from the sensor engineering sub-discipline were constantly tailoring and creating "workarounds" to accommodate new ideas for domain specific research. Below a sensor engineer constrained by existing protocols describes creating an alternative delivery mechanism for data:

> *"Some of the researchers ask me to get different data from different instruments. I have managed to make a special program for them in our data loggers in a way that new data doesn't go into the [existing] database but into a different place. Because we cannot touch how the database has been structured. So they are able to get the data but it doesn't go to the main webpage."*

In this example, the context of production has shifted from the "point" of collection to the point of access. In a sense, the data were only valuable after this shift occurred, and the constraints of the protocols in place spurred innovation that increased the value of the data across a network of researchers.

Documentation about the context of production also adds value to data, both in terms of increased discoverability and more accurate appraisal by secondary parties. Ocean and coastal modelers stressed that there was a need for thorough documentation of datasets from repositories. Formal metadata, however, was rarely enough to make a dataset trustworthy for re-use. Most modelers noted that either spatial coverage amenable to their model's grid or informal "word-of-mouth" reputation about quality were most important when deciding between comparable datasets.

This informal appraisal was true for instrument generated data as well as data gathered by field campaigns. Without the "being-there" of fieldwork, modelers sought context in every imaginable detail – from the weather during a plane's flight pattern, to a satellite's serial number, to irregularities in taking mooring data in the open sea. In the latter case, one modeler noted the importance of establishing context with a scientist in the field:

> *"…you have to go back to the data gatherer and ask them, "What's this value? This doesn't seem to be right. Do you remember what happened? Did a shark hit your boat or something?" …the quality control doesn't exist really well. So one has to work back and forth with the data collector."*

This same sentiment was echoed in the volcanology group. When using data collected by others, researchers emphasized the need for "a full repertoire of data," as one volcanologist put it, as well as the ability to replicate the experience of "being there" in the field. Many expressed a strong need to "create context" by continually consulting field notes or photographs, and, whenever possible, to speak with the original collectors before reusing their samples. One researcher expressed serious concerns about scientists who did not make the effort to understand the context of production, saying,

> "*[non-field researchers] get a bunch of data. And they've never seen them in the field, they've never seen the rocks in the field, they never saw how they fit in the system."*

This ties understanding the context of production to the ability to understand the actual "system" of ESS. We heard repeatedly that data are neither trustworthy nor justifiably usable without understanding how and why they were collected or created.

Clearly, capturing sub-discipline specific context is necessary to support secondary analysis or re-use. However, supporting personal interaction between data producers and users is also essential. In some cases, the rarity or value of data forces unlikely relationships between otherwise disparate groups, and enabling these interactions are especially important in a collaborative domain like ESS.

From Ecologies to Data Economies

It has been previously suggested that the emergence of a data-intensive paradigm in science is evidence of a larger, more sophisticated "ecological" approach to transferring, managing, curating, and preserving data (Choudhury, 2010; Smith, 2010). But as Vertesi and Dourish (2011) point out, a "data economy" is perhaps a more appropriate metaphor for describing the current environment of data-intensive work.

Similar concepts like "knowledge economy" have been invoked since Taylor first promoted scientific management at the beginning of the 20th century (Drucker, 1969), but what is meant by a data economy here is more specific to data practices, where systems are designed in a patchwork process to interoperate between small groups, or diverse sub-disciplines like those in ESS. As we saw in the analysis above, our participants consistently noted that *knowing* context was highly valuable when using their own data, but

equally important to their work in re-using data was the ability to *create* or *discover* the context of production.

Part of the challenge then for a networked data economy is to design interoperable systems that capture sub-disciplinary data practices as they are enmeshed in a larger-scale or domain-wide context of production. In an ESS setting, where most studies are only capable of being conducted through discovery and use of others' data, this means accounting for the context of production by explicitly documenting workflows and provenance, and by dynamically or statically linking tabular data to field notes, processing scripts, simulations, photographs, physical specimens, and other research products necessary for *creating* context.

Technical systems capable of creating context are not natural or pre-assembled constructions; they are purposefully and uniquely tailored to scientific inquiries within the domains they are meant to coordinate. Hence, many of these systems require that values inherent to a sub-discipline are "designed-in," such that, "the technological infrastructures that we introduce to each new information economy context must respect and enhance – or at least, not directly challenge – the processes by which the data they handle gains currency and value: including those specific to the context of data production" (Vertesi and Dourish, 2011, p. 541).

The social approach to domain analysis provides a comprehensive understanding of ESS as a whole domain, but this macroscopic view is only made possible through a thorough the investigation of individual sub-disciplinary data practices. At this level of analysis, it is clear that documenting context allows data to, as Vertesi and Dourish put it, "gain currency and value" in a data economy.

The Functional Approach: Resilience through Use

In our study of the data practices of ESS sub-disciplines, we also found that participants frequently described a phenomenon in which data that are repeatedly or iteratively used gain in value over time. In some cases this increase in value was a result of transforming the data into a more reliable or trustworthy product through debugging or cleaning. In other cases increased value was a result of applying data to a novel context that shed light on other domains of scientific – or even societal – interest.

For instance, instrumentalists consistently noted the need to make data accessible to exceptionally diverse groups – from discipline-specific research teams, to triathletes, or even search-and-rescue professionals. The variability in use by these groups influenced the way data were normalized and packaged for dissemination:

> *"We have people who are participating in triathlons…and they want to know about the water temperature and want to know about patterns. We've had Search and Rescue teams download our data to be able to predict what will be going on…fishermen will request data to look at trends…We also have industry people, need to know what the typical water level will be so they can get their boat in there."*

The path of data from sensor to scientist is here extended to include a larger population, which necessarily taxes both the data and the data producer. Formats, file transfers, and "finished products" are necessarily reconsidered when the audience of the sensor data becomes more diverse. The content of the data then demonstrates value by withstanding this expansion of use and adapting to reconfigurations that provide both value and wider accessibility.

In domain specific instances of data use and re-use we often noted that particular requirements for factors like frequency, accuracy, or precision were reconfigured between groups of researchers. In the following example, data shared by one community are improved when a second team of researchers worked closely with the original data collectors. Sensor engineers developing network sensors for use in the Brazilian rainforest provided a dataset using an average or "block" calibration to a group of researchers at the University of *São Paulo* who then compared the dataset with their own temperature observations, eventually improving the individual instrument level calibration. The sensor engineers could then communicate the particulars of this adjustment back to those that originally generated the data, and they were in turn able to recalibrate their own work and present more accurate data to the public:

> *"What we did for [Group Name] is that we were collecting the data and then we were processing it.... But then at the end of the project they went back and made this recalibration...up to 0.5 degrees Celsius difference [from] what we call a block calibration, which is something that was applied to all the hundred temperature sensors... they did this very arduous task of actually going and calibrating each one."*

This iterative cycle of processing-analyzing-improving-reprocessing is common in most scientific pursuits, but what the above scenario illustrates is the way in which networked researchers improve data by exposing them to broader contexts of use. In a sense, no improvement in the temperature calibration would have been considered had the data not been shared. And it was this back-and-forth transfer of data that lead to more accurate measurements by forcing the data gatherers to painstakingly improve the data through re-calibration.

Data quality improves not just through normal, repetitive use, but also by being stressed or taxed by application to novel problems. Stratigraphers have what might be thought of as "circuits of data" – wherein various forms of data are transferred between spreadsheets, tables, publications, and sometimes back to spreadsheets once again. It is not simply the act of extrapolating numerical data from printed graphs that improves data through "context creation" (as described

in the previous section) – it's also that these extracted data are then compared to other datasets to more accurately "triangulate-in" on what information a time signal is conveying, thereby improving the overall efficacy of future analysis. This process of comparison between different time signals (isotopic, astronomical) is central to stratigraphy work in general, but is also what makes its data so robust and valuable within the domain of ESS.

Similarly, ocean and coastal modelers discussed the idea of iteratively tuning a model to "reality" – where exposing the model to field data or coupling the model with a new physical system (e.g. tidal dynamics) forced them to seek new, often very diverse data sources to reanalyze, or even recalibrate interpolated data. The stress of tuning a model improves output data for the researcher, but importantly also has the potential to contribute to data archives by building collections of gridded data for reference and re-use.

Ocean and coastal modelers often referred to the process of performing inter-comparisons, or validating separate branches of a model with highly reliable, well-known datasets of similar temporal and spatial coverage. This process iteratively improved a model's ability to accurately forecast or simulate a given phenomena, but at the same time checked the "reality" of the methods used in collecting this observational data. Like the stratigraphers, this work is performed on diverse data sources – often comparing output from ocean models against atmospheric models – leading to more robust and valuable collections of processed data that can be shared and re-used in future inter-comparison projects.

These cases are just a few examples of what was expressed across ESS: data sharing is not just important for regenerative scientific work but is also crucial to improving the quality and value of data. When we discussed data sharing with participants who conducted fieldwork (Soil Ecology, Volcanology and Stratigraphy), most of their reluctance to share data wasn't for fear of being scooped, or undermined by competitors – it seemed to be a feeling that their data were too messy, or an insistence that their data were too specific in scope to be used meaningfully by another researcher. In a sense, they considered their data too *fragile* for re-use.

But when we discussed using, appraising, or discovering data that is created by others, in particular with researchers who routinely re-use data (Ocean and Coastal modelers, Sensor Engineers and Stratigraphers), we heard how important reputation was in trusting data. And as one modeler explained, reputation was established through "data referendums" – the vetting of quality was established most firmly with data that are tested, well shared and well debated amongst experts in a variety of research settings.

Throughout the sub-disciplines of ESS, we also found that value isn't necessarily static or self-determined. The data practices described above indicate that data improve in value through repeated use, and also through application to new or novel contexts. At first blush, this notion seems obvious – as the saying goes, "many eyes make for better seeing." However, it's important to note that widely used data don't just improve in reliability or quality – they actually become *more valuable* and in a sense, *less fragile*.

Anti-Fragile Data

As Nassim Taleb explains (2012), most languages lack an antonym for the word "fragile." Robust is often suggested, but this implies a kind of simple brute strength, and fragility doesn't necessarily imply that a system or good is weak, but instead that it is brittle, and suffers structurally when directly stressed. For instance, a wine glass is fragile in the sense that even if it were to survive a fall from a dinner table, it would be irreversibly weakened by that impact, and would never regenerate or recoup its lost strength.

As we've observed in ESS, a fragile dataset is one that cannot be improved through recalibration, or is not worthy of being painstakingly re-engineered from plotted graphs. The lack of a "fitness-for-use" here inhibits the data from generating new or improved understanding.

The opposite of fragile data would be a dataset that not only withstands deterioration over time, but also benefits from being broadly used. Instead of simply being resistant to external forces, these datasets can actually gain density through different kinds of use, exposure to new contexts, and by being stretched to accommodate broader audiences.

In ESS, one example of this resilience-through-use is shown by researchers' painstaking improvements of their data through a process of sharing and iterative calibration of measurements. However, the property that allows these kinds of data to gain in value through repeated use- therefore, becoming less fragile- seems to lack a formal moniker.

By mathematically modeling fragility in terms of path-dependent payoffs, Taleb demonstrated that the inverse of fragile is convexity, or what he explains might best be thought of as "anti-fragile" (2011). An anti-fragile entity will respond positively to unpredicted, unplanned, and unexpected uses. The same is true for goods that are capable of being valuable in contexts that stretch far beyond their patterned or particular use.

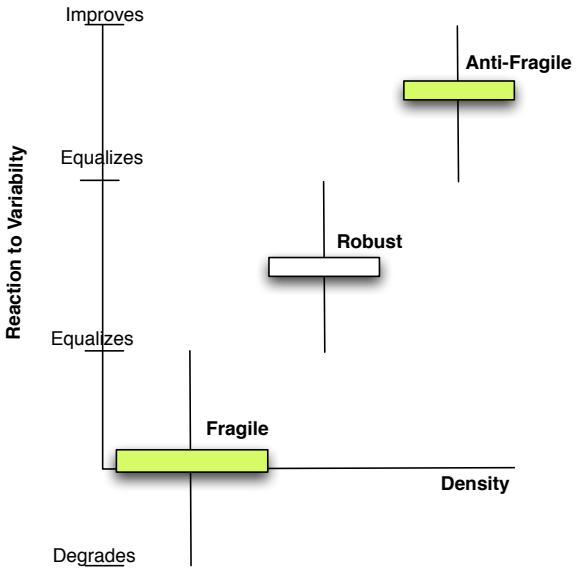

Figure 1. Upper and lower bounds of systems and goods. As goods become more dense the reaction to variability improves.

So on the one hand, goods and systems that are susceptible to breakdown and deterioration are fragile – like wine glasses or, from our examples above, soil sediment samples. These fragile goods maintain structural integrity by avoiding stress. On the other hand, we have goods and systems that are anti-fragile – like skeletal systems, or from our own ESS analyses, ocean models. These types of objects benefit structurally in terms of both strength and resilience though variations in use and stress.

THE CYBERINFRASTRUCTURE PERSPECTIVE

If, as our ESS examples show, data gain strength through exposure to use, recalibration, normalization, interpolation or even assimilation, then there are important implications for the technical systems meant to preserve and provide long-term access to these data products.

Traditional information systems are designed to treat data as fragile goods—that is, they have an upper bound of "not breaking" and a lower bound of degradation (see Figure 1). In the case of digital datasets, for example, the upper bound would be persistent access in a repository system and the lower bound would be degradation of the files integrity, such as susceptibility to digital bit-rot. In systems oriented to fragile data the best we can hope for is that a data collection persists, and in the worst case there could be loss of intelligible access to the content. Now, if the goal of cyberinfrastructure is to connect disparate knowledge bases in hopes of producing new knowledge (Atkins, 2003), then it seems antithetical to treat data – the currency in which cyberinfrastructures interoperate – as a fragile good.

If access to data that doesn't degrade is the design goal for cyberinfrastructure then the benefit of data being re-used or reanalyzed for a new purpose will always exist outside the information network. To put it more simply, if the upper bound or best case scenario for data is "not breaking" then how will interdisciplinary scholars ever harness the networked capability of a cyberinfrastructure to produce new knowledge? Building networked systems around fragile data essentially compounds the complexity that cyberinfrastructure development is currently envisioned to address – namely isolated collections not usable or discoverable by scientists attempting to cross traditional disciplinary boundaries.

TOWARDS ANTI-FRAGILE SYSTEMS

How to design effective and efficient information systems has been a central question in IS for decades (e.g. Taylor, 1986; Bates, 1999). In the data-intensive practice of contemporary science work, our research indicates that *efficient* systems are those that accommodate the context of production alongside *effective* systems that reliably preserve and transfer data across increasingly complex, interdisciplinary networks.

In our social and functional domain analyses of ESS, we saw that capturing the context of production at the sub-discipline level was crucial for establishing trust, sharing resources, and motivating further data analysis. We also observed that highly valuable data actually gained value as they were used in diverse contexts or transferred across a network of diverse actors. Of crucial importance then for future work supporting interdisciplinary data practices is the development of efficient and effective systems that are designed to function beyond the traditional notions of a static dataset so as to respond and evolve new datasets. We propose that it's only by designing cyberinfrastructures and information systems that are capable of accommodating, and attracting anti-fragile data – those that benefit from the stress of use, computational innovation, and variability in methodological approaches – that the transformative potential of widespread data sharing and re-use will be fully realized.

ACKNOWLEDGEMENTS

This research was supported by NSF Grant #OCI-0830976. We wish to thank Dr. Melissa Cragin for leading the collection of these data, her work the Data Conservancy Data Practices group and her insights on earth systems sciences.

REFERENCES

Asrar, G., Kaye, J. A., & Morel, P. (2001). NASA research strategy for earth system science: Climate component. *Bulletin of the American Meteorological Society*, *82*(7), 1309–1330.

Atkins, D. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the NSF blue-*

ribbon advisory panel on cyberinfrastructure. Retrieved from: http://hdl.handle.net/10150/106224

Bates, M. J. (1999). The invisible substrate of information science. *Journal of the American Society for Information Science*, *50*(12), 1043-1050.

Bates, M.J. (2002), "Speculations on browsing, directed searching, and linking in relation to the Bradford distribution", in Bruce, H., Fidel, R., Ingwersen, P. and Vakkari, P. (Eds), Emerging Frameworks and Methods: Proceedings of the *4th International Conference on Conceptions of Library and Information Science* (CoLIS4), July 21-25, Seattle, WA, Libraries Unlimited, Greenwood Village, CO, 137-49.

Choudhury, S. 2010. Data curation: an ecological perspective. *C&RL News* 7 (4): 194-6.

Chubin, D. E. (1976). The conceptualization of scientific specialties. *Scientometrics*, 12(5–6), 373–379.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023.

Cragin, M. H., Chao, T. C., & Palmer, C. L. (2011). Units of evidence for analyzing subdisciplinary difference in data practice studies. *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, 441–442.

Drucker, P. (1969). *The age of discontinuity: Guidelines to our changing society*. New York: Harper and Row.

Ellis, E. C., & Ramankutty, N. (2008). Putting people in the map: anthropogenic biomes of the world. *Frontiers in Ecology and the Environment*, *6*(8), 439-447. doi:10.1890/070062

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, *19*(3-4), 355–375. doi:10.1007/s10606-010-9117-8

Frohmann, B. (2004). *Deflating information: From science studies to documentation*. Toronto: U of Toronto Press.

Fry, J. (2006). Scholarly research and information practices: a domain analytic approach. *Information Processing & Management*, *42*(1), 299-316. doi:10.1016/j.ipm.2004.09.004

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: domain-analysis. *Journal of the American Society for Information Science*, *46*(6), 400–425.

Lawton, J. (2001). Earth system science. *Science*, *292*(5524), 1965.

NSF (National Science Foundation). (2007). *Cyberinfrastructure vision for 21st century discovery: report to the NSF cyberinfrastructure council*. Retrieved from:http://www.nsf.gov/pubs/2007/nsf0728/index.jsp/nsf11001/gpg_2.jsp#dmp

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). The analytic potential of scientific data: Understanding re-use value. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–10.

RIN (Research Information Network). (2008). *To share or not to share: Publication and quality assurance of research data outputs*. Main report and annex: Detailed findings for the eight research areas. Retrieved from: http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs

Schellnhuber, H. J. (1999). "Earth system" analysis and the second Copernican revolution. *Nature*, *402*.

Smith, M. (2010, December). *Managing research data at MIT: Growing the curation community one institution at a time.* Presented at the 6th Annual International Digital Curation Conference, Chicago, IL.

Taleb, N. (2011). Antifragility — or— The Property of disorder-loving systems. In J. Brockman (Ed.), *What scientific concept would improve everybody's cognitive toolkit? World question center, Edge.* Retrieved from: http://www.edge.org/q2011/q11_3.html

Taleb, N. (2012) *Anti-fragilty: Things that gain from disorder*. Random House, New York.

Talja, S., & Maula, H. (2003). Reasons for the use and non-use of electronic journals and databases: A domain analytic study in four scholarly disciplines. *Journal of Documentation*, *59*(6), 673-691. doi:10.1108/00220410310506312

Uhlir, P. F. (2010). Information gulags, intellectual straightjackets, and memory holes: Three principles to guide the preservation of scientific data. *Data Science Journal,* 10, 1-5.

Vertesi, J., & Dourish, P. (2011). The value of data: considering the context of production in data economies. *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 533–542.

Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1), 5–16.