

# Libraries, Process, and Data

Anna Gold

California Polytechnic State University

San Luis Obispo, California, U.S.A.

akgold@calpoly.edu

## ABSTRACT

The metaphor of knowledge as a product is found throughout contemporary discourse. It can be argued that sometimes product, and at other times process, will be a good, useful, or even the only way something can be viewed. But when one looks at process, one ignores product; and vice versa. This is a fundamental human problem, not an oversight, and as a metaphor echoes Heisenberg's uncertainty principle that suggests we can never know position and momentum simultaneously, but are limited to knowing one or the other. I illustrate briefly how a focus on product dominates discourse on core library issues including scholarly communication and library assessment. I then explore further the implications of this focus for the emerging practice and theory of data curation, despite widespread acknowledgement of both data production and data curation as processes. I conclude by urging greater attention to the possibility and challenge of reshaping practice in data curation around attention to process.

## Keywords

Product, process, assessment, data curation, knowledge, uncertainty

## KNOWLEDGE AS PRODUCT

The metaphor of knowledge as a product is found throughout contemporary discourse. It appears in the idea of education as a product to be consumed by students, and in the trending metaphor of students as products. It figures prominently in library practice and discourse. As part of a prevailing mental model about knowledge, libraries, and universities, this metaphor tends to go unexamined (Senge, 1990). Yet it has significant implications for where we place our attention, how we set priorities, how we describe what we do, and how we assign value to what we do in education and research. Deliberately reflecting on this mental model or metaphor can help generate new ways of

framing our choices, suggest alternative priorities, and redirect our attention. Reflection on mental models takes into account broader system dynamics, and thereby opens opportunities for changing the models on which learning is based. This is the notion of "double-loop learning" (Argyris and Schon, 1978).

A vision statement produced in 2009 by the American Association of Universities and the Association of Research Libraries in partnership with two other organizations is "The University's Role in the Dissemination of Research and Scholarship — A Call to Action." The document begins by stating:

*"The **creation of new knowledge** lies at the heart of the research university and results from tremendous investments of resources by universities, federal and state governments, industry, foundations, and others. **The products of that enterprise** are created to benefit society. In the process, **those products** also advance further research and scholarship, along with the teaching and service missions of the university. Reflecting its **investments**, the academy has a responsibility to ensure the broadest possible access to **the fruits of its work** both in the short and long term by publics both local and global. "*

*"Faculty research and scholarship represent invaluable **intellectual capital**, but the **value** of that capital lies in its effective dissemination to present and future audiences. Dissemination strategies that restrict access are fundamentally at odds with the dissemination imperative inherent in the university mission." [Emphasis added] (p. 1)*

This language is not remarkable. It offers a progressive vision of expanded access to scholarly information. Yet the industrial and capitalist metaphors in this statement are striking. A primary mission of libraries - sharing research - is framed in terms dominated by industrial production: products, investments, work, intellectual capital, and value.

Even the dramatic changes brought about by the emergence of digital, networked communication and scholarship are presented as having generated "new kinds of digital products" (p. 2). The metaphor of knowledge as a product constrains this "call to action" to universities even on a topic as sweeping as the "dissemination of research and scholarship." As a result, the range of actions envisioned to enhance the dissemination of knowledge is limited to

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

This work is licensed under the Creative Commons Attribution NonCommercial 3.0 Unported License. To view a copy of this license, visit [http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en\\_US](http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en_US)



strategies for distributing traditional and new “products” to as wide a market, at as low a cost to consumers as possible.

### Long Shadow of the Industrial Revolution

Until a few decades ago, in the pre-digital but post-Gutenberg age, the circulation of knowledge and ideas by a growing global industry involved the reproduction of books, journals, and other printed works in tangible forms: ink on paper. This industry overlaid and increasingly formalized many other forms of scholarly communication that have persisted throughout history: conversation, negotiation and co-creation of ideas in break rooms, cafes, over dinner, on walks, at conferences, in classes and seminars, and lectures.

The metaphor of knowledge as product is contradicted today by everyday practices and behaviors of scientists, many of whom correspond directly and informally with others through blogs that reveal their research and work as it progresses; others openly produce and share multiple drafts and versions of their work with others, inviting comment and collaboration along the way to producing final “products.” But in today’s discussion of scholarly communication these activities are overshadowed by attention to the tangible products in circulation - books, articles, texts and papers. We silently equate these records - “products” - with knowledge.

The dominance of the knowledge-as-product metaphor in higher education has metastasized into areas of education that might have seemed immune: education itself is a product consumed by students. Students, too, are products: raw inputs to a system whose efficacy is measured by such factors as the consistency and rapidity of outputs (graduation rates, retention), the economic viability of those outputs (income of graduates, job placement of graduates), and the competitiveness of one university’s graduates against others’ according to these and other standardized measures (Cosgrave 2010).

Typical of this discourse is a 2010 article on the first National Higher Education Productivity Conference, sponsored by the Lumina Foundation for Education. The conference was aimed at spreading “ideas that will increase the number of college graduates at institutions...at a lower cost to both governments and students” (Kelderman, 2010). A national movement bolstered by new legislation and rules from several levels of government and regulatory bodies (federal, regional, state, local) calls for assessment and accountability in education through comprehensive institutional “analytics” that track student progress towards achieving “outcomes.” The assessment movement, which could be centered on human development, is frequently overshadowed by the knowledge-as-product metaphor.

Generating these tangible records, and their digital surrogates, is considered essential for scholars to establish their “productivity” as a scholar; and universities are ranked based on their aggregate scholarly productivity (see for

example the Faculty Scholarly Productivity Index from Academic Analytics).

Nor is the broad domain of research and education alone among human enterprises in experiencing commercialization and commodification. This trend is notable across many areas of human effort, from arts to health to personal services to food to entertainment and recreation. The trend toward commodification means assigning exchange value to products in the marketplace.

The logical result of accepting the commodification of scholarly production (including peer review) is suggested - in all seriousness I fear - by a faculty member who writes:

*“If academic work is to be commodified and turned into a source of profit for shareholders and for the 1 percent of the publishing world, then we should give up our archaic notions of unpaid craft labor and insist on professional compensation for our expertise, just as doctors, lawyers, and accountants do.” (Gusterson, 2012)*

Gusterson goes on to assert: “The life of the mind is not billable.” But the commodification of knowledge has evolved hand in glove with the idea of intellectual property, an idea that only became commonplace at the height of the industrial revolution, resulting in increased concern with managing and pricing exclusive property rights in those products.

In an atmosphere where the commodification of intellectual work has become pervasive, it becomes harder to recall that the contribution of an individual to knowledge may not correlate neatly with measures of “productivity.” The dominance of industrial metaphors for the production of knowledge is something of a problem for team science generally: it may be difficult and even impossible to accurately document individual credit or responsibility for collective work. The relationship between “credit” for intellectual work, and ownership of the resulting intellectual “property” is also not without its strains. The divergence of credit and property is the basis, indeed, for much university policy and for the extraordinary profit margins of some scientific publishers. Harvard Medical School notes some of the complexities involved:

*“The conduct of a scientific experiment or other research project has many components, including formulation of a hypothesis, development or application of methods, collection of data, analysis of results and creation of a public description of the work. To the degree that contributions to any of these components require not only technical skills but also intellectual input, they are appropriately recognized by authorship. However, authorship does not imply any legal ownership of an idea, method, research materials or data.”*

### **Libraries and the Production of Knowledge**

As institutions closely tied to and expressing dominant cultural models of knowledge, it is natural that the focus of libraries for at least several centuries has been on collecting, describing, arranging, and sharing the “products” of knowledge. Until very recently, the measure of a great research library in the U.S.A. has been the number and uniqueness of the products they have amassed. For example, only since 2006 has the Association of Research Libraries (ARL) begun to offer alternative library rankings based on levels of institutional investment rather than on extent of “holdings” (ARL, 2006; Thompson 2006).

Much of the distress in academic libraries about flat or diminishing library budgets stems from concern for the impact of decreased buying power on libraries’ continued ability to amass - or at least lease - these products. The seemingly endless escalation of the costs of knowledge products, and the multiplication of new products (journals, databases, value-added services), undermine the ability of libraries to maintain their role as primary providers of access to knowledge products. At the same time, some publishers demand that libraries maintain a given level of expenditure, as if publisher’s products were like a utility such as a cable service that forces consumers to buy 200 channels they don’t need or use, in order to watch their favorite hockey team.

As producers, publishers naturally seek to expand the market for their products, both to end-users generally and to specific groups such as alumni who have moved outside the circle of authenticated access. Copyright fees continue to climb, reducing the cost-effectiveness of library resource sharing through interlibrary loan; and publishers embrace acquisition models where purchases are crowd-sourced or driven by individual customers rather than by librarians.

Some librarians have responded to these pressures by creating programs to provide scholars with low- or no-cost publishing infrastructure that could, in theory, replace commercial production options. This is laudable; they are well-placed to do this, as managers of nonprofit service organizations in long-lived institutions; and librarians understand some aspects of publishing technologies pretty well. But at this point, library-owned knowledge vehicles are like small hybrid cars - well-suited to special niches of the overall ecology of transportation systems. But just as smart cars are not poised to replace our collective capital investment in SUV’s or big rigs, library-published journals are not close to replacing *Nature* or ScienceDirect.

### **Libraries and the Production of Students**

An emerging example of how academic libraries are responding to the new production-focused assessment discourse is the movement in the profession to measurably demonstrate that libraries contribute to the production of students. The gold standard in this endeavor is evidence that libraries contribute to student grade point averages,

retention in programs, degree completion rates and time to degree, and success in job placement.

While most data collected in assessment efforts like these are correlative, libraries are nevertheless pleased if they can establish that students who use the library regularly or conduct searches in library databases have higher grade point averages, and complete their degrees in a more timely way.

In one instance, the University of Wollongong undertook a project to analyze library usage by students with data about student academic performance, revealing “a strong correlation between students’ grades and the use of information resources the library provides” (Cox and Jantii, 2012). A related initiative is the Library Impact Data Project in the UK, also seeking to establish a positive relationship between student success and student use of library resources. Not surprisingly, their work suggested lower library use by students in fields of design and engineering: “We have found that usage is low in areas such as Art and Design and Computing and Engineering. Is this OK?” Apparently they think not. Rather than reflecting on what these findings suggest about the distinctive learning and development processes characteristic of these disciplines, they instead conclude that libraries need to collect more products to serve these students (Stone, 2012).

In the United States, the University of Minnesota has been at the forefront of formal, comprehensive studies that link libraries with the production of successful students. The Library Data and Student Success project aims “like any unit on campus” to make a quantitative case for the value of the library to student success. In published articles and presentations this project has reported positive correlations between student grade point averages, retention, and use of library resources (Soria et al, 2013).

Since 2009 the Association of College and Research Libraries (ACRL) has been leading a grant-funded multi-year initiative to support the demonstration of library value through metrics. The initiative produced a major report in (Oakleaf, 2010), whose language in the opening paragraph of the executive summary is quite apt:

*“Academic libraries have long enjoyed their status as the “heart of the university.” However, in recent decades, higher education environments have changed. Government officials see higher education as a national resource. Employers view higher education institutions as producers of a commodity — student learning.”*

Oakleaf lays out a research agenda for developing evidence of library impacts on ten institutional outcomes: student enrollment, student retention and graduation, student success, student achievement, student learning, student experience, faculty research productivity, faculty grants, faculty teaching, and institutional reputation.

An ACRL senior strategist has called the continuing challenge to demonstrate library value “the top issue facing our profession today” (Malenfant, 2013).

The discourse on library assessment takes place in an environment dominated by monetized expressions of value as “return on investment,” where educational leaders are increasingly focused on the business proposition of how to stay afloat, by increasing student through-put, increasing cost-effective enrollment practices (e.g. increasing international and out of state students), and decreasing the per capita cost of the education experience through MOOCs, online and hybrid learning, increased class sizes or more teaching by adjuncts.

The very nature of this discourse causes institutions to pay attention to and invest in strategies that reduce the net cost of producing students. An alternative discourse focused on learning and personal development as a process might turn libraries’ attention to their role in providing certain kinds of experiences, such as those described by George Kuh as “high-impact educational practices” (Kuh, 2008). Among the practices identified are undergraduate research, capstone projects, service-based learning, collaborative projects, learning communities, and first-year seminars. In contrast with discourse centered on “returns on investment”, the focus of high-impact practices is not on throughput but on transformation - learning. High impact processes are as likely to be collective processes as individual ones. Interestingly, research has suggested that these practices themselves lead to increased rates of student retention as well as student engagement. Further, some evidence exists that high-impact learning practices contribute to the success of underserved and first generation students (Swaner and Brownell, 2008).

The literature on library support for high-impact practices is relatively slight. Oakleaf notes that “many libraries actively support these practices. However, many libraries do not collect, document, and communicate evidence about impact of their support of high-impact practices.” How might library practices be documented and measured? To begin with libraries need to become familiar with what research suggests are the most effective high-impact practices, and to expand and document their own contributions within those practices (Finley, 2008). Libraries might also partner with faculty to develop methods to collect stories of personal transformation and change from students. Such stories could provide models not only of outcomes but also of processes that may serve as a guide or model for others.

A library that not only supports but facilitates, inspires, and partners to offer opportunities for students to experience high-impact learning practices might well be said to be contributing to the quality of the learning processes experienced by the students (Rodriguez, 2012).

## Digital Reawakening

While much professional discourse in higher education is dominated by the metaphors of knowledge as a product, and students as products, a powerful counter-metaphor of knowledge as a process now coexists and competes within that discourse. The digital revolution in how we share ideas and knowledge has reawakened our culture to two realizations: first, that knowledge is not only – in most cases – an arbitrarily chosen end product but *also* a process; and second, that every person in that process plays a role not only as a consumer of knowledge but *also* as a participant and performer in that process.

Although the points on a curve representing process could be approximated by the calculus, the mathematics for every participant and performer in this process has yet to be invented. This is one reason why the notion of process as described in this paper remains elusive. Quantifying the human element, whether through a product-dominated language or in an “unsettled mathematics” remains a task to be completed (Tao, 2009).

The digital revolution has not caused artifacts of knowledge to disappear from our experience, but it has given us a collective experience that knowledge is not only a consumable product. In Thomas Jefferson’s words:

*“He who receives ideas from me, receives instruction himself without lessening mine; as he who lights his taper at mine receives light without darkening me.”*  
(Jefferson, 1813)

In the digital age, knowledge-as-product no longer holds a monopoly on how knowledge is understood. The commodification of knowledge is undeniably a dominant part of the economy of higher education. But a powerful democratization and diffusion of education took root with the rise of the World Wide Web; exploded into a powerful global force with MIT’s high-stakes bet on OpenCourseWare; and flourishes today in forms that include Wikipedia, MathOverflow, and Hathi Trust.

We are not only witnessing an unprecedented opening up of access to learning and knowledge to more participants, but we are also reawakened to learning itself as a process, not a product.

## DATA AS PRODUCT AND PROCESS

The realm of digital data is an important arena for understanding the implications of recognizing knowledge as a product *and* as a process.

Initially the explosion of data production in the last decades has drawn attention to the new potentials of computational science to contribute to scientific knowledge. Advances in computational science have implications for the tractability of hard problems, given the capabilities of high performance computing, networked and parallel processing. They have related but distinctive implications for modeling and visualizing knowledge through simulations and

visualizations that are capable of supporting both the exploration of hypotheses, and pattern-finding. More recently, computational developments have led to new fields of research built around newly available data as evidence (such as sensor networks and the digital sky survey). A “fourth paradigm” for scientific research was proposed in a work honoring Microsoft’s visionary Jim Gray (Hey, 2009) laying out a new type of “data-intensive science” that moved beyond three pre-existing paradigms of empirical, theoretical, and computational science.

Many in the world of data-intensive and eScience have understood that one of the most significant impacts of this perspective is to refocus attention on the processes, or life cycles, of science. This entails shifting attention from the end product of scientific work, to documenting the decisions and context from which the research arises, and feedback loops throughout the research process. Unlike published textual narratives, data by its very nature lends itself to being moved around, filtered, added to, visualized, and linked with other data as part of a knowledge design process. A metaphor of data flow, including the ideas of upstream and downstream phases of that flow, illustrates the importance of decisions long before there is a knowledge “product” to be managed.

Yet the dominant initiatives developing library roles that support data-intensive science continue to emphasize data as a product, as objects that can be fixed by schemas, ontologies, and controlled vocabularies; unique identifiers; citation practices; and other standards that in Gray’s words “objectify knowledge,” beginning with “basic things like units, and what is a measurement, who took the measurement, and when the measurement was taken.” (Gray, 2007). Even as Gray envisions a desired future world of scientific communication in which data is self-describing and can be understood across domains and disciplines, Gray’s aim appears at some level to contradict his vision.

As the flows of data through their life cycles are studied, it has become more clear that digital data is not “objective” in any transcendent or permanent sense. Rather it is the outcome of agreements and decisions that have been used (with a range of skill and accuracy) to imagine and execute the use of instruments and practices that identify, gather, and record data. Its interpretation and long-term management are, equally clearly, the result of design decisions and processes that have purposes and goals. Thus even beyond the issues raised by data cycles and flows, the explosion of data science and data management in the world of information and libraries represents a fundamental challenge to the idea of knowledge as a product, revealing that data, however strictly described and constrained, is also evidence of a process - however hidden from view. At present there is no well-understood way to coax this evidence out of hiding.

It follows that a vision of a world of interconnected data, translatable and transparent across disciplines, time, space, and authors, must be complemented by an increasing understanding of the narrative nature of data, one that stretches the apparently fixed narratives of the textually encoded knowledge world, and opens up our collective ability to “tell” the data in many alternative ways.

Indeed, data literacy implies the ability to understand and create visualizations of data that expose the fluid nature of knowledge, its dependency on choices, on continuities and discontinuities in space and time. In a data-rich world, the excitement and interest aroused by effective data visualization is at least partly due to how visualization exposes, through nearly infinite options and choices, how data is always connected to a context. Teaching data visualization means teaching through example that knowledge is not only a product but also a process, and that we are players, or potential players, in how that process unfolds.

A special issue of *Nature* on scientific publishing calls attention to the impact of new NSF proposal guidelines that expand the definition of research products to encompass data as well as publications:

*“Until recently, data have been considered a second-class citizen in the science and publishing world, and that’s all about to change,” says Michener. ... A step in that direction came this year, when the NSF altered its proposal guidelines by allowing researchers to list ‘products’ that they have created, such as data sets and software, instead of just publications.” (Monastersky, 2013)*

The transformation is one that moves publishing “away from narrative...as the major output of research.” Yet fitting any data process to the Procrustean bed of “product” is complicated, as the article also describes. Best practices for achieving that fit are being developed at research libraries like Johns Hopkins University and Stanford University. Ironically, what is needed to transform process to product is itself a process: a “curation” process.

Data curation involves the processes of storing, together with the research data products (data, unique and permanent citation number), data procedures and processes, including those that guard the files from obsolescence and degradation, and files that are used to collect or process data.. The ongoing nature of data curation as a process is illustrated by cost models that propose to use a set portion of research funding to sustain data curation processes. National data management initiatives in several countries also represent major investments in sustaining long-term processes.

In the spectrum of these activities supporting data curation and data science, libraries are focused on those that reflect the deep imprint of a product perspective. These activities

include data description (metadata) and data citation practices; and data provenance.

### Metadata and Citation

According to Altman (2012), citation plays a role in assuring long-term use and reuse as well as providing “unambiguous chains of provenance” to “specific, fixed versions of data” as well as to quantitative measures of downstream usage and impact essential to incentivizing data sharing by researchers. Summarizing the findings of a workshop on the principles of data citation, Altman notes that citation of data is essential both for discovery of that data and for attribution of credit “to all contributors” by other users downstream.

Initiatives to make data citable in a consistent and unambiguous way include DataCite, a global consortium and also a service for assigning DOI’s and metadata to data sets. The complexity and manual dependencies of metadata and citation standards for data are also a concern, and researchers are working to develop more automatic metadata schemes that leverage identifiers for persons (ORCID) and organizations or institutions, for related articles (DOI’s), as well as data objects (URI, DOI, Handles) (Qin et al., 2013).

In the same presentation Altman asks, “so when do things go wrong” in short and long-term data curation for reuse, replication, and for open data generally? Altman’s answers begin with several flavors of researcher failure, including failure to manage data during research, failure to capture knowledge required for replication/interpretation, and “failure to capture tacit knowledge required for understanding.”

However, the researcher is being tarred unfairly with the brush of “failure” when charged with failing to “capture tacit knowledge required for understanding.” While tacit knowledge can be “told” in part with narratives, and in retrospect, it may well evade capture. It is encoded in and part of complex human processes, and it may not be possible to make it explicit.

According to some who study collective intelligence, “the stuff that matters” never makes it to documentation, but can best – and maybe only – be conveyed face to face (Pentland, 2007). In this sense, only as long as the scientist lives and can be interrogated can “reasonable detail” be acquired about the data.

Unlike mathematical and logical systems that are theoretically subject to Turing’s “halting problem” (the undecidability of whether a computer program will halt), human beings, as we know, all halt, eventually. There is, of course, no solution to the problem of death. Once the expert “halts,” so too does most of the “stuff that matters.” The database, however richly adorned with metadata, becomes a mass of questions and questions about questions.

One approach is to encode “the stuff that matters,” not directly in the data or database, but as a structured narrative. This may sound suspiciously like a journal article, and Phil Bourne has long claimed that journal articles provide essential narratives about science, and should be linked to the data.

There may be ways of making the scientific narrative more open to interrogation, whether by person or by machine. For example, systems can be developed that support scientists in documenting their questions and choices. Jane Hunter has proposed something like this in her notion of “publication packages,” consisting of:

*“...selective encapsulation of raw data, derived products, algorithms, software and textual publications...encapsulating expert knowledge.”*  
(Hunter, 2006)

Note in this passage the two words: “selective” and “expert.” The notion of selectivity points to the notion that that we can’t encapsulate or capture all the data, or all the expertise. At the same time, it may be possible to use mark-up languages within narrative documents to flag choices, such as descriptions of failed paths and ambiguous results. So there are some interesting technical possibilities for describing processes rather than outcomes.

### Provenance

Bowker addresses the aspect of “context” and its importance for understanding data:

*“We need to retain the context of development of a given database in reasonable detail; the political and social and scientific contexts of a set of names and data structures are all of interest. I emphasize reasonable detail here: a perfect archival system is a chimera.”* (Bowker, 2006)

The prevailing information paradigm leads to efforts, such as the PROV standard created by the W3C, to create strict descriptions of the context of data in the form of standard descriptions of data provenance.

The goal of creating these standard descriptions is to make it possible to exchange, query, and otherwise compute and process these descriptions: in other words, to turn messy contexts into data products. In the vocabulary of the emerging W3C standard for describing the provenance of data, that context includes what processes were executed, and by whom, in what role, for what use, etc.:

*“Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. ... PROV enables one to represent and interchange provenance information ... In addition, it provides definitions for accessing provenance information, validating it, and mapping to Dublin Core.”*



Missing from PROV is the concept of the provenance of provenance data - any reflexive awareness that data about the provenance of data is itself a design, with its own context and narrative of decisions and goals.

Further, as it stands today, the PROV standard reduces the process-based temporal, spatial, and human context of data to a set of descriptive facets. In order to support handling a process as though it were a product, the narrative features of process are stripped away. Provenance indirectly describes process by describing the identity of its source.

The goal of documenting the provenance of data is ambitious, aiming towards the goal of reproducing the research process itself. According to Wikipedia,

*“Scientific research is generally held to be of good provenance when it is documented in detail sufficient to allow reproducibility. Scientific workflows assist scientists and programmers with tracking their data through all transformations, analyses, and interpretations. Data sets are reliable when the process used to create them are reproducible and analyzable for defects.”*

In other words, reproducibility of data as a product requires very special and complex metadata that “captures” the processes that yielded the data. Yet,

*“What is needed is a record of processes as well as a record of facts. However, processes and facts cannot be in principle disentangled, so we are never going to have a perfect data set wrapped in complete metadata. Moreover, the processes that we need to record in order to ensure the viability of data in the long run do not constitute an easily enumerable set.” (Bowker, 2006)*

Some argue that the purpose of data curation is reproducibility. Clifford Lynch implies this, writing that:

*“The scientific record... should make enough data available, and contain enough information about methods and practices, that another scientist could reproduce the same results starting from the same data...”*

Yet Lynch goes on to note that:

*“... the ideal of reproducibility for sophisticated experimental science often becomes problematic over long periods of time: reproducing experimental work may require a considerable amount of tacit knowledge that was part of common scientific practice and the technology base at the time the experiment was first carried out but that may be challenging and time consuming to reproduce many decades later.” (Lynch, pp. 177-183, in Hey, 2009)*

Several medical researchers comment that “[m]any

landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models” and that

*“[i]n the complex data sets now being generated by genomics and proteomics, the nature of reproducibility is changing...The basic meaning of scientific knowledge is shifting, and the community must learn to deal with it in ways that go beyond simple semantics.” (Begley and Ellis, 2012)*

Constructivists suggest that no matter how well it has been documented, data will always remain an incomplete representation of decisions and choices. As Gitelman observes, data needs to be imagined as data in the first instance, and this process of the *imagination of data itself* entails an interpretative base: “Every discipline and disciplinary institution has its own norms and standards for the imagination of data.” (Gitelman, 2013; see also Pickering, 1984, for a canonical account of how scientific knowledge emerges from scientific practice.)

## DATA PROCESS AS DATA STORIES, QUESTIONS, AND SCAFFOLDING

Is there a process-centered alternative to product-oriented data curation? While process-centered alternatives are still in their infancy, some efforts are worth noting, among them Altmetrics.org (Priem et al., 2010).

One is the development of alternatives to formal citation, based on measuring the impacts of science and data through “altmetrics,” aggregated from multiple sources, many of them semantic. One example of an altmetric is the “ImpactStory,” providing contextualized information about data and science and their impact:

*“Citation measures are narrow; influential work may remain uncited. These metrics are narrow; they neglect impact outside the academy, and also ignore the context and reasons for citation.” (Galligan and Dyas-Correia, 2013)*

The reframing of products as “versions” may be another bridge to a more process oriented perspective on data. This orientation has recently been expounded by Herbert Von De Sompel (2013).

Another perspective is the suggestion of Brett Stalbaum that there is a continuum of relationships that we have with data that go far beyond relationships of ownership, authorship, or use. Rather, Stalbaum suggests we rethink our relationship with data in terms of processes that include participation, prediction, and exploration:

*“[T]he role that data plays ... represents a progression in the cultural role of data from relatively static descriptor, through that of active semiotic agent, and onto data itself being a type of unexplored and*

*uncertain context in which developing well formed questions is the primary, and very interesting, problem."*

Stalbaum goes on to suggest that databases can be thought of as "question spaces" – the space of all possible questions – and he acknowledges that this is a daunting prospect:

*"[T]he problem of not having well formed questions about vast data sets is in fact one of the most provocative and unexplored problems facing humanity as our ability to collect data outpaces our ability to process it and derive new knowledge from it."*

The centrality of relationships and process, of inquiry and context, is difficult to encode but resonates compellingly with an understanding of knowledge not only as a product, but also as a process. The corollary to this is the opportunity to participate in that process.

A primary reason to curate data, as for curating other digital and tangible artifacts, is not to accumulate, control or manage products. Rather the work of curation is a form of participation in the process of knowledge: libraries and their collections are both systems and scaffolding for knowledge processes. The goal of that work, like other "scaffolding" work, is to help us think.

Data curation is a creative enterprise responsible for producing a question space from scientific choices and measures of the world that enable both new narratives and new questions about the world. As we navigate the fractal shores of data we will come to understand that curators of data are involved in an effort that is fundamentally creative, full of human expression and choice.

On the one hand this should humble us because the data and the databases that we curate will always and for many different reasons be incomplete and uncertain. On the other hand it should inspire us. The choices made by data curators – participants in the knowledge process - truly matter. Their imaginations and their choices, together with the questions and choices of researchers, help constitute a world that they will change.

## CONCLUSION

This article begins to touch on the complexity and difficulty inherent in shifting our attention on knowledge as a product to its complementary dimension as a process. A deeper research program is planned that will describe and probe what is gained in shifting to a process metaphor in libraries. The familiarity and ubiquity of the product metaphor for knowledge as well as the inherent limitations of the English language and its grammar to describe process, make the product metaphor for knowledge both tractable and seductive. But it can be argued that sometimes product, and at other times process, will be a good, useful, or even the only way something can be viewed. When one looks at process, one ignores product; and vice versa. This is a fundamental human problem, not an oversight, echoing

Heisenberg's uncertainty principle that suggests we can never know position and momentum simultaneously, but are limited to knowing one or the other.

## REFERENCES

- Academic Analytics (2013). *Faculty Scholarly Productivity Index*. Retrieved April 23, 2013 from <http://www.academicanalytics.com/Public/WhatWeDo>
- Altman, M. (2011). *Data Sharing and Data Citation, September 16, 2011*. Retrieved April 23, 2013 from <http://www.slideshare.net/drmaltman/data-sharing-data-citation>
- Argyris, C. & Schon, D. A. (1978). *Organizational Learning: A Theory of Action Perspective*. Reading, Mass.: Addison-Wesley.
- Association of Research Libraries et. al. (2009). *The University's Role in the Dissemination of Research and Scholarship — A Call to Action*. Washington D.C.: Association of Research Libraries.
- Association of Research Libraries (2013). *ARL Ranking*. Retrieved April 16, 2013 from [http://www.arlstatistics.org/about/arl\\_index](http://www.arlstatistics.org/about/arl_index)
- Begley, C. & Ellis, L. (2012). "Drug development: Raise standards for preclinical cancer research." *Nature* 483, 531-533.
- Bowker, G. (2006). *Memory Practices in the Sciences*. Cambridge: MIT Press.
- Cosgrave, R. (2010). October 5, 2010, "Do you see your students as products?" Retrieved April 15, 2013 from <http://tertiary21.blogspot.com/2010/10/do-you-see-your-students-as-products.html>
- Cox, B. & Jantti, M. (2012). "Discovering the impact of library use and student performance," *Educause Review Online*, July 17, 2012. Retrieved April 15 2013 from: <http://www.educause.edu/ero/article/discovering-impact-library-use-and-student-performance>
- Finley, A. (2008). *Assessment of High-Impact Practices: Using Findings to Drive Change in the Compass Project*. Washington, D.C.: AAC&U.
- Galligan, F., and Dyas-Correia, S. "Altmetrics: Rethinking the Way We Measure", *Serials Review* 39, 56-61.
- Gitelman, L., ed. (2013) *"Raw Data" is an Oxymoron*. Cambridge: MIT Press.
- Gray, J. (2007). "Jim Gray on eScience: A Transformed Scientific Method." In: Hey, T. et al (2009), *The Fourth Paradigm*. (Based on the transcript of a talk given by Jim Gray to the NRC-CSTB1 in Mountain View, CA, on January 11, 2007.)



- Gusterson, H. (2012). "Want to change academic publishing? Just say no." *Chronicle of Higher Education*, September 23, 2012.
- Harvard Medical School (n.d.). *Guidelines for Attribution of Credit and Disposition of Research Products*. Retrieved April 29, 2013 from <http://hms.harvard.edu/content/guidelines-attribution-credit-and-disposition-research-products>
- Hey, T. et al. (2009). *The Fourth Paradigm: Data Intensive Scientific Discovery*. Microsoft Research. Retrieved April 23, 2013 from <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>
- Hunter, J. (2006). "Scientific publication packages – A selective approach to the communication and archival of scientific output." *International Journal of Digital Curation* 1, 33-52.
- Jefferson, T. (1813). *Letter to Isaac McPherson*. Retrieved April 4, 2013 from [http://press-pubs.uchicago.edu/founders/documents/a1\\_8\\_8s12.html](http://press-pubs.uchicago.edu/founders/documents/a1_8_8s12.html)
- Kelderman, B. (2010). "With revenue drying up, educators look to productivity as the way to serve more students," *Chronicle of Higher Education*, November 15, 2010.
- Kuh, G. (2008). *High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter*. Washington DC: AAC&U. Retrieved April 15, 2013 from <http://www.aacu.org/leap/hip.cfm>
- JISC (n.d.). *Library Impact Data Project*. Retrieved April 13, 2013 from [http://www.jisc.ac.uk/whatwedo/programmes/di\\_informationandlibraries/emergingopportunities/lidpphase2.aspx](http://www.jisc.ac.uk/whatwedo/programmes/di_informationandlibraries/emergingopportunities/lidpphase2.aspx)
- Malenfant, K. (2013). "Demonstrating the value of the library and librarians: The top issue facing our profession today." Michigan Academic Library Council 2013 Spring Workshop on Imperatives of Library Leadership. (Flint, MI, March 15, 2013).
- Monastersky, R. (2013). "The Library Reboot." *Nature* v. 495, March 28, 2013, pp 430-432. Retrieved April 23, 2013 from [nature.com/scipublishing](http://nature.com/scipublishing)
- Oakleaf, M. (2010). *The Value of Academic Libraries: A Comprehensive Research Review and Report*. Chicago: Association of College and Research Libraries. Retrieved April 29, 2013 from [http://www.ala.org/acrl/sites/ala.org.acrl/files/content/issues/value/val\\_summary.pdf](http://www.ala.org/acrl/sites/ala.org.acrl/files/content/issues/value/val_summary.pdf)
- Pentland, S. (2007). *Communications Forum, Collective Intelligence, October 4, 2007*. Retrieved April 20, 2013 from [http://web.mit.edu/comm-forum/forums/collective\\_intelligence.html](http://web.mit.edu/comm-forum/forums/collective_intelligence.html)
- Pickering, A. (1984). *Constructing Quarks: A Sociological History of Particle Physics*. Chicago: University of Chicago Press.
- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto* (v.1.0), 26 October 2010. Retrieved April 29, 2013 from <http://altmetrics.org/manifesto>
- Rodriguez, D. (2012). "Answering questions about library impact on student learning," blog posting, April 4, 2012. Retrieved April 16, 2013 from <http://www.inthelibrarywiththeleadpipe.org>
- Senge, P. (1990). *The Fifth Discipline: The Art and Practice of the Learning Organization*. New York: Doubleday/Currency.
- Soria, K., Fransen, J. & Nackerud, S. (2013). "Library use and undergraduate student outcomes: new evidence for students' retention and academic success." *portal: Libraries and the Academy* (forthcoming). Retrieved April 23, 2013 from <http://purl.umh.edu/143312>.
- Stalbaum, B. (n.d.). "The landscape and culture: Data as mediator, modes of engagement." Retrieved April 29, 2013 from <http://www.c5corp.com/research/landscapeculture.shtml>
- Stone, G. (2012). "What should be our focus for future work on the LIDP?" December 18, 2012. Retrieved April 29, 2013 from <http://library.hud.ac.uk/blogs/projects/lidp/>
- Swaner, L. & Brownell, J. (2008). *Outcomes of High Impact Practices for Underserved Students: A Review of the Literature*. Washington DC: American Association of Colleges and Universities.
- Tao, T. (2009). "Feynman's lectures online," July 15, 2009. Retrieved August 13, 2013 from <http://terrytao.wordpress.com/2009/07/15/feynmans-lectures-online/>
- Thompson, B. (2006). "Some alternative quantitative library activity descriptions/statistics that supplement the ARL logarithmic index." Houston: Texas A&M University. Retrieved April 17, 2013 from [http://www.libqual.org/documents/admin/2012/Thompson\\_2006\\_Some\\_Alternative\\_Quantitative\\_Library\\_Activity\\_Descriptions.pdf](http://www.libqual.org/documents/admin/2012/Thompson_2006_Some_Alternative_Quantitative_Library_Activity_Descriptions.pdf)
- Von de Sompel, H. (2013). "From a version of record to a version of the record," CNI Spring Membership Meeting, April 2013. Retrieved April 23, 2013 from <http://www.cni.org/news/video-van-de-sompels-plenary-from-the-version-of-record-to-a-version-of-the-record/>