June 4, 2008

The island of our knowledge, the shore of our ignorance

Anna K Gold
"As the island of our knowledge grows, so does the shore of our ignorance."

Thank you (Dean Unsworth; Melissa) – title of talk – reflects the theme that we’re learning about something that’s still largely unknown – and in other words – we find that:

“the more we know, the more there is to know”

What I have planned for this morning is a bit more abstract than much of the Institute has been, and is in some ways an extended meditation on data curation and its meaning for libraries. It’s going to get a little far out at the end of my talk, and I hope that will be OK and even a little fun.

I’ll start with a little context, and then go on to talk in more detail about four problems for libraries when it comes to applying our traditional skills and knowledge to the issue of data curation.

Next, I’d like to pose three questions that I hope can offer some new and perhaps provocative perspectives on the roles libraries have with respect to data.

Then I’ll wrap up and we should have plenty of time for conversation together.

Context and overview
Critical reflection - four problems
Three provocative questions
Wrapping up

So to begin with – let’s acknowledge that large claims are being made for the importance of data curation to the future of research libraries.
SLIDE 7:

This quotation taken from the ARL Task Force Report on E-Science, is illustrative:

"Given the core library mission of preservation and the well-established role in the provision of access to information sources, there is considerable potential for library involvement in data curation." (Agenda for Developing E-Science in Research Libraries: ARL Joint Task Force on Library Support for E-Science Final Report & Recommendations, Nov. 2007 [PDF])

SLIDE 8:

We are all here today in part because we have some stake in these claims and hopes.

SLIDE 9:

Here at the Institute this week you’ve shared your experience, and learned about tools and approaches to various aspects of data curation.

SLIDE 10:

Institute participants and others here this morning are probably also drawn to this topic by intellectual curiosity, as well as the conviction that understanding the challenge of data curation is becoming essential for libraries.

I consider myself one of these curious people, and I too am convinced that data curation is an important challenge to the future of libraries.

What I hope I can contribute today is to open up the conversation to some broad speculation and questions.

SLIDE 11:

Although I’m not an expert in data curation. I have been interested for a long time in the relationship between human systems and machine systems.

SLIDE 12:

Also, as a library manager I’ve felt it was important to understand the transformational challenge of data-intensive science on future directions of librarianship.

Most recently when I was at MIT I formed a data study group composed, initially, of science and engineering librarians. It was through this group that I met Melissa, and we’ve had a number of opportunities to work and talk together since then. I also participated (on behalf of MIT Libraries) in the ARL E-Science Working Group whose report was released last November.

SLIDE 13:

So all of us here are in the middle of a journey of professional transformation.
We’ve been getting experience with practical, experimental, and collaborative work in the area of data. We’ve been finding people to share experiences with, finding ways to describe and share our experience and ideas, to identify the new skills and knowledge needed to work with data, and to assess the scalability and success of our efforts.

These experiences and activities are what the GSLIS program in data curation, and this Institute, are all about. There is very little I can add to what you’ve already experienced and learned, in the way of practical experience, reflection or synthesis.

So what I’d like to do today is a little different. I’d like to step back and offer my own version of a “critical reflection” on data, with the hope that it may inspire others to reflect similarly.

**SLIDE 14:**

**Critical reflection – four problems**

**SLIDE 15:**

Why is critical reflection important?

Without critical reflection, we are tempted think of the data curation challenge for libraries as a mapping or matching game. Our temptation is to lift up the template of our traditional skills, mission, values, and capabilities, and fit these onto digital data as though data were just another genre.

I say this is a “temptation” because I think trying to apply this template to the challenges of data curation will quickly overwhelm us.

This may sound heretical to some of you, because research libraries have been making the case that there’s a goodness of fit between our traditional skills, mission, values, and capabilities, and the demands of data curation.

**SLIDE 16:**

Indeed the ARL E-Science Task Force made the following claim:

“A case can be made that research libraries already have existing capacity and expertise that they can bring to bear to support e-science. By virtue of their experience in service and data management, and, for many, their mission, they are capable of advising and developing infrastructure to support the needs of scientists working in a cyberinfrastructure-enabled environment.”  (p. 13)

It’s probably fair to call this claim a bit tentative. We are all in the process of testing it. But if we find it problematic, I don’t think we need to panic, because I actually don't think we need to stake our claim on such mappings.

I believe that we need to use the challenges of data curation, and this ‘heretical’ departure mentioned above--to help us stake out claims elsewhere. I’ll try to offer some ideas on that “elsewhere,” though at this point I have more questions than answers.
I’ll begin by describing four problems that I see with mapping our old practices and skills to data curation.

**SLIDE 17:**

Here are the four problems I see:

1. **Mass** (how much data there is, and where it is)
2. **Abundance** (how various data is)
3. **Networks** (connections among data and people)
4. **Ontology** (what data has to do with the world)

**SLIDE 18:** (transition only)

**SLIDE 19:**

**Problem (1) Mass:** One estimate I’ve heard recently is that the amount of data produced doubles every year. A new report written in March 2008 by IDC and published by EMC says this about the volume of data produced:

**SLIDE 20:**

“With a compound annual growth rate of almost 60%, the digital universe is growing faster and is projected to be nearly 1.8 zettabytes (1,800 exabytes) in 2011, a 10-fold increase over the next five years.

**SLIDE 21:**

We've see similar statements about scientific data. This one is representative.

**SLIDE 22:**

So there’s not only a huge amount of data but also a question of data density – how much data there is relative to available space.

Again, from the 2008 IDC report,

“The amount of information created, captured, or replicated exceeded available storage for the first time in 2007. Not all information created and transmitted gets stored, but by 2011, almost half of the digital universe will not have a permanent home.”

**SLIDE 23:**

All of this is leading to the diffusion of data across what are being called data “clouds,” managed by major outfits like Google and IBM, Amazon, and Microsoft.

**SLIDE 24:**

So here’s my conclusion: While Libraries have been able, collectively, to directly manage the collection of virtually all publications of interest to research, it’s beyond doubt that **libraries will not be able to map their historic collection**
management role to all of data. Libraries will never be able to directly manage all the mass of data of interest to research. Others will do this, to the extent it can be done at all.

SLIDE 25: (transition only)

SLIDE 26:

Problem (2) Abundance: Abundance in one sense is the opposite of scarcity. Institutions whose mission and programs have been based on scarcity of information, will find that mission tested in an age like this of information abundance. This includes libraries.

In this simplest sense, libraries that have played a critical role providing access to scarce information will not be able to map this traditional role of providing access to a world where data are abundant. So we may want to ask how scarce or abundant, in this sense, the data is that we’re curating.

Another meaning of “abundance” is “diversity” or “multifariousness”. While digital data are all bits and bytes, the abundance of types of data including their purposes, data structures, standards, and the like is extreme.

SLIDE 27:

From the IDC/EMC March 2008 report:

“The diversity of the digital universe can be seen in the variability of file sizes, from 6 gigabyte movies on DVD to 128-bit signals from RFID tags. Because of the growth of VoIP, sensors, and RFID, the number of electronic information “containers” — files, images, packets, tag contents — is growing 50% faster than the number of gigabytes. The information created in 2011 will be contained in more than 20 quadrillion — 20 million billion — of such containers...Dealing with the digital universe is not a technical problem alone.”

SLIDE 28:

Liz Lyon, in the report she authored last year called “Dealing with Data,” identified these as just some of the kinds, purposes, categories, and types of data.

So we have many different types of data and many different types of containers for data.

Still another aspect of data abundance is that data may not be self-similar. The same data may, like isotopes in chemistry, act differently and effectively be different when stored in different databases.

SLIDE 29:

My conclusion is that, considering abundance in this meaning of “multifariousness,” the experience of libraries in “controlling” information recorded in traditional media (books, journals) is unlikely to take us very far with data. Libraries' traditional strategies of applying standardized description and identifiers may not be generally applicable to controlling this diversity of data.
SLIDE 31:

**Problem (3) Networks:** Networked technology is already changing data access and reuse in the same way that it has revolutionized access and reuse of text, images, and video.

SLIDE 32: (Swivel)

For example Gapminder, which is based on Trendanalyzer technology, Swivel, Data360, and OpenWetWare, are each examples of the use of web-enabled social networks to manage and reuse data.

Data 360 writes on their web site:

> "Updating, analyzing and commenting on the data which defines our world is a massive challenge. We are seeking out other curious and analytical individuals... who want to provide objective information and commentary about their current situation, real problems and real progress." (Data360)

This networking will almost certainly change both the data itself and people’s relationship to data. Networking will increase not only the distribution of data, but also increase awareness of data.

SLIDE 33: (OpenWetWare)

Also, the network is self-reinforcing, as we’ve seen with both text and video: increasing the public reward and recognition for the production of data will probably encourage data reuse, correction and feedback.

SLIDE 34: (Palimpsest)

Here’s another variation on the data network: As reported by Wired magazine in January of this year, Google’s Palimpsest project aims at providing very massive, networked open source data storage for scientists, complete with annotation features. And where the digital network becomes awkward for massive data transfer, Google has joined their data network via the real-world sneakernet of UPS trucks and hard drives. [ref: http://blog.wired.com/wiredscience/2008/01/google-to-provi.html]

SLIDE 35:

My conclusion is that the broad adoption of unmediated peer-to-peer data collection, sharing, re-use, and critique, along with the emergence of widely-acquired data literacy skills, **calls into question whether the traditional library role in selection and mediation of access to content can be mapped usefully onto the mass and abundance of data and the immense networks that will connect them.**

SLIDE 36: transition only
Here’s my last problem, and in some ways the most abstract but I think also the most interesting.

The problem (4) is one of ontology. When we’re dealing with data, what relationships do we believe the data have to people, and to the world?

I’m not talking about the problem of epistemology – how knowledge is acquired, what people know, or what is “truth.” This is a different problem: what relationship we believe data has to the world.

There’s an informal ontology many of us use, that I’ll call “naïve” or “uncritical.” It is grounded in the idea is that there is a real world, and there is real data, and the data corresponds to the world, with the correct correspondences being made through indexing.

In this naïve view, data is a commodity that translates pretty directly into knowledge. The more data we have, the more we know about the world. Collect it all in one place and you “have” knowledge.

As I say, the correspondence view of data acknowledges the importance of making explicit the linkages between the data and the world. But its focus is on capturing data that in turn captures reality.

Now, even if we know that data doesn’t really “capture” reality, this way of thinking about data pervades the way we talk about it. So what’s the alternative?

Another informal ontology we hear a lot about is that data doesn’t correspond, even indirectly, to the world. Instead data is seen as a construction of human ideas and choices.

(Among the well-known proponents of this perspective in our field are Leigh Star and Geoffrey Bowker, co-authors of Sorting things out; Bowker is also the author of the recent “Memory practices in the sciences”).

As Bowker writes,

“There is no such thing as pure data – as Hacking (1995) has shown, all categories come ‘under a description,’ and data comes in a dizzying set of categorical bins. You always have to know some context.” (Bowker 2007 p. 116)
SLIDE 42:

A reasonable name for this perspective is “constructivist.” As in the correspondence perspective, the relationship between the data and the world must be made explicit.

And as in the constructivist perspective, that work is once again done by metadata (indexing, description, etc.).

<ALT>

SLIDE 43:

Liz Lyon again, expresses this view, saying that:

“Data can be viewed conceptually as a social construct evolving from a theoretical basis, and many layers of subsequent interpretation act upon and transform the data.” (Lyon)

And Bowker writes,

“The problem of what data to retain in order to keep a data set live (sic) is a metadata problem.” (Bowker p. 122)

SLIDE 44:

The constructivists actually make an excellent case for data curation and for metadata that goes beyond naïve “indexing” to encompass description and coding to express how the data and world theory are related.

As outlined by Bowker here, this is not so easy to do. It may require nearly superhuman, complicated efforts.

“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, p. 177)

Many people thinking about these things recognize these as major challenges. But it’s common for them to conclude that we must work towards conquering these challenges, even while acknowledging the path will be difficult and the product imperfect.

After all, if there is no alternative, one soldiers on.

SLIDE 45:

The conclusion I’ve come to however is a little different: it’s that **libraries need an alternative ontology to inform how we think about representing the relationships between people, data, and the world.**
Why? The naïve view of data simply doesn’t hold up to the constructivist critique; yet there are problems with the constructivist ontology too.

SLIDE 46:

Specifically, in the constructivist perspective, the “world” has disappeared. It’s been replaced with a “world theory.” The constructivist perspective omits the world as a reality independent of our constructions, because epistemologically, that world cannot be directly known.

But as I see it, the constructivist ontology gives up too easily on the idea that data has some degree of ontological independence from human constructs.

SLIDE 47:

It appears to me to accept that we humans live in a cave of perceptions that we cannot leave, but can only describe with greater and greater levels of detail.

And, notably, it appears to omit the possibility that our perceptions and actions modify that independent reality.

I think that we need an ontology that allows for the world to exist independently of us while also allowing that our perceptions and actions actually modify the world. We can call this a “participatory” ontology or as I have here, a “constitutive” one.

SLIDE 48:

This perspective proposes that the person and the data work together in a process that creates (constitutes) both realities and potentialities in the world.

Those of you who are familiar with the work of Bruno Latour will see a similarity to his “actor-network” theory. Actor-network theory allows that the objects we create have independent “agency” in the world, even if they owe their existence to our actions.

SLIDE 49:

In terms of data, another way of thinking about this is that data neither correspond to reality, nor are they projections of theory, but instead they represent choices that shape reality, including the possibilities of further choices.

This is not merely a metaphorical idea; it’s grounded actually in the ontological perspective that the physicist John Archibald Wheeler referred to memorably as “It from Bit.”

SLIDE 50:

As Wheeler wrote,

"every particle, every field of force, even the space-time continuum itself – derives its function, its meaning, its very existence entirely – even if in some
I’m going to come back to this idea as it relates to data curation in a few minutes.

3. Three provocative questions

But now I’d like to pose three questions. I hope these will provoke some new ideas, both critical and reflective, about data and data curation.

SLIDE 53:

Question #1: what about the database?

I’ve been thinking about the fact that in talking about data curation there may be a tendency to focus on the data, as if the data existed in some uncontained form separate from the database.

This got me thinking that in a traditional library, a librarian may have an interest in words, but the object of library curation is the book – the result of someone’s assembly and presentation of those words into a work, or narrative.

Analogously, I thought, that while scholars are interested in words, and scientists in data (to oversimplify greatly), it is worth thinking about whether it is really the data that interests the data curator, or rather the database.

So what is a database? Well, it is a structured collection of data. Unlike the book – a structured collection of words – the database as a creation may be finished, or not: is it the same database, whether you add or remove data from it?

And, the database is an “artistic” thing too: databases have to be “designed,” and the way they are designed also expresses an ontology – the relationships among things, people, and information.

The database thus is both a concrete thing, and an abstract thing: both an entity, and a “work.” A database is inherently composite, and can be decomposed: it includes not only data but also data structures, computer programs, and logic. Databases use different models and different programs to organize and access data. These in turn make certain inquiries and processes possible, and others not possible.
Lev Manovich, a professor at UC San Diego of visual arts, published an influential 2002 book on *The Language of New Media*. In it, he calls "database" one of two forms of new media. He claims that database is a cultural form that stands in contrast with narrative:

He writes:

"As a cultural form, the database represents the world as a list of items, and it refuses to order this list. In contrast, narrative creates a cause-and-effect trajectory of seemingly unordered items (events)." (219, 225)

Manovich says that narrative and database compete to make meaning of the world. Narratives – order, pathways, including interactive or branching pathways – can be and often are imposed or layered over the database.

Manovich also argues that the database underlies all narrative. His point is that the computer's logic, PLUS the data creates the ability to produce endless variations of elements.

Thus transforming input to create new output “becomes the logic of culture at large,” and the database becomes the structure whose encoded choices shape the narratives that are possible to construct using the data.

Why is this important? For me, it makes me ask the question, in thinking about curating data, what attention is needed to the database, and also what attention is needed to the narrative?

SLIDE 55:

Several new media theorists, including Katherine Hayles, have questioned whether narrative and database are actually “opposed” as Manovich says, arguing rather that the two forms require each other:

Because database can construct relational juxtapositions but is helpless to interpret or explain them, it needs narrative to make its results meaningful. Narrative, for its part, needs database in the computationally intensive culture of the new millennium to enhance its cultural authority and test the generality of its insights. ... [D]atabase catalyzes and indeed demands narrative’s reappearance as soon as meaning and interpretation are required.


SLIDE 56:

Brett Stalbaum, also at UC San Diego, is another new media theorist. Stalbaum suggests that databases can be thought of as "question spaces" – the space of all possible questions – and he acknowledges that this brings daunting challenges:

“ the 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.”

"http://www.c5corp.com/research/databasellogic.shtml" and
"http://www.c5corp.com/research/landscapeculture.shtml"

SLIDE 57:

So some answers to this question, “what about the database?” remind us that the database and the data are not the same; that database and narrative stand in some sort of interdependent relationship; that metadata is not narrative; and that the database as a “question space,” is greater than the sum of its data.

SLIDE 58:

Question #2: what’s the trouble with databases?

The constructivists suggest that data, no matter how well documented is an incomplete representation of decisions and choices:

Jonathan Freedman, another new media critic, writes:

To celebrate the branching, rooting, rhizomic, proliferating quality of database—to celebrate database as a kind of autonomous form, rooting and branching by a logic of its own—is ... to downplay the inclusions, exclusions, choices that have gone into the making of databases and hence to occlude the possibilities for questioning those choices. Not to get too Frankfurt school about it, but the seeming conditions of our freedom—our increasing access to a world of information—only conceal our greater constraint.” (PMLA 2007)

SLIDE 59:

But what if we leave aside the constructivist issues and consider the status of the data and the database in mathematical terms?

Well, data can be reduced to a little bit of arithmetic, and as such are part of a larger formal - logical - system. Gödel proved that some of these formal systems are incomplete.

SLIDE 60:

So bear with me here.

I’m going to take a leap, and conjecture that databases are formal systems of this type, and to the extent they are consistent, they may be incomplete.

One reason for my thinking this is that people researching whether it’s possible to have a generalized optimization program for databases, especially for very large databases, consider the problem undecidable: this is a dilemma similar to Turing’s halting problem, which in turn is related to Godel’s first incompleteness theorem.

So some formal systems, including databases, that are consistent, may not be complete. And we will never know the correctness of certain kinds of data – even if we try to shoe-horn a larger system into the data through metadata.
Further, a query in a database is really about the data, not about what the data represents.

To be able to query the data and get a result that is about the world requires that the data include an analog for every entity in the domain of interest as well as every relationship that it participates in.

This is an ideal state in which the database is a true analog of the domain and no choices have been made. [see Levesque and Brachman, 1985]

So, databases may be incomplete. We don’t have to be constructivists or refer to the social construction of the data to come to that conclusion. It’s in the math as well as in the distinction between the data and what the data represents.

Helen Couclelis, a geologist at UC Santa Barbara, has said something similar in her 2003 article on geographic information systems. Couclelis writes:

“It turns out that there is a surprising number of things we cannot know (or questions we cannot answer) that are not the result of imperfect information. Forms of not knowing are pervasive in domains as diverse as mathematics, logic, physics, and linguistics, and are apparently irreducible. This being the case it may help to explore how these realms of ignorance may affect our efforts... I argue for accepting that uncertainty is an intrinsic property of complex knowledge, and not just a flaw that needs to be excised.” (p. 165)

So the trouble with databases is that they may be incomplete, and they are lossy.

Question #3: do questions create reality?

Earlier I put up a brief quotation from the physicist John Archibald Wheeler expressing the theory known as “it from bit.”

Quantum physics interpreted by Wheeler suggests that every aspect of reality derives from “bits” – binary choices, answers to questions. The question creates reality, it does not “find” it. One name for this is “digital ontology.” I think another name for it could be “constitutive” – the ontology I suggested above. Wheeler referred to the implication that we live in what he calls “a participatory universe.”

The idea is that not only data, but the reality of the physical world doesn’t exist until an observation or measurement – a choice – has been made. In a quantum physical sense, the world results from such observations and measurements.
This suggests – as I was hinting just now - that **queries**, or **choices** - deserve greater centrality in a data curation system. If the questions asked literally create data – results - then we should foreground the questions asked, and the choices made, as much as the results (data).

**SLIDE 66 (Bowker):**

One way of acknowledging the importance of these choices is to encode them as metadata. Bowker for example argues that, though the result will be imperfect, this is exactly what is needed:

“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 183-184)

Yet metadata is also data, so you may have an infinite regress started here. Doesn’t the metadata itself require documentation of context with “reasonable detail”?

**SLIDE 67:**

Also, 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 (Sandy Pentland, Communications Forum, Collective Intelligence, October 4, 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 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 this problem of human 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.

But possibly there are ways we can address the problem once we recognize it.

**SLIDE 68:**

Harking back to the database-narrative debate above, 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. In fact Phil Bourne at UC San Diego who is a key figure in the Protein Data Bank, has claimed that journal articles provide essential narratives about protein structures and should be linked to the data.
SLIDE 69:

There may also 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 within “scientific publication packages”. (Hunter, 2006)

Note in this passage the two words: “selective” and “expert.” The notion of selectivity echoes the notion that data is inherently lossy. We can’t in fact “encapsulate” or capture all the data, or all the expertise.

Or, 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.

SLIDE 70

But I think that people involved in data curation can also think about human solutions. I have one last philosophical stretch to put you through! Here it is.

The philosopher Giorgio Agamben has developed Aristotle’s distinction between “zoë” – bare life – and “bios” – qualified life.

In a very oversimplified sense, zoe, or bare life is life stripped of political or civic life.

Bios, or Qualified life is political life: life with options and the possibility of choices.

Here’s the connection to data. Many of us know first hand the “bare” feeling of being judged on the basis of only the data others have about us.

To the extent that a database and its data are separated from living human selection and expertise, one might say that the data is “bare.”

It may seem inevitable that the databasing of our culture – whether social or scientific - brings with it the risk of losing the “bios” dimension.

SLIDE 71

But what if we imagine data curation as having a fundamental role of establishing a human chain of custody over databases, carrying forward not only the data (and metadata) but the human system with its political dimension of choice?
4. Wrapping up

To recap:

I’ve proposed that libraries need to be critically reflective about data, and not assume that mapping our old mission, role, and skills to data is going to work.

I’ve identified four problems in particular with that mapping: mass, abundance, networks, and ontologies.

I’ve proposed three questions to help get us beyond the notion that in dealing with data or databases, we are dealing with something that can be complete.

Data begin as an expression of choices and we add new choices to data when we create (or add) metadata, load the data into a structure, or choose the programming language used to query the data.

While all of this may not change what you do tomorrow or next week, I hope that it begins to suggest that one of the reasons that work with scientific data is so compelling and interesting is that it represents a massive challenge to libraries to begin to think about how impossible it is, really, to curate data or databases in the way we curate objects or even narratives.

But that’s OK. The reason we curate data is not to control or manage the data itself, but to participate in creating a kind of knowledge scaffolding.

We alter the world by creating data archives, and the goal of that work, like other “scaffolding” work, is to help us think. [Andy Clark – quote]

The scaffolding we create is as full of choices, gaps, and unanswered questions as it is of experience and measured events. Yet, as scaffolding for future choices, it is meaningful.

Sometimes it is said that libraries are moving into the business of digital data curation because of the critical risks of total data loss; and because it’s urgent for the survival of libraries that we find ways of mediating and providing access to data despite its abundance.
Even without discounting these reasons, I think there’s a third reason that is also compelling.

The data we curate forms a landscape that simultaneously constructs islands of knowledge even while every increase to those islands also increases the unanswered questions that form “the shore of our ignorance.”

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 I believe we will come to understand that curators of data are involved in an effort that is not only technical in its challenges, nor is it merely a “service” to the scientific community.

SLIDE 79

Rather data curation seems to me work 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.

SLIDE 80

On the other hand this should inspire us. The choices you make as data curators participating in “qualified life” truly matter. Your choices help constitute a world, and they will change it.

SLIDE 81 (blank – thank you & discussion)

Readings:


Online preprint at: HYPERLINK "http://vv.arts.ucla.edu/AI_Society/manovich.html"

Torkel Franzen, Gödel’s Theorem: An Incomplete Guide to its Use and Abuse.

Oracle article: “Oracle Magazine White Papers - Solving the SQL Tuning Problem” http://docs.oraclewhitepapers.com/oraclewhitepapers/solving-the-sql-tuning-problem/?pg=24&pm=1&u1=friend&sub_id=Cz7EV0k6i60fd