

Thoughts on Scale and Complexity

Abby Smith

When sitting down to solve a problem, it is good to imagine the size, shape, and dimensions of the problem; then to articulate the problem as precisely as possible, taking particular care with words that bear the burden of defining (and hence, solving) the problem. I offer a few thoughts on how to imagine the scale and complexity of digital data; a few thoughts on the language we use to describe the problem; and a few thoughts on whether existing frameworks and practices can help us—or not.

Scale and complexity: What is the problem?

The organizers of this workshop define the “scale and complexity of data” issue primarily as technical challenge. What about the imaginative challenge of scale and complexity: our inability to conceive richly and objectively of any particular scale and degree of complexity much beyond the human scale? This inability is probably an innate feature of our minds, given adaptation to the specific parameters of our environment and our physical size. We may study phenomena on the scale of galaxies, viruses, and mesons, but those scales are not native to us. I have observed that extremely large and extremely small scales are often experienced subjectively as confusing, confounding, and at times paralyzing; so too, are extreme levels of complexity and interdependence. At the same time, I have observed that such scales and levels of complexity can also be experienced with awe, wonder, curiosity, and the other affects that motivate scholarly inquiry. This last fact leads me to believe that the problem of imagining scale and complexity in the digital information environment is a tractable one. But we have not gotten there yet, as scientists and scholars. And once we do, we need to translate that imaginative capacity into operational knowledge and management skills. For, as the organizers of this workshop wisely remind us, the fate of data-driven scholarship will ultimately lie as much in the hands of managers, administrators, and funders as in those of scientists and scholars.

So, imagine this: you are a responsible manager of reputable stewardship organization and are asked by a member of your governing body to explain how the organization will respond to the recently released report by IDC, “The Expanding Digital Universe,” (http://www.emc.com/about/destination/digital_universe/), a report the member has skimmed. This report forecasts that by 2010 there will be 988 exabytes of data in our world, up to 70% of which will be user-created. (I take that report as an example somewhat arbitrarily. Any publication that mentions petabytes and exabytes could be substituted here.) How precisely do you explain, objectively, reasonably, to both your funders and your staff what such a reality would mean for your organization and its mission? What does that mean for a natural history museum director, a collection development librarian, museum curator, or a data scientist? How can they not feel confused, confounded, even a tad paralyzed when contemplating this news, at least in the context of everyday decisions about what to collect, preserve, to provide access to, and what resources they will need?

For this is what is at stake: if we are serious about data-driven scholarship, then we must respond decisively but wisely to the imperative to collect, to curate, to preserve, and to provide access to the content that we believe, to the best of our abilities, will

have enduring value. And we need to do so now. Thus the challenge of scale and complexity becomes a matter of urgency in our political economy. I hope that this challenge will be addressed technically, but I do not believe that we can get to a technical solution until we can wrap our heads around it imaginatively.

One more point: it's helpful to remember that in the digital world, content wants to be machine-readable, not human-readable. When most of the historical record (presently in analog form) is in machine-readable form, the problems of migrating content repositories into a digital environment and deciding which born-digital content to collect and how will appear ever so much smaller. People will be able to see what they are now only asked to imagine.

And another point: worrying about what researchers five decades hence will want is a waste of time at best, a dodge at worst. Let's spend our time addressing today's research problems, and tomorrow's.

So, one challenge before us is about vision:

Can we create a frame of reference that is meaningful for people making decisions today and tomorrow about what digital content to collect and commit resources to?

You say data, I say content, let's call the whole thing off

Actually, let's not. I'd rather we all work together, each of us tackling the part of the problem most important to us. I like the definition of "data" found in the NSF cyberinfrastructure vision report: "...data are any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data." The term "content," more frequently used among the general public, the creative industries, and scholars of human culture, literally means "the substantive or meaningful part" of a created work (American Heritage Dictionary, 4th edition). By extension, in digital parlance "content" refers to that which is created when data (that is, data points) are combined in a specific context with the purpose of creating meaning, order, or significance. Content is always characterized by specific behaviors and, loosely speaking, some structure—a certain level of organization or of fixity—that results in such familiar forms as...music, webpages, geo-mash-ups, visual patterns, simulations, etc. "Subterranean Homesick Blues," for example, comprises a lot of data points—sound waves—but taken together in the specific way that Bob Dylan devised a few decades ago, emerge as "content," that is, a song. (A trendy way of putting this would be to say that content is an emergent form of data, but I am sure researchers who actually need to use the word "emergent" in very specific research contexts would prefer that we leave that word alone, and so I shall.)

Why does this matter? Just because I'd like to see insights and solutions to the scale and complexity problem in the scientific and engineering domains be accessible to those in other domains, especially the general public.

One more point, a distinction with a difference: content can be copyrighted. Data, in the sense of facts, cannot be.

And a point about that 70%: people love creating content! Content is meaning. It is expression. It is how we experience ourselves and connect with others.

So here's another challenge for us:

Let us agree that data-driven scholarship, and content-driven scholarship are the same thing, and get more people working on the problem.

Lessons learned: Is there any continuity between yesterday, today, and tomorrow?

The biggest problem posed by the scale of digital content production is figuring out what needs to be collected and what does not. For very good reasons, most long-standing stewardship institutions—social science data archives, research libraries, art and natural history museums—have focused on the technological and personnel challenges demanded by the migration from an all analog environment to a hybrid analog-digital environment. They are particularly concerned about the implications of these changes for their long-term sustainability and their short-term business models. At times they act as if they do not understand the decisions about what born-digital content to collect will be determinative both of their long-term sustainability and their short-term business models. Okay, so the scale of production of relevant content for institutions has ballooned completely out of proportion to the resources available to capture and manage them. And okay, collecting always involves a commitment of resources both presently, and into the future, so these decisions are not easy. But without getting the collecting bit right, a lot of resources will be wasted. (And a lot of valuable content will be lost.)

Would it help to know that the problem posed by the scale and complexity of information for content selection is by no means a new problem? Of course, as a historian, I'm inclined to answer that question "yes." To take one example: libraries have always made trade-offs between comprehensiveness of subject matter coverage and depth of coverage, even the largest library in the world. Of the Library of Congress's collections, numbering well over 130 million items, about a third are bound volumes. A very large portion of the remainder are manuscript archives, most of them personal archives. Even if the scale of any individual collection is familiar to us—the scale of a single individual's life—we still encounter the problem of determining how much documentary evidence is enough to create a meaningful and authentic record of a phenomenon—the phenomenon in this case being a person. In all domains of scholarship, there's a very real question about how much data are needed to have something—a species, the bathysphere, a historical event, Hamlet—authentically represented. Deciding where the boundaries of any person's life begin and end is conceptually very similar to deciding what are the boundaries of, say, a webpage. People commonly say "Well, a webpage isn't just a page, it's all the links to the page." By this, I believe they mean that what makes a webpage have value as a tool of expression, information, or communication is that it links to other pages, that those pages become part of the communication in toto, and if you leave out too much of all the stuff that makes the webpage meaningful to begin with, you might as well not save the webpage at all.

This is precisely the boundary problem that archivists have faced for a long time when determining the collecting scope for an individual human life. Humans, and certainly noteworthy humans, tend to be interesting because of all the people, places, things, and historical events that they are connected to. What happens when you have a wonderful subject, such as John Muir or Gifford Pinchot, who had wide correspondence during long lives with other wonderful subjects? A good archivist would want to collect as much from that subject and about that subject as possible under one roof. But that is not possible. Given that, what you, as a researcher, would want to do is to be able to have access to the content in all the other collecting institutions that hold materials relating to subjects with whom Muir or Pinchot corresponded, and then, of course, people with whom those subjects corresponded.

And that's just the correspondence link. What about that person who, as a historical agent, such as Muir or Pinchot, was widely involved in the growth of the environmental movement, public policy, the institutions of higher education they attended and taught at, the numerous domiciles in which they grew up and lived, the interesting organizations that they were associated with, all the trips they took, etc.?

How is it that archivists are not constantly stymied by these boundary definitions? That's a long story. But they don't always get it right, and their solutions have extreme limitations because they're dealing with physical artifacts. But I do believe that part of solving the scale problem lies in grappling directly with the boundary problem, that is, the scoping problem in collecting. One straightforward way of solving the scale problem, one that should be pretty obvious in the Web 2.0 environment, is to have more people/organizations collecting. The more people/organizations we have collecting, the more finely we can parse the boundary problem. The fewer people/organizations we have collecting, the greater is the pressure on the ones that do to settle boundary problems both at very large scales and very small scales. Working through communities of practice, with each community taking on responsibility for content that is most valuable to them, could go a long way to capturing more, rather than less, content. Of course, this solution would achieve our goals for content sharing only if the collections were open and interoperable. Again, a topic urgent for our political economy and one that, fortunately, is a topic for another breakout group.

So our third challenge: reach out to other communities of practice.