

THOUGHTS ON SCALE AND COMPLEXITY

Abby Smith

What we talking about?

I wrote up a position paper that had some thoughts on complexity and scale, and most of what I talked about was my understanding of what data or content is (or are, in the case of data). I also put forward some ideas about the challenge of selecting and collecting large amounts of data as being a boundary problem. However, knowing that everyone read the position papers carefully, I will not summarize the points I made. I'll take advantage of Ron's invitation to be provocative and ask 4 questions that are quite simple. I take it that, among friends, no question is too stupid to ask. I will assume that the implications of the questions are both philosophical and technical.

First, a note about the unfortunate word "science." In some languages, there is one word that encompasses all reason-based inquiry, such as *wissenschaft* in German and *nauka* in Russian. While these words are usually translated as "science" in English, they really embrace the natural sciences and humane letters together and without prejudice. Thus speakers of these languages avoid the dreadful linguistic choice that we, in English, are forced into when we cannot simply use one word for all domains of reason-based inquiry. Would that were not so. In the following I am not making such distinctions, no matter which word I use.

Question #1

First, is complexity always a feature of scale? On the one hand I mean: are large-scale phenomena necessarily more complex than small-scale? On the other, I also mean: does the piling up of data, amassing more and more data, lead inevitably to complexity? Sometimes, in the real world, we see the opposite effect. Crime-solving, for example, operates on the principle that the more good information, the better, and eventually, if you have enough of the right data, things become simple again. This is true of all intellectual challenges that are essentially a form of puzzle-solving. (Note I say "good data." Bad data are false friends.) I think of this as the Sherlock Holmes model of scholarship. We're trying to make sense of or account for something and are looking for those pieces of information which let us understand the phenomenon.

Abby Smith
NSF/JISC workshop on data-driven science
April 17-19, 2007

Question #2

The second question I have is: is there really no way to reconcile scale and expressiveness—“expressiveness” being another way of describing complexity. I’ve been witness to more arguments than I care to remember about the relative merits of different metadata schema, and they generally go like this. One side argues that the more generalized, or high-level, metadata are, the better, because the underlying data can be easily found, easily made useful to different communities of users for different purposes....the virtues of Dublin core, etc. Then there is the other side, arguing just as vehemently that the more detailed and specific the metadata, the better, and the underlying data will have a greater share of the universal virtues of being easy to find, easy to use, etc. Both groups mean “better for specific purposes,” they just have different purposes in mind. So I ask you, are these really irreconcilable differences? Why can’t we live with a “it just depends” kind of position, in the quantum model? Imagine data in a repository that behaves in a quantum fashion: if you ask one question of the data, they are revealed as the equivalent of a live cat; and if you would ask the data it a completely different question, in a different frame work, they are revealed as the equivalent of a dead cat. But unless and until you ask the question, the data live in both states.

Question #3

Okay, here's another question: What is our ultimate purpose here? What are we looking for with all this data? Is it to find patterns—patterns from which emerge order, harmonies and complementaries, and meaning? Are we looking for relationships among the data that are vivid, clear, rational, and of course—if you are seeking funds as well as patterns—innovative. I tend to think pattern-finding is an important goal, although maybe not uniquely important goal. But it is often the hallmark of science and scholarship in general to gather up enough information to make a meaningful generalization about a phenomena. We hope that the more data we pile on, the more likely we are to see a pattern emerge from the data themselves which we would not be able to see if we look at just a few instances of data/the phenomenon. Obviously, this can happen at any scale. On the one hand, looking at one flower reveals various levels of complexity within that unitary phenomenon, called a flower. On the other hand, if we look at all flowers, we can see that the complexity and details of a single flower are generalizable upwards to comprise a set of attributes common to all flowers. People who study historical human phenomena deal with this scale dichotomy all the time. When we look at an individual life and its history, we see certain features. Can we generalize any of them to cover all lives, biological or social? When a historian decides to tell a story about an individual life, she proceeds along the line of selecting things that make the life typical of others at the time, and, of course, distinctive from

Abby Smith
NSF/JISC workshop on data-driven science
April 17-19, 2007

others. In one hand, it is the distinctiveness of the individual, every difference from all the other individuals in their cohort, that makes the biographical subject worthy of a biography. But if the individual were too distinctive, it would not be scientifically interesting. So I conclude, that patterns are extremely important in detecting, if not actually creating, meaning.

A further note on this question: It seems to me that many of the questions that we will address in the next two days have to do with the instruments of observation, loosely speaking, we are employing: which lens are we looking through, the microscopic or the telescopic? Are we looking at large-scale phenomena, small-scale phenomena, or are they, from our point of view, essentially the same? Alex Szalay demonstrated quite vividly last night that instruments very often becomes the determinant of scale. And note that each community of users or domain can in many ways be distinguished by the specific focal length they choose—the miniaturists and the globalists, the biographers and the prosopographers, the virologists and the cosmologists, and so on.

Question #4

Finally: does “data-driven” imply some determinism? Certainly the word “driven” would indicate that. I don’t think of data as having any intrinsic meaning—it depends entirely on the context in which we place them. The context is, to a large degree, the interpretation.

So, am I correct in thinking that our collective goal here is to find ways, primarily through technology, to ensure that data do not become trapped by the context that they were first discovered in and in which they find meaning? We would like to see data in open repositories to which many different communities can come, extract, and recombine according to their own devices, etc. This is certainly one meaning of the term open data—not just freely accessible, but open to lots of different interpretations. The political question then becomes: who has the privilege of gathering the data, marking it up, and using it first? Because the assumption is that the community that marks up the data is the community that owns the data, or at least prioritizes their research uses. It is not clear to me—though it may be clear to you—the data can live in these open repositories in their “native format,” that is, completely decontextualized. But decontextualized in a way that makes it very easy for people to discover them, to put in a new context, recombine them with other of data in new contexts, and create new meanings or new knowledge. How could that be possible? Can we make data open for many types of use and reuse?