Striving toward ad hocracy in dataland

Ben Li
Department of Information Processing Science, University of Oulu
PL 3000, 90014 OULUN YLIOPISTO.
Oulu, Finland
banji.li@oulu.fi

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.
ACM 978-1-4503-1051-2/12/02.

Abstract
This paper sketches the data-intensive collaboration problem in long-term ecological research and more general cases in order to present an alternative perspective to infrastructure problems. It argues that the predominant view from infrastructure neglects opportunities offered by diverse potential collaborators and their data. It concludes that our tools must facilitate, rather than hinder, emergence of broad data-intensive collaboration paradigms.

Author Keywords
Big data; social science; information infrastructure

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human factors, cyberinfrastructure, standards

Introduction
Discussions about data-intensive collaboration focus on the problem of using information and communication technologies (ICTs) to collectively analyze large volumes of information. Presumably, *collective* and *large* refer to vertical, horizontal, and temporal scales of integrated efforts that exceed scales in common practice. Also presumably, the interesting parts of the collaborative sense-making problems are not those solved by waiting for raw computer storage, processing, and bandwidth to increase exponentially through usual product development and engineering. Google, high-energy particle physics, and genetics research all show that low-level technical and social challenges around collectively gathering, accessing, and analyzing large data can be handled effectively in well-defined problem domains.

The interesting problems must then consider new dimensions of science for which few acceptable solutions exist. In particular:

> a) How could we develop accessible and sustainable infrastructures for large-scale collaborative sense-making in heterogeneous domains?
>
> b) How could we consider scaling routine small and individual data practices and assumptions to large and collective efforts?

This paper sketches the big data problem in long-term ecological research (LTER) to contrast it to a more general case in order to present an alternative perspective to the infrastructure parts of the big data problem. It argues that the predominant view from infrastructure neglects opportunities offered by diversity of potential collaborators and their data. It concludes that we should provision our tools to facilitate, rather than to hinder, emergence of broad data-intensive collaboration paradigms.

Big data infrastructure in ecology

Data could loosely be defined as any information available for analysis. In LTER, data can be large in terms of time-series collected, the size and number of sites studied, and the variety of phenomena measured. It was thought that recognizing that such data existed would lead to data-intensive collaboration, and hence methods were devised to share it [7]. But far from a “data deluge” [3], ecology appears to experience a continued data drought. With online news, e-mail, and shopping, ICTs must aggregate and filter data (via search engines, on-line forums, social media, meta-communities like Reddit, etc.) enabling us to gain value without considering thousands of individual heterogeneous data items. There are no comparable complaints about sorting through 10,000 tables about tropical trees found by general search engines in order to find the ten most useful data sets. The biggest challenge appears to be unchanged, namely: How can data sharers make explicit the tacit details and knowledge required to create a data set, so that potential users can access that data?

Big artificial barriers

The current infrastructure to formalize LTER data seems to hinder sharing than support it. To be accepted at a small number of repositories, data sets and meta-data must meet strict content and format requirements of one of several data archival and discovery infrastructures. The several-hundred-page Ecological Metadata Language (EML) specification [6] is popular in the U.S. Popular in Europe, the “Simple Darwin Core” [12] controlled vocabulary derivative of Dublin Core that discards most of the “Darwin Core” standard for describing biology-related data. In the rest of the world, GBIF attempts to reinvent and provide a parallel infrastructure for publishing data on the web [1], partially based on both EML and controlled vocabulary approaches. The few tools that implement such standards require users to learn about access controls, LDAP security, data representation and encoding, server administration, GIS, pointers, relational database structures, XML style sheets and translations, and other technologies. In short, the focus on web-based tools for LTER data management requires LTER participants to also become web services providers specializing in Drupal deployments, Tomcat servers, virtualization, and (in countries outside the US) circumventing political and accounting barriers on the public Internet. This is thought to be a more sustainable solution than those developed at sites of information gathering and use over several decades of successful practice.

By contrast, the world’s largest collaborative data infrastructure, the Internet, is designed to let anyone use, contribute to, or make mistakes on any part of the stack without prior approval. Unlike the monolithic designs of ecological collaborative infrastructures—which can include everything from transport, authentication, security policy, and database schema in one bundle—the Internet’s independent encapsulation layers enables any data to be published, indexed, and discovered in almost any way without requiring the standard or the publisher to anticipate all potential kinds of sharable information or uses.

Follow the big tools?

Even as LTER adopts Internet-based ICTs to manage and share data, it struggles to replicate the solutions and practices brought about by such tools. That few data sets have been published despite the decades-old ability of data-handling tools to export widely compatible CSV, HTML, Excel, or SQL file formats with trivial effort—and despite widespread availability of content management systems—suggests ICT shortcomings have not been a main barrier to data-intensive collaboration.

The LTER collaborative sense-making infrastructure (SMI) attempts to impose top-down standards to stimulate grassroots support. Rather than enable emergent sustainable data-intensive and collaborative science or information infrastructure designs, LTER SMIs centrally and strictly enforce post-hoc data presentation rules. They seem to misunderstand key benefits and requirements of large data sets to tolerate diversity and small errors. In true ecological fashion, more recent competitors like GBIF have scavenged the best parts of the EML and Darwin Core and combined them with simple tools that enable researchers to copy and paste existing data into instrumented spreadsheets that encourage but do not force standards compliance. Although EML and controlled vocabulary advocates have endeavored for over a decade, it was GBIF users who recently first published a collaborative “data paper” in a journal [1], replicating the form of traditional journal articles but focusing on describing data.

Elsewhere, the past fifteen years in biology have advanced data-intensive collaboration where instruments and empirical phenomena impose *de facto* high-level standards on data morphology (gene sequence banks, distributed proteomics, etc.). Social networking sites and search engines allow participants to see more of the networks in which they’re interested, and have low costs of entry and hence present low risk. By contrast, EML is costly. Every LTER participant already knows what everyone else is doing. And the academic system provides no particular reward to data altruists. New community members might benefit from a groomed view of the network, but that provides little value to users guarding existing data. LTER SMIs appeal to neither established nor new members.

Collaborative insight

In many ways, LTER stakeholders have adopted web tools (but not necessarily web principles) hoping for a data deluge and new tools to discover and mobilize connections in those data. However, modern Internet collaborative sense-making tools arose only after content was available, and they built on practices and content from existing publication and search infrastructures such as Gopher, WAIS, Archie, Veronica, etc.

Apple’s new speech-based search interface, Siri, represents the latest mainstream collaborative SMI. It offers value by bringing narrow slices of large data sets close to geographic, temporal, and conceptual sites of potential use. Siri also provides disagreement, ambiguity, and choices of interpretation, all of which LTER SMIs strictly design against based on small data assumptions about data cleanliness.

Siri is both a data-intensive collaborative sense-making success and failure. It is sold as an oracle, yet it is severely limited by both available technology and by data supplied by humans [4]. Shortly after its release, Siri was heavily criticized for suggesting pro-choice organizations to users seeking information about abortions [5], thereby apparently imposing social values on users. Critics pointed out that Siri is programmed to be funny or snarky when responding to some (hopefully) humorous questions [13], and concluded that answers to abortion questions were similarly manipulated. But Siri only gathers information from Google, Yelp, Wolfram Alpha, and other search engines and databases [11] containing different kinds of information gathered using different approaches by or from the public. This is not a new kind of collective sense-making activity, even among ICTs. Usenet, FidoNet, mailing lists, etc. have all gathered and offered collections of local intelligence. The major differences now are the real-time access and parallel use of diverse sources, as evidenced by the successes of Twitter, foursquare, etc.

Assumptions and failure

Despite returning easily identified incorrect results in the correct topic area, critics maintain that “Everyday Apple doesn't fix this, women are getting hurt” [9], as though users forget how to use search engines or how to talk to people to find information because Apple released a convenient product. Others claimed that: “More disturbingly, Siri would not respond to pleas for help for sexual assault or rape clinics, and services for emergency contraception” [8], even though no reasonable expectation could be made of Google to respond to such “pleas for help”.

Siri’s stumbles expose our assumptions and expectations about how our values translate into silicon. We have not abandoned existing real-time emergency and medical services information infrastructures. Yet, consumers use the same SMI as Siri to amplify their unmet expectations into non-productive rage at SMI designers without understanding the design or technology supposedly at fault [10]. Clearly, providing large-scale (near real-time) collective SMIs compounds engineering challenges with new kinds of risks and considerations that are not simply scaled up versions of smaller challenges.

How would we rationally consider scientific and social accountability and responsibility in collaborative decision-making at larger scales? No individual can possibly understand everything between cell phones used to gather individual data up through to inter-continental storage and processing clusters. At the very least, our current concept of scholarly publication must expand to consider other types of authors, contributors, and possibly non-human agents in big data management, as well as the processes they undertake with data and reasoning. Presently, our system of scientific publication seeks to actively discard the human contribution at almost every step on the assumption that our software contributions are unnecessary to the reuse of science and data.

“Big” expectations

Failures of large-scale SMIs are only spectacular because SMIs work nearly perfectly most of the time. Indeed, the point at which we start to notice a system’s failures rather than its successes might be that which distinguishes an infrastructure’s emergence as more than a potential ICT. Rare failures highlight or undermine our assumptions about SMIs. When BlackBerry outages simultaneously disrupt distributed communication among legislators and business leaders for hours worldwide, all they can do is ask the mechanical Turk’s operators to come forth to take responsibility for their misdeed of not fully anticipating a very small number of bad states in hideously complex dynamically interconnected distributed systems!

Such failures provide at least empirical tools to identify collaborative SMIs, if not necessarily the insights to understand them. We commonly become outraged upon discovering in retrospect that some organization had recorded several separate data items about some individual’s malicious intent but failed to act, expecting that more data always implies better results. Yet we also become outraged that our rights have been violated upon discovering that organizations collect, analyze, and act on data pre-emptively through law enforcement or marketing. This again reminds of the importance of documenting and understanding use and intent.

With large organizations and governments (major drivers of ICT research and development) being bound by social expectations and legislation *not* to link and analyse the biggest existing heterogeneous data sets about people, what motives and resources exist elsewhere to develop SMIs and tools?

An alternative view

Let us further consider the development of collaborative SMIs for big data in settings other than those that inherently produce well-managed data. In other words, let us consider a framework supporting (but not necessarily “designed” for) emergence and plurality when and where they occur. Some supportive milestones stand out on the path to this point.

First recall that Google did not start by indexing all data. It first developed a search and indexing capability that simultaneously

allowed individual data items to be discovered, but also enabled the users to work with groups of related data items (related by the users' arbitrary theory embedded in search words). Only after Google devised an effective way to present hundreds of millions of web pages on a screen that ignores most of them did they then expand to other kinds of data, treating them as web pages to be searched and bundled according what they are (statistically) about. And only then did the profit motive manifest, as a sense-making tool for advertisers. Real-time Google tools did not appear until more than a decade after Google was established [2]. Similarly, Wolfram Alpha is able to mine data to infer the meanings of parts of factual questions as a way to test and refine its commercial offerings. There is no comparable facility among the LTER tools to easily discover that several data items concern the same thing except by mathematically testing strings of species names or GIS coordinates against each other. This may be due to an insufficient corpus of accessible data against which to deploy such discovery tools, but absent incentives to publish such as effective meta-search tools, few data will be published to enable the discovery tools. Instead of waiting for a killer app for large-scale heterogeneous data, we might consider how to discover, aggregate, and abstract from the scant data that we can freely access now, to provide hands-on experience to guide our thinking.

Second, in the usual case, we construct the information system artifact, but in few cases, the information infrastructures construct us. Sometimes it makes sense to optimize for the aggregation and abstraction tool's perspective. Using logical web page structures makes information more accessible to both Google PageRank and humans. But we do not yet know if it will be sustainable to conform our concepts and data 'free' infrastructures such as Apache Hadoop (derived from Google's MapReduce) or BigTable that are designed for specific purposes other than scientific collaboration. It is clear that when we blindly assume the infrastructures' perspective of the world, small technical failures could be multiplied into large social failures. Thus, the inability of researchers to publish data that imports cleanly into SPSS or Octave or some database is viewed as a social failure of the scientist to produce conformant data, rather than a technical failure of the tool to deal with data as it's produced in real life. Instead of perpetually blaming the users or designers of misunderstood or misused collaborative or data infrastructures, could we try to generalize from working examples of clean and unclean data-intensive collaborations? Could we also enable all data-intensive collaboration stakeholders to sustain their most important SMIs just as the physicists have done?

Third, we're already great at erecting barriers against collaborative data-intensive sense-making. Academic publisher search engines explicitly exist to facilitate collective sense-making, but also explicitly prevent the vast majority of non-institutional humans, practitioners of science, and gatherers of data from accessing most published data. The most interesting questions are interesting because they make or propose unexpected connections with high surprise value (and often at high social and professional risk), but our approach to ICTs practically discourages work from that direction. We explicitly demand our ICTs to minimize surprises by suppressing unexpected outputs as "errors". We algorithmically apply statistical techniques developed to understand direct relationships among narrow-scale and well-behaved linear phenomena in simplified systems, to broad scale systems in order to find ways to dismiss unfamiliar phenomena in hideously complex emergent dynamic systems. Instead of demanding data to be unambiguous and uniform, could we remove limits to our thinking by defining data or knowledge problems in terms of ambiguous human processing in addition to deterministic algorithms and discreet outputs amenable to textual and numerical processing?

Finally, there is no immediate benefit from over-provisioning for sustainable data management when it is not predictable if and when particular big data infrastructures and processes will succeed, just like it cannot be predicted if and when an article or web page might be read. But dormant data are required to provide contrast to exciting data, and successful data-intensive

collaborations require most data to be dormant most of the time. Thus, good design patterns for collective SMIs must also be good risk-taking patterns that invest in potential future use of data produced currently. Notice that many modern sense-making infrastructures (e.g., transistors, mobile phones, GPS) were devised by entrepreneurs who had not set out to build infrastructures, and who could not envision the full extent of the potential of their work to change others' practices. Notice also that it's difficult to ascribe certain finality to modern collaborations whose web pages and outcomes live on in databases and caches. We could continue the default outcome-focused view of data-intensive collaboration (to which even the largest currently successful data-intensive collaborations strive). But could we also devote a small portion of our vast ICT resources to record the thoughts and processes of those who produce and use data (rather than just documenting the workflows), so that such thoughts and logics may be captured and reused as data themselves?

In short, information infrastructures for big data, like many other endeavors, could benefit from a middle-up-down approach. Currently, externalized scientific data is concentrated around a small fraction of those who practice science, whether via highly selective publication processes or via capital-intensive instruments required to conduct data-intensive science. As we seek to expand data-intensive collaboration to an explicitly low-stakes model in which everyone can contribute (incrementally, perhaps not unlike at Wikipedia), we must discard assumptions that managed and formalized information infrastructures and processes are necessarily better than unmanaged alternatives. Collaborators generally do not care whether their information systems are managed, they require an infrastructure that "just works", something that so far has been denied in LTER SMIs and more generally. This may mean explicitly encouraging competing and contradictory standards and practices to reflect diverse views of the world held by our global researchers. It may also mean sacrificing some detail and much treasured irreplaceable historic data.

Conclusions

Our fundamental relationship with ICTs has cast them in the role of memory assisting and labor saving devices that provide abstractions that help us to better understand the subject of the data. In the era of siloed data, rarely did we let ICTs make decisions for us, except perhaps when the sight of automated monthly bank statements limits lifestyle choices. In law, medicine, and mass media—settings that facilitate individual collective sense-making through ICTs—we readily understand choices offered since such ICTs are developed with a particular kind of user in mind. In spirituality, sustainable decision-making principles are encoded as generalized patterns.

But we cannot model or build for all the varieties of cognition or reasoning possible with heterogeneous data in different settings. There is no universal data solvent that would be less unwieldy than the vast data to be dissolved. We cannot sustain the rules and infrastructures to address heterogeneous data in silicon, we must rely on interdisciplinary people.

The automated interconnection of data presents a novel opportunity for ICTs to become more than fancy containers and calculators that operate on information we supply. Interconnection may enable ICTs to present ongoing original stories from the multiple sources and perspectives required to collaboratively comprehend big data. The Internet design embeds and encourages a mechanism to routinely revise itself through suggested new drafts of standards that enable stakeholders to thrive in contradictions and mistakes.

Our (potential) data-intensive collaborative tools must not continue to be designed, nor themselves design, against this kind of opportunity.

Acknowledgements

This work was funded by the Academy of Finland. It was made possible by informative discussions with the staff, students, and visitors at the University of Oulu's Oulanka Research Station, Helsinki University's Kilpisjärvi Biological Station, and the Taiwan Forestry Research Institute.

References

[1] Global Biodiversity Information Facility. First database-derived 'Data Paper' published in journal. Nov. 28, 2011. http://www.gbif.org/communications/news-and-events/showsingle/article/first-database-derived-data-paper-published-in-journal

[2] Google. Relevance meets the real-time web. Official Google Blog, Dec. 7, 2009 http://googleblog.blogspot.com/2009/12/relevance-meets-real-time-web.html

[3] Hey, T. and Trefethen, A. "The Data Deluge: An eScience Perspective", John Wiley & Sons, Ltd, 2003.

[4] Honan, M. Siri Is Apple's Broken Promise. Gizmodo, Dec. 5, 2011, http://gizmodo.com/5864293/siri-is-apples-broken-promise

[5] Hopper, T.Know-it-all iPhone app draws a blank on abortion. National Post, Nov. 30, 2011, http://news.nationalpost.com/2011/11/29/iphone-users-find-voice-assistant-feature-to-be-suspiciously-bad-at-finding-abortion-clinics/

[6] Knowledge Network for Biocomplexity. Ecological Metadata Language (EML). http://knb.ecoinformatics.org/software/download.html#eml

[7] Michener, W. K, Brunt, J.W., Relly, J.J, Kirchner, T.B., Stafford, S.G. Nongeospatial Metadata for the Ecological Sciences. Ecological Applications, 7,1 (Feb., 1997), 330-342.

[8] Mr. Banana Grabber. What's the Deal with Siri? The Abortioneers, Nov. 27, 2011, http://abortioneers.blogspot.com/2011/11/whats-deal-with-siri.html

[9] Ngak, C. Siri's abortion glitch critics still waiting for a fix. CBS News, Dec. 5, 2011, http://www.cbsnews.com/8301-501465_162-57336644-501465/siris-abortion-glitch-critics-still-waiting-for-a-fix/

[10] Steinmetz, F. Oh, Siri, You Uptight Prude: The Speed Of Rage. Ferrett, Dec. 5, 2011, http://www.theferrett.com/ferrettworks/2011/12/oh-siri-you-uptight-prude-the-speed-of-rage/

[11] Sullivan, D. Why Siri Can't Find Abortion Clinics & How It's Not An Apple Conspiracy. Search Engine Land, Dec. 1, 2011, http://searchengineland.com/why-siri-cant-find-abortion-clinics-103349

[12] Taxonomic Database Working Group. Simple Darwin Core. Biodiversity Information Standards, http://rs.tdwg.org/dwc/terms/simple/index.htm

[13] Very Siri. Siri, where do I hide a body? Oct. 16, 2011. http://www.verysiri.com/2011/10/siri-where-do-i-hide-a-body/

About the author

The author has held various stakeholder roles in data and knowledge management as a data manager in a national social science research network, as a research and development manager in an international ICT firm, as a scholar studying research management and governance in large open source projects and organizations, and as a policy analyst for a regional government. The author's current research interests connect e-democracy, international governance, knowledge management, and information infrastructures.