Environmental Systems Research

RESEARCH

Open Access

Data sharing practices, information exchange behaviors, and knowledge discovery dynamics: a study of natural resources and environmental scientists

Yi Shen[*]

Abstract

Background: This paper presents a deep-dive examination of the cross-boundary data practices of natural resources and environmental scientists in the context of Virginia Tech's institutional visioning and strategic development efforts. The goal is to understand scientists' actual data information behaviors, their communication and exchange dynamics, and their knowledge discovery mechanisms for effective and productive data sharing and reuse. A focus group and multiple individual interviews were conducted using critical incident, story telling, and scenario building techniques.

Results: The results reveal the subtle importance of interpersonal communication and interactive discussion in deciphering nuances, discovering novelty, and revealing insights in data, all of which enable productive exchange and effective reuse. In the new transformative and disruptive research environments, novel discoveries are catalyzed by scientific knowledge, driven and inspired by research curiosity and creativity, and enabled by unique and rich data collections.

Conclusions: As such, an integrated view of social and technical factors must be figured into the holistic design of data repository, discovery, and learning system. Libraries have significant roles to play to advance both social and technical infrastructures of a research data ecosystem in a strategic, targeted, and synchronized fashion.

Keywords: Cross-domain and grand-impact scientific research, Data sharing practices, Holistic data curation, Information exchange behaviors, Information infrastructure and data network, Knowledge discovery dynamics, Natural resources and environmental scientists, Transformative and disruptive research environments

Background

Natural resources data from various monitoring sources are becoming bigger, faster, and diverse than ever before. Transforming these data into scientific evidence for policy making and business opportunities requires enhanced capacities to discover, access, and integrate diverse data in a distributed context (Research Data Alliance 2015). To support data discovery, reuse, repurpose, and integration in creative ways to address new questions or grand challenges, it is essential to understand how scientists communicate, exchange, and interact with data to create benefits and add values that impact business practices, government policies, and scientific knowledge. This paper presents a study of cross-boundary data instrumentation and knowledge discovery dynamics in natural resources and environmental science where Big Data is powering transformative changes and where knowledge gaps and policy opportunities reside.

In particular, this research investigated the intellectual data work of scientists whose research centers on forest resources and environmental conservation. Based at Virginia Tech (VT), the research participants include all members of the Center for Natural Resources Assessment and Decision Support (CeNRADs) and other

*Correspondence: yishen18@vt.edu
Virginia Polytechnic Institute and State University, 560 Drillfield Dr., Blacksburg, VA 24061, USA

scientists who approach critical natural resources issues from many diverse angles and disciplinary perspectives. Besides their disciplinary ties, these scientists are all active agents participating in various Interdisciplinary Graduate Education Programs or Global Change Center, which constitute unique venues of innovation. In this current research, this focused group of subjects offers the flexibility to examine cross-boundary data opportunities through a disciplinary lens and to study interdisciplinary data scholarship with a practical area focus and common problem space.

Such efforts were further grounded in the larger institutional context of Virginia Tech's Beyond Boundaries and Destination Areas initiatives. The Beyond Boundaries initiative looks at the long-range visioning and strategic investment of the University to "address complex problems that transcend economic, geographic, social, and spatial boundaries" and to "advance Virginia Tech as a global land-grant university" (The Virginia Tech beyond boundaries 2015). The Destination Areas initiative intends to identify and build large-scale, boundary-crossing, trans-disciplinary areas of core strengths that will differentiate and distinguish the University as a destination for global talents (The Virginia Tech Office of the Executive Vice President and Provost: destination areas 2016). These movements define the changing landscape and future trajectory of the University that require a radically different and disruptive data approach. To effectively design and implement a data infrastructure, which aims to mobilize and cross-fertilize the institutional talents and expertise and to facilitate knowledge sharing for collaborative action, it is critical to first chart the practical steps needed for a balanced and integrated development.

By adopting the critical incident methodology and creative scenario-building approach, this study explored a wide array of pressing data questions spanning multiple scales, ranging from individual field-specific data discovery and reuse scenarios, to institutional boundary-transcending data initiatives, to international data-enabled collaboration and citizen science engagement in the global data ecosystem.

The goal is to understand scientists' actual data information behaviors, their communication and exchange dynamics, and their knowledge discovery mechanisms for effective and productive data sharing and reuse. Here data information behaviors refer to behaviors of seeking and finding information that describes, explains, contextualizes, locates, or represents data for effective use and robust reuse. More specifically, this research aims to explore the following questions:

1. *What are the application needs and reuse scenarios related to cross-boundary data scholarship?*
2. *What data communication and translation mechanisms are most effective in supporting the booming exchange and reliable use of data?*
3. *How can research data flow and revolution in disruptive innovation and transformative environments be supported?*

The main contributions of this study reside in its focus on the human-centered process of knowledge-driven, curiosity-inspired, and data-enabled scientific discoveries. Through the deep examination of natural resources and environmental scientists' data communication and sharing mechanisms, this study highlights the different stages of scientists' data work and exchange dynamics, and suggests an integrated view and holistic design of data repository, discovery, and learning system in the transformative and disruptive research environments.

Literature review

Addressing global issues of sustainability and resilience requires exploring myriad natural resources topics including forestry, wildlife, conservation, natural resource ecology, environmental policy, and geospatial analytics. Such research increasingly requires integrating large amounts of diverse data across scientific disciplines to deliver policy-relevant and decision-focused knowledge. This knowledge will be useful for the society to respond and adapt to global environmental changes, to manage natural resources responsibly, and to grow our economies sustainably (Gurney 2016). Numerous actions that have been taken to design data exchange platforms, to equip people with the necessary skills, to draft and implement relevant data policies, and to coordinate all these efforts (for examples, DataOne, AmeriFlux, NEON, and Research Data Alliance are among these efforts). The National Science Foundation's Big Data Regional Innovation Hubs have identified the management of natural resources and its impacts on habitat planning and hazards as priorities of research and innovation (National Science Foundation 2015).

Natural resources and environmental scientists traditionally work with large data sets and over time have developed their own ways to handle scenarios involving massive data. Current developments in Information and Communication Technologies (ICT) and Big Data science potentially provide innovative and more effective ways to do so (Lokers 2015). The technology landscape of this field is becoming more diverse and quickly expanding. Not only have temporal, spatial, and spectral resolutions of traditional sensors increased dramatically, but new sensing array paradigms have been created, leading to ubiquitous sensing and providing massive data sources (North Carolina State University 2011).

"With forests, environment, and water more than ever at the forefront of ongoing discussions of ecological, social, and economic well-being at regional, national, and global levels" (Virginia Tech News 2016), it is timely to study data mechanics and flow in this excitingly comprehensive and increasingly integrative area of research (National Research Council 2001; National Science Foundation 2003). This area provides essential opportunities to analyze knowledge discovery and innovation dynamics at the science-society and science-policy interfaces. Such research will provide valuable insights to data professionals and information agents who are seeking to develop their organizations' technical and human infrastructures for domain-connecting, boundary-breaking data sharing and exploitation.

The Virginia Tech College of Natural Resources and Environment represents an exemplary site as a leading program in natural resources and conservation in the United States. Within the college, the Center for Natural Resources Assessment and Decision Support (CeNRADs) is uniquely positioned given its multidisciplinary, translational nature of work that involves collecting, mapping, repurposing, and integrating large, diverse data sets from sources such as the U.S. Forest Service, U.S. Geological Survey, Virginia Department of Conservation and Recreation, and Virginia Department of Forestry etc. The Center's research incorporates and integrates multiple subject areas to study market conditions, landowner behaviors, policy implications, business decision-making, and natural resources management. As noted, CeNRADS aims "to improve the collective capacity of companies, agencies, and natural resources scientists to thoroughly assess the complex dynamics of changing land use, resource conditions, ecosystem services, and markets" (The Virginia Tech Center for Natural Resources Assessment and Decision Support: Research Projects 2015).

It has been widely recognized that sharing research data encourages collaboration and multiple perspectives, and supports knowledge exchange between researchers working in different disciplines (Kowalczyk and Shankar 2011). However, by sharing, researchers may feel that they are relinquishing control over their data. They may have concerns such as others will discover errors in the data, or reach contradictory conclusions using the data, or possibly misuse the data (The University of North Carolina at Chapel Hill and The University of Edinburgh 2016). They may also worry about being "scooped" by other researchers (Gewin 2016). So the question is how to effectively enable data sharing and reuse while overcoming these concerns and obstacles. Drawing from scientists' own scholarly communication dynamics and data information behaviors, this research offers a unique perspective on how to develop effective communication mechanisms and interaction channels to enable productive data sharing while alleviating the associated concerns and problematic consequences.

To date, much academic research about scientific data work has focused on individual fields or domains or disciplinary differences (e.g. Hampton et al. 2013; Kelling et al. 2009; Herold 2015; Borgman et al. 2007; Birnholtz and Bietz 2003) and has rarely attended to data mobilization in support of domain-transcending and boundary-crossing signature areas of a university. No research has particularly addressed collaborative data reuse and knowledge discovery in the context of institutional-wide strategic visioning and program planning efforts. Such institutional efforts especially aim to revolutionize the academic enterprise and forge large-scale, grand-impact areas of development that combine core strengths of a land-grant research university to address significant societal challenges. To fill this gap, this current research investigated a key VT area of strength in natural resources and environmental development situated in the larger context of the VT Beyond Boundaries movement and Destination Areas development. It depicted and illustrated scientists' engagement with different data sharing mechanisms and their perspectives of creative reuse scenarios. By doing so, this research formulated pathways towards developing an integrated socio-technical system of data infrastructure.

Methods

This research was conducted through a focus group interview with the CeNRADs team and multiple individual interviews with other scientists. The CeNRADs group was first contacted and asked to provide suggestions on other interview candidates. Using a snowball sampling approach, the following interviewees were then identified and contacted. A total number of six natural resources and environmental scientists at VT were interviewed during December 2015–April 2016. This qualitative sample size permits the deep, information-rich, and case-oriented analysis of a focused group of subjects with clearly identified study objectives, and meets the qualitative research design and sampling recommendations of Creswell (2007). The VT site is exemplary in this particular field of scientific research, and may offer implications applicable to other sites of similar institutional portfolio and research profile.

Each interview lasted from 1 to 2 h. All interviews were audio-recorded and fully transcribed. The qualitative interview data were then analyzed with open coding and axial coding to discern contexts, gather insights, and draw patterns on the scientists' data work.

This study implemented a carefully designed interview instrument incorporating critical incident, story telling, and scenario building techniques that call special attention to self-reflective practices of scientists. The deployment of critical incident technique (Flanagan 1954) allows the construction of typical scenarios of user behaviors and significant experiences when they interact with various technologies, including data and information systems. The use of narratives and story telling technique helps contextualize and place the critical incidents in specific scenarios or actual cases. Different from the open-ended, retrospective description of critical incidents, creative scenario-building on the other hand provides forward-thinking, prospective outlook of events and new opportunities.

Using a combination of these techniques, this study captured rich and vivid images of the dynamic data scholarship and interactive knowledge inquiry of the scientists. Asking the participants to describe practical examples and encouraging them to perform creative thinking, this study identified both actual cases and creative scenarios to determine the unique data use and reuse mechanisms, value propositions, intellectual prospects, and future opportunities in this integrative knowledge space. These results are all situated in a transformative and disruptive research environment that leverages scientific evidences while harnessing economic interests and societal benefits.

Findings

Knowledge discovery: scientific curiosity and data richness

In this research, the interviewees were asked to describe incidents when they were making requests and seeking responses in data for specific tasks, but produced unique results that had not been pre-programmed or anticipated. This was followed by another question probing what has catalyzed the new discoveries.

As a result, one participant described a real-life example of data discovery process guided by the pursuit of global level analysis.

> *"I'm working with a colleague and we're interested in how the leaf area of a tree scales with the amount of nitrogen that the tree has. We had data on arctic systems but didn't have data on tropical systems. I knew that a colleague had measured the amount of leaves basically in a forest down to the tropics. I said 'I bet they also measured nitrogen,' so I emailed them and they said 'yeah, but we never looked at the data before.' So I took the data and plotted it on the same graph as the arctic data and found a very clear global relationship that we were looking for. That tropical data was already almost 10 years old, it's pretty much dead, kind of old data. So we were able to look at it in a slightly different way and found something neat. [So what triggers this discovery is] a desire to develop a global level analysis, so you play around your own data, your own site, and then wonder how this scales, you find colleagues who collected in a totally different ecosystem and location, you like 'hey, do you have this kind of data? You can be a co-author etc.'"*

Another respondent believed that the capability to work with data is important, but emphasized that a scientist's fundamental knowledge of a field and his/her innovative drive to push the envelop on new ways of data exploration catalyze novel discoveries.

> *"I think certain ability to work with data is what's really important. But people that are incredibly reliant on just other people's software and purchases will have fewer opportunities to do this, because it [new discovery] happens as you're pushing the envelope on new ways to look at the data, whether be visualization or analysis. When you start to push on that for specific task, you're going to find other things emerging that you didn't anticipate. So it's certain amount of data facility, in other words, just being able to work with it, write programs, visualize it, and understand what you're seeing… The fact that you can look at something, you have enough knowledge of your ecosystem or whatever science you're doing, you can recognize something is cool, it just emerges."*

A third respondent highlighted that novel discoveries are rooted in scientists' deep thinking, knowledge accumulation, and creative ideas. Rigorous and robust data analysis is directed by clearly defined goals and properly formulated questions.

> *"It takes a good creative thought about how you can look at the data, you have to know what you're looking at and how you can look at it. It's not just as easy as visualizing it and things popping out at you. I think it's more likely that the person who's really thought about it in quite a bit detail [and] quite a depth for many years in their career would think about it and say, 'I want to search for certain patterns to help you understand some relationship that may be going on.' It could still be quite a complex analysis, maybe they have to transform the data or summarize the data in some way, filter it, screen it, factor things in and out, and then in the end they can come up with some observations that reveal something interesting and relevant. [It] not necessarily just pops right out, [but] still requires quite a bit thought, creativity, and efforts. Because these data sets do have variations, a lot of variations, some are natural variations, some are measurement errors,*

> *sampling errors, or other types of errors, so a person has to be very careful about the questions they ask. But on the other hand, if they formulate the question properly, they can find the answer despite the noise."*

The scientists' responses indicate that fundamental knowledge of a field and scientific curiosity are the drivers for new discoveries. Different from the overwhelmingly popular data-driven discussion of knowledge discovery, the current respondents believed that Big Data and data visualization is not the magic wand. Instead, it still takes the creative thoughts and spontaneous actions of researchers to ask the right questions, conduct relevant search and inquiries, and transform data in a rigorous and innovative manner to make reuse cases feasible and viable. New discoveries often belong to scientists who bring both historical understanding and fresh perspectives to a field when exploring data.

On the other hand, unique and rich data collections inspire and enable new discoveries. Often they attract different researchers with varied knowledge and diverse capabilities for mining. They also support the applications of multiple theoretical and analytical approaches. These data withstand continuous testing and facilitate cross-validation of different perspectives and methodologies, especially by allowing researchers to constantly compare and contrast their findings and results. Such points are exemplified in a respondent's comments below.

> *"I think that's generally true now that we have a very large set of [LegacyTree] data, and since about July of last year we started doing more scientific analysis of the data that we collected with some specific goals in mind, and we found that some aspects of the problem that we're studying turned out to be more complicated or different than what we had thought or anticipated. What led us to do then is to search a different body of literature that we weren't really expecting to search, and we found that other researchers, in some cases many years ago, had studied this problem... in a slightly different way, more from an anatomical and physiological perspective, while we were looking at it more from an empirical perspective, a statistical, mathematical perspective... so we leaned something about the bark of the trees that we didn't expect."*

> *"A couple of things. One obviously is the existence of this new collection of data that has never before been compiled. We sometimes say we have the largest collection of felled tree measurements in North America, which is true, and because this is a unique data set, it results in discoveries that would not probably have been made without that data. And the other is just having good, competent people and also some good objectives to pursue. In the pursuit of certain objectives, we oftentimes find other things that we weren't exactly looking for. Those are catalysts for new discovery in our case."*

To sum up, new discoveries are catalyzed by scientific knowledge, driven and inspired by research curiosity and creativity, and enabled by unique and rich data collections.

Creative scenarios for data reuse

To construct visionary scenarios, the participants were asked to think creatively about data and describe what new data collections, repurposed existing data, or new approaches to data they would consider as appropriate for their research questions of interest. This interview technique allows the participants to think beyond existing boundaries and constraints, to seize new opportunities, and to chart possible new scientific discourses or research trajectories for innovation.

Consequently, the scientists advocated new ways of collecting, thinking, interacting, and working with data using new model approaches. In their views, as the natural environment changes so rapidly and new, advanced measurement technologies are being deployed, researchers should not be constrained by old analytical models and conventional ways of thinking. Instead, new approaches, systems, and models that allow novel views of new types of data should be developed. These points are indicated in the following comments and examples given by the participants.

> *"There's no doubt in forestry now a lot of data that are being collected are novel, using a lot of measurements from airplanes and satellites. These are opening up new avenues for discovery. Also [we] have electronic sensors [that] can measure a lot of things impossible to measure by any tools we had in the past. They use very high-definition imagery, three-dimensional imagery, and multi-spectrum imagery, parts of the spectrum are not even visible to human eyes. So those sort of things will no doubt have major impact going forward."*

> *"Why can't we build approaches, models, systems that allow users new ways to look into the world? I think that's actually one of the biggest constraints. We have all these new ways to get data, [but] we're trying to put them in old systems of data utilization, that's actually a huge issue now."*

> *"One other problem is that as we model, especially with process models, there are all these parameters*

> *that are so painful to collect. Can you imagine somebody climbing up there [and] putting in little instruments on chunks of the trees? It will be a nightmare, and as a result, the number of data points around the world where we really know is very sparse. So as we start to do these kinds of models, we need to know how those parameters are changing as the environment changes. Right now one of the big questions is there's so much more CO_2 in the atmosphere than there was even 15 years ago, we don't really know how much that's changing. So we keep using what we knew about the system, but maybe that's not right anymore, because the undergirding assumptions are now no longer the case. Same with nitrogen availability... All our system understandings are based on measurements. The world is changing so rapidly that in many cases we need new data associated with those changes."*

> *"Forestry is changing these days. I mentioned wood products that come from forests, they've been very standardized for many years. Now there are many mills starting to use wood for electricity, from wood pallets. So there are new products like biomass. When you harvest a tree, usually all branches just fall on the ground and they leave them to rot, now people are starting to collect these branches and chip them up, and use that as fuel to generate electricity. [But] we don't have any data on how many tons of branches come from forests. So we're going to need data collections on new wood products coming from the forests. We would like better price data too."*

Correspondingly, the interviewees were then asked to think broadly and creatively about their data, and describe the possible scenarios of their own data being reused by other researchers either across multiple fields or within broad disciplinary areas to answer different types of research questions. In response, they described prospective reuse scenarios or actual application cases, as shown below.

> *"We've been so focused on wood supply in forests, but we've always known, kind of in the back of our minds, that this data could be valuable for somebody who studies water quality... We know that the data we will be creating could be used for water analysis, for wildlife habitat analysis, and maybe for other analyses that we haven't thought about. So we're eager to collaborate to develop new client base."*

> *"Mostly I think what happens with us is people reuse the way we do it rather than the data themselves. Say we create a new algorithm or way to evaluate change in time series, more likely people will want to try that on their own data. The data we're sharing are actually code."*

> *"I think that's really a strong possibility that's already starting to happen. We have people in engineering or in different areas outside of forestry [asking for our data]. They may be interested in diversity or ecosystem health or sustainability. It's hard for me to even imagine all the different people who might ultimately end up using the data... Oftentimes we see citations from a range of disciplines: mathematics, computer science, engineering, remote sensing, atmospheric science and oceanography, climate science, many different fields. Oftentimes, forests are just a small part of a very large system, like earth system, terrestrial system, or atmospheric system. The researchers will need some data or information just to account for that one part of the system. So they look for some information. In the future I am thinking that they could just go to this website to download the data and use whatever information they need to account for it."*

These reuse possibilities and actual cases demonstrate the tremendous values and great dimensionalities of natural resources data for studying a wide array of pressing questions regarding environmental diversity, ecosystem health, and sustainability.

In fact, these scientific data producers are highly enthusiastic for their own data to be reused by others for large-scale global synthesis efforts. This is exhibited by scientists' intentional efforts to build on and enrich existing data sets, as shown in the following example.

> *"My colleagues and I are working on this forest tower site, it's our dream for it to be used in someone else's global synthesis effort. So [when] someone is looking at global pattern and photosynthesis of plants and things like that, where you need to pull in all these data from different sites, if our site shows up in one of these analyses, that would be great. Or someone could use our site to build or tune their models, because once you establish a field site that has these types of measurements, other people want to do measurements there as well. So for example we have airplanes fly over and look at ecosystems, they could have done it anywhere but they want to do it at our site because of our existing data sets."*

Scientific drivers for data reuse

When asked about the main drivers for data reuse, the respondents highlighted the strengths of combining, synthesizing, and cross-validating diverse data sets to generate scalable, rigorous, and reliable scientific results as well as produce unique and critical insights. According to the respondents, data reuse and analytics increasingly drive the research agenda, but are obviously dictated by the availability and accessibility of data as well as the quality and clarity of documentation. The researchers explained these points from different angles below.

> *"With respect to main driver, clearly strength in numbers, this is why people do meta-analysis, it allows you to go more regional studies or continental or global studies, [and it] allows you to cross check and verify with independent data sets."*

> *"Because you are doing some larger-scale synthesis to look at the global pattern, and trying to build a global or large-scale understanding of processes. To do that, you need to synthesize everyone else's observations. So you need to put their data into a database, get them authorship, and look at the patterns."*

> *"I think sometimes one of the drivers is actually the existence of the data themselves, it gives people ideas that they maybe never would've thought about. When they see there's this resource available, they will think about and come up with ideas… people do this all the time, they'll search large databases to try to discover patterns. So just the fact that the data is available [allows] people to come up with creative ideas to mine the data for various patterns, and some of those are actually quite interesting."*

> *"I guess the presence and clarity of the documentation [is the driver]. I know from my perspective it is important. Because we haven't gotten to the point yet where we have any real reuse of our data by other people, but I speculate it's going to hinge on how easy it is for them to load this data. So whatever format we created, like an R database or ASCII text files or GIS layers, we have to make it easy for them to load and access, and have it well documented."*

Communication mechanisms for data sharing

With regard to community practice, the scientists highlighted the major mechanisms for data sharing and reuse. In particular, one interviewee complimented the emergence and abundance of multiple data sharing platforms including institutional, domain or disciplinary-specific, publisher or publisher-related, scientist-hosted, as well as other types of data repositories and archives. These platforms diversify and enrich the traditional sharing mechanism, which is mainly through personal contacts and direct requests.

> *"For the research community, the traditional mechanism has been direct contact communication between researchers, that's been historically no. 1 way for people to share their data. Now that's starting to change, there are more and more data that people are submitting with their journal articles, so that other people can download data from journal websites, or agency websites like NASA, or some repositories like [legacytreedata.org], and also many public databases like maps or GIS or census data. In the past, it would be very hard for most people to get that data, we had to have very good contacts in the government or some universities."*

However in current practice, a lot of data sharing is still happening via direct contacts instead of through formal digital networks or repository platforms. In fact, personal contact and interpersonal communication often acts as a direct channel to bridge cross-system information flow or fill gaps of data sharing not being realized by formal repositories or archival systems. These are described below.

> *"University professors can actually be quite protective of their own data. The federal agencies are interesting, because if you just look in the federal directory and find a person's name and then call them to ask for their data, they don't have to give it to you. But if you have the right contact, maybe you met someone who knows that person, sort of informal channel, then it's more likely that you can share their data. Still there is a lot of data sharing that just has to come from direct contacts."*

> *"Just knowing whom to ask for, so we kind of have to network to the right person. I think that helped with the introduction from a third-party… the University of Maryland was doing research on forest mapping and they had all this data that was going to be publicly available eventually, but this mutual friend, a researcher in forest service who knew them and knew our need, introduced us in an email, that personal contact broke open the door and then they were happy to share. So yes, sometimes it's just getting access to the person that can make that happen."*

So despite the multiplication of sharing platforms, interpersonal connection and communication still plays significant roles in enabling researchers to effectively understand and productively reuse data.

Data translation for effective reuse

With many identified problems in data preservation such as inconsistent data management, incomplete data storage, and insufficient documentation (Shen 2016), one key question is: how do researchers seek and gain a practically good understanding to make effective use of data?

To address data comprehensibility, the scientists reiterated that face-to-face interaction actually helps them gain a better understanding of data. Human intervention and interpretation is necessary to decipher nuances in data for effective reuse. In-person conversation also helps uncover tacit knowledge and reveal rich details of data in ways that standardized metadata schemes cannot achieve when describing data and representing knowledge. A respondent gave the following examples.

> *"Frankly the data management part of our [cross-institutional collaboration] project is run by each individual institution. There are five institutions, so they each has its own sort of protocol for data management… A classic way we would [bring diverse data together and] work with modeling is to have the person who collected the data sit in the office with the person who is writing a model to help work through the data…by working with the investigator who created the data set."*

> *"Another trick is that even though data may be freely available, it really helps to talk to the person who actually collected the data. So for example, I am working with a person on integrating a study from the 90s to 2000s where they pumped CO_2 into a forest to measure how much faster it grows. I was able to get the data off a paper basically, in the supplemental material, but I have a colleague who worked at that site where the data was from. My model is just not doing well fitting this data, and he said, 'yeah, cause this one plot had this one weird thing in it,' it just happened to be a little bit more productive than the other plots, and that would not have necessarily been clear from the data. But talking to him, he was like, 'yeah even though we called it paired, it might not be.' They know the nuances of the data and so that can be really helpful."*

In addition, when asked how existing data standards and documentation schemes may have prohibited the processes of data discovery and analysis, the same interviewee recounted the subtle importance of interpersonal communication in discovering novelty and revealing insights in data.

> *"That's also where talking to the actual researchers helps, cause you force your data into someone else's format, but 'what do I not know about this data that you can help me with?'"*

Boundary-crossing data integration and prospect

According to the scientists, interdisciplinary data fusion and collaboration are becoming an integral part of their research practices. One respondent gave the following example.

> *"I have a collaborator in Hawaii who is using the data testing engineering signal calibration. He's testing measurements that are made from an airplane with high-resolution laser, and he knows about the properties of the instrument that has certain amount of noise in measurements. So he's testing the sensitivity of the instrument to the quality of the measurements. Since we have more measurements than he had in the past, he wanted to use our data, so we shared with him."*

In another example, the CeNRADs team described their cross-disciplinary integration of geospatial, economic, social and behavioral data.

> *"The sample for those landowners was defined spatially, we have parceled data from different counties that show landowners who are on different parcels. The first question was who owns forest, so we overlaid the parcel data with our land cover data to find where the forested parcels are. Then we surveyed those landowners. We cannot tie a landowner's response back to spatial location due to privacy and all that. But we can tie it back to the county, so it is once again brought back to spatial location. We know that landowners in these counties in this region of Virginia responded together, so we tend to take the results from the landowner survey to summarize it by regions of Virginia, and then use it in agent-based models. We use those responses at that region to show how willing landowners in different geographic regions are to sell timber."*

Also, greater opportunities exist with crowd-sourced data gathering by engaging citizen scientists. With the possibility of integrating "human sensor" and other sensors into natural resources observations and street mapping, a plenitude of data exist to support data fusion from heterogeneous sources while also leveraging social sensing and semantic enrichment by the interested public.

The inclusion of citizen science data can provide tremendous opportunities for complex system modeling and interactive data exploration, as shown in the following example.

> *"There are crowd sourced street maps. It will be interesting to compare the distances from [wood] harvests to mills using the open street map versus our resource. My naive hunch would be that those streets that are not in that database are infrequently traveled by rural people who may not be networked, and those may be the very roads that are dependent on for the wood trips. What if every logging truck has a GPS in it recording and reporting..."*

However, participatory crowd sensing also has many challenges. To proceed, the most prominent first step is to define data collecting procedures and documenting guidelines for citizen scientists to ensure the quality, consistency, and continuity of crowd-sourced data, as indicated by a respondent below.

> *"The challenge there is the consistency of [data] quality, how they provide consistency or continuity so that you can integrate all that data and understand what inference you can make from that."*

Furthermore, by thinking broadly on a global scale, the scientists also described the prospect of conducting international data-enabled research and collaborations.

> "*I expect that at some time there'll be a global collection, [so] we can do investigations where we can search for patterns on a wider scale or in different climate systems, like tropical versus arctic where trees grow, we can see how things like climate affect various properties."*

> *"There was hope early on that whatever we do here could be replicated internationally, because we're trying to answer questions like, is our use of wood sustainable? That's an international question. It all depends on location. I'm not sure how hard it would be to go international, [but the key] is data availability. We depend on having land cover data, forest inventory data, market data, and so we could go anywhere those data are present. But I just don't know to what extent those data are present in maybe developing countries."*

But there obviously exist significant barriers to globalization efforts. These include differences in industry and manufacturing standards, protocols, and measurements. There are also cultural barriers, which are further compounded by communication obstacles. The demands for large-scale storage and high-performance distributed computing across national boundaries are inherently major challenges as well. Notably, it was once again emphasized that having the ability to engage in interpersonal conversation with the people who know the data more intimately is critical for deciphering differences and bridging exchanges in international contexts. These are described below.

> *"Instruments from different manufacturers, protocol differences, these are barriers to collaboration. You either have to find big repositories where people can bring the data together, or you really have to find truly robust easy-to-use mechanisms by which people can link to and access other data in a process. The problem with that inherently is, the larger the data sets become, the bigger the problems you got, cause you really then need to push the algorithms to the data rather than the data back to the algorithms."*

> *"Obviously units, cause we all have different units, and cultural differences in sharing, and another is just finding the right person to talk to. You know how important it is to talk to the people who know the data more intimately. I know whom to talk to here, but I don't necessarily know whom to talk to in other countries."*

Above all, interdisciplinary and international data flow and assimilation, possibly enriched by crowd-sensing data ingestion, have great prospects for grand discoveries, but obviously face many technical challenges and social barriers that have yet been sufficiently addressed.

Interpersonal data communication for effective exchange and productive engagement

Having a strong sense of value in data, the researchers believed that data presentations, conferences, and publications could serve as effective communication channels and exchange mechanisms that will lead to active discovery and productive engagement. According to their experiences and expectations, having carefully crafted and purposefully designed presentations about data can effectively drive conversations, build synergies, and boost interests for data sharing and reuse. These scientists thus advocated a move towards active research data communication and requested a space for interpersonal exchange and critical discussion of data.

> *"Now I've gone to many different universities and conferences, and I make a presentation explaining to people the development of this new database [LegacyTreeData], this repository, then people call me or I call them, they say, 'we have some data in our files.'*

> *They then send it to me, and we add to the legacy data."*
>
> *"I think it helps to have some sort of visibility so that people know about it, not only they can get it, but they know they can get it. So it helps to have some publications, presentations at conferences, and symposia to announce the availability and advertise it."*
>
> *"The publication itself is a data set [that] becomes a citable piece that you can put on your CV, other people can cite it, and you get credits for this. There are a few journals that are dedicated towards basically publishing and reporting on data sets. It could be good if journals start expanding the scope of their journal articles, instead of having to always publish original research, maybe publish original data sets, or publish novel data or compilations."*
>
> *"Another [venue] could be conferences where the focus of the conferences is on database design or data sharing... or maybe a place where interdisciplinary meeting of people who have large collections of data or repositories is possible. [It could be] anywhere you have opportunities for people to come together, communicate, and learn about each other's work, and get credits so they can justify the effort."*
>
> *"I think we would have to make an effort, for example, to make a packaged presentation of what data we have, so that we could go to and show water researchers and let them understand how that data could be of use to them. That's going to take some communication steps that we haven't made yet. [It's important to find] the time for us to sit there and with intentions, 'okay let's create a message to other researchers about what data we have and how it might be used,' just that other researchers being aware that there might be a chance they could use that data... I gave a presentation at the University of Georgia and the person who came up to me first was a wildlife researcher, he said, 'you know your data could be used to study habitat,' and I said 'yes,' and he said, 'come to Georgia, do this model, and we will use your data for habitat.' So there is the conversation that really just started and it's exciting."*
>
> *"We could build a data set showing results of our data in terms of map 30 years into the future...and then show other researchers what they could do with this kind of data, [such as in] water quality models, habitat models. I would take that message to colleagues here but also elsewhere, make presentations at conferences, and get interests from outside of the wood-using industry. It takes a concerted communication effort."*

These responses indicate that face-to-face dialogue and information exchange represents a critically important means for researchers to register personally and engage interactively with data while uncovering details and gaining insights in intellectual work. It has the potential to illuminate implicit facets and latent factors embedded in research and scholarship that might otherwise not get noted.

Data network across educational and research enterprises

The scientists also discussed how data stewardship and network activities should be aligned with the Interdisciplinary Graduate Education Programs (IGEP) and the inter-connected Destination Areas at VT.

> *"We're heading toward an active computing model, so basically people can be more hands-on, able to work with data, write programs, do visualization, and incorporate across social sciences. I think there's going to be a tremendous role for the library, because in the end [it's] just some common assets that can be utilized across the educational enterprise."*
>
> *"Like our IGEP, we have such a diverse group, we got economists, space physicists, engineers, and that's good. The problem is that [we are] so diverse, there're people doing ionosphere physics, and people looking at whether leaves coming off a tree affect housing values. [So what we really need is something that] truly is integrated and helpful."*
>
> *"It's a foundation underpinning all these, data stewardship activities [are] associated with all Destination Areas, and tie them into good data stewardship is probably ideal."*
>
> *"When it comes to these Destination Areas, we are talking now about integrating knowledge and expertise from a variety of different perspectives... I think in terms of integrating different subject areas and expertise, data network can be a major driver."*

To move toward a digital future that effectively showcases all facets of institutional scholarship, it is essential to create an integrated record of the scholarly work by the institution. Sharing, using, and reusing data in a holistic manner across the University's disciplinary strengths is essential for cross-domain investigations to address the growing complexity of examined phenomena, which often reside at the boundaries of established sciences. The development of signature destination areas will find a rich, varied, and continuously growing collection of datasets, offering tremendous values but also posing daunting challenges. Data librarians and information agents shall become an integral part of this development with dedicated efforts to making an impact with data. This is particularly needed when researchers are hesitant to commit or still in doubt when it comes to data networking, as indicated by a respondent.

> *"We're definitely thinking about it. But we've been a little slow, mostly because we don't want to make a wrong turn and make a big investment in something that's not going to be very helpful."*

To support their decision-making, libraries could help conduct the necessary socio-technical assessment and evaluation to determine the feasibility of building a cross-disciplinary data network for certain groups or faculty clusters. Libraries could also take the initiative to develop a capability model to facilitate data infrastructure self-evaluation. This will assist any targeted group of researchers to self-identify their current and desired levels of provision in each essential area of data management. It will also enable cross-disciplinary research centers or programs to self-evaluate the feasibility of data network system development, to validate their existing repository platforms and functionalities, and to identify gaps in their provisions. By providing a framework to inform, direct, and promote discussion among stakeholders, we can deliver real decision support.

Another way for libraries to help is to curate data around key variables that characterize the local institution's academic approaches, areas of strength, and information using habits. As requested by an interviewee, it is necessary to show the availability of data collected across campus around certain variables to support meta-level data discoverability.

> *"[I will be interested] if I see soil carbon data from a colleague of mine… So maybe knowing that they are collecting this kind of data would be useful. What data streams are being created by colleagues? What are the variables being collected on campus? These could be interesting. Once you know who is measuring different things, you might be able to go talk to them about their data. Maybe [we can call upon] the faculty members being involved in a Destination Area and require them to submit to a database about the variables that they are collecting. So that, instead of having to query, you know who's measuring soil respiration on campus, then you can contact that person."*

Discussion and conclusions

Data communication and sharing mechanisms

Informal channels such as interpersonal communication and personal connection provide an important mechanism and safe environment for researchers to discuss ideas, identify potential collaborators, and exchange data. Knowing whom their audiences are and having personal interactions help build up trust and stimulate personal bonds that often lead to openness and sharing. An interpersonal discussion of data also helps provide context, declare assumptions, disambiguate terminology, clarify jargons, elaborate data structure, and explain variables in a timely, direct, and professional fashion. Such nuanced information is often hard to capture in a public-domain, general-purpose data archive or repository. The interactive, dynamic nature of personal conversation could help researchers quickly pinpoint issues, ask questions, exchange thoughts, gain understanding, reach agreement, and identify points of mutual interest for effective data sharing and productive collaboration.

In this respect, libraries should act as the primary anchors for community engagement by providing interactive platforms for faculty and scholars to vividly present, demonstrate, explain, and discuss their data in front of colleagues and fellow researchers. It could take the form of a data forum, seminar, or conference to support cross-pollination of ideas and networking across disciplinary boundaries. Such a platform gives its participants the opportunity to present and witness cutting-edge data scholarship in a focused and interactive setting. It should be flexibly large enough to be diverse and lively, but small enough to allow for extensive interaction and intensive discussion.

A data presentation and discussion platform organized and mediated by libraries offers opportunities for faculty and researchers to demonstrate openness and willingness to share data while also maintaining a sense of control, engagement, and ownership. Such awareness exists in terms of knowing who might be using the data, how the data could be re-used or re-purposed, and what the data might be used for, as well as the possibility of forming co-authorship and even developing future collaborations. These issues are often expressed as concerns or disincentives when researchers consider sharing data in a public domain repository system where they feel exposing data means losing control of data. This situation can be mitigated when social bonds and personal ties are being forged during interpersonal conversations and intimate discussions about data.

Data presentations, seminars, or conferences should aim at bringing together researchers from different scientific fields to make sense of what is being gathered, to discover the contextual meaning of data, and to discuss novel scenarios, approaches, and applications of data. In this sense, technical infrastructure and system development is not the sole solution for promoting data sharing and open access. Social factors need to be figured into a changing dimension of data culture.

Stages of data work and exchange mechanisms

Different stages of data work call for different communication channels, exchange mechanisms, and sharing platforms. Personal contacts are important when researchers enter detailed discussion and perform actual work on data. Having a data sharing technical system is useful for searching and identifying who has collected what data that might be of interest, which may subsequently lead to in-depth personal discussion about data and reuse possibilities.

To support actual use, effective reuse, and meaningful repurpose of data, having a data repository in place is far from sufficient. It is useful during the exploratory stage of research when scientists are identifying problems, formulating questions, and searching for information, usually at the starting point for discovery.

After scientists locate the data of interest, personal conversation and discussion with the original data producer(s) or collector(s), if possible, are always helpful by revealing first-hand information and subtle details of the data collection, including the fundamental assumptions, processing history, and native context that are essential for rigorous research inquiries.

Dynamic interpersonal discussion can address rich nuances and reveal exclusive insights about data while tailoring to specific questions of interest and reuse scenarios as proposed by a potential secondary user. Judgment of data suitability and reusability for a certain scenario can then be made during this process.

Once researchers have established a good understanding of the data, a technical database or repository platform once again becomes useful when they are sharing or collaborating on data. This process is normally accompanied and steered by frequent interpersonal discussions among peer collaborators. Across the lifecycle of scientific discovery, technical and social systems for information exchange and knowledge inquiry are inextricably interwoven while playing uniquely different but complementary roles at different stages of research.

Human-centered process: knowledge-driven, curiosity-inspired, and data-enabled discoveries

As the broader society is calling for data-driven technologies to generate scientific and societal benefits as well as economic insights, it is after all humans that drive the whole process. Big Data research and discovery is built upon solid rationale, fundamental knowledge, and deep thinking of scientists and scholars. It is driven by newly conceptualized questions and previously unexamined angles that researchers come up with. It is the enthusiasm and curiosity of dedicated researchers in scholarly inquiry and broad impact that liberates new possibilities for data.

With the proliferation of computing techniques and data analytics, algorithms increasingly have the ability to exert impact on ethics, politics, and economics through automated decision-making and interpretation of Big Data (Lustig et al. 2016). Underpinning these technology-level features and functionalities are robust conceptual, logical, and theoretical reasoning and knowledge framework that drive sciences. As we celebrate exciting new tools and novel algorithms in machine learning, artificial intelligence, data mining, and data science, we need to examine and address the human factors and discovery dynamics underlying the analysis and interpretation of Big Data. Data use and reuse is multifaceted and must consider the complex interplay of social and technical systems at work to enable greater analytics. This requires us to explore the dynamic interactions and trade-offs between purely data-driven methods that potentially require very large datasets versus theory-driven approaches that rely on prior knowledge and existing problem structure. By articulating the interconnected nature of social and technical decision-making at the heart of Big Data (Big Data 2016), we will be able to strike a balance between the socio-technical developments of a research data ecosystem.

An integrated view and holistic design of data repository, discovery, and learning system

Data reuse and repurpose calls for innovative ideas to retrieve, filter, and integrate data from a large number of diverse sources. With the availability of a huge amount of data that are of high-dimensionality and inter-linkage, traditional databases and search mechanisms cannot satisfy data discovery requirements. In the Big Data paradigm, researchers explore creatively and learn progressively while searching, cross-referencing, and making sense of data and information.

Data repository systems should not be pictured as isolated tools in support of search and retrieval to satisfy immediate data needs. Rather, data search and reuse should be configured as part of a larger, complex environment, in which humans observe, learn, and discover while interacting with data objects and information content. Scholarly inquiry and discovery dynamics along with scientific thinking and sense making processes must be understood and figured into designing and developing research data ecosystems. This calls for research in areas such as interaction monitoring, discovery performance, and reuse experiences.

In periods of rapid sociotechnical changes, we need to adopt new ways of thinking about data sharing and information exchange among researchers and between communities. New research enterprises are going beyond boundaries and benefit from organic and unorthodox methods. Creative and transformative research agents are seeking and creating innovative methods to break new ground while producing rigorous insights. The new disruptive environments challenge us to study the dynamic flow, mobilization, conversion, diffusion, and convergence of data in the ever changing and constantly morphing communities of practice.

At the boundaries of multiple domains, data must be reconciled to support a holistic analysis to address grand challenges, and to ensure promising practices and transformational opportunities for both education and research. This requires a proactive, context-sensitive approach to holistic curation that involves requirement engineering, ontology refinement, and synthesis design. To support the new types of interaction, we must advance data modeling, information organization, and knowledge representation for an integrated discovery and leaning system to ensure the rigor and relevance of data curation.

Future research

In order to develop truly user-centric systems, we need to grant scientists and scholars both control in speculative design and empowerment in participatory design of data infrastructure and knowledge network. This requires us to decide how these self-reflective practices of scientists and scholars can be incorporated and embedded in the functionalities of new digital systems. It also requires us to determine how to accommodate differences in carefully designed spaces to support consensual decision-making and collaborative knowledge creation. To better understand the sociotechnical configurations of collective intelligence, we need to investigate the complex interplay of human interactions and distributed cognition as a way of describing underlying data mechanics and relations.

Building upon the current understanding of research data ecosystems, further research should also examine how different domains govern, organize, and maneuver the management of shared data resources and how fast evolving multidisciplinary, interdisciplinary, and transdisciplinary fields disrupt, negotiate, and transform this process. These continuous research efforts are critical for library data programs to advance both social and technical infrastructures in a strategic, targeted, and synchronized fashion.

Authors' information

YS is Associate Professor and Research Environments Librarian at Virginia Tech. YS holds a Ph.D. degree in Information Studies from the University of Wisconsin-Madison. From 2011 to 2013, YS was CLIR Postdoctoral Fellow in the Digital Research and Curation Center at Johns Hopkins University. During 2010–2011, YS worked as a Post-Doctoral Researcher in the School of Engineering Education at Purdue University. Currently, YS leads research environmental assessment efforts and contributes to problem solving with regard to researchers' data needs and emerging scholarly practice challenges. YS also investigates cross-disciplinary data management, preservation, and integration. You can find the author's work at https://johnshopkins.academia.edu/YiShen.

Acknowledgements

The author acknowledges the support of Virginia Tech's Open Access Subvention Fund to cover the publication cost.

Competing interests

The author declares that they have no competing interests.

Consent for publications

The author has obtained consent to publish from all interview participants.

Ethics approval and consent to participate

All appropriate human subjects procedures were approved and followed under VT IRB-15-968.

Received: 24 October 2016 Accepted: 25 January 2017
Published online: 06 February 2017

References

Big Data (2016) Special issue on social and technical trade-offs. http://www.liebertpub.com/lpages/big-data-cfp-social-and-technical-trade-offs/155/. Accessed 1 Sept 2016

Birnholtz JP, Bietz MJ (2003) Data at work: supporting sharing in science and engineering. In: Proceedings of the 2003 international ACM SIGGROUP conference on supporting group work, Sanibel Island, 9–12 November 2003

Borgman CL, Wallis JC, Enyedy N (2007) Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. Int J Digit Libraries 7(1):17–30. doi:10.1007/s00799-007-0022-9

Creswell JW (2007) Qualitative inquiry and research design: choosing among five approaches. Sage Publications, Thousand Oaks, p 126

Flanagan JC (1954) The critical incident technique. Psychol Bull 51(4):327–359

Gewin V (2016) Data sharing: an open mind on open data. Nature 529:117–119. doi:10.1038/nj7584-117a

Gurney R (2016) Science of the (near) future: its power and requirements. In: The Association of Learned and Professional Society Publishers (ALPSP) seminar: standing on the digits of giants, Temple Place, London, 8 March 2016

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH (2013) Big data and the future of ecology. Front Ecol Environ 11(3):156–162. doi:10.1890/120103

Herold P (2015) Data sharing among ecology, evolution, and natural resources scientists: an analysis of selected publications. J Librariansh Sch Commun. 3(2):eP1244. doi:10.7710/2162-3309.1244

Kelling S, Hochachka WM, Fink D, Riedewald M, Caruana R, Ballard G, Hooker G (2009) Data-intensive science: a new paradigm for biodiversity studies. Bioscience 59(7):613–620. doi:10.1525/bio.2009.59.7.12

Kowalczyk S, Shankar K (2011) Data sharing in the sciences. Ann Rev Inf Sci Technol 45(1):247–294. doi:10.1002/aris.2011.1440450113

Lokers R (2015) Big data challenges and solutions in agricultural and environmental research. In: The agricultural information management standards (AIMS) webinar, 17 December 2015. http://aims.fao.org/capacity-development/webinars/webinaraims-big-data-challenges-and-solutions-agricultural-and. Accessed 1 Sept 2016

Lustig C, Pine K, Nardi B, Irani L, Lee MK, Nafus D, Sandvig C (2016) Algorithmic authority: the ethics, politics, and economics of algorithms that interpret, decide, and manage. In: Proceedings of the 2016 CHI conference

extended abstracts on human factors in computing systems, San Jose, 07–12 May 2016
National Research Council (2001) Grand challenges in environmental sciences. National Academy Press, Washington D.C.
National Science Foundation (2003) Complex environmental systems: synthesis for earth, life, and society in the 21st century. NSF Report No. 03–27. http://nsf.gov/geo/ere/ereweb/ac-ere/acere_synthesis_rpt_full.pdf. Accessed 1 Sept 2016
National Science Foundation (2015) Establishing a brain trust for data science. http://www.nsf.gov/news/news_summ.jsp?cntn_id=136784&WT.mc_id=USNSF_51&WT.mc_ev=click. Accessed 1 Sept 2016
North Carolina State University (2011) Data-driven science. https://facultyclusters.ncsu.edu/clusters/data-driven-science/. Accessed 1 Sept 2016
Research Data Alliance (2015) E-infrastructures & RDA for data intensive science—pre-RDA plenary workshop: track 4 research data infrastructures for environmental related societal challenges. https://rd-alliance.org/plenary-meetings/sixth-plenary/programme/research-data-infrastructures-environmental-related. Accessed 1 Sept 2016
Shen Y (2016) Strategic planning for a data-driven, shared-access research enterprise: Virginia Tech research data assessment and landscape study. Coll Res Libraries 77(4):500–519. doi:10.5860/crl.77.4.500
The University of North Carolina at Chapel Hill and the University of Edinburgh (2016) Coursera MOOC course: research data management and sharing. https://www.coursera.org/learn/data-management. Accessed 1 April 2016
The Virginia Tech beyond boundaries (2015) http://www.beyondboundaries.vt.edu. Accessed 1 Dec 2015
The Virginia Tech Center for Natural Resources Assessment and Decision Support: Research Projects (2015) http://cenrads.cnre.vt.edu/research.html. Accessed 1 Dec 2015
The Virginia Tech Office of the Executive Vice President and Provost: destination areas (2016) http://provost.vt.edu/destination-areas.html. Accessed 1 Sept 2016
Virginia Tech News (2016) https://vtnews.vt.edu/articles/2016/01/012816-cnre-sullivandepthead.html. Accessed 1 Feb 2016