Designing a Data Commons for Urban Big Data

Steven P. French and Camille Barchers

Abstract

Infrastructure systems and smart buildings are rapidly joining the Internet of Things and evolving into advanced cyber-physical systems. As a result, massive amounts of data that characterize the structure and function of urban areas in minute detail are being generated. However, these data are often fragmented and managed by a variety of public agencies and private corporations. As a result, they are not readily available to the urban research community. This paper lays out a strategy to develop a data commons that would collect, curate and distribute Urban Big Data to support research on infrastructure systems and how they interact with the human populations they support.

S. French (Corresponding author) • C. Barchers
School of City and Regional Planning, Georgia Institute of Technology, Atlanta, Georgia 30332-0155, US
Email: steve.french@design.gatech.edu

C. Barchers
Email: cbarchers3@gatech.edu

1. Introduction

As we enter the first urban century, complex interdependent infrastructure systems have been developed to support human habitation. In the US and other advanced nations these systems are rapidly joining the Internet of Things (IoT) and evolving into cyber-physical systems (Gartner, 2013). Data from these instrumented systems are layered on the extensive base data that cities and counties have developed over the past three decades in the form of relational databases and geographic information systems (GIS) (Drummond and French, 2008). These instrumented systems now provide a detailed, real time depiction of urban metabolism that tracks the consumption of resources and energy and the generation of waste. As a result, extensive data are being produced on the performance of these systems under both normal conditions and under periods of severe stress caused by natural, technological and intentional hazard events. These infrastructure systems interact dynamically with human activity through social, economic and political systems and this newly available data provides a unique opportunity to investigate the dynamic interactions between human and physical systems.

The data on urban infrastructure systems can be combined with massive amounts of cell phone location data, social media postings, transit access card swipes, drone and surveillance video and credit card transaction records. These unstructured data depict the activities of urban residents. Together this rich confluence of data provides a dynamic, comprehensive view of the functioning of the city and the activity patterns of urban populations. The combination of instrumented urban infrastructure data with social media and transaction data is known as Urban Big Data. Urban Big Data provides a truly unique opportunity to investigate and understand the dynamic interactions between urban residents and built environment systems (Boyd and Crawford, 2012). However, these data are often fragmented and controlled by a variety of public agencies and private corporations. As a result, they are not readily available to the urban research community. This paper lays out a strategy to develop a data commons that would collect, curate and distribute urban big data to support research on infrastructure systems and how they interact with the human populations they support.

To exploit these new opportunities, the urban research community needs to develop a strategy on how to tap this rich new source of data to support its investigations. Private corporations are launching similarly large-scale big data initiatives. For example, IBM is developing systems to use big data in decision-making for many of their corporate clients (Davenport and Dyche, 2013; Perret, 2014). Connecting previously separate types of infrastructure data will support investigation of the complex interactions among previously separate infrastructure systems.

To develop a strategy for collecting and distribution urban big data the Civil, Mechanical, Manufacturing and Innovation Division (CMMI) of the U. S. National Science Foundation (NSF) held a workshop on February 6-7, 2017 in Arlington, Virginia. The authors organized and coordinated this workshop that included engineers, material scientists, urban planners, data scientists as well as publishers, librarians, and representatives of relevant professional organizations. This interdisciplinary group reviewed the experience of several successful data repositories in the physical sciences (e.g., EarthCube) and identified the data needs and opportunities facing the urban research community. The workshop developed a strategy for using, storing, and sharing urban infrastructure data and began to define specific research projects that can lay the groundwork for a data commons platform. It produced a set of recommendations on the methods and techniques for collecting and curating large volumes of urban data, including software platforms to make data discoverable and useful to the urban research community. This data commons is intended to become a part of NSF's Cyber-infrastructure. The workshop assembled several interdisciplinary teams to develop prototype platforms for sharing this type of data. This paper will build upon the results of this workshop to describe a strategy to create a data commons for collecting and sharing Urban Big Data to support the next generation of urban infrastructure research.

While civil and mechanical engineers sometimes conduct lab experiments on infrastructure systems (e.g. the Network for Earthquake Engineering Simulation), a vast amount of data that describes the characteristics and ongoing performance of urban infrastructure systems is collected and maintained by public agencies (e.g. state transportation agencies, local water and sewer authorities) or private companies (electric power utilities, telecom companies) that operate those systems. To conduct research in this area requires getting access to already collected data and addressing the privacy and security issues associated with using that data. While the workshop focused on the needs of the CMMI research community, manyn

of the strategies developed in this workshop should be useful to a wider set of disciplines.

2. The Benefits of Sharing Data

A number of other disciplines, including Astrophysics, Earth Sciences and Genomics, have realized significant benefits from sharing data within their the research communities. Sharing data within a research community has been found to lower research costs by reusing available data, increasing the rigor of scientific research, and providing enough data to support machine learning and other techniques that depend on large volumes of data. This workshop developed a strategy to help CMMI realize similar benefits by collecting, curating and sharing data to support its research mission.

The astrophysics and structural biology communities now routinely share data collected by large instruments and by individual investigators. It is routine for biologists to upload data on the structure of new molecules when the paper describing the result is submitted for publication. Large astronomical databases, such as the Sloan Digital Sky Survey and the Hubble Space Telescope Archive, have demonstrated that a large community of users who are not directly connected with the investigators, who acquired the original data, can effectively use the data. This increases the impact of these instruments and improves the return on investment to support them. These communities have shown that widespread sharing of scientific data is both possible and effective. Thus, in order to realize the full potential impact of Big Urban Data, it will be necessary to explore methods to incentivize the collection, curation and sharing of data.

Much of the cost of doing research on urban infrastructure systems goes toward collecting data. The data are generally quite fragmented and controlled by a number of different public and private entities (Hissan, 2012). There are legitimate security and privacy concerns with releasing this data, so acquiring the data is often difficult, time consuming and sometimes impossible. The creation of a data repository or data commons would make this data much more readily available to the research community. This would lower the cost and the barriers to entry for doing this type of research and allow investigators to focus their efforts on analysis rather than data acquisition. The repository should be designed to address security and privacy concerns and, therefore, be a trusted dissemination site for data owners. It would also relieve data owners from the burden of providing

data to the research community. The data commons would function much like a library for urban infrastructure data.

One of the main benefits of creating a shared data repository is increasing the volume of data available to any single researcher. By aggregating larger volumes of data in a repository, researchers will be able to use innovative analysis techniques, such as machine learning or graph analysis, that are not possible with more limited sets of data. This will allow researchers to investigate the complex interactions and interdependencies among infrastructure systems. Combining this data with detailed social media and transaction data will provide a basis for understanding how urban residents impact infrastructure systems, and how the performance of those systems impact the activity patterns of urban residents. Many of the most important discoveries about the function of urban areas are likely to come from connecting previously separate streams of data. For example, better understanding the complex interactions among previously separate infrastructure systems (e.g., water and energy) can support the design of more sustainable solutions (French, Barchers, and Zhang, 2015)

3. Workshop Results

Workshop participants were divided into five breakout groups and asked to address key issues related to developing a shared data repository. The results of those discussions are summarized below. The five breakout groups were:

- Sustaining a Data Repository
- Incentivizing Data Sharing
- Innovative Data Creation and Fusion
- Metadata Schema and Resource Discovery
- Using Data Management Plans

Sustaining a Data Repository

Building a data commons requires long-term strategic planning to create a sustainable platform. To develop a successful repository it will be important to develop a sustainable business model. The Data Research alliance has done a study of the revenue streams of a number of data reposito-

ries. The report is available at https://www.rd-alliance.org/final-report-income-streams-data-repositories.html

It is important to understand the business requirements of stakeholders to know what they want and need from a data repository. These characteristics should be built into the repository from the start. This will require surveying the community and end-users as a part of the design process. This will help determine when and why researchers will be willing to share data. It is also important to clearly understand the value added proposition for users, both those who contribute data and those who use it. This will be especially important if there are fees associated with contributing or accessing the data.

It is also important for the data commons to have a clearly defined scope. This should identify what will be included in the repository and what will not. Sustainability and incentives are related. There must be incentives for data sharing (e.g., credit for data citation, improved research performance, access to new data). Principal investigators are primarily focused on research, so new participants will be required to focus on the development of a data repository.

There is considerable diversity in data produced and consumed by the CMMI research community. One data model may not fit all the users. It may be best to consider a federation of data repositories that meet the specific needs of different communities.

There are three distinct business models that have ben used to support data repositories:

- Users pay to access the data, and
- Users pay to store their data in the repository
- A central organization supports the repository.

While a repository that is supported either by users of data contributors is appealing, neither of theses options is likely to be a viable business model for the long term. To be effective a data repository needs to be supported by a stable funding source, probably from a single funding agency. Funding for such a repository by an agency, such as the National Science Foundation, can be justified by the cost savings of data collection on individual research projects, by ensuring more thorough analysis of data that is

collected by individual projects and by enabling new discoveries from fusing heterogeneous data sources.

A successful repository will require a governance structure for the long-term. The repository design must consider system reliability and on-going support. The user community must have trust that the repository will be long lived and not go away. The required data management plans (DMP) may be a method to get NSF funded projects to contribute data to a shared repository.

Given that similar data access discussions are occurring across multiple research areas, there is an opportunity for multiple NSF programs (and non-NSF programs) to pool their resources to launch a combined effort to develop a data repository or federation of repositories.

Incentivizing Data Sharing

Sharing data is common in many research communities, yet within the urban research community this is not the case. The objective of this group is to identify mechanisms that can be used to make data sharing the norm in this community. There should be approaches that reward Pinvestigators for sharing data.

There are a number of reasons why the research community should support data sharing. Perhaps the most compelling is that data sharing can improve individual and group productivity. A large amount of time and energy is expended in data collection in many research projects, leaving limited time to analyze the data collected. Even when the data is fully analyzed, new hypotheses or approaches may emerge at a later date that would suggest new ways to analyze the existing data. Also, the larger amounts of data available by pooling data across many projects can enable machine learning and other advanced analysis techniques that can increase research productivity. This increased productivity is potentially a strong incentive to encourage data sharing.

There are a number of possible incentives to encourage data sharing within the research community. These included a competitive advantage in funding, data citation, and showcasing outstanding data sharing examples. In addition, researchers could be given credit for the data resources they make available through citation and recognition or through awards from professional or academic societies. Universities could also recognize data

sharing as a scholarly contribution in the promotion process. Peer recognition can be another strong incentive for data sharing.

Rules can also create a framework that requires data sharing. This included mandates from funding agencies, assured security of data, embargos on access to data until researchers have published their results and requirements to use existing data to validate new models developed by research projects. Funding agencies can mandate data sharing as a condition of funding. These mandates would have to be monitored to insure compliance. Similarly, publishers can mandate that authors make their data available as a condition of publication. Ultimately, the research community has to agree on acceptable norms that govern behavior in that community.

Innovative Data Creation and Fusion

Large-scale fusion of data from heterogeneous sources is creating a new way of doing science (Batty, 2013). . Urban infrastructure systems and smart buildings are being monitored continuously by imbedded systems, mobile sensors and increasing by the cell phones functioning as citizen sensors. In addition, social media postings (Facebook, Twitter, FourSquare, etc.), surveillance cameras, drones, cell phone location data, license plate readers, transit access cards and credit card transaction records provide a dynamic view of human behavior that can be connected with the performance of the city's infrastructure systems and their performance (Hasan et al, 2013). A great deal of this Urban Big Data includes either a time stamp or geo-location (Crampton et al., 2013). These two items will be key to fusing the wide variety of infrastructure data with detailed human activity pattern data.

Analytics is evolving into a new way of creating data rather than just analyze data after it is collected. One of the key benefits of combining large amounts of data from multiple sources is the ability to see new patterns and relationships that may not be apparent within a single project or data set. Access to large fused data sets supports machine learning and graph analysis (Few, 2009; Cuzzocrea, 2011). This is one of the most promising aspects of moving to a shared data model.

Historically, we have studied critical infrastructure with limited, imperfect data. Urban infrastructure systems are rapidly joining the Internet of Things (IoT) as instrumentation is added to transportation, water and sewer systems and to electric power grids. This system-level data can be com-

bined with human behavior data drawn from social media to better understand urban metabolism and activity patterns. However, this data is often incomplete or in private hands.

The nature of analytics is changing and fusion techniques are evolving to support data creation rather than simple post-collection analysis. Data fusion methods can be applied to large-scale sensor networks and Internet of Things. There are special requirements for the geospatial data to account for new forms of geospatial data collection, including drones, social media and surveillance cameras. Taking full advantage of these new analytic techniques will require a combination of domain knowledge and computational expertise.

There are significant privacy concerns when dealing with high resolution remote sensing, social media or travel data. We must develop better algorithms to prevent re-identification from linked data, for example identifying vehicle identification numbers (VIN) from motor vehicle data. We need to have the ability to link data, yet preserve privacy of VINs and other information. We need to create multi-disciplinary approaches to studying privacy that include data science, social science and legal scholars. Privacy concerns must be addressed as a part of the data commons platform.

Metadata Schema and Vocabulary for Resource Discovery

Data drawn from a variety of sources will inevitably include differences in vocabulary, metadata and data naming conventions. To be successful a data commons requires a common set of metadata. This is key to making the data easily discoverable and understandable to users. Ontologies that bridge the differences in data naming conventions are key to building a successful data repository. Developing a metadata schema and requiring data to conform to it will be necessary to support robust search and data discovery.

New forms of metadata may be required to support unstructured data such as video and social media data. Again, time and location tagging are the key to making the data discoverable and linkable with all the relevant data that describe urban systems at a particular time and place.

Using Data Management Plans and Existing Data Centers

The Data Management Plans (DMP) that are required as a part of all NSF proposals can play a significant role in attaining the goal of sharable and discoverable data for all CMMI programs. Data Management Plans can we evaluated as part of the review of any new proposal. However, his will require specific criteria for evaluation versus a simple compliance approach. To be effective these criteria will need to be similar to those used to evaluate Intellectual Merit and Broader Impacts.

Considering data sharing as an evaluation criterion would require a culture change among reviewers and NSF program managers. Past performance in data sharing from earlier projects could be evaluated much as publication of results is currently. Previous work should document dissemination of data. Investigators with a strong track record of wide data dissemination would be given credit much as publications are now. Data sharing could be considered as a part of Broader Impacts rather than as a separate criterion.

Currently, domain experts with little expertise in data sharing are writing and evaluating Data Management Plans. More specific criteria are needed to evaluate Data Management Plans and reviewers and investigators would need training to implement this culture change. To be effective NSF would need to develop a way to monitor and enforce Data Management Plans.

To encourage data dissemination NSF should require proposals to specify funding to make project data public. Investigators would need to be provided guidance on the costs associated with data sharing. NSF can facilitate data sharing if it develops a data repository (or repositories). This would also lower the cost and increase the effectiveness of data sharing. NSF may need to provide supplemental funding to make data public. Post grant awards like Research Experience for Undergraduates (REU) funding or supplemental funds as part of the grant request from a pool within CMMI or the Engineering Directorate could be used to support data sharing.

2. Challenges to Building a Data Commons

The amount of data available on infrastructure systems is rapidly growing, and the existence of a central data repository would greatly benefit the urban research community. However, there are many challenges to initiating, expanding, and maintaining such a data repository. Due to the sheer breadth and depth of the data available for a repository, there is concern regarding data accuracy and avoidance of duplication. Therefore, systematic maintenance and screening of data is necessary, as well as efforts to clean and validate data, and maintain data currency in order to handle data evolution on timescales much shorter than data retention periods within the repository.

Initiating a central data repository for the CMMI research community will necessitate taking advantage of existing, long-running repositories in an effort to reduce duplication of data that may already be available to the CMMI research community. Prior to initiating a central data repository, existing data repositories will need to be examined in order to better understand data standards and protocols for archiving, linking, and generating metadata. It may also be necessary to look to existing data repositories for platform developers and ways to incentivize data suppliers and users.

A cultural change regarding data ownership and overcoming biases towards sharing data within the CMMI research community will be crucial in attracting data suppliers and users of the repository. It is vital that the repository is able to gain recognition and citations in order to further ensure credibility and attract more data suppliers and users to the repository. Preserving and ensuring confidentiality when necessary, for locational and identification purposes, will also be crucial when attracting data suppliers. Gaining access to proprietary sources of data, and overcoming data security limitations will also be necessary in order to provide users with unique data sources, which will aid in attracting and maintaining users.

Ensuring that the CMMI research community has access to a reliable and robust repository will be important in attracting and maintaining users. Search and fusion techniques, such as cross-indexing heterogeneous data sources will need to be developed to ensure that the repository is user-friendly and contains an ample amount of data that is organized efficiently to improve searchability. Balancing access to raw data sources with desired analytics and varying computational needs will further ensure user satisfaction of the users of the repository.

Although there are numerous challenges to initiating, expanding, and maintaining a data repository, it is certainly an undertaking that will prove beneficial to the CMMI research community. Protocols for maintaining data accuracy and credibility will need to be established, and existing, long-running data repositories will likely need to be examined to ensure adherence to standards. Further development of ontologies to better use heterogeneous data sets will be necessary to improve efficiency of the repository and attract and maintain users. These efforts will likely aid in shifting the perception of data sharing, thus making the data repository a vital part of the CMMI research community.

3. Alternative Repository Models

The workshop reviewed several alternative models of existing data repositories. One example is Earthcube. This repository supported by the NSF Geosciences Division contains a wide variety of geosciences data that is shared across the earth science research community. The National Institute of Standards and Technology has developed and supports a materials data repository. (See https://materialsdata.nist.gov/dspace/xmlui/) The Urban Big Data Centre at University of Glasgow is a leader in combining transportation, infrastructure and social media data to understand the structure and function of urban areas. Citrine is an innovative start up company that collects materials data and applies machine learning to address materials research questions for industrial customers. Participating researchers can store access data without cost, while revenues from industrial users support the repository.

Workshop participants were asked to consider three alternative repository models:

- A data repository that covers all of CMMI
- An urban infrastructure repository
- A federated repository that links a number of existing repositories.

While a CMMI date repository is theoretically feasible, the heterogeneity of the research community and variety of data types would make it difficult to build coherent and easily searchable data repository. A repository that focuses specifically on urban infrastructure appears to be a more real-

istic short-term objective. Such a data commons would begin by incorporating data produced by NSF-funded projects. A next step would add infrastructure data available from other federal agencies, such as the U. S. Army Corps of Engineers, the Department of Homeland Security and NIST. A larger, multi-agency federated repository is a more ambitious target, but would benefit from the experience gained through building the more focused repository within CMMI.

4. Strategy for Building an Urban Big Data Commons

The results of this workshop suggest that there are definite benefits to creating a data commons to support the CMMI research community. The primary benefit will be increasing the efficiency of research by thoroughly analyzing all available data. By aggregating larger volumes of data in a shared repository, researchers will be able to use innovative analysis techniques, such as machine learning that are not possible with more limited amounts of data.

The workshop recommended a phased approach to developing a robust data sharing strategy for the CMMI community. In the short term (next 12 months), CMMI should fund a research effort to develop a prototype data repository. Such a project should include active researchers from the domains represented within the CMMI research community and data scientists who have expertise in developing repositories in other domains. The project would build on this workshop to delve more deeply into the questions that the research community wants to answer. Based on that analysis the project team would develop a schema for the repository and a robust set of query and data fusion tools to assist users in finding and creating useful data to support their research interests. This initial attempt would be comprised of data within the CMMI research community. The data management plans (DMP) for new projects should require that data produced by the project be included in the repository. This activity would be considered a part of the research effort and budgeted. Current projects and those completed within the last three years would be able to apply for supplemental funding to prepare their data for inclusion in the repository. The initial project would include an assessment of the usage patterns of the repository and determine its strengths and weaknesses.

Based on the user experience lessons learned from this initial prototyping exercise, the medium term (next 2-3 years) would develop a more ex-

tensive repository designed to serve the infrastructure and natural hazards communities. This repository should be designed as a federation that draws on existing repositories run by other agencies. In this type of federated repository the emphasis would be on developing ontologies that bridge the semantic difference in the schemas of the separate repositories. CMMI should develop a repository that specifically addresses the needs of the infrastructure and natural hazards research communities. This repository would combine the infrastructure data included in existing repositories, such as the Natural Hazards Engineering Research Infrastructure, NIST Center for Risk-Based Community Resilience Planning, and the Argonne National Lab Resilient Infrastructure Initiative repositories. Like any federated approaches the focus would be on developing robust ontologies to combine these independently developed repositories and develop a software platform to support data query, fusion and analysis. If possible, a team that includes domain experts as well as data scientists should develop this repository. Joint funding with Computer and Information Science and Engineering Directorate (CISE) should be explored.

In the long term, CMMI should look toward developing a large federated data repository with other NSF divisions, other federal agencies, state and local government agencies and private utilities and companies. Such a massive, searchable database would open new avenues of inquiry to researcher across a number of disciplines and significantly increase the speed and scope of scientific discovery within the urban research community.

References

Batty, M. (2013). *The New Science of Cities*. MIT Press.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provacations for a cultural, technological, and scholarly phenomenon. *Information, communication & society, 15(5)*. 662-679.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: situating 'big data'and leveraging the potential of the geoweb. *Cartography and geographic information science*, *40*(2), 130-139.

Cuzzocrea, A., Song, I. Y., and Davis, K. C. (2011). *Analytics over large-scale multidimensional data: the big data revolution!* Paper presented at the Proceedings of the ACM 14th international workshop on data Warehousing and OLAP.

Davenport, T. H., and Dyché, J. (2013). Big data in big Companies. International Insti.

Drummond, W. J, and French S. P. (2008). The Future of GIS in Planning: Converging Technologies and Diverging Interests. Journal of the American Planning Association 74:2 (Spring).

Few, S. (2009). Now you see it: simple visualization techniques for quantitative analysis: Analytics Press.

French, S., Barchers, C. and Zhang, W. (2015). Moving beyond Operations: Leveraging Big Data for Urban Planning Decisions. Computers in Urban Planning and Urban Management. July 11, 2015. Boston, Mass.

Gartner (2013). Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020. 12 December 2013.

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing* (p. 6). ACM.

Hinssen, P. (2012). Open data, Power, Smart Cities How big data turns every city into a data capital Retrieved from http://datascienceseries.com/assets/blog/Greenplum-Open_data_Power_Smart_Cities-web.pdf

Perret, R. (2014). *Why Infrastructure Matters for Big Data & Analytics*. Presentation to IBM. Retrieved from http://www.ibmbigdatahub.com/presentation/why-infrastructure-matters-big-data-and-analytics.