RS-Cities Proof of Concept #1 Plan and Justification: Initial IRD perspective

Aaron Birkland (IRD)
Cornell University

April 7, 2011

Abstract

This is the first of several proofs of concept related to data practices of social and environmental scientists involved in urban vulnerability studies. It will focus on issues surrounding the synthesis of quantitative data for statistical analysis from heterogeneous tabular input data. These issues will be explored in part by developing a demonstration application for producing a novel derived datasets based upon selection and transformation criteria provided by individual researchers. Concerns related to sementics, heterogeneity, derivation, and provenance are likely to produce a set of requirements and insights that will be relevant to current Data Conservancy infrastructure research and development at large.

1 Background

Resilient and Sustainable Cities (RS-Cities) is a research effort at NCAR which aims to investigate the relationship between urban vulnerability and climate change. The Data Conservancy social science team at NCAR are collaborating with RS-Cities researchers to in turn investigate *their* data practices, both as a specific project and as an archetype of particular kind of emerging research. This effort intends to produce a series of technological pilots and proofs of concept that investigate areas in which cyberinfrastructure can address scientific and data management challenges of social science researchers. To that end, the Data Conservancy Infrastructure Research

and Development (IRD) team is also involved in order to assist in the implementation and extraction of broad infrastructure requirements resulting from this work.

As a significant driver of this collaboration, the teams are launching the Urban Resilience Observatory (URO) initiative, which aims to provide an environment and a set of tools to aid researchers and policy makers with an interest in climate and urban vulnerability related research. This Urban Resilience Observatory effort will focus on the main overarching question:

> How can cities take an active role in mitigating and adapting to climate change?

As a step toward that goal, the initial activities of the URO will adopt a working research question:

> What are the dimensions (e.g., hazards) and determinants (e.g. age, gender, social networks) of a vulnerable/resilient situation across urban areas to the impact of climate hazards such as temperature change?

This scientific research question then leads into the initial area of technical inquiry of the first series of proofs of concept:

> How to integrate and synthesize disparate heterogeneous data for social scientists such that they can be useful in a meta-analysis and meta-framework across approaches, and useful for the purpose of decision support.

These questions will be explored by developing and evaluating a series of applications or tools which will constitute the URO.

There are at least three tools presently conceived for the URO exemplar: a data synthesis tool, a meta-analysis tool, and a decision support system. All three tools have the potential to produce one or more proofs of concept in order to demonstrate the theoretical soundness, utility, and feasibility of aspects within each tool.

The *data synthesis* tool intends to provide a means by which researchers can select, integrate, and transform heterogeneous spreadsheet data in order to support subsequent analysis. This tool may itself be divided into two sub-parts: quantitative synthesis and qualitative (or conceptual) synthesis. The former is perhaps the most straightforward, involving mostly numbers

and ranges of values such as temperature, population density, death rates, etc. The latter typically involves coded values derived from analysis of surveys or literature.

The *meta-analysis* and *meta-framework* tools intend to provide a means by which researchers can classify, partition, or compare studies and research artifacts. These synthesis tools (particularly qualitative synthesis) can be seen as a providing prerequisite functionality which provides some of the building blocks involved in performing a meta-analysis.

Finally, the *decision-support* tool aims to incorporate the results of a particular meta-analysis and present a set of visualizations of observed and predicted effects according to a particular model. Unlike the other two tools, this is aimed at an audience of policy makers and their related research or decision-making staff. It also targets researchers of issue-driven science who seek a conduit for their research to inform public policy.

There are many activities involved in the RS-Cities and DC collaboration which have not been mentioned here. The scope of this document is limited to the background and technical details of the proofs of their concept, and their potential impacts on broader infrastructure research and development within the Data Conservancy.

2 Definition

This initial proof of concept aims to demonstrate technical concepts underlying the proposed *quantitative* synthesis tool for social and environmental science data. A target date of the third Data Conservancy all-hands meeting (June 20) has been set for the completion of a demonstration application that is able to synthesize a subset of data types, selection, and transformation parameters that are involved in urban vulnerability analysis. Successful completion of this proof of concept shall inform the conceptualization of future full-scale pilots, as well as determine a basic understanding of the role and requirements of Data Conservancy infrastructure in support of such a tool.

The basic design and functionality of the synthesis tool will be determined by the the research needs expressed by RS-Cities scientists. Table 1 provides an example of the kinds of inputs and output parameters that may be pertinent to this tool. The precise set of input and outputs used in the proof of concept have yet to be chosen, but will will aim to representative of scientific interest and relevance

Input			Output
Indicator	Dimension		Parameters
Temperature	Hazard	Synthesis	Granularity
Population Density	Exposure		Region
Age, Gender	Sensitivity		Subset
Income, Education, Social networks	Adaptive Capacity		Categorization
Mortality, Morbidity	Impacts		Temporal

Table 1: Examples of input types and output parameters of the quantitative synthesis tool

Data will fall into one of at least five categories within a conceptual model of urban vulnerability (such as Hazard, Exposure, Sensitivity, etc). Each concept has its own set of relevant measures (e.g. temperature, population density).

Most (if not all) of the input data to the quantitative synthesis tool will be tabular in nature, in the excel spreadsheet format. The tool will categorize this data into the appropriate concepts and measures prescribed by the model. With this in place, the tool will allow selection, integration, and transformation of the data according to a set of user-supplied parameters. The output of this process will be a novel derived dataset in the form of a file containing tabular data.

The output data will be suitable for visualization or statistical analysis within a variety of tools (SPSS, for example). This analysis will occur by outside means— the primary value of the synthesis tool lies in producing the derived dataset itself.

A representative subset of concepts, measures, and parameters will be chosen to implement as part of this proof of concept. This subset shall be chosen by their capacity to illustrate the science and engineering challenges inherent in this kind of tool. In addition, this subset will be bounded by the pragmatically assessing the time and effort required to produce a working demonstration by a June 20 deadline.

3 Expected Infrastructure Outcomes

Over the course of the research, requirements gathering, and implementation of the proof of concept, there a few initial areas that seem likely to produce implications for Data Conservancy infrastructure at large.. While the scope of the proof of concept may be defined to cover only a subset of these areas (or discover new ones), particular attention will be spent generalizing the problem statements where they occur.

3.1 Semantics

The tool will be extracting data from particular rows, columns, or cells of tabular data based upon their type/domain. To do so requires either making assumptions about the structure of a given spreadsheet, or providing access to appropriate metadata describing the semantics of row, column, or cell values.

Determining the problem space will require characterizing the techniques and conventions typically encountered within the relevant social science communities— which might range from highly structured and well-described to completely ad-hoc. Likewise, this collaboration may identify opportunities to define approaches or best practices which may be adopted by the community.

This process of characterizing, representing and leveraging the semantics associated with the data will evaluate the capacity for the data conservancy infrastructure to facilitate the use or development of analysis tools. Does the data conservancy provide relevant services or models to aid in conveying the necessary semantics where they are present, and provide a means for coping with their absence (e.g. provide a straightforward means for manual or semi-automated annotation)?

3.2 Heterogeneity

Urban vulnerability research, particularly of the quantitative synthesis type being explored in this proof of concept, involves analyzing heterogeneous data from a number of domains. Different parameters, units, and ontologies are used between and within communities. This heterogeneity is usually resolved through manual labour. If the semantics of the underlying data are adequately characterized, automated integration becomes possible.

IRD has discussed issues such as unit conversion and normalization as part of feature extraction and/or the query framework. This proof of concept may explore the appropriateness, feasibility, and role of data conservancy services that could in performing integration tasks such as required by this tool.

3.3 Derivation and Provenance

There are two kinds of provenance that may be explored within this proof of concept. The first is 'source' provenance, concerned with tracking the specific set of data used to produce an integrated dataset. The second is 'process' provenance, concerned with tracking the selection, faceting, or transformation parameters used.

The tool will likely have its own set of requirements for managing provenance for the benefit of the scientist user. If we consider the case where a derived dataset is deposited in a Data Conservancy System (DCS) alongside the source data, there are questions related to the representation of this provenance within the DCS model. We may examine the value proposition of representing the two types of provenance in the DCS, and determine what representation or access capabilities would be necessary for an application to use these capabilities for its own purposes.

4 Process

Design, development, and evaluation of the proof of concept will be performed as the result of a collaborative effort of multiple teams and individuals:

Requirements and Use Cases The ultimate source of requirements will be RS-Cities scientists Patricia Romero-Lankao and Hua Qin. Lynne Davis from the NCAR social science team will lead the extraction of use cases and requirements that will determine the form and function of the proof of concept application.

Development Peter Alston from the NCAR social science team will provide user interface design and implementation. Aaron Birkland from IRD will provide the design and implementation of the synthesis tool logic.

Evaluation Lynne Davis and the NCAR, RS-Cities teams will evaluate the effectiveness of the prototype in addressing challenges faced by urban vulnerability researchers. Aaron Birkland and the IRD team will evaluate the the feasibility and generalizability of infrastructure requirements or questions that result from this work.

The term "proof of concept" is used to imply a certain number of properties which will characterize the progression of this effort:

- A core set of functionality or problem statements will be defined. Development effort which does not directly address one of these issues will be minimized, and may result in mocked up or hard-coded surrogates for certain components.

- The proof of concept may or may not use existing Data Conservancy software or services in its demonstration application. That being said, it is important that the role of existing or proposed DC infrastructure in an expanded pilot is understood and articulated by the end of the proof of concept.

- Simplifying assumptions may be stated and refined as the proof of concept progresses, especially where they describe plausible conditions that may be met only through defining new requirements or capabilities on Data Conservancy infrastructure. An understanding of how reasonable these assumptions are should be achieved by the end of the proof of concept.

The proof of concept team is currently drafting a timeline and set of design goals/requirements for this effort. An initial set of three publications has been chosen in order to frame the specific datasets, science questions, and examples of heterogeneity that will be demonstrated by June 20. The scope of this proof of concept has been limited to the extent that its completion is feasible, yet exposes a satisfactory number of challenges relevant to infrastructure and scientific practice. This balance will be actively maintained and adjusted throughout the entire proof of concept time period, while keeping in mind the overarching questions and areas of inquiry that motivate the Urban Resiliency Observatory effort.