A Distributed Infrastructure for Scalable Data Democratization and Visual-Analytics for the Masses

Executive Summary:

The Semantic Web is coming to fruition in our time, and will most likely become key to the foundation of the future Internet ecosystem. The World Wide Web is transforming into a giant living database; a rich ecosystem of data providers, service providers, and interactive applications. Weave is a web-based collaborative interactive visualization and analysis software framework developed here at the Institute for Visualization and Perception Research (IVPR) at the University of Massachusetts Lowell. Weave is currently in use by about a dozen large public and government organizations and will be released to the public at the end of March 2011 when its use is expected to grow tremendously.

Our proposed activity concerns the design and implementation of the next generation Web infrastructure for Weave: a federated system for collaborative data dissemination, integration and interactive visualization based on Semantic Web technologies and existing Weave components. We have designed a data model which will enable development of key Weave features which have been significant road blocks up to this point: data provider federation, extensive structural and descriptive metadata, and the ability to interactively navigate federated hierarchical data cubes.

The proposed development will serve as an enabling technology for achieving a long term vision of the Open Indicators Consortium, the primary stake holders and contributors to Weave development, namely the notion of a *data commons*: a global Internet-based public resource into which anyone can deposit data, and which anyone can query for any purpose [1]. We plan to realize this consortium vision as a distributed web of Weave servers, each owned and operated by the organizations publishing the data. Although our present goals are driven by interactive web-based collaborative visualization and data analysis, we seek to build a general infrastructure that enables *any* community to effectively pool their data resources and build a comprehensive web of public data for any purpose.

Proposed Research:

Weave is a system in which the server is responsible for exposing data to a web-based visualization client and an administration console. The Weave server comminicates with the and client and administration console through a Web API we have defined for accessing and managing data tables and geometry collections. On many occasions throughout the development of Weave, we have encountered recurring roadblocks in implementing Weave to it's full potential. There have been many moments when someone would imagine a feature which would be ideal to have at the moment - such as "Now that I've seen the data for the States, I want to zoom into the counties of Texas!" - but when the developers were confronted with these possibilities, it became clear that some new structures in the *data model* for Weave needed to be introduced.

Weave is the third generation of systems developed at IVPR with the similar overarching system architectures and feature sets - visualization tools capable of brushed selection and probing across *tabular data sets*. A simple data table model has been a basic assumption in Weave from it's inception, therefore it permeates its architecture and design. Implementation of certain desired features which require a *more elaborate fundamental data model* are stalled for this reason. Though as a team we have great desire to implement these "slightly out of reach" features as soon as possible, it appears that changing the fundamental Weave data model would require changes in a great many locations throughout the code. It would also impact the way in which data is imported and described, therefore would require extensive

development in the Weave Administration Console as well.

The research problem we intend to address is that of how the basic functionality of interactive data visualization and analysis tools can be augmented with a comprehensive yet practically oriented data model and data dissemination and management infrastructure. We have a plan to resolve many persistent issues with Weave which would drastically improve it's usefulness and usability and are critical to the formation of a Weave-based data commons. Our plan addresses the following:

1. *Data provider federation and dynamic data integration:* "Here I see financial statistics published on one site and demographic data on another. I want to plot them side by side in Weave." Currently Weave can integrate data from multiple sources, but only when the strings used as row and column identifiers match (and the unit of measurement is the same) across all data providers involved, which is often not the case and difficult to maintain when it is. These three requirements lead to elaboration of the next three points:
2. *The universal identification of entities*: "I have a data set for population of US States using FIPS codes, and another for income using state abbreviations. I want to plot these together in a single Weave visualization." Weave currently uses unqualified strings as entity identifiers, and has no built-in mechanism for supporting entity equivalence mappings. A scheme for both universal identification and equivalence declaration is needed for this functionality.
3. *The universal identification of attributes (i.e. "measures" and "indicators"):* "Server A provides "population" for the counties of Texas, while Server B provides "popTotal" for the counties of California. I want Weave to plot the union of both columns, because they mean the same thing." Currently, Weave only keeps track of column meanings using unqualified strings, and no attribute equivalence class mechanisms have been established.
4. *The ability to integrate data sets which express the same attributes using different units*: "I have a data set for population in millions for European countries, and for population in thousands for African Countries. I want to plot these side by side in Weave." Currently, Weave has no metadata regarding units and cannot utilize unit conversion factors.
5. *The expression and traversal of hierarchical relationships*: "Now that I've seen the data for the US States, I want to zoom into the counties of Texas." Weave supports visualizations at both levels of detail, but there exists no metadata that links them together. Such metadata is needed to implement hierarchical data drill-down interactions within the Weave user interface, which is a major long standing goal of the Weave team.
6. *The expression and navigation of multidimensional aggregated data*: "I'm looking at a bar chart of obesity by US States. I'd like to pivot the visualization so each bar represents obesity per year instead of per state." Weave can produce both visualizations, but the data must be pre-projected from it's full data cube form. Most Weave data sets are most naturally modeled as hierarchical data cubes [2], but Weave has no infrastructure to support metadata structurally describing how data table contents fit into a larger scheme of hierarchical data cubes.
7. *The inclusion of data provenance information in the data visualization interface*. "This data seems a bit sketchy. Where did it come from? When was it published? Who published it originally?" Weave currently does not support provenance metadata in it's visualization interface, though we have recently developed support for storing Dublin Core metadata elements [3] associated with data tables through the Weave Administration Console.

Though the list of features above may seem to span a wide range of research topics, from our

perspective each issue represents a useful and in-demand feature for Weave which would bring us leaps forward in terms of our capabilities. Our proposed activity involves applying Semantic Web and database technologies to tackling all issues listed above.

Semantic Web technologies [4] address data store modeling, federation and integration. We have identified the Semantic Web platform as viable route to solving the above listed challenges as follows: 1: RDF graphs can be correctly merged; 2 and 3: the platform is built on the URI universal identification standard; 4, 5 and 6: we can build ontologies (or utilize existing ones) for units and hierarchical data cubes, and 7: we can use an existing standard such as Dublin Core for descriptive metadata.

Some of our goals are solved simply by using existing standards. To solve the others, we have designed a data model for federated hierarchical data cubes with associated units. We plan on using Semantic Web technologies to express, store and query all structural and descriptive parts of the data model, which includes the following core entity classes: *data set, data table, data table column, dimension, level, member, measure, quantity, unit* and *aggregation function*.

Many entity classes in our data model are derived from standard concepts in OLAP [2], but are designed explicitly to support federation and integration. Most of the data relevant to Weave users is best modeled abstractly as a *hierarchical data cube* - a structure which contains numeric *measures* aggregated over multiple independent hierarchies called *dimensions*. Entries within a dimension hierarchy are called *members*. Each *member* belongs to a *level* in a dimension hierarchy. These entities will be defined universally and referenced by data table metadata to qualify their contents.

Many meaningful data cube dimensions have already been defined within the Semantic Web, but not explicitly as such. For example, the GeoNames database (geonames.org) defines a hierarchy of named geographic regions, which could be leveraged to define the data cube dimension representing spatial regions. Wherever possible, we hope to seek out existing hierarchies and taxonomies within the Semantic Web and leverage them to define *dimensions*.

The part of the data model describing *measures* has additional detail not present in most OLAP systems concerning the *quantity* and *unit* of a *measure*. For our purposes, a *quantity* represents a class of comparable *units*. “Currency” is an example of a *quantity*, whereas “US Dollars” and “Euros” are examples of units. Based on this representational foundation, we plan to incorporate automatic normalization by unit conversion into Weave for dynamic data integration purposes.

The remaining tasks in realizing our proposed goals involve developing a bridge technology for allowing users to specify, using our data model, how their original tabular data is to be precisely interpreted. We plan to create user interfaces in the Weave Admin Console for import of two distinct kinds of data tables: 1: tables where rows correspond to RDF Resources, and 2: tables where rows correspond to *members* or intersections thereof (we call these *data cube fragment tables*). These two cases require very different treatment. We plan to address the first case with the assistance of an existing relational-RDF mapping systems such as the D2R project [5].

For the second case, we plan to develop a metadata management user interface for our data model within the Weave Administration Console. Importing data cube fragment tables implicitly involves manual and permanent definition of new instances of all classes in our data model. These structural metadata definitions will over time form an interlinked universal data cube metadata skeleton from which the actual contents of the data tables will hang. Our proposed plan will enable such an overarching metadata web to be established on a global scale, which will semantically link together all public data sets published using instances of the Weave server. The entirety of this universal data cube web can then be interactively explored using Weave and integrated tightly within the Weave user experience.

Footnotes:

1. Here we use the term *data commons* in same sense in which Richard Cyganiak used it during his 2008 interview published at the following URL: http://www.semantic-web.at/1.36.resource.252.oa-growing-data-commons-from-meaningful-bits-and-pieceso.htm
2. The *hierarchical data cube* data abstraction is described especially well in the 2002 InfoVis paper "Multiscale Visualization Using Data Cubes" by Stolte, Tang, and Hanrahan. On Line Analytical Processing (OLAP) is an area of data warehousing and business intelligence concerning hierarchical data cubes.
3. The Dublin Core Metadata Initiative (dublincore.org/) has published a standard set of descriptive metadata elements called the Dublin Core Metadata Element Set published at the following URL: http://www.dublincore.org/documents/dces/
4. RDF (Resource Description Framework) and OWL (the Web Ontology Language) are two Semantic Web standards released by the World Wide Web Consortium (W3C). RDF is a versatile data model which is fundamentally a *semantic graph* (a graph with labeled edges). OWL is a language for defining and publishing RDF vocabularies (i.e. schemas, data models) and inference rules. In RDF, each entity (or "resource") is identified by an internationally unique URI. *Linked Data* is a term coined for a subset of the Semantic Web focusing on using dereferenceable URIs as resource identifiers. A dereferenceable URIs is a URL which, when accessed, delivers (typically) an RDF/XML description of the resource, which may contain references to other dereferenceable URIs. In this way, a web of interlinked data is formed, termed the *Linked Data Cloud*.
5. The D2R Project is a project by Chris Bizer and Richard Cyganiak enabling SPARQL queries to be executed against relational databases. Information on this project can be found at the following URL: http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/