Emerging From the Quagmire: Building Expert Systems technologies for the Social Sciences

*By Robert Wozniak **

With the acceptance of processable metadata and the exploding growth of today's online data storage capacity, current stateless, largely context-free http- or cgi-driven extraction interfaces are quickly proving inadequate for traversing the vast amounts of online social science information. This paper explores ways of taking advantage of the latest technology for the discovery and access to ever-growing amounts of social science data as they are explored for the development of the NHGIS project at the Minnesota Population Center at the University of Minnesota.

Before the web, people could only go to experts who understood the data they were interested in. They described what they were after, using what terminology they were capable of, and left it to professionals to translate their request into a language the data extraction system understood. Putting an extraction process on the web, while relieving the burden on the professional, has simply shifted the burden of expertise onto the user. Without the guidance of a domain expert, users are only able to rely on the informational content displayed on their computer screen. Users risk spending their time scouring through a quagmire of documentation (sometimes with little context) and overwhelmed by seemingly inexhaustive and often times irrelevant lists and options.

Domain experts understand the ontology of their domain and can effectively draw the necessary (even common sense) inferences and deductions from a user's request to make a data extraction. It is this intellectual property that is missing in the vast majority of current online data extraction systems. Difficult hit-or-miss keyword searches and large selection lists are the norm today. But as data grows in size, comprehension and complexity, this approach becomes a hindrance. It is of paramount importance that organizations and domain experts take advantage of current technology and incorporate as much domain knowledge as possible within their search systems. Such advances will accommodate an ever-broadening user base confronted with an ever-growing amount of social science data.

Tomorrow's web-based solutions offer the means of democratizing access to data as well as interactively assisting users in understanding social science data and methodologies. Leveraging the development of the DDI, rule-based grammars for middle-tier processing, and xslt-driven interface and documentation generation, the web can be used as a pedagogic device to assist both novice and expert users in compiling meaningful social sciences data in a highly dynamic, personalized and intuitive way. This democratizes access in the best possible way: first, by accommodating both novice and expert level usage; and second, by offering the means by which the novice can expand and improve upon their knowledge of social sciences and quantitative research to become, should they so choose, a domain expert themselves. NHGIS is the Minnesota Population Center's first step towards making this next generation of web-enabled technology a reality.

The National Historical Geographic Information System (NHGIS) Overview

The NHGIS will make accessible the aggregate U.S. census data for all available census years between 1790 and 2000. There's over a terabyte of data with over 300,000 variables for these years, all of which we propose to make accessible online, with real time data views and downloading, to students, policy analysts, journalists and academic researchers. Simplifying access to this complex data so that these users will not need specialized training to make use of it is of crucial importance.

The United States summary census data are the primary source of statistical information about growth and change of the American population. The great bulk of these data exist in machine-readable form, but they are largely inaccessible. The over a terabyte of data covering the period 1790 through 2000 exist or are in preparation, but they are scattered across dozens of archives and stored in incompatible formats on CD-ROM, magnetic tape, or paper. Only a small fraction of these data are available on the Internet, and even those offer only primitive documentation and extraction tools. Moreover, census summary data cannot be effectively exploited without clear definitions of each geographic unit, but high-quality electronic boundary files exist only for the 1990 census year.

Technological change presents an unprecedented opportunity to make these data readily available for social science

research. Bringing the complete census within reach of social scientists will unlock the potential of two centuries of data collection, and will stimulate research in economics, history, sociology, geography and other fields.

The project consists of three major components: data and documentation, mapping, and data access.

- The data and documentation component gathers all extant machine-readable census summary data; fills holes in the surviving machine-readable data through data entry of paper census tabulations; harmonizes the formats and documentation of all files; and produces standardized electronic documentation according to the recently developed Data Documentation Initiative (DDI) specification.

- The mapping component creates consistent historical electronic boundary files for tracts, minor civil divisions, counties and larger geographic units.

- The data access component creates a powerful but user-friendly web-based browser and extraction system, based on the new DDI metadata standard. The system provides public access free of charge to both documentation and data, and presents results in the form of tables or maps.

This project was in part conceived as an online tool to relieve the burden of data archivists at the Machine Readable Data Center of MN (Minnesota) from conducting a request for an extraction of the aggregate census data in person or over the phone. As a result the situation lends itself to a traditional *Expert System* development scenario. But it does so without requiring the construction of such a system for the whole of social sciences, nor the whole of that part of the social sciences that lends itself to quantitative research. The restriction of work to well-defined domains within the social sciences as well as the availability of expertise in these domains, make this kind of approach to problems, like those faced with projects like the NHGIS, possible.

In general, *Expert Systems* software performs tasks otherwise performed by a human. In particular, for the NHGIS project, the software will function as a component of an online data extraction engine that encapsulates higher level *knowledge* about the domain of U.S. aggregate census data for the purposes of efficient exact as well as approximate data discovery over large data sets. The NHGIS middle-tier is designed to abstract out, make explicit, distribute and leverage *expertise* of its domain as opposed to automating manual procedures via the traditional development of algorithms. This is one of the key differences between knowledge-based systems like the NHGIS compared to current conventional data extraction engines.

What behooves one to build such a system?

The decline in the cost of data storage during the last five years, as well as the exploding growth and availability of the Internet, make it both possible to maintain the entire body of machine readable census data online as well as dramatically slashing the cost of access to that data for the end user. But while the storage and the port of access to this data improve, the *mode of access*, the underlying *data discovery mechanisms* employed for this access, must necessarily evolve to improve accessibility to this enormous data store for an ever-widening user base.

As the thirst for social science data as well as the storage capacity for this data grow hand in hand, we are faced with the peculiar problem of effectively attaching a drinking straw to a fire hydrant. As a result, brute force and algorithmic methods of pruning search space for discovering data may prove too inefficient or otherwise cumbersome. This situation is more complicated due to the symbolic nature of the metadata as opposed to the ordered or quantifiable nature of data most algorithms apply.

But it's not simply a question of methods we employ but also a question of how these methods are structured. Current methods are procedural in the sense that they are hardwired into the process logic of the middle-tier. As such they don't lend themselves as easily to the old 'Plug and Play' type scenario where rules can be manipulated at a higher level, untangled with the inner plumbing of a system's process logic, then dropped in and out of the process logic as necessary. They necessitate the work of programmers who translate the higher-level business logic of some expertise into lower level machine code that is then hand woven into the fabric of systems process logic. It is in this sense that the business logic of many systems can be considered "hardwired". This affects the dynamics of a system, such as its ability to grow, shrink, and adapt.

Systems of the size and complexity of the NHGIS could benefit from a modular design of the middle-tier that accommodates the quick prototyping of the business logic of the system in a non-procedural way. This necessitates the ability to easily add, modify and delete or disable business rules, which, in turn, require tools for the construction of these rules that accommodate usage by *domain experts* in addition to their systems' programmers. For example, we would like the ability to allow our domain experts to say:

"Actually this kind of data for this geography in 1960 doesn't exist, the system need not concern itself with this data at this level, in fact it need not concern itself with this class of data, at this level and all levels beneath it for all years until 1980."

They would then be able to use a tool to prototype a rule that states just that and drop it into the system for further testing. This *declarative approach* to rule specification are what expert systems technologies allow as a short cut

that can reduce both time and cost. That is, we can ignore the procedural aspects of such a rule, the reinvention of the loop, since an *expert system framework* takes care of that for us, allowing us to concentrate on the logic of the problem as opposed to the logic of the underlying implementation. In other words, the logic of the middle-tier more closely models the logic of the expert that defines that middle-tier. This approach also encourages tighter development and test cycles, allowing one to develop the system's intellectual infrastructure incrementally in the same manner the knowledge that governs a search is obtained incrementally through experience by domain experts.

With over a terabyte of data, 300,000+ variables, real time, online data viewing and downloading and a commitment to accessibility for users of all experience levels, this ability to keep the logic that governs search and presentation of this data in an explicit, higher level form is essential to the middle-tier component not only for its maintainability and modifiability over time but also the testing of its correctness, completeness, and consistency during development.

The NHGIS Knowledge Base

While rules govern the procedure of search over the data, a *knowledge base* represents the structure and content of that data and is precisely what a *rule base* depends on for satisfying some search criteria. The *NHGIS Knowledge Base* describes what entities exist in the data, it describes what these entities are, their properties and attributes as well as relevant relations that exist among them. In other words, the *knowledge base* consists of a high level, machine computable specification of what domain the NHGIS project ranges over. It deals with defining characteristics of identity and partial identity, it makes explicit a conceptual containment of terms into set theoretic and taxonomic structures, it defines a term's attributes or properties, it makes these relationships computable, effectively producing the means by which we can impart semantics to these terms and definitions. In some respects the *knowledge base* is like an online thesaurus.

Information of this sort exists in any extraction system at some level but the relationships between entities in these systems are implicit in either the layout of the metadata or hardwired in the procedural code that uses that metadata. The *knowledge base*, on the other hand, makes these relationships explicit and computable for the extraction process. By making the relationships explicit, the system gets closer to the semantics of the metadata, since it uses the semantic relations to prune search space, build interfaces or morph a search criteria, etc.

These semantic relations, for the purposes of the NHGIS project, borrow from lexicography, set theory, and philosophy and include:

Synonymy

In general, a definition of synonymy states that for two words, if a property exists such that the substitution of one word for another does not change the meaning of the sentence in which that substitution occurs, then the words can be considered synonymous. For NHGIS, we define synonymy as the property that exists between variables where the substitution of one for the other does not change the data for which those variables refer. This property is especially important for questions of comparability of variables across time.

Hyponymy and Hypernymy

The relationships of Hyponymy and Hypernomy classify entities into a hierarchical categorization of classes and instances, they denote in set theoretic terms to what set an object belongs and the attributes it may therefore inherit. For instance, the variable "poverty status" belongs to the set "population characteristic" and may therefore inherit much of its identity from the definition of the term "population characteristic". In this example, "poverty status" is a hyponym of "population characteristic" and "population characteristic" is a hypernym of the variable "poverty status". We can use these relationships to broaden the search to include terms of the same class as terms in the input, to assist the dynamic construction of interfaces, and in the translation of input.

Meronymy

This is a part/whole relationship that decomposes an entity into its component parts or the "stuff" from which that entity is made. In the *knowledge base* we take apart composite entities much like a car mechanic takes apart the engine of a car. While some of these entities are not exhaustively decomposable into a numerable set of atoms, like a car engine can be completely dismantled, many are, and it is the composition of these atomic elements of an entity that often times define that entity itself. For example, the "United States of America" is composed of a numerable set of states, which in turn are composed of a numerable set of counties, etc. until you reach an atomic building block from which the composition of higher level entities are composed. In some cases, like the entity "the United States of America", this composition comprises its functional definition.

Decomposition may occur along many lines for some entities in an ontology, but for the purposes of the NHGIS, those lines are often times evident if not already defined. For example, the U.S. Census Bureau's hierarchical decomposition of geography for the 1990 summary files describes this kind of relationship (figure 1).

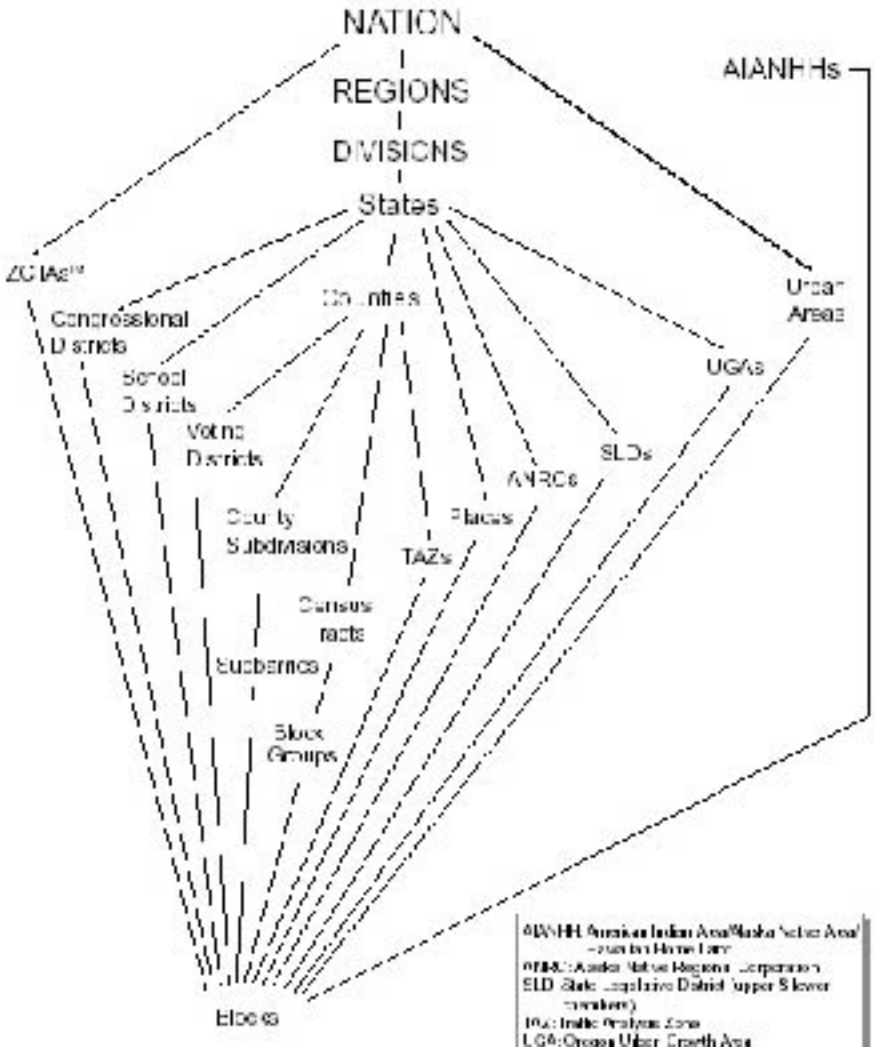

Figure 1

We can use this relationship to help determine variable availability for different geographies. For instance, if "income" and "education" variables are not available at the block level the system could use a *proximity rule* to expand the search to include the next geographic level up (or down) and deduce that these variables exist at that level instead. The system could then offer this result to the user with an explanation as to why it reached that conclusion.

Other lexicographic relationships we are exploring for use in the NHGIS Knowledge Base include:

Antonymy Similarity Polysemy

Origins of the NHGIS Knowledge Base and the Knowledge Acquisition Bottleneck

What keeps the development of an *ontological knowledge base* feasible, how can it be done and employed with a system as practical as an online data extraction system? How do you not get bogged down acquiring the knowledge for the ontology? I want to close with a few words addressing what's called the "Knowledge Acquisition Bottleneck" and how we propose to handle it for NHGIS. While many papers and talks have been given to address this problem, only a couple of points are mentioned here as they pertain to the project.

• Most if not all of the ontology for our project, as well as much of the rules that govern that ontology, already exist in the Census Bureau's technical documentation for the summary tape files and other forms of metadata.

• Much of this metadata can be parsed and put into the ontology automatically with scripts and software. The problem then becomes one of devising a clever parsing scheme to handle a document as opposed to mining deeply ingrained, non-systematized intelligence from domain experts.

These two facts go a long way to alleviating the knowledge acquisition problem and are something that many *knowledge engineers* do not have the opportunity to leverage. In fact, given the insurmountable complexities inherent in knowledge acquisition, the absence of this metadata would have been cause enough for us to reconsider.

Building a rule-based ontological knowledge base for any domain cannot be considered a trivial task but nor can the development of a middle-tier business logic for a project as large and complex as the NHGIS. We think, however, that our approach best models the intellectual infrastructure we need to incorporate into the NHGIS to successfully mine its data. The addition of this better model in turn solves some of the complexity of the development of the middle-tier since it allows for shorter-term quick prototyping as well as longer-term ease of maintenance and extendibility. In the end, it is our hope that this approach will prove beneficial not only to the NHGIS project but to the development of tomorrow's web-based solutions for the Social Sciences in toto.

* Paper presented at the IASSIST Conference, Storrs, CT, June, 2002. Robert Wozniak, Minnesota Population Center, wozniak@pop.umn.edu.