The GovStat Ontology

Maria Cristina Pattuelli, Stephanie W. Haas, Ron T. Brown, & Jesse Wilbur
School of Information and Library Science, University of North Carolina at Chapel Hill
100 Manning Hall, Chapel Hill, NC 27599-3360
[pattm, haas, browr]@ils.unc.edu, jdwilbur@email.unc.edu
http://www.ils.unc.edu/govstat/

Abstract

The goal of the GovStat Project is to create an integrated model that facilitates access and use of U.S. government statistical information. The vocabulary support team within the GovStat Project is developing the Statistical Interactive Glossary (SIG) and the GovStat Ontology as online vocabulary tools for supporting users seeking statistical information in government websites. This paper describes the functions and structure of the GovStat Ontology.

1. Introduction

The GovStat Project (http://www.ils.unc.edu/govstat/) is a joint project developed by the University of North Carolina Interaction Design Lab and the University of Maryland Human-Computer Interaction Lab. The goal of the project is to create an integrated model that facilitates access and use of U.S. government statistical information. The final perspective directing the project is the creation of a unified Statistical Knowledge Network (SKN). The primary research areas under investigation are metadata, interface design, and vocabulary tools (Marchionini, Haas, Plaisant, Shneiderman, & Hert, 2003).

People with limited statistical knowledge often have problems finding the statistical information they need and understanding what it means and how to use it. The vocabulary support team within the GovStat Project is developing the Statistical Interactive Glossary (SIG) and the GovStat ontology as online vocabulary tools for supporting users seeking statistical information in government websites.

2. Purpose

The purpose of the GovStat ontology is to supply semantic support for the SIG. The SIG is an enhanced glossary of statistical terms that users of federal statistical websites often need to understand in order to find or use the information they seek. Relevant characteristics of the statistical glossary include:

- Content. The selection of terms and the content of each explanation are aimed at providing a basic level of statistical literacy. Links to more complete or more technical explanations, such as those found in agencies' technical documents, can be included in the terms' presentations or ontology entries. The content of the explanations may include definitions, examples, brief tutorials, demonstrations, interactive simulations, or combinations of these elements.
- Context specificity. Explanations will be provided at different levels of specificity. When a user invokes help from a term (e.g., by clicking on it), the most specific explanation for the term and context are offered. If there is no explanation appropriate for a specific context, then a more general explanation is offered. The default is a set of context-free, "universal" presentations that can be invoked from any location.
- Format. Explanations in a variety of formats, including text, text plus audio narration, still images, animation, and interactive simulations will be provided.

The GovStat ontology supports the design and deployment of the SIG explanations in a number of ways. As an organizational tool, the ontology provides support for constructing and presenting explanations.

- The hierarchical structure of the ontology identifies related terms, including terms that are synonymous, broader, or narrower. Inheritance of taxonomic relationships between concepts supports the provision of context-specific presentations. When a user invokes help from a term that does not have an explanation tailored for that specific content, a more general explanation can be drawn from a more general term in the ontology.
- Semantic relations among concepts suggest opportunities for combining related concepts into a single more comprehensive explanation, such as a tutorial. For example, the *part-whole* relationship between *sample* and *population* suggests that an explanation of *sample* should include a mention of the population from which a sample is drawn.
- Once a way of explaining a concept has been established, then definitions or examples of subclasses of the concept can follow the template, with minor adjustments. Templates streamline the creation of additional presentations for other subclasses or for additional contexts. For example, explanations for *adjustment* can include a template that illustrates the general notion of smoothing statistics to remove predictable variation. Explanations of subclasses of *adjustment,* such as *seasonal adjustment* or *age adjustment,* can also be incorporated into this template.

As a navigation tool, the ontology provides the user with a means to navigate through statistical and agency-specific terms and definitions linked in a network of relationships. It can be manipulated directly as a standalone tool that offers the user a view of the domain coverage and the scope of the service. Used as an exploratory device, the ontology may help to increase the user understanding of statistical terms by browsing the semantic network of the concepts and facilitating serendipity.

3. Scope

The GovStat ontology is tailored for performing specific tasks in the domain of statistics. It reflects the scope of the SIG, being generally limited to those terms and concepts that a user may encounter on the agency websites. The exception to this is the occasional need to include concepts to bridge semantic gaps between target concepts. The GovStat ontology is an application-dependent and user-specific type of ontology (Fernandez Lopez, 1999; Zuniga, 2000).

4. Structure and content

The GovStat ontology represents terms and their relationships, and also informs the content and context of a term's presentations. The two categories of relationships used in the GovStat ontology are taxonomic and domain relationships. The taxonomic relationships are the partial ordering relations: *is-a* and *part-whole*. The *is-a*, or subsumption relation, is the basis of taxonomy and it is the most common relation for modeling concepts (Guarino & Welty, 2002). Examples in the GovStat ontology include:

```
Seasonal_adjustment Is-a Adjustment (Figure 1)
Age_adjustment Is-a Adjustment (Figure 1)
```

The *part-whole*, or mereological relation, can be of various types. An example of the *part-whole* relation in the GovStat ontology is:

```
Sample Is_part_of Population (Figure 2)
```

According to the classification proposed by Winston, Chaffin, and Hermann (1987), the relationship between sample and population would be considered a "portion-mass" or "slice-cake" relationship. The other category of relationships represented in the GovStat ontology is that of domain relations. These are typed relationships between terms which are able to express rich semantics. Examples in the GovStat ontology include:

    Smoothes (Figure 1)
    Is_an_estimate_of (Figure 2)

The final structure of the GovStat ontology is likely to be a "forest" (Sowa, 1984) or a family of trees, each expressing specific aspects or facets of the domain of interest (Smith, 2002) rather than a wide taxonomy composed of a single tree.
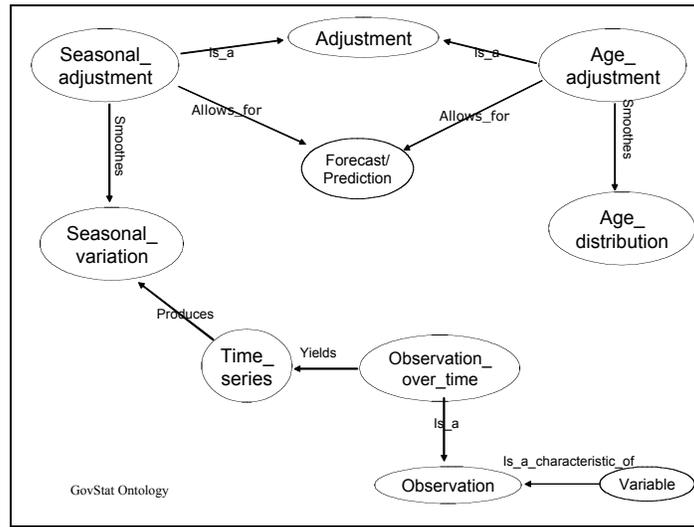

Figure 1. Portion of ontology that represents *Adjustment.*
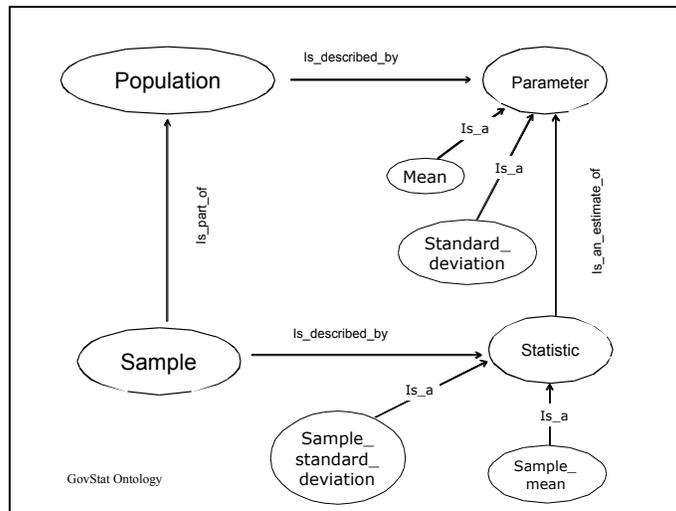

Figure 2. Portion of ontology that represents *Sample* and *Population.*

5. Development

Ontology development is an iterative process composed of a series of activities. The methodology adopted for the development of the GovStat ontology includes the following activities:

- **Specification**
- **Knowledge Acquisition**
- **Conceptualization**
- Formalization
- Implementation
- Integration
- Evaluation
- **Documentation.**

The activities in bold represent those completed or underway.

6. Acknowledgements.

7. References

Fernandez Lopez, M. (1999). Overview of methodologies for building ontologies. In Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends. CEUR Publications, 1999. *Intelligent Systems*, 16(1):26-- 34, 2001.

Guarino, N. & Welty, C. (2002). Evaluating ontological decisions with Ontoclean. In *Communication of the ACM*. 45:61-65.

Haas, S. W., Pattuelli, M. C., & Brown, R. T. (in review). Understanding statistical concepts and terms in context: The GovStat Ontology and the Statistical Interactive Glossary. Submitted to the *ASIST 2003 Annual Conference*.

Marchionini, G., Haas, S. W., Plaisant, C., Shneiderman, B., & Hert, C. (2003). Toward a statistical knowledge network. In Proceedings of *dg.o2003* .

Smith, B. (2002). Ontology and information systems. Retrieved December 3, 2002 from http://ontology.buffalo.edu/ontology.doc.

Sowa, J. F. (1984). Conceptual structures: Information processing in mind and machine. Reading, MA: Addison Wesley.

Winston, M. E., Chaffin, R., & Hermann, D. J. (1987). A taxonomy of part-whole relations. *Cognitive Science,* 11:417-444.

Zuniga, G. L. (2000). Ontology: Its transformation from philosophy to information systems. In Papers from the Second International Conference (pp. 187-197). ACM Press.