

The GovStat Ontology: Technical Report

Purpose

The GovStat ontology is a knowledge structure intended to semantically support the online Statistical Interactive Glossary (SIG). While SIG provides enhanced definitions of statistical terms in context, 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 will help users identify related terms, including terms that are synonymous, broader, or narrower. Glossary explanations will be offered at various levels of specificity and the ontology will provide a device for linking those different levels of explanations. Inheritance of taxonomic relationships between concepts will support the provision of context-specific presentations. For instance, if a term does not have an explanation tailored for a specific context in which it appears, a more general explanation can be drawn from a more general term.
- 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.

General Characteristics

The GovStat ontology is a domain-specific ontology tailored for performing specific tasks. Domain ontologies are focused on modeling specific areas of interest or domains. The conceptual domain represented by the GovStat ontology is statistics. However, only a limited portion of the statistical domain will be addressed based on

the tasks the ontology will be performing. Essentially, the GovStat ontology reflects the scope of the SIG, being limited to those terms and concepts that a non-expert in the statistical domain 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. The task to be performed or supported by the ontology has a great influence on the design of the ontology.

Methodology

There is a great variety in the way ontologies are created, and an ongoing discussion in the ontology community about the best practices for ontology development. One of the greatest challenges in constructing an ontology is the lack of formal standards or consensual methodology. Nevertheless, we identified a series of processes that should be addressed in developing the GovStat ontology which include:

- **Specification**
- **Conceptualization**
- Formalization
- Implementation
- Integration
- Evaluation
- Maintenance
- **Documentation**

The GovStat ontology is now at the beginning of its life cycle. The activities in bold indicate processes completed or in progress.

Content

When starting to develop an ontology, it is highly recommended to consider existing ontologies in the same or similar domain (Noy & McGuinness 2001). Existing ontologies can then be refined, extended, or simply used for mapping purposes. As for our ontology project, a number of online libraries of ontologies have been examined, including the Ontolingua Server (<http://ontolingua.stanford.edu>),¹ WebOnto (<http://kmi.open.ac.uk/projects/webonto>),² and DAML Ontology Library (<http://www.daml.org/ontologies/keyword.html>)³. Also, the "Ongoing Ontology Project" (<http://www.lsi.upc.es/luigic/ON-TO>), a rich collection of ontology projects, has been reviewed in order to identify ontology projects related to ours. Unfortunately, it does not appear that ontologies on statistics have been developed or made publicly available yet.

Nevertheless, I didn't start to collect information from scratch. A preliminary source of knowledge was provided by a vocabulary of over 60 terms which has been the

¹ Developed by the Knowledge System Laboratory (KSL) at Stanford University that, among other services, provides access to a library of ontologies.

² Server freely available to the ontology engineering community. WebOnto contains over 100 ontologies accessible and browsable.

³ Ontology library hosted by the DARPA Agent Markup Language (DAML) Program.

basis for the SIG. This vocabulary was not meant to be a comprehensive or definitive collection of terms, but a growing and flexible one. To supplement it, I have also consulted a number of online and printed statistical glossaries, dictionaries, manuals, and tutorials. This activity was extremely useful for identifying possible semantic discrepancies, for better understanding the meanings of the terms, and for discovering the semantic proximity of terms and relationships among concepts.

As a starting point in developing both the SIG and the GovStat ontology, we have selected groups of concepts because foundational (e.g., *Sample-Population*) and semantically challenging (e.g., *Age adjustment*, *Seasonal adjustment*, *Distribution*). So far, the ontology has modeled near 30 concepts. Additional concepts integrating the initial vocabulary are: *Variable*, *Multiple variable*, *Observation*, *Observation over time*, *Formula*, *CPI*, *Forecast/Prediction*.

Structure

During the conceptualization phase, the structure of the ontology has been defined by modeling selected clusters of terms around key concepts and their relations and by identifying the terms representing those concepts and relationships.

The concept organization of the GovStat ontology is based on two categories of relations: taxonomic and domain relations. Diagrams of the conceptual schemas modeled so far are provided in Appendix A, Figures 1-5. The diagrams are in the form of labeled directed graphs where the nodes indicate concepts and the arcs indicate binary relationships.

The taxonomy is traditionally the central part for most ontologies and the only one for some. The taxonomic relationships are “partial ordering relations” of the type *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. Examples in the GovStat ontology include:

Mean	Is_a	Parameter	Fig.1
Standard_deviation	Is_a	Parameter	Fig.1
Seasonal_adjustment	Is_a	Adjustment	Fig.4
Sample_mean	Is_a	Statistic	Fig.1
Sample_standard_deviation	Is_a	Statistic	Fig.1
Age_adjustment	Is_a	Adjustment	Fig.4
Observation_over_time	Is_a	Observation	Fig.4
CPI	Is_a	Index	Fig.3

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

Sample	Is_part_of	Population	Fig.1
--------	-------------------	------------	-------

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 contextual relations. These are typed relationships between terms which are able to express rich semantics. Examples in the GovStat ontology include:

Population	Is_described_by	Parameter	Fig.1
Sample	Is_described_by	Statistic	Fig.1
Sample	Is_composed_of	Observation	Fig.2
Statistic	Is_an_estimate_of	Parameter	Fig.1
Statistic	Is_described_by	Sample	Fig.2
Variable	Is_a_characteristic_of	Observation	Fig.2,4
Multiple_variable	Combines	Formula	Fig.3
Index	Is_calculated_by	Formula	Fig.3
Seasonal_adjustment	Smoothes	Seasonal_variation	Fig.4
Seasonal_adjustment	Allows_for	Forecast	Fig.4
Age_adjustment	Smoothes	Age_distribution	Fig.4
Age_adjustment	Allows_for	Forecast	Fig.4
Observation_over_time	Yields	Time_series	Fig.4
Time_series	Produces	Seasonal_variation	Fig.4
Distribution	Has	Central_tendency	Fig.5
Distribution	Has	Variation	Fig.5
Central_tendency	Is_estimated_by	Mean	Fig.5
Central_tendency	Is_estimated_by	Median	Fig.5
Central_tendency	Is_estimated_by	Mode	Fig.5
Mean	Is_an_average_of	Variable	Fig.5
Mean	Synonym_of	Average	Fig.5
Variation	Is_estimated_by	Range	Fig.5
Variation	Is_estimated_by	Standard_deviation	Fig.5
Variation	Is_estimated_by	Variance	Fig.5

So far, the GovStat ontology is composed of separate small tree structures with potential intersecting nodes (e.g., *Variable*). It is very likely that the final structure will be a *forest* (Sowa 1984) or a family of trees, each expressing specific aspects of the domain of interest rather than a taxonomy composed of a large single tree.

Formality

The GovStat ontology will most likely be implemented as a small *light-weight* ontology. This means that the level of formalization would include concepts, taxonomic relations among concepts, and association between concepts. This is the level of formalization most common among the majority of ontologies. The tasks that the GovStat ontology is intended to perform will probably require only minimal or no

axiomatization. A light-weight ontology can basically be implemented by all the ontology editors currently available (Staab et al. 2000).

References

Brown, R.T., Wilbur, J., Haas, S.W. & Pattuelli, M.C. (2003). The GovStat Statistical Interactive Glossary (SIG). *Proceedings of the National Conference on Digital Government Research, dg.o2003. Digital Government Research Center*, pp. 322-323.

Haas, S.W., Pattuelli, M.C., Brown, R.T. & Wilbur, J. (2003). The Understanding statistical concepts and terms in context: The GovStat Ontology and the Statistical Interactive Glossary. *Proceedings of the Annual Meeting of the American Society for information Science and Technology*, pp. 193-199.

Noy, N. & McGuinness, D.L. (2001). Ontology development 101: A guide to creating your first ontology. Retrieved October 23, 2002 from <http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html>

Pattuelli, M.C., Brown, R.T. & Wilbur, J. (2003). The GovStat Ontology. *Proceedings of the National Conference on Digital Government Research, dg.o2003. Digital Government Research Center*, pp. 355-358.

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

Staab, S., Erdmann, M., Mädche, A., & Decker, S. (2000). An extensible approach for modeling ontologies in RDF(S). Paper presented at Metadata ECDL 2000 Workshop on the Semantic Web, September 21, 2000, Lisbon. Retrieved October 11, 2002 from <http://www.ics.forth.gr/isl/SemWeb/PPT/1>

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

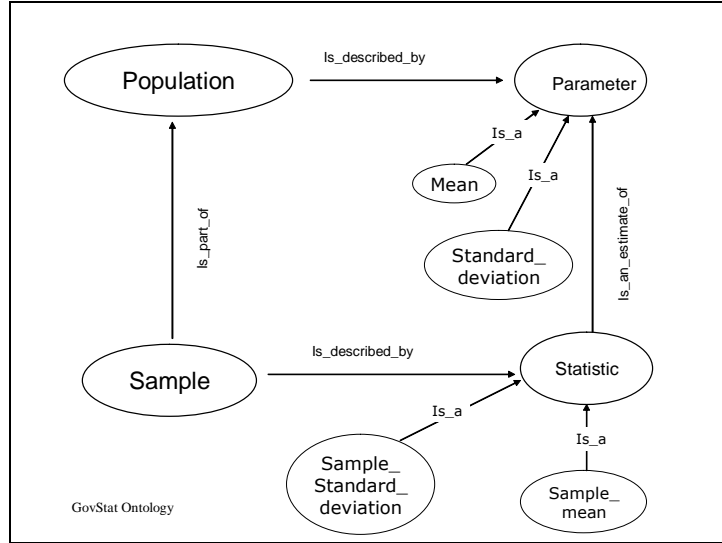


Fig. 1

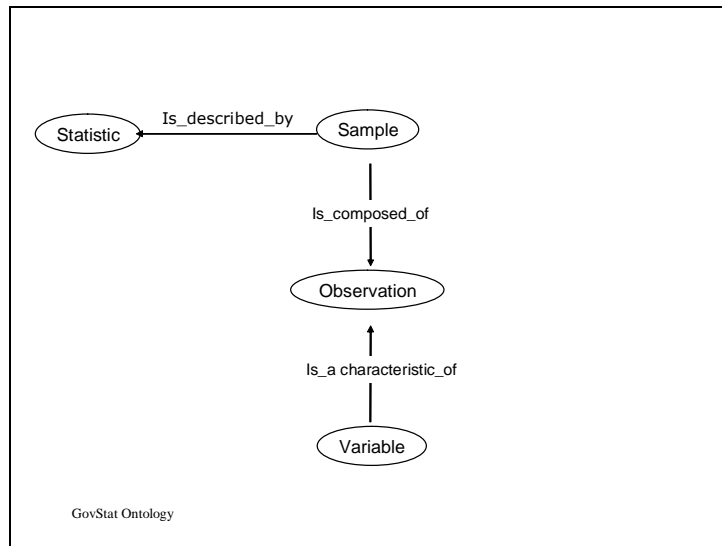


Fig. 2

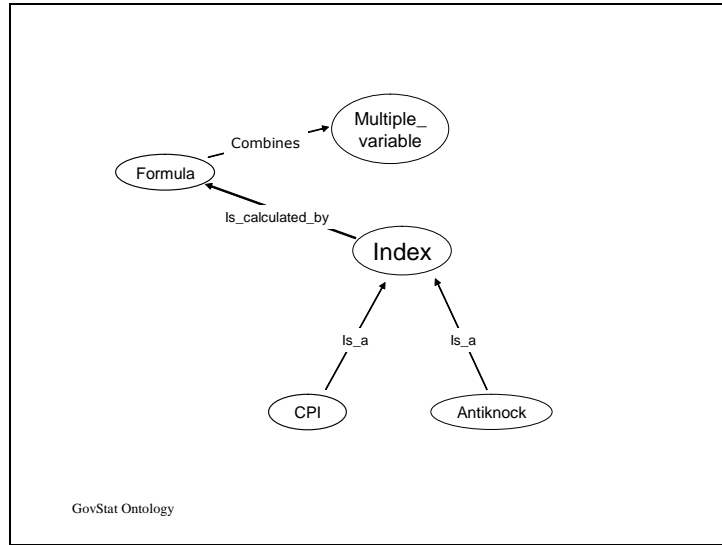


Fig. 3

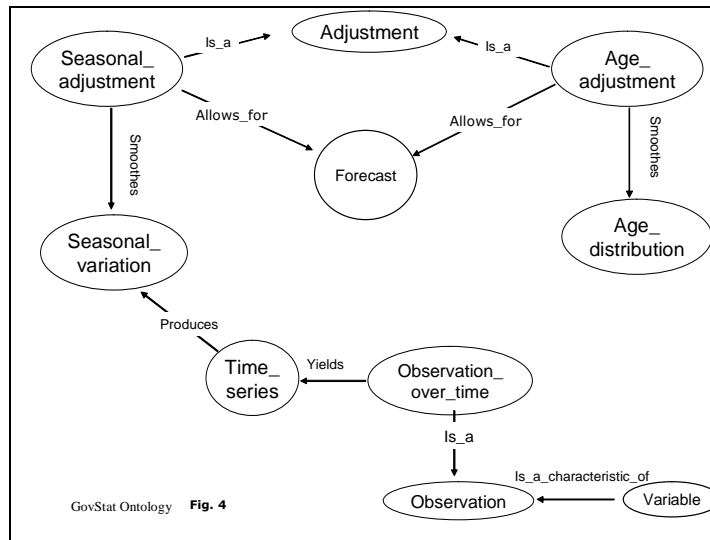


Fig. 4

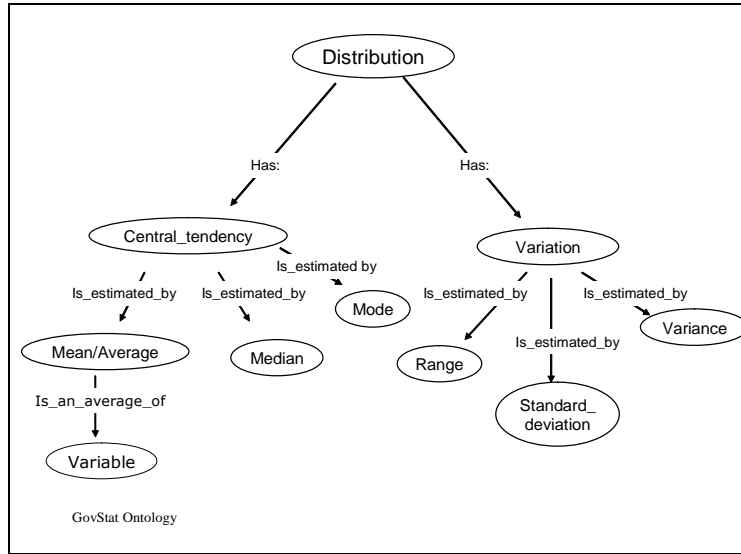


Fig. 5