

Experiences With Semi-Structured Data and XML

Wednesday, April 02, 2003 – John Elrick, Project Manager, Fenestra Technologies

Overview

Fenestra Technologies began work with XML in 1998 as part of an R&D project with the Bureau of Census. The goal of this project, Generalized Instrument Design System or GIDS, was to streamline and accelerate the process of creating survey instruments for the collection of statistical data.

During the ensuing years, Fenestra has utilized XML quite successfully as a medium both for persistent storage of object instances as well as a platform agnostic means of transferring data between disparate sources.

Although the common literature refers to XML as being "semi-structured" by definition, Fenestra's primary use of XML has been to utilize data that are very structured - that structure being defined by schema and by design documentation.

Fenestra's work was concentrated in statistical areas where accuracy was paramount. Therefore, it was an essential component of the work that precise information be available from domain experts as to the context and meaning of the data.

Persistent Storage of Object Instances

One use of XML by Fenestra has been as a simple mechanism for persisting object instance information. Fenestra has successfully used XML for this concept for purposes ranging from application configuration information to form design models.

One of the advantages to using XML for this purpose has been to create a text based representation of objects that is language neutral.

Fenestra's work included the development of a more robust version of XSL:FO (Formatting Objects), termed XSL:SFO (Survey Formatting Objects), which was created for the GIDS project. Fenestra is currently in the process of converting the GIDS applications to use the new SVG (Scalable Vector Graphics) standard, which will increase the flexibility and usefulness of those applications.

Data Transfer Between Disparate Sources

During the GIDS project, Fenestra worked with other third party providers, notably Oracle. By utilizing XML as a transfer medium, the GIDS suite was able to both read and write to a database without being tied to a specific database design. The advantage of this approach was that Oracle and Census were able to evolve their database without significant interference to the parallel development of the user applications.

This design also permitted Fenestra to adapt the existing software easily for a project with IBM and Census to produce typographically rich reports for the 2000 Population Census.

Conclusion

In conclusion, Fenestra Technologies has made extensive use of XML for persistence and data transfer. In practice, XML has served more as a structured data repository than as semi-structured data.

It has also been Fenestra's experience that there can be benefit gained by understanding the context of the data through direct interaction with domain experts rather than through inference.