

# Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



## Issue Statement: Data Interoperability Issues

S. A. Miller  
SIL International

Exchanging data between people and applications has always been problematic. First is the problem of format:

- Windows, Linux, Macintosh, and other operating systems have different file systems.
- Different database management systems (DBMSs) have different data storage mechanisms. Most are relational, but some are object-oriented, and others are hierarchical. Despite the SQL standard, each relational DBMS has its own proprietary language extensions.
- Operating system and database vendors frequently put out new releases of their software, causing version mismatch issues. I just spent six weeks upgrading our software to a new database version.

Second is the problem of transfer:

- The Internet has resolved many of the speed problems we had as recently as ten years ago. However, many rural areas of the world, including my parents' home in the United States, do not yet have DSL service.
- The amount of data that can be transferred varies, depending on the network or media used.

Among linguists, these issues are often exasperated:

- Linguists often maintain their data in documents which are written with various word processors and spreadsheet applications.



- Linguists often live in areas of the world where Internet service is limited, or simply does not exist.
- While SIL's Shoebox has been the tool of choice among many linguists, no one standard of "Standard Format Markers" (SFMs) exist. Last year I met a linguist who made up her own "standard" markers.

XML has solved many of the transfer issues between software applications. However, XML can be structured many different ways. Software developers must code software to accept each data schema they want to work with.

Transfer speeds are an infrastructure problem, and won't be solved by a conference such as this. The problem of a standard data schema could be. To my knowledge, SIL's UML class model is the most extensive, and could be used as a base for such an effort. An XML format for exchanging lexicons and dictionaries called LIFT has also been proposed.

