Application of Interpretation Standards in Complex Data Models

R. Mark Ward, FamilySearch (USA)

Abstract

Normalization is a common process utilized for databases, but what about structured and unstructured data contained within them or flat file systems? How does an organization standardize the interpretation of data types such as localities, names, dates and other controlled vocabularies? "Interpretation" in a global environment where structured and unstructured data exists intersects the practices of L10N, translation, transliteration, and data structure methods into a system which can accommodate indices and searches; across both closed and federated systems.

FamilySearch has spent over 50 years collecting, grooming and applying standards-based place, name (given, surname), date and controlled vocabulary data to its digital genealogical collection. This includes over 9 million distinct names for published places, over 12 million person names and a dozen calendaring systems.

With more than 3.3 million rolls of microfilm maintained by FamilySearch in a granite mountain vault outside of Salt Lake City and another 127 million digital records being captured via scanning projects every year across the world, the need to standardize the content to enable search and retrieval is critical.

The challenge is to provide relevant standards-based content that matches the diversity of geography, culture, language and history; thus creating a very complex data model. For example, for localities, a specific latitude and longitudinal place may have different names due to political upheaval (St. Petersburg, Leningrad, back to St. Petersburg), different languages due to conquering nations and even physical transition of a locality (Valdez, Alaska after the 1964 earthquake). Considering person names and moving to character-based languages in Asian collections, the level of difficulty and challenge increases in order to provide a roman transliteration of place and name content for non-native speakers. The ability to identify a name and it's pieces (prefix, title, given, surname, suffix, etc.) is important for search and understanding the cultural, variant, and linguistic relationship of person names.

Thus, simple tables are insufficient to address the multi-dimensional needs of standardizing the world's genealogically relevant information without incorporating complex data relationships.

This paper will discuss FamilySearch's use of an acyclic graphical model with its place and locality data to address historical and political diversity within jurisdictional hierarchies of the world, and how standardization processes for name piece identification, transliteration and date conversions provide the basis for strong search indices and the ability to federate information across diverse environments.