Semantics for Scientific Data:

Smart Dictionaries as Ontologies

Syd Hall, Nick Spadaccini, Doug du Boulay and Ian Castleden, Crystallography Centre and Department of Computer Science and Software Engineering, University of Western Australia, Crawley, 6009, Australia (syd@crystal.uwa.edu.au)

The Challenges

Phenomenal growth in scientific databases over the past decade, as well as the enormous disparity in data expression across long-standing and important taxonomic collections poses serious challenges to existing data handling methodologies. These will undoubtly be met with an array of approaches - common data protocols, better collaboration, global ontologies and active knowledge bases. Some scientific disciplines are advanced in removing the communication and access barriers, and there is a general recognition that a much higher level of data semantics is a key objective.

Existing data handling approaches apply semantic knowledge (meta-data) which is encoded as part of highly customised software (e.g. existing database, procurement, and inventory systems) lack the generality or extensibility needed for the easy addition of new data structures or methods. The major challenge met by our research is the development of a generic object-oriented approach to coalescing dictionary meta-data and instantiated data into executable processes that will underpin active knowledge bases.

A Generic Dictionary Approach

We embed machine-executable meta-data into dictionaries as simple text attributes capable of representing complex data relationships. These "smart" dictionaries provide the knowledge framework for generating data manipulation and interpretation tools targeted at local data needs. The presence of extensive meta-data in a dictionary affects future archival practices, in that only non-derivable data (i.e. measurements, etc.) need be archived - the rest can be generated from current knowledge. The effect of this paradigm shift on database management systems will be ubiquitous. In science, data can be arranged and interpreted according to derivation dependencies and the semantic content of dictionaies will proffer a level of flexibility and generality that is unattainable with current approaches.

Progress So Far

Our research is aimed specifically at designing a generic knowledge-base model using the Star File (Hall, 1991; Hall & Spadaccini, 1994), and at developing supporting tools. A Star File contains data that are textual, loosely structured and self-identifying. This work will extend the recent prototype dictionary efforts (Spadaccini, Hall & Castleden, 2000) of the authors. Applications of the Star File have been widely used for a decade. In chemical science, a domain-specific version of the Star File, the crystallographic information file (CIF) (Hall, Allen & Brown, 1991) is used extensively for publication and database purposes. In biological science, the macromolecular CIF data file mmCIF (Bourne et al., 1997) has been adopted by the Protein Data Bank (PDB), the Nucleic Acid Database (NDB) and Macromolecular Structure Database group at the European Bioinformatics Institute.

The scope of Star File data is enhanced considerably when individual items are defined as meta-data stored as a collection of data attributes in dictionary files. The allowed attribute types represent the definition language of the Star dictionary (DDL), and two DDL versions (Hall & Cook, 1995; Westbrook & Hall, 1995) are in current use. The development of a prototype version of a relational expression language dREL (Spadaccini, Hall & Castleden, 2000), as part of a new dictionary language StarDDL, is the basis for the current research program.

The development of dREL and StarDDL prototypes has shown that the precision of data definitions is enhanced significantly by specifying relationships between items as symbolic expressions that can be used to compute derivative data values. In particular, this work demonstrated that Star dictionaries are made much richer semantically when the attribute set is extended to include stronger typing and executable methods. The StarDDL differs significantly from other dictionary languages that are used solely to validate the structure and content of a data file (e.g. DTD in XML). A StarDDL dictionary may be compiled into executable dictionary objects that can be injected with specific data instantiations (i.e. particular data within a file) so that related items are dynamically linked through the dictionary methods. This is a dictionary approach which is well suited to knowledge retention and reuse. The approach does, however, incorporate other languages and data handling approaches when they complement the application of StarDDL dictionaries. Although XML has no intrinsic method functionality, it is used to interface our dictionaries to other computing languages and to off-the-shelf editing/browsing software.

Objectives and Significance

The objectives of this project are directed at the most serious deficiencies in existing data handling methodologies. Most archived data in science are unsuited, and even inaccessible, to modern access tools. Biological taxonomic data are a case in point. There is an enormous and continuing effort to capture biological-species information in the many museums, herbaria and universities around the world, with almost as many databases archiving taxon-based descriptive data. There are currently CODATA and OECD (Edwards et al., 2000) efforts within the GBIF program to coordinate and integrate the coding standards, such as DELTA (Dallwitz et al., 1992), used in these collections so as to provide new data structures better suited to systematic query methods. This interest also reflects a need for on-line sharing of data across disciplines - such as the integration of taxonomic data derived from morphology with molecular and genome sequence data.

The molecular structure data in the Nucleic Acid Database (NDB) at Rutgers University, New Jersey, and the taxonomic botanical data in the Western Australian Flora (Paczkowska & Chapman, 2000) and FloraBase database at the WA Herbarium in Perth, are of special importance to the project. They are excellent exemplars of data which must be interoperable, via consistent protocols, with facilities at other sites and in other countries, and therefore provide ideal test data for our research.