Betsy Rolland
Independent Study Proposal
LIS 600
Summer 2006
Dr. Terry Brooks

1. Abstract/General Description:

This Independent Study will examine XML schema mediation through the use of ontologies. Organizations are increasingly storing and exchanging data in the XML format. Frequently, related data will be stored in different XML files with different, yet similar, XML schemas. XML schema mediation is a technique for finding ways to integrate the data from those different schemas. Sometimes that means simply understanding that, for example, the “Name” field in one schema is the same thing as the “Author” field in another. More frequently, however, the data in one or both XML documents needs to be manipulated somehow. For example, one file may have “Last Name, First Name” in one field while the other file has separate fields for “Last Name” and “First Name.” In order to be integrated, either the “Last Name, First Name” data will need to be split or the “Last Name” and “First Name” data will need to be concatenated.

One of the main techniques for dealing with schema mediation is to use ontologies that describe the domain and allow for translation among XML files or between an XML file and the ontology. This method allows for a standardized way of looking at the data and ensures that mediation will be consistent.

A data schema is an abstraction of an information structure. The information structure this Independent Study will cover is the genetics databases of the Bio-Mediator project, run by Dr. Peter Tarczy-Hornoch in the UW Biomedical and Health Informatics department. The purpose of Bio-Mediator is to allow scientists to run a single query that gathers information from various genetics databases. As each database has only a portion of the information a research scientist might need, having the ability to automatically aggregate data is of paramount importance. For example, a researcher can use Bio-Mediator to query for all genetic information about diabetes. Bio-Mediator then translates that query into languages that each database understands and returns a standardized result set in XML.

One of the key aspects of Bio-Mediator is the way in which it uses XML to mediate between different data schemas. Each research database has its own way of organizing its data and each genetics project is collecting different, yet related, information. The strength of Bio-Mediator is that its architecture utilizes an ontology to simplify the schema mediation. Rather than mediating among a wide variety of schemas, Bio-Mediator simply mediates between each individual schema and the ontology itself. I will be working with a group working on SNP (Single Nucleotide Polymorphism) which deals with genetic polymorphism in the human genome. I will be working directly with this research group to create their mediated schema that will interact with the established Bio-Mediator ontology.

2. Learning Objectives:

Through this Independent Study, I will learn how to create a mediated schema based on an ontology.

3. Textbooks and/or Resources Required

- Books:

    - XML for Bioinformatics by Ethan Cerami
    - Essential genetics : a genomics perspective by Daniel L. Hartl, Elizabeth W. Jones
    - Fundamental genetics by John Ringo
    - Owl: Representing Information Using the Web Ontology Language by Lee W. Lacy
    - Practical RDF (Paperback) by Shelley Powers
    - XML hacks by Michael Fitzgerald
    - The XML schema companion by Neil Bradley
- Websites:
    - Bio-Mediator Project Homepage http://www.biomediator.org/
- Software:
    - Altova MapForce
    - Stylus Studio
    - XML Spy

4. Activities

- Research ontologies and how they are used in XML schema formation.
- Research how XML schemas are created and what are the best practices for mediating between and among them.
- Research how scientific projects utilize XML for exchanging and aggregating data.
- Work with the Bio-Mediator team to understand fully the information architecture of Bio-Mediator.
- Create a mediated schema for the SNP Bio-Mediator project.

5. Expected Outcomes/Project Deliverables

- Delivery to Dr. Brooks of a webpage demonstrating all work completed, including the stylesheet that mediates between the schemas and ontology and documentation produced for the Bio-Mediator team.
- Project demonstration of mediated schema for the Bio-Mediator team.

6. Evaluation/Assessment Methods

- Delivery of demonstration webpage to Dr. Brooks
- Achievement of satisfactory mediated schema for the Bio-Mediator team