XML-Enabled Wide Area Search in Bioinformatics (XEWA) - Position Proposal

Genomica Corporation, a company involved in providing integrated genomic software tools and publicly available reference data, is particularly interested in the ability to query and retrieve data from multitudes of diverse publicly available data providers. Areas of particular interest for our company include data to support genetic analysis and pharmacogenomics (e.g. physical maps, most importantly "Golden Path" sequence maps, genomic annotations, SNPs and other polymorphic markers, Loci, etc.). Our previous experience in querying and obtaining data from a multitude of existing public sources can be summarized as follows:

1. Semantics of publicly available data is frequently unclear
2. Data ambiguities exist in the data as there is frequently a lack of appropriate constraints
3. Data formats and content are frequently poorly documented
4. Data formats and content change over time
5. Consistent query interfaces are usually not present and requires resolution of above characteristics.

In order to make these various sites truly useful to the scientific public at large, without each user requiring tremendous investment in resolving the described issues, an attempt to "structure" the query interfaces and provide descriptions to the data is required. These include querying capability as well as return data formats and/or representations. In addition, perhaps the following scientific realities should be considered and/or addressed:

1. It is difficult to get agreement to standards (e.g. ask scientists what a "Locus" is and you will get as many answers as there are scientists).
2. Even if standards are agreed to, how does one encourage widespread acceptance? An associated issue is what is the cost to the various sites to conform to standards?
3. Science is a rapidly evolving field where new data types and semantics appear frequently. A sufficiently flexible representation is required to adapt to these ever-evolving requirements. A very flexible representation allows divergent representation of similar concepts requiring additional mapping or meta-data layers – perhaps limiting acceptance due to associated complexity.
4. Conformance to the decided approaches should be relatively easy as data may be maintained by biologists (i.e. individuals without extensive computer science backgrounds).

We are interested in participating to discuss, provide, and test technology and scientific solutions to the above problems, particularly in the ability to interact with diverse sites in a relatively straightforward manner and retrieve consistent data. We favor a system with some of the following general characteristics:

1. Ensuring sharable semantics
2. Constraints ensuring data quality
3. Simplicity of use
4. Using "accepted" technologies to provide increased likelihood of conformance
5. Defined, yet extensible interfaces – we feel that overly generalized systems tend to be difficult to understand even if they are extremely flexible

6. Meta data driven transformation ability. It is unlikely that everyone will accept certain standards – how do we convert from one data format to the other?
7. Systems supporting evolutionary development & standards

Structured XML-based ontologies and associated technologies appear to satisfy many of the described criteria. We look forward to discussing these issues in the context of pragmatic real world solutions.